If and How Many 'Races'?
The Application of Mixture Modeling to World-Wide Human Craniometric Variation

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Bridget Frances Beatrice Algee-Hewitt
December 2011

Copyright © 2011 by Bridget Frances Beatrice Algee-Hewitt
All rights reserved.

Dedication

This Is Just To Say[1]

I have eaten
the plums
that were in
the icebox

and which
you were probably
saving
for breakfast

Forgive me
they were delicious
so sweet
and so cold

Thank You, Mark.

[1] William Carlos Williams, 1934

Acknowledgements

First, let me take this opportunity to express my deep appreciation to my doctoral committee members, Lyle Konigsberg, Richard Jantz and Andrew Kramer in the Department of Anthropology and, Karla Matteson, in the Department of Medical Genetics. Without their helpful guidance and genuine interest throughout all of my time at the University of Tennessee, even across many state and continental lines, this dissertation would not have been completed. I would also like to extend my many thanks to Tsunehiko Hanihara, from the Kitasato University School of Medicine, and to Richard Jantz and Lee Meadows Jantz, in the Forensic Anthropology Center at the University of Tennessee, for kindly providing the skeletal data that has made this project possible; I am equally grateful to Geoffrey McLachlan, in the Department of Mathematics at the University of Queensland, for sharing with me, without hesitation, his mixture software in advance of its public release. The generous support of the National Science Foundation, the Social Sciences and Humanities Research Council, and the William M. Bass Endowment made my graduate research, and this project, possible.

The contributions, both personal and profession, of the scholarly community in the Department of Anthropology at the University of Tennessee deserve special mention. Lyle Konigsberg not only introduced me to the Bayesian tradition, but he reminded me of my roots as the granddaughter in a string band family. The value of his interest, in all likelihood, cannot be over-estimated. Andy Kramer has served as a kind and thoughtful mentor who has given me, time and again, the opportunity to expand my teaching and research skills, broaden my area of knowledge, and to work towards becoming a well-rounded anthropologist. Richard Jantz

oversaw my first research project in morphometrics and was patient and interested even when I was keen to discuss articles on capturing shape differences among fly wings, apples, and mouse mandibles. Murray Marks gave me the freedom to pursue my interest as a teacher, which undoubtedly has contributed to my growth as a researcher. Lee Meadows Jantz provided me with the exposure and experience to understand the fundamentals of the field and the confidence to share my skills in the classroom and the lab. Finally, I would like to thank Rebecca Taylor and Ann Carson for being colleagues and friends, especially, by entertaining my often randomly distributed questions and accommodating my data needs.

I am forever indebted to my parents, Janet and Boyd Algee, for their support in all things and in all ways. It is from their example that I have learned what it is to have unrelenting ambition and achieve lofty goals and it is through their compassion, interest and nurturing that I am, I hope, developing into my best self. I also owe much of my interest in learning to my sister, Hannah, whose high expectations for my breadth and level of knowledge has made me work hard to meet the bar. Most importantly, I owe this dissertation, and all my successes in life, to the symbiosis that I share with my significant-other, Mark Algee-Hewitt. He has helped me find the "probabilities" in everything.

Abstract

Studies in human cranial variation are extensive and widely discussed. While skeletal biologists continue to focus on questions of biological distance and population history, group-specific knowledge is being increasingly used for human identification in medico-legal contexts. The importance of this research has been often overshadowed by both philosophic and methodological concerns. Many analyses have been constrained in their scope by the limited availability of representative samples and readily criticized for adopting statistical techniques that require user-guidance and *a priori* information. A multi-part project is presented here that implements model-based clustering as an alternative approach for population studies using craniometric traits. This project also introduces the use of forced-directed graphing and mixture-based supervised classification methods as statistically robust and practically useful techniques.

This project considers three well-documented craniometric sources, whose samples collectively permit large-scale analyses and tests of population structure at a variety of partitions and for different goals. The craniofacial measurements drawn from the world-wide data sets collected by Howells and Hanihara permit rigorous tests for group differences and cryptic population structure. The inclusion of modern American samples from the Forensic Anthropology Data Bank allows for investigations into the importance of biosocial race and biogeographic ancestry in forensic anthropology. Demographic information from the United States Census Bureau is used to contextualize these samples within the range of the racial diversity represented in the American population-at-large.

This project's findings support the presence of population structure, the utility of finite mixture methods to questions of biological classification, and the validity of supervised discrimination methods as reliable tools. They also attest to the importance of context for producing the most useful information on identity and affinity. These results suggest that a meaningful relationship between statistically inferred clusters and predefined groups does exist and that population-informative differences in cranial morphology can be detected with measured degrees of statistical certainty, even when true memberships are unknown. They imply, in turn, that the estimation of biogeographic ancestry and the identification of biosocial race in forensic anthropology can provide useful information for modern American casework that can be evidenced by scientific methods.

Table of Contents

Introduction: Problems and Prospects 1

Chapter 1: Race 17

What's The Use of Race? 17

Towards a Definition of Race 20

Race in Early Anthropology 43

Race at the Advent of Genetics 83

Chapter 2: Human Variation 101

Race in the Genomic Age 101

Race and Human Biological Variation 110

Human Craniometric Variation 117

Classification and Quantitative Methods 133

Forensic Anthropology and FORDISC 156

Engaging the FORDISC Debate 190

Detecting Structured Variation: A Mixture-Clustering Approach 232

Statistical Fallacies: Revisting Classification 242

Genetic Clustering: STRUCTURE and Race 258

Chapter 3: Finite Mixture Model-Based Clustering 296

Classification and Clustering 296

Mixture: Development 304

Mixture: Theory 328

Mixture: Stature Example 333

Defining the Finite Mixture Model 341

Validating the Mixture Solution 369

Implementation: Mixture Software 374

Advantages of Finite Mixture Methods 377

Chapter 4: Materials, Methods and Analysis 380

Project Rationale 380

Global Patterns of Craniometric Variation: The Howells and Hanihara Samples 385

The Forensic Significance of Race: The Forensic Anthropology Data Bank 407

Sample and Model Assumptions 415

Size Correction 419

Missing Data Treatment 422

Variable Selection 430

Diachronic Trends and Secular Change 436

Tests of Structured Variation: Howells' World-Wide Data Set 446

Tests of Structured Variation: The Combined Howells and Hanihara Data Set 569

Tests of Structured Variation: Howells' "TEST" Data Set 590

Tests of Race: The Forensic Anthropology Data Bank 608

Validation Tests: Supervised Classification and Information Visualization 649

Conclusion: A New [Re]view of Race and Human Variation 678

List of Works Cited 712

Vita 814

List of Tables

Table 1: Available model parameterizations (reproduced from Fraley & Raftery, 2011, p.6).... 349

Table 2: Sample composition of Howells' craniometric data set 387

Table 3: Craniometric variables selected (after Howells 1973)..... 390

Table 4: Combined Howells and Hanihara data set, sample composition by sex and region..... 402

Table 5: Combined sample composition of Howells' base (1989) and "TEST" (1995) data set .. 406

Table 6: Missing data patterns 423

Table 7: Available model parameterizations (Fraley & Raftery, 2011, p.6) 450

Table 8: Top 10 models in order..... 453

Table 9: Classification table 459

Table 10: Quantiles of uncertainty 461

Table11: Allocation rates..... 468

Table 12: Cluster memberships by population (females) 472

Table 13: Cluster memberships by population (males)..... 473

Table 14: Classification table 504

Table 15: Allocation rates..... 508

Table 16: Quantiles of uncertainty 509

Table 17: Cluster composition by population contribution 511

Table 18: Classification success and model uncertainty 518

Table 19: Summary of EMMIX analysis 526

Table 20: Classification table 526

Table 21: Quantiles of uncertainty 526

Table 22: Cluster memberships by population (females) 527

Table 23: Cluster memberships by population (males) 528

Table 24: Allocation rates 530

Table 25: Summary of sex-specific clustering 544

Table 26: Top 10 models in order 554

Table 27: Cluster memberships by population (pooled sexes) 555

Table 28: Allocation rates 558

Table 29: Quantiles of uncertainty 559

Table 30: Statistical properties of select mixture analyses 560

Table 31: Top 10 models in order 574

Table 32: Allocation rates (imputed data set) 577

Table 33: Allocation rates (case deleted data set) 578

Table 34: Quantiles of uncertainty 578

Table 35: Model statistics for the Howells and Howells and Hanihara data sets 581

Table 36: Cluster memberships by population 582

Table 37: Summary of EMMIX model statistics 587

Table 38: Cluster memberships by population (pooled sexes) 588

Table 39: Top 10 models in order 591

Table 40: Allocation rates and uncertainty by model 593

Table 41: Classification table 595

Table 42: Cluster memberships by population 597

Table 43: Cluster memberships for the "Prehistoric" specimens 601

Table 44: Implied classification rates for select series 602

Table 45: Mixture DFA results using Howells' (1989 and 1995) data sets 605

Table 46: Quantiles of uncertainty ("Paleo" specimens excluded) 607

Table 47: U.S. Census (2010) and FDB (2008) race statistics 611

Table 48: Model selection 5 races 613

Table 49: Allocation rates 614

Table 50: Quantiles of uncertainty 614

Table 51: Cluster memberships by race 616

Table 52: Top 10 models in order and EMMIX model 622

Table 53: Allocation rates 623

Table 54: Quantiles of uncertainty 623

Table 55: Top 10 models in order and EMMIX model 629

Table 56: Allocation rates 630

Table 57: Quantiles of uncertainty 630

Table 58: Populations by U.S. Census and geographic labels 642

Table 59: Top 10 models in order 643

Table 60: Cluster membership by race and sex 644

Table 61: Allocation rates 645

Table 62: Quantiles of uncertainty 646

Table 63: Summary results of supervised cluster analyses 658

Table 64: Test data classifications (equal priors) 659

List of Figures

Figure 1: The annual percent distribution of publications referencing mixture-based methods (1980-2010). 316

Figure 2: The by-subject distribution of publications referencing mixture-based methods (1980-2010)..... 318

Figure 3: Histogram of the combined adult male and female statures from the William M. Bass Donated Skeletal Collection 336

Figure 4: Two component mixture density plot for the combined adult male and female statures from the William M. Bass Donated Skeletal Collection..... 338

Figure 5: Boxplot of the individual adult male and female statures from the William M. Bass Donated Skeletal Collection, when sex is "known." 339

Figure 6: Density plot of the individual adult male and female statures from the William M. Bass Donated Skeletal Collection, when sex is "known." 340

Figure 7: Model selection by BIC 453

Figure 8: Coordinate projection showing uncertainty 461

Figure 9: Uncertainty profile 462

Figure 10: Pair-wise scatterplot matrix showing clusters 469

Figure 11: Coordinate projection showing clusters..... 470

Figure 12: Random seed projections showing clusters 470

Figure 13: Cluster 1 composition..... 475

Figure 14: Cluster 2 composition..... 476

Figure 15: Cluster 3 composition..... 478

Figure 16: Cluster 4 compostion..... 479

Figure 17: Cluster 5 composition..... 480

Figure 18: Cluster 6 composition..... 481

Figure 19: Cluster 7 composition..... 482

Figure 20: Cluster 8 composition..... 483

Figure 21: Cluster 9 compostion..... 485

Figure 22: Cluster 10 composition..... 486

Figure 23: Cluster 11 composition..... 487

Figure 24: Cluster 12 composition..... 488

Figure 25: Cluster 13 composition..... 490

Figure 26: Cluster 14 composition..... 491

Figure 27: Cluster 15 composition..... 492

Figure 28: Cluster 16 composition..... 493

Figure 29: North Japanese, South Japanese, Guam 496

Figure 30: Ainu, Andaman, Easter Island, Philippines, Tolai 496

Figure 31: Zulu, Dogon, Teita, Bushman, Tasmania 497

Figure 32: Santa Cruz, Peru, Arikara, Moriori, Eskimo, Buriat..... 498

Figure 33: Berg, Norse, Zalavar, Egypt 499

Figure 34: Anyang, Hainan, Atayal 500

Figure 35: Pair-wise scatterplot matrix showing clusters 505

Figure 36: Coordinate projection showing clusters..... 506

Figure 37: Random seed projection showing clusters 506

Figure 38: Coordinate projection showing uncertainty 507

Figure 39: Uncertainty profile 507

Figure 40: Continental clusters..... 510

Figure 41: Dogon, Teita, Zulu, Australia, Tasmania, Tolai 512

Figure 42: Arikara, Santa Cruz, Peru, North Japan, South Japan..... 513

Figure 43: Ainu, Andaman, Anyang, Atayal, Philippines 514

Figure 44: Berg, Norse, Zalavar, Egypt 515

Figure 45: Eskimo, Buriat, Bushman..... 516

Figure 46: Dogon, Zulu, Teita, Australia, Tolai, Tasmania 532

Figure 47: Ainu, Bushman..... 534

Figure 48: Eskimo, Buriat..... 534

Figure 49: Peru, South Japan, Santa Cruz, Moriori, Mokapu, Guam..... 535

Figure 50: Anyang, Santa Cruz, Easter Island, Mokapu, Hainan, Andaman 536

Figure 51: Berg, Norse, Zalavar, Egypt 536

Figure 52: Clusters 1, 3, 8 and 9 538

Figure 53: Clusters 4 and 5 539

Figure 54: Continental clusters (females)..... 542

Figure 55: Continental clusters (males)..... 542

Figure 56: Scatterplot of size and sex (MCLUST G=16) 543

Figure 57: Scatterplot of size and sex (EMMIX G=10) 543

Figure 58: Model selection by BIC (females and males respectively) 546

Figure 59: Coordinate projections of the best female and male model respectively..... 546

Figure 60: Pair-wise scatterplot matrix showing clusters (females) 547

Figure 61: Pair-wise scatterplot matrix showing clusters (males)..... 548

Figure 62: Best female model chosen by BIC 550

Figure 63: Best male model chosen by BIC 550

Figure 64: Model selection by BIC 554

Figure 65: Random seed projection showing clusters 556

Figure 66: Pair-wise scatterplot matrix showing clusters 556

Figure 67: Coordinate projection and data profile showing uncertainty 559

Figure 68: Cluster 2 memberships; Cluster 3 (Ainu) 563

Figure 69: Cluster 8 memberships; Cluster 6 (Berg) 564

Figure 70: Cluster 10 memberships; Cluster 12 (Easter Island) 564

Figure 71: Cluster 4 memberships; Cluster 9 (Santa Cruz) 565

Figure 72: Cluster 5 memberships; Cluster 1 (Bushman) 565

Figure 73: Cluster 13 memberships; Cluster 11 (Tasmania) 566

Figure 74: Cluster 7 (Andaman); Cluster 14 (Buriat); Cluster 15 (Eskimo) 566

Figure 75: Continental clusters 567

Figure 76: Pairs plots for missing data (females); imputed values in black 571

Figure 77: Pairs plots for missing data (males); imputed values in black 572

Figure 78: Model selection by BIC for data sets with missing values treated by imputation and by case-wise deletion respectively 574

Figure 79: Pair-wise scatterplot matrix showing clusters (imputed data set) 575

Figure 80: Pair-wise scatterplot matrix showing clusters (case deleted data set) 576

Figure 81: Coordinate projections displaying misclassifications for the imputed and case-deleted data sets respectively 579

Figure 82: Uncertainty profiles displaying misclassifications for the imputed and case-deleted data sets respectively 579
Figure 83: Cluster 4 584
Figure 84: Cluster 5 584
Figure 85: Cluster 7 586
Figure 86: Model selection by BIC for 12 and 57 variables respectively 592
Figure 87: Coordinate projections showing clusters for 12 and 57 variables respectively 593
Figure 88: Coordinate projections showing uncertainty for 12 and 57 variables respectively ... 594
Figure 89: Cluster composition for the "Prehistoric"sample (12 variables) 602
Figure 90: Cluster composition for the "Prehistoric"sample (57 variables) 603
Figure 91: Model selection by BIC ("Paleo"specimens exluded) 606
Figure 92: Coordinate projection showing clusters ("Paleo"specimens excluded) 606
Figure 93: Coordinate projection showing uncertainty ("Paleo" specimens excluded) 607
Figure 94: Coordinate projection showing errors (filled circles) for target data using mixture DFA 607
Figure 95: Model selection by BIC 614
Figure 96: Coordinate projection showing clusters 615
Figure 97: Pair-wise scatterplot matrix showing clusters 615
Figure 98: Coordinate projection and data profile showing uncertainty 616
Figure 99: Black, East Asian, Hispanic, Native American, White 617
Figure 100: Coordinate projection showing clusters 621
Figure 101: Model selection by BIC 622
Figure 102: Coordinate projection showing clusters 623

Figure 103: Pair-wise scatterplot matrix showing clusters 624

Figure 104: Coordinate projection and data profile showing uncertainty 624

Figure 105: Black, Hispanic, White 628

Figure 106: Cluster composition by race 628

Figure 107: Model selection by BIC 629

Figure 108: Coordinate projection showing clusters 631

Figure 109: Pair-wise scatterplot matrix showing clusters 631

Figure 110: Coordinate projection and data profile showing uncertainty 632

Figure 111: Cluster composition by race 633

Figure 112: Coordinate projection showing clusters 634

Figure 113: Scatterplot by size and sex (all races) 638

Figure 114: Scatterplot by size and sex (American Blacks and Whites) 638

Figure 115: Scatterplot by birth year and sex (American Blacks and Whites) 639

Figure 116: Model selection by BIC 643

Figure 117: Coordinate projection showing clusters 644

Figure 118: Pairwise scatterplot matrix showing clusters 645

Figure 119: Coordinate projection showing uncertainty 646

Figure 120: Data profile showing uncertainty 646

Figure 121: Cluster composition by race 648

Figure 122: American Indian, Asian, Pacific Island, White, Black, Other, Admixed 649

Figure 123: Mosaic plot of Howells' data set (populations by color shade; U.S. Census races by color range) 666

Figure 124: Mosaic plot of Howells' data set (continents by color shade) 667

Figure 125: Mosaic plot of Howells' data set (sex by color shade) 668
Figure 126: Mosaic plot of Howells' data set (MCLUST 16-component model; cluster by color shade) 669
Figure 127: Mosaic plot of Howells' data set (EMMIX 10-component model; cluster by color shade) 670
Figure 128: Mosaic plot of FDB data set (5 races; races by color shade) 671
Figure 129: Mosaic plot of FDB data set (MCLUST 7-component model; cluster by color shade) 672
Figure 130: Mosaic plot of FDB data set (EMMIX 5-component model; cluster by color shade) 673
Figure 131: Mosaic plot of FDB data set (American Blacks and Whites; races by color shade) . 674
Figure 132: Mosaic plot of FDB data set of American Blacks and Whites (MCLUST 7-component model; cluster by color shade) 675
Figure 133: Mosaic plot of FDB data set of American Blacks and Whites (EMMIX 4-component model; cluster by color shade) 676
Figure 134: R Color Chart 677

Introduction: Problems and Prospects

"It does not require a trained anthropologist to classify a range of Englishmen, West Africans and Chinese with 100% accuracy by features, skin color, and type of hair in spite of so much variability within each of these groups that every individual can easily be distinguished from every other."(Wright, 1978, p. 439)

Sewall Wright, in his 1978 text, stakes a claim on the subject of human biological variation and its phenotypic expression that many today would still consider obvious. Nonetheless, the project of classifying modern humans by physical characteristics using quantitative methodologies has experienced a lengthy and difficult history in biological anthropology. Understanding the distributional patterns of human biological variation through time and across space remains a well-recognized, and essential, step towards revealing past population structure, demographic history and the trajectory of modern human evolution. The use of contemporary, world-wide skeletometric data in phenotypic trait analyses to (1) isolate biological differences among broadly sampled geographic, cultural, and genetic groups and to (2) access the nested hierarchical substructuring that underlies the morphogenetic relationships observed among populations that have been, otherwise, traditionally morphoscopically and metrically defined, has been variably and often contentiously received. The distillation of any presumed patterns of morphological [dis]similarity down to the level of the individual and the translation of these quantitative signatures into the language of biogeographic ancestry and biosocial race continues to generate heated debate. For this problem of classification validity, Diamond (1994) has clearly explained that, "[t]here are many different, equally valid procedures for defining races, and those different procedures yield

very different classifications…Faced with such differing classifications, many anthropologists today conclude that one cannot recognize any human races at all" (p. 84). There are, in short, two different ways of "knowing" race: one that permits scholars to understand and identify population differences, ancestry, ethnicity and race, and another that does not. A perceived gap exists, therefore, in contemporary anthropological thinking between what can be conceptualized intuitively based upon social constructions of human diversity and ideas of the self and others, and what can be tested and corroborated empirically using best practices and following accepted standards of scientific rigor.

Concerns over the scientific legitimacy of classification-based studies remain active, in part, because the phenotypic characters that drive population-specific patterns, such as those isolated above by Wright, have long served as the basis for the classic definition of race. In the larger biological community, race continues, for many, to be synonymous with the archaic "race concept." As such, it is often associated with the socio-political programs that defined the eugenics movement and enabled the abuse of human rights in the past, as well as the modern agendas that continue to reinforce long-practiced discrimination, and, ultimately, sanction the acts of violent persecution that are still ongoing in the present. Understandably, there exists, therefore, considerable anxiety over the continued use of race in Science. This fear of an intrinsic and inevitable scientific racism has led, in consequence, to a division in opinion among researchers across disciplines. While, for example, many biomedical scientists still argue for the utility of race as a valid proxy for ancestry or population structure in experiment design, and many clinicians routinely consider race when making decisions for medical treatment, equally as many practitioners and theoreticians alike dismiss race as biologically meaningless, lacking a legitimate evidentiary basis or objective reality. Researchers in human biology are left with the need to negotiate a place for race within scientific discourse on human variation and define its relationship to the similar

grouping concepts of ancestry, ethnicity, and population, which are themselves often confused and interchanged in common practice. Duster (2006) has succinctly stated the problem of race in recent history: "[t]he unenviable task has been to try to walk a tightrope – to figure out a way to effectively deploy in research the concept of race (or population groups designated by race) without endowing 'race' with a false sense of genetic essentialism" (p.3).

Within the field of physical anthropology, the critique of race has moved beyond an "assault on the race concept"(Brown & Armelagos, 2001, p. 34) to an inquiry into the validity of any study that uses morphological variation as a way to explore relationships among human groups and extract potentially valuable information on population history, contact and change, and individual identity. It is argued by many "mainstream" anthropologists that any biological information captured by contemporary concepts of either biosocial race and ethnicity or even biogeographic ancestry, are necessarily confounded by the promotion of disciplinary-specific ideologies, the imposition of personal beliefs and expectations, and the application of seemingly robust but, in actuality, flawed quantitative methodologies. Although there is a tacit recognition among most critics that "racial thinking is not necessarily synonymous with *racist* thinking"(Keita & Kittles, 1997, p. 534), there is an explicit emphasis placed upon the dangers that come with supporting a taxonomic or systematic approach to human population diversity; that is, adopting what Keita and Kittles (1997) have called the "racial" point of view. A perceived subjectivity in the analytical (visual and statistical estimation) procedures that have become standard protocol in skeletal biology has provided for many critics the arguments necessary, first, to call into question the validity and practical utility of craniofacial classification and, then, to use these methods of analysis as talking-points through which they can reiterate concerns over the misuse of race.

Traditionally, when assessing population relationships or when contributing to an individual identification, physical anthropologists have relied upon a tool-kit of estimation techniques, often including morphoscopic trait lists that delineate the skeletal attributes common to a particular ancestry or race, linear distance measurements and computed angles, and straightforward multivariate statistics for analysis. Although these methods have been shown to be both informative for studies of group discrimination and biodistance and profitable in their practical use in the field, their successful application is often heavily influenced by the experience level of the individual researcher. The quality of the results produced frequently depends upon the ability of the practitioner to properly employ the techniques of interest and "guide" the analysis towards the most appropriate outcome "by choosing the populations against which to classify the unknown" (Freid et al., 2005, np). Along with knowledge and practice also comes the need for exposure, as it is essential for the practitioner to have a working familiarity with skeletal samples that represent the groups that are not only relevant to the individual or assemblage of remains under evaluation, but also cover the breadth of possible variation in the features or forms that are key to the chosen methodology. More conventional approaches to pedagogy and limited engagement in professional development often permit situations in which techniques are learned and repeated based on preference and familiarity and without considering associated levels of method and observer error. Moreover, when these approaches are evaluated purely for their theoretical framework, a potential procedural problem arises from the fact that most of these traditional methods are based upon a presumed, direct correlation between the phenotype that is under evaluation and the genotype that is driving its expression, despite limited testing of this assumption and, often, an inadequate understanding of the consequences of its violation.

Given the mounting concerns over both the theoretical foundation of these methods and the questionable implementation of their techniques, a common view in contemporary anthropological discourse is that the results of any craniofacial classifications, whether they speak to patterns of population history or inform casework in human identification, represent nothing more than "forced allocations" (Keita, 2007). As population-informative traits must be typically prescreened for their "known" ability to signify membership, it has been said that "in practice, racial classifiers have to select the genetic traits and morphological features that support their preconceived notions of race"(Brown & Armelagos, 2001, p. 35). Many critics have contended, therefore, that there is, in reality, "an astounding mismatch between actual cranial variation and the variation modeled by racial constructs." They have concluded, in turn, that races are cultural artifacts because "the races we get are little more than a function of the trait or traits we use" (Armelagos & Van Gerven, 2003, p.61). These views have been further extended to the extreme case of their generalization, so that it is now commonly argued that there is no possibility for an unbiased explanation of human variation. For even the most rigorous of analyses involving either global ideas of human population [sub]structure or the more focused detection of among-group distinctiveness for the purpose of classification, "the problems in racial classification become insurmountable" (Brown & Armelagos, 2001, p. 34). Given these views, it should not be surprising that the most significant disagreement, fueling this "great race debate" (Gill, 2000), has emerged in the context of ancestry estimation and the use of race in forensic anthropology. In this regard, MacEachern (2000) has written that, "Anthropology has been involved in making racial distinctions on the basis of biological characteristics through much of its history. This is still an important pursuit in forensic anthropology at least, although a significant goal of this latter research appears to be the substantiation of law-enforcement folk taxonomies"(p. 359). This resistance to personal identification in medico-legal casework has held especially fast over the last decade, despite the fact

that the practical value of estimating a set of unique biological parameters (e.g., sex, geographic ancestry, age) for human remains identification and the utility of the controversial FORDISC software (Ousley & Jantz, 2005) has been clearly and consistently demonstrated in the contexts of medico-legal casework, mass disaster recovery, and international human rights projects. This lack of resolution between perception and practice in biological anthropology is still best captured by Sauer's (1992) now well-known statement on race and human morphological identification, in what has become a well-known forensic anthropological paradox: "if races don't exist, why are forensic anthropologists so good at identifying them?"

Although 20 years have already passed since Sauer (1992) first posed his question on race, the issue of calibrating a measure of personhood, that is based upon some predefined set of biosocial identifiers or proxies (such as ancestry, race, ethnicity), which are themselves dynamic in their social-evolutionary process, remains a topic of significant interest. Moreover, reaching some consensus or arriving at a best-practice solution continues to grow in importance in the wake of strengthened evidentiary standards and new, more sensitive, methods. The human biological sciences are charged, therefore, with formulating an active response to the rapid emergence of new technologies and the changing demands of research foci in the context of 21st century concerns. Any field or area of research that takes human variation as its topic must now engage the growing expectations for innovation in scholarship that have been placed upon current practitioners, who are operating in a digital, genomic, and computationally intensive age. An increasing number of anthropologists are coming to realize, therefore, that there is a great and immediate need to develop new techniques and make steady improvements to preexisting methods that use biological variables to interpret the patterned distribution of human variation and reliably estimate, with some known or computable level of accuracy, the demographic parameters that

contribute significantly to building an individual biological profile. The tension that exists between the anthropologists who use craniometric variation as a means to access past population structure in phenotypic distance studies and estimate individual ancestry in forensic contexts and those scholars, who dismiss any discussions of group differentiation as artifacts of personal ideological positions, does not discredit further study of population differences. Instead, it stresses the importance of reevaluating the question of identifying biologically meaningful patterns of human variation, using alternative techniques. For craniometric research, skeletal biologists should be adopting protocols that support best-practices when capturing differences between individuals and among subgroups and interpreting this observed variation. Consideration should be given to the choice of acquisition methods, the type of data that they yield, and the quality of fit with the information requirements of the particular question(s) at hand. The craniofacial data obtained should also be considered in relation to both global trends and specific populations-at-large, using robust statistical approaches that have high power, low error, and reduce the risk for bias, which may be introduced by requirements for the addition of *a priori* information. By isolating unique phenotypic traits that characterize *an* individual as *the* individual or as a member among *these* individuals, with a probability of association and some accepted measure of statistical significance, partitions of biological variation and individual classifications can be reliably obtained, classification overlaps and multi-membership assignments can be related to questions of interaction and admixture, and the biological meaning and practical value of these results can be critically assessed and externally validated. In turn, the conclusions obtained can carry the weight of good scientific practice. Population and personal identifications can and should be made, therefore, with the support of some accepted, methodologically transparent and interpretable standard of scientific certainty.

The project presented here is largely methodological in its emphasis, as it engages these theoretical and procedural concerns as they relate to the global distribution of human craniofacial variation and the estimation of individual identity in forensic casework. It is believed that there exist new and potentially useful methodologies that remain to be explored in craniometric studies for scientifically detecting the morphogenetic boundaries between populations and for determining if the nested hierarchical substructuring, that is critical to the assumption of neutrality for craniofacial morphology, can be revealed from phenotypic data alone, in the absence of comparative reference populations and *a priori* identifiers, like biosocial race or biogeographic ancestry. The robust model-based clustering methods of finite mixture analysis (McLachlan & Peel, 2000) have yet to be (1) adequately tested for their applicability to craniometric studies focused on detecting latent population [sub]structure and (2) used as a means of evaluating or validating the theoretical assumptions on group distinctiveness that underlie the conventional human identification methods, which are largely reliant upon supervised classification using metric variables. This project proposes that finite mixture methods can be profitably employed to address key questions of group membership using quantitative traits, and, specifically, craniometric data, and to resolve ongoing concerns over the role of population and personal identifiers in biological anthropology. The analyses implemented here and the discussions that either justify this dissertation's approach or contextualize its findings are structured around a suite of interrelated questions, which are taken to be points of interest that are fundamental to the overall goals of this research:

1. In the absence of predefined group-identifiers, is it possible to allocate "unknown" individuals with high probability into clusters in which all members share more common relationships (by biogeographic ancestry or biosocial race) than individuals assigned to other clusters?

2. Do descriptors, such as "geographic ancestry," "race," or "ethnicity," capture any of the phenotypic differences between these clusters?
3. Are these differences both statistically and biologically meaningful and, if so, useful in determining and defining the morphological variation that exists among globally distributed groups?
4. Can similar approaches be applied with success to questions of group variation, at the level of the American population-at-large, in forensic identification contexts?
5. Do model-based clustering methods offer a greater or different understanding of the patterns of biological distance between sub-populations than supervised clustering or classification methods, like traditional discriminant function analysis?
6. Can new methodological approaches resolve the common concerns over "forced allocations" and, in turn, alleviate the issues related to the scientific merit of estimating individual group-membership in forensic contexts and, in turn, meet the practical and evidentiary needs of the medico-legal community?
7. Does the power of differentiation using mixture modeling vary given the qualities of the data under analysis (metric traits, sample size, breadth of geographic sampling, missing data), the preprocessing decisions made (imputation, treatment for sex and time depth, and correction for morphological size), and the number and choice of variables used (all-available measures, the most statistically significant predictors or the most biologically meaningful and volumetrically representative subset)?

While the intention of this dissertation is to address these questions of methodology in detail, this goal cannot be fully achieved without also responding to the philosophical issues that cloud a full

understanding of the patterned distribution of modern human craniometric variation and its relationship to the personal, bicultural identifiers of race and ancestry. Together, these concerns over theory and quantitative method in physical anthropology as well as the challenges of merging the views of scientists and society have necessarily shaped this project's research design. For these reasons, a more holistic approach has been adopted here to address the debates over (1) the presence of latent, but detectable, population structure, (2) the reality of biological race, and (3) the utility of the popular metric identification methods that rely upon the statistical analysis of craniofacial traits using supervised classification techniques.

This project takes up race as its first topic of discussion. It takes on an extensive review of the history of the word and the nuanced and often ambiguous "concept." It considers as well the historical impact that race has had upon research trends in the biological sciences. Chapter 1 of this study attempts to provide a clear and consistent definition of race in current scholarship. On account of the many varied views on its use and the many degrees of its acceptance as an interpretable objective phenomenon, this chapter, ultimately looks backwards to the early beginnings of skeletal studies as a means to access the significance of race in contemporary thought and practice. The use of race and its evolution in meaning is subsequently traced throughout the developmental trajectory of the field of physical anthropology. This historical overview begins with a discussion of Boas' groundbreaking anthropometric research into the effect of environmental change on population-specific signatures and the influence of his work on the utility of craniofacial shape in the reconstruction of population history. Similarly, it considers the important roles of Hrdlička and Hooton in the formulation of what has become known as the "typological tradition" in physical anthropology, the establishment skeletal collections for comparative studies, and in the codification of the trait "lists" that would influence the classification criteria used in

the morphoscopic identification of races, populations and ancestry groups throughout much of the 20th century. This review also surveys the socio-political milieus out of which the anthropological ideas of biological race emerged and under whose influence the views on the "origins" of the races, biological determinism, and human diversity grew and changed. This discussion ends with the confluence of several critical events in anthropological theory, all of which took inspiration from the agents of the Modern Synthesis and the developing field of population genetics. These include the call for the "new" physical anthropology by Washburn, the promotion of clinal variation as conceptualized by Livingston, the introduction of populationist thinking as supported by Howells, and, perhaps, the most influential of these events, the advent of modern molecular anthropology and the subsequent rejection of race that Lewontin's work on global human genetic diversity inspired. As this revolutionary era of genomic anthropology marks a breaking point in how race is perceived and human variation is studied, the changing trends in biological anthropology over the last 40 years are treated separately, in Chapter 2, as a new historical period, which is dominated largely by the critical debate over the quantitative assessment of human variation.

Since the rejection of race first gained its cultural and scholarly capital in the early 1970s, studies of human variation have been diverging along two, largely diametrical, lines; each capturing the opposing views on the reality of human biological differences, their relationship to group identifiers, and the practical utility of the contemporary interpretations of race and ancestry. In order to understand how this current project's finite mixture approach differs from the standard discriminant function practices and to anticipate how it may come to resolve many of the common concerns for quantitative approaches to estimating affinity and identity by revealing cryptic population structure and implementing model-based classification, it is necessary to also review the trajectory of contemporary

thought on race. This project identifies the strengths and weaknesses in the methods used to detect among and within group distinctiveness. It presents the prevailing opinions on group membership results produced by the classification-based studies that are founded upon perceived patterns of human morphology at both the global and local levels of variation. It evaluates the evidence that has fueled the debate over the utility of the FORDISC software for the estimation of ancestry in modern American casework. Finally, it also explains the tenuous position negotiated by the relatively new and growing subfield of forensic anthropology within the larger discipline of anthropology, as it straddles the divide between a rigorously theoretical science and a field focused on practical application with immediate medico-legal concerns and critical human rights implications, both of which demand the use of race. To this last point, this project discusses in detail the current trends in interdisciplinary and, especially, molecular research and it calls attention to the studies that are using cluster analyses to renew discussions on "what's the use of race?" (Whitmarsh & Jones, 2010) in what is now the "genomic age" (Koenig, Lee & Richardson, 2008). The potential that these new genetic clustering approaches hold for the improved study of human variation in biological anthropology is stressed, and special emphasis is placed upon the effect that these methods may have on the ways in which forensic anthropology is both practiced and received by the anthropological and larger human biological communities. Moreover, the methodological similarities between these new mixture-based techniques, as they are applied to genetic data and most often realized within the Bayesian clustering framework of the STRUCTURE software (Pritchard, Stephens & Donnelly, 2000), and the model-based clustering methods advocated here are highlighted. Finite mixture analysis is also compared to the more traditional distance and classification-based procedures used to quantify cranial variation and aid in human identification.

In order to appreciate the statistical framework that underlies these new model-based clustering approaches to molecular data, as well as the form of these techniques implemented in this project, it is essential to also engage in both a historical discussion and a mathematical review of the methodological basis for finite mixture methods. Chapter 3 begins the discussion of mixture modeling by, first, contextualizing its unique model-based approach within the larger body of work on pattern recognition: addressing clustering and classification as a logical analytic choice from a human behavioral perspective, tracing the rapid growth in popularity of cluster analysis in recent years and emphasizing the prominent position that it has gained within the current compendium of science literature, and explaining the statistical qualities of the unsupervised, computationally intensive, clustering methods that especially recommend their application to biological data and their use for questions of affinity and identity. A brief history of statistics, that focuses on the key developments in probability theory, is given in order to trace the emergence of finite mixture methods within the context of Pearson's early "method of moments," its mid-century evolution in light of maximum likelihood estimation, the realization of its full potential after the critical introduction of the Expectation-Maximization algorithm by Dempster, Laird and Rubin in 1977, and the changes that it has undergone more recently, given the advances in computational power, which have enabled the creation of efficient and widely accessible statistical software. The statistical theory and the assumptions that serve as the foundation for the generalized form of these finite mixture techniques are discussed in detail. The popular iterations of model-based clustering methods are also presented, as they are integrated within several free-access or open source software packages, including EMMIX (Peel & McLachlan, 1999; McLachlan et al., 2011), MCLUST (Fraley & Raftery, 2011) and the statistical computing environment R (R Development Core Team, 2010), in order to explain the practical implementation of the concepts that are fundamental to this project's craniometric analyses and to offer best-practice recommendations for the expanded use of these

mixture-clustering methods in biological anthropology. A test-case, using a data set of well-documented modern human statures (Meadows Jantz, 2008) is presented as an example of the steps taken in implementing the model-based clustering procedures and the analytic potential that this finite mixture approach offers for skeletometric data.

In Chapter 4, these finite mixture methods are realized in a multi-part statistical study. The analyses conducted here seek to, simultaneously, (1) demonstrate the utility of model-based clustering as a robust statistical technique for craniometric research and suggest its potential generalization for other population-based questions that are of interest to biological anthropologists, (2) test for the presence of latent, but statistically, biologically and anthropologically meaningful, population structure across a variety of craniometric data sets, (3) address the use of probabilistic classification as it relates to world-wide partitioning of human variation and the identification of individual membership, (4) resolve questions of best-practice methodology by providing differently derived evidence on the effective "blind" identification of population-specific differences and intergroup relationships using craniofacial traits, and, ultimately, (5) speak to the information value of population designations, biogeographic ancestry and biosocial race in contemporary anthropological discourse and forensic practice. This project makes use of three well-documented sources of human craniofacial measurements: Howells' (1996) 28 population and TEST craniometric series, Hanihara's (2008) broadly sampled, global craniometric data set, and a modern American craniometric sample extracted from the University of Tennessee's Forensic Anthropology Data Bank (Jantz, 2008). These widely-known and reliably used resources collectively permit large-scale among-group analyses or tests of "population structure" at a variety of partitions (e.g., at continental, regional, local levels, through time and between the sexes) and to differing investigative goals (i.e., revealing globally distributed patterns or producing forensic

classifications) using mixture-based clustering. These data sets are also combined with demographic information obtained from the United States Census Bureau (2010, 2011) as a means to contextualize these craniometric samples within the range of racial diversity represented in the American population-at-large.

The methods of finite mixture analysis implemented here are presented as promising alternatives to the supervised clustering techniques, and especially the standard discriminant function analyses, that have been traditionally employed in population studies using craniometric traits. This project also investigates other complementary techniques that require, to differing degrees, the input of prior information. The newly developed force-directed graphing techniques (Aurenhammer, 1991; Di Battista et al., 1999; Kauffmann & Wagner, 2001) and finite mixture-based (partial and fully) supervised classification procedures are specifically considered here (McLachlan, 1992; McLachlan & Peel, 2000; Fraley & Raftery, 2002). These methods are implemented as means to further explore the questions of population [sub]structuring and the (biosocial) reality of race using both approaches and data types that are familiar to biological anthropologists (discriminant function analysis and distance measures) but which are applied in previously untested ways. They offer, therefore, the opportunity to investigate this project's tests data set further under objective validation criteria and to introduce additional, potentially more powerful, techniques for user-guided analyses.

While the discussions of the materials, methods and analyses, that together make up Chapter 4, are annotated with explanations of the test-specific results produced and commentary on their significance, Chapter 5 provides a summary account of this project's findings. It is argued that these findings support (1) the presence of clearly detectable population structure for all three test data sets, (2) the

identification of a meaningful relationship between statistically inferred biological differences and the socially defined concepts of populations, geographic ancestries, and races, (3) the utility of finite mixture methods for cluster analysis projects and questions of biological classification and its ability to resolve many of the concerns commonly expressed in the context of the standard statistical discrimination techniques, and (4) the validity of employing supervised or user-directed methods as investigative and confirmation tools, when they are appropriately applied and when their limitations are acknowledged. The cluster and classification solutions are further presented in terms of the external historical and molecular evidence on the patterns of population [inter and intra] relationships as revealed by contemporary genetic data, and consideration is given to the dominant theories of modern human evolution, dispersal, and divergence. The practical value of these findings as they relate to the concepts of biogeographic ancestry and modern American biosocial identifiers is also discussed. Special emphasis is given to the appropriate place of race in craniometric studies and, especially, to the validity of its use in forensic anthropological discourse. This project concludes that race is meaningful: in practice, reporting race can make important contributions to the initial estimation stages of medico-legal casework and, in theory, understanding the relationship between race and human variation is critical to the interpretation of human biological data.

Chapter 1: Race

What's The Use of Race?

"A popular statement now is 'There is no such thing as race.' I wonder what people are going to think when they hear this. They would have to suppose that the speaker, if he were dropped by parachute into downtown Nairobi, would be unable to tell, by looking around him, whether he is in Nairobi or Stockholm. The visible difference between different populations of the world tells everyone that there is something there. We had better be prepared to explain what is there, and why, before we discuss what it does or does not mean." (Brues, 2009, p. 78)

The ambiguity in this "something," which is all at once omnipresent, inherently identifiable, yet inadequately explicable, well represents the problem of the race concept in current anthropological discourse. Perhaps more than any other concept in physical anthropology, race has held fast to its influential position in the growth and continued development of the discipline. It has not only remained resilient to the changing trends in social thought, but it has also retained global cultural currency through its diverse reception within popular culture and scholarly circles. Despite the many, often strident, claims that "the biological conception of race is philosophically and scientifically dead" (James, 2008, np), it is "biologically meaningless" (Schwartz, 2001, p. 1392), and that the reality of race is but a common sense 'truth' (Diamond, 1994, p.82), there remains an interest in race. It is *because* these opinions on the non-existence of race are repeatedly voiced that discussions of race continue. When any statement is made for or against the concept, a voice on race is heard. At the core of this scholarly debate lie arguments made in response to concerns about the utility of race in contemporary thought,

method, and theory. What is the social significance of race? Can arguments be successfully made for a "feedback loop between cultural and biological factors" (Edgar, 2009, p. 59-60), and, if so, can they ratify the biological study of race? Does this "racial thinking" or "racial point of view" (Keita & Kittles 1997) tacitly imply racism or actively support its pseudoscientific misuse? Does Hirschfeld's argument (1996), that the collective knowledge of race is the product of a special-purpose cognitive competence for understanding and representing human groups, differ substantially from the "concept of race"? Is it possible to also find genetic or morphological meaning in race and can this meaning be useful to anthropologists engaged in scientific discourse? What can race contribute to global studies of human biological variation, and more specifically to the pedagogy and practice of individual identification in medico-legal contexts?

It is perhaps worthwhile here to restate the question that both Sauer (1992) and Kennedy (1995) have posed in the context of this debate: what contribution does race make to forensic anthropological theory and method if, in reality, it does not exist? In thinking through the questions above, it becomes apparent that they have yet to receive a fully satisfactory answer. Even the most basic questions on the "meaning"of race continue to go unresolved, leading to further concerns over the methods and meanings of identification. What are the qualities, traits, characteristics or contexts that guide perceptions of difference and similarity? Why and how are they influential? Do they even exist outside of the mind of the analyst? Can they be isolated, quantified, and validated scientifically? If these factors, and, similarly, the perceptions of race that they generate, do not exist, what alternatives remain for describing and analyzing variation within the human species? As Lieberman and Kirk (2004) have asked "Now that you have taken race away from us what will fill the void?" (p. 139).

The enormity of these questions, both in their number and in the breadth of their implications, demonstrates how the reach of the race concept extends far beyond the purview of past and present anthropological perspectives. Race continues to actively exert its influence upon any field that takes up the study of human variation. It is a lasting phenomenon because it occupies a liminal space between a mythology and an uncomfortable truth. The effect of this enduring quality of race on anthropology has already been observed by Armelagos and Goodman (1998), who wrote: "With remarkable resilience, race, which emerged as a typological non-evolutionary folk taxonomy, became a scientifically recognized unit of evolutionary change" (p. 359). It is perhaps on account of this uneasy negotiation between conventional wisdom and scientific evidence that such a ubiquitous, loaded, and potentially powerful term as race has been called upon time and again as a key explanatory tool for describing human variation. Yet, for the same reason, it has simultaneously resisted any consensus towards a consistently used and universally accepted definition (Brace, 1982; Molnar, 1975). Race is instead given meaning only through its interpretation as a "cultural construct" (Smedley, 2007). This is problematic because, as Smedley (2007) has cautioned, "this is not a definition or even a characterization of 'race,' but an assertion about the scholarly or existential domain in which we can best examine and explain the phenomenon of race" (p. 1). Any change in "domain" can, therefore, produce a different perspective on race, so that its use, meaning and reception are always already in a state of flux and its significance is, in turn, largely context-specific.

As the theoretical and practical implications of race are both complex and profound, any resolution to its questionable role in scientific discourse must begin by first establishing the race paradigm. The intention here is not to reconcile what Gill (2000, np) has called "the great race debate" but, instead, to contextualize it. To this end, the discussion that follows attends to the meaning(s) of race, by clarifying

its uses in the study of human biological diversity with special attention to the historical and philosophical factors that have contributed to its complex relationship with population-based research in the field of biological anthropology. Only in coming to an understanding of the role race has played in the field of anthropology over the course of its disciplinary development is it possible to fully appreciate how this long and difficult history has shaped contemporary thinking on human variation. In turn, such an understanding will aid in uncovering the reasons that underlie its continued influence on the reception of quantitative studies that take up the concepts of biosocial race and biogeographic ancestry in discussions of group-specific morphologies, population structure, and individual forensic identifications.

Towards a Definition of Race

"Human plurality - bodily, culturally, physiologically - has long been a seductive object of study going back to classical history." (Whitmarsh & Jones, 2010, p. 3)

When Livingstone (1962) first made his seminal arguments for the [re]interpretation of human "biological variability" in terms of "clines" against the identification of "discrete packages labeled races" (p. 279), he framed his discussion within the general context of biology. If race is approached now, through the same lens, it can be shown that, for a concept lacking in a scientific basis and for which criticisms have been long standing, it continues to generate substantial scholarly activity (Futuyma, 1998). Although Livingstone observed already in 1962 that a "growing minority of biologists in general are advocating a similar position with regard to such diverse organisms as gackle, martens and

butterflies" (p. 279), recent queries of research databases covering the most current literature in the life sciences present a much hazier picture on the utility and reality of race than might be expected for scholarship today, given Livingstone's observations. A simple search of the BIOSIS Previews database (BIOSIS, 2010), with "race"in the search-field "topic,"returned over 15,000 publication records for the last 5 years alone (January 2005-August 2010). Almost 50 years after Livingstone's pronouncement, these results suggest that race continued to remain, for discussions of among-group differences, "overworked as an explanation of this variability" (Livingstone, 1962, p. 279). When the query parameters are further refined to produce a subset of examples that are arbitrarily drawn from all available hits but constrained to the same year of publication (in this case, 2001), the results returned from a second BIOSIS search reveal that the percentage of articles devoted to "race"remains consistent. This suggests that race is not only omnipresent in current scientific discourse but also sees regular use as a widely accepted explanatory concept. Despite the sampling constraints imposed upon the second query, race still clearly appears to cut across the boundaries of many fields in biology. Furthermore, in expanded searches that considered race alongside other standard scientific concepts that could be equally used in the explanation of "difference," race was consistently flagged as the term having the highest frequency of use out of the full series of keywords considered. What is perhaps even more interesting is the way in which race is used in these publications. Despite its omnipresence, the "race concept" appears to defy a standard usage or a consistent definition. This irregularity in meaning even persists when "race" is considered without the added complication of the prefix, "human." For example, Rehfeldt and Gallo (2001) treat race in terms of its cladistic and ecological components in their discussion of the provenience of Douglas fir trees; Vicente and co-workers (2001) consider race in the context of ecological adaptations for the pathovar, *Xanthomonas campestris;* Jiggins et al. (2001) link race to speciation in their study of the emergence of two sister butterfly populations, *Heliconius*

melpomene and *cydno*; while Stone and colleagues (2001) consider two ecotypes of marble gallwasps, *Andricus kollari,* as reasonable examples of races even though they do not represent independent lineages and are known to hybridize. The meaning of race is further muddied by circularity, and, so, confusion, when the goal is a standard definition of the term. Consider here the common ties that bind and intertwine the meanings provided for race, subspecies and ecotype in the most recent (7th) edition of the *Dictionary of Genetics* (King, Stansfield & Mulligan, 2006). Race is defined as "A phenotypically and/or geographically distinctive subspecific group, composed of individuals inhabiting a defined geographical and/or ecological region, and possessing characteristic phenotypic and gene frequencies that distinguish it from other such groups. The number of racial groups that one wishes to recognize within a species is usually arbitrary but suitable for the purposes under investigation." A subspecies is "1. A taxonomically recognized subdivision of a species; 2. Geographically and/or ecologically defined subdivisions of a species with distinctive characteristics." Finally, an ecotype is considered a "Race (within a species) genetically adapted to a certain environment."

From the flexibility and interdependence of these terms and their uses, it becomes apparent that a race, in non-human biology at least, can be identified or delimited from "others" using various criteria, many of which intersect, correlate or even overlap. Possibly the most important of these relationships is the frequent conflation of race with subspecies, as this may introduce confusion when distinctions are made between groups based upon breeding population rules (see Keita et al. [2004] who address this issue for conceptualizing human variation). The application of race, nevertheless, seems to remain comfortably undefined, allowing for its criteria for meaning or interpretation to be a function of the individual research subjects, goals and contexts.

In the case of human biology, the definition of race is further complicated, and potentially obscured, by the interplay between public perception and scientific methodology. This becomes particularly important for biomedical research, clinical practice and ultimately the state of public health. The risks involved have been frequently stated: a racial determinism persists when genetic studies are brought into contact with patient populations and race is used uncritically to inform the creation and implementation of health care plans (Braun, 2006; Frank, 2007). While even a conservative estimate suggests that the word "race"appears in 80% of recent health-science articles (from reviewing recent biomedical literature in subjects as diverse as epidemiology, genetics, nursing, and health services), the meaning of the word is rarely defined, its sense as a larger concept is poorly (if ever) stated and the research-specific intent that underlies its use is frequently left unexplained (Anderson & Moscou, 1998; Drevdahl et al., 2001; Comstock et al., 2004; Sankar et al., 2007; Gravlee & Sweet, 2008). Coming to terms with race and the associated race concept is equally, if not especially, challenging in anthropology, given its four-field approach and the implicit position of race within these areas of study. Here, two factors contribute significantly to the race problem in anthropology: (1) what *explicitly* do anthropologists mean by race and (2) can any meaning be reliably applied to human infraspecific , or among-group, variation (Keita et al., 2004; Keita & Boyce 2001). For both anthropologists as a whole, as well as for individual subspecialists, a distinction must be negotiated between race in the social context (i.e., as a means of self-identification that is essential to the production of culture, the perpetuation of collective identity and the support of individual group memberships) and in the context of human biology at the level of *Homo sapiens sapiens*. As current research is frequently located at the crossroads of disciplines outside of the traditional focus of physical anthropology (including statistics, biomedical sciences, legal studies and political science, demography, sociology and history), it has become essential to understand the nuanced meaning of race in these often disparate areas of study whose scholarly

foundations, trends in research, and collective disciplinary goals can be very different. Sensitivity must also be shown towards the role of race within the unique history of the American experience: as a "genetic melting-pot," a country founded upon the backs of immigrants, with a complex history of slavery and segregation and recurrent events of migration and mixture (Crow, 2002; Feldman, Lewontin & King, 2003).

An awareness of the diverse and frequently discordant interpretations of race as well as its broader, lay, significance within the global social and scholarly communities is also necessary for any thoughtful appreciation of the meaning of race when physical and cultural anthropology intersect. Take, for example, the complexity of identity for the human identification work performed in the context of war, genocide, and post-colonialism. Specialists in the various sub-fields of physical anthropology must also contend with the sub-disciplinary fissuring among those who self-identify as, e.g., skeletal biologists, bioarchaeologists, and/or physical or forensic anthropologists. Such scholars can and often do hold memberships in multiple academic circles, and, as a result, each presents a mixture of many theoretical and philosophical views that have been shaped by these interactions. The meaning of "race,"the race concept, or a racial category and any current opinions on if and how these formulations can be used are all inevitably variable, even among North American anthropologists. Keita and colleagues (2004) rightly state that race "engenders much discussion, with little agreement between those who claim that 'races' are real (meaning natural) biological entities and those who maintain that they are socially constructed" (p. S17.) This discordance in opinion can be related, in part, to the problem of semantics. Race is either not defined at all or it is defined inconsistently with its referents shifting given the context. Moreover, it is often confused with colloquial uses or its "lay concept" which "postulates the existence of near uniform groups of individuals that can be identified by a few externally visible traits such as skin color"

(Long & Kittles, p. 468; Keita & Kittles, 1997). Race may also refer to a wide range of arbitrarily defined and malleable human groupings: from social, religious, political, linguistic and geographical groups, to anatomical races based upon phenotype or some "biological" proxy for genotype. These classifications are often equally rooted in ideas of appearance, ability and behavior, but they are rarely supported by evidence of genetic transmission and often show patterns that are inconsistent with expectations for heritability (AAA, 1998; AAPA, 1996, 2009). Because of the variability in its criteria, the definition of race can be easily changed so that it becomes evidence of an evolving social-state. Brodkin (1999), for example, has discussed the transition in the concept of "Jew" from the rigid ethnic classification to a socio-religious identity, as Jewish, within the broader category of White. Soong (1999) has demonstrated the diversity and fluidity in [self]identification and racial memberships with comprehensive lists of individual options for classification in Latin America. Similarly, Parra and coworkers (2003) have further articulated this relationship among race, color and genetic inheritance. By looking at Brazil specifically, Parra et al. (2003) show that race is equivalent to color, that ancestry is not a direct criterion for color, and that color offers little information on genetic heritage. They note that color is identified along a range, with consideration given to skin pigmentation, hair, and facial structure, so that individuals with a similar genetic make-up (e.g., siblings) may not share these same traits or be classified within the same group. This approach to race is likely unfamiliar and possibly unsatisfying to a modern American public, to whom ideas of ethnicity, ancestry or heritage can be both a source of pride and a justification for repression. The current mass-marketing of "family"stands as a testament to contemporary trends in the collective concern for self-identity within a communal American history through (1) the growing number of "family tree" search tools (e.g., the LDS Church's FamilySearch Internet Genealogy Service which provides searchable, global access to the International Genealogical Index, the entire Ancestral File, the SourceGuide, and the Family History Library Catalog to millions of

users), (2) the popularity of online family networking services that integrate with social media sites that "*help you find and stay connected to your relatives*" (e.g., the 60 million registered users of the FamilyLink application for Facebook and Twitter), and (3) the celebrity origins spin on reality television (e.g., the "Who Do You Think You Are?" series, which premiered on NBC on March 5, 2010). A study conducted in 1995 found that, even before the widespread access to the internet in American homes and the popularization of web-based genealogy sites (such as the multi-national for-pay search engine, Ancestry.com), some 113 million adults in the United States (that is, four out of ten individuals) claimed to be at least somewhat interested in their family history (Howells, 1999). A more lasting example may be found in the "one-drop-rule," which implies a quasi-permanency in an "inherited" bio-social assignment. The rule dictates an individual racial identity based upon traceable minority heritage and the governing system of social hierarchy (Hickman, 1997; Frederickson, 2002). This is a rule that even the President of the United States seems obliged to follow. Barack Obama has said of himself that "I'm an African American, but I am somebody, like many African Americans, who has all kinds of stuff in him...You should have seen Thanksgiving, we were like the United Nations...But I self-identify as an African American. That's how I am treated and that's how I am viewed and I'm proud of it." (Bacon, 2007, np).

From these examples, it becomes possible to see how the concept of race can be easily broadened to represent a worldview, so that race may be understood as "a culturally structured, systematic way of looking at, perceiving, and interpreting" reality (Smedley, 2007, p. 18). It is this latter type of explanation that has fueled arguments against a biological basis for population differences even while it has led to the codification of a radically diametric view of race as a willful social construct. This can be observed, for example, in the "fact" published in *The Stanford Encyclopedia of Philosophy* (under the entry on

"Race") that "humans themselves choose (consciously or unconsciously) which physical characteristics constitute a racial group." The collective cultural experience of inequality, injustice and racism in the forms of segregation, apartheid, slavery, ethnic cleansing and genocide, hate speech and hate crimes, and the continued need for affirmative action should make clear that the cultural definitions of race are just as problematic as any biologically-based explanation. Not all humans have the choice in membership that the statement above would imply. Denying a definition of race in anthropology, therefore, does not *ipso facto* signal the end of racism (Mukhopadhyay & Moses 1997; Lieberman & Kirk 2004). Smedley's (2007) "worldview" of race, however, also allows scholars such as Gravlee (2009) to focus on its "central tenet" that humans can be naturally grouped into a few biological subdivisions that are called races, and that these races are presumed to be unchanging, well-delimited, and homogenous in their make-up. Definitions based on racial theories, like this tenet, also ultimately imply that the phenotypes or characteristics used to distinguish one race from another reflect deeply rooted or innate biological differences (Keita & Kittles, 1997; Banton, 1998). Although this meaning of race can be (and is typically) criticized for its essentialist and out-dated perspective on human variation, Gravlee argues for a more nuanced, or a shift in, the critique of biological race that does not include its complete dismissal. Following this worldview approach, he (2009) states that, in fact, "this definition should not be taken to mean that race is merely a bad idea" (p. 47).

The above are but a few specific examples of the complications of interpreting preexisting, contemporary, and evolving definitions of race and the difficulties in making sense of a concept that is all at once ideologically charged, often abused or poorly used, and inscribed with meanings that are both arbitrary and obtuse. Moreover, this problem is not limited to scholarly discourse, but also informs the interactions between public governance, individual identity and anthropological research. Although

the United States Office of Management and Budget's (OMB) classification system has become the standard for medical research, its terminology has changed with social evolution through time (U.S. Census Bureau, 2010). Since 1900, twenty-six different racial terms have been used to identify people by populations in the U.S. Census (AAA, 1997). For the U.S. Census 2010, there are 14 minimum designations for race. An optional 15th, form-fillable category is also included for the identification of any unlisted "other" race. Furthermore, respondents are encouraged to indicate affiliations with multiple races by selecting one or more boxes. Despite the considerable value of census data, the analytic utility of these categories (that is, their "scientific" relevance or significance) to demographic studies outside of the traditional census goals remains poorly addressed. Directive 15 of the OMB (1997) states only that "...the racial and ethnic categories set forth in the standards should be developed using appropriate scientific methodologies, including the social sciences" and "the racial and ethnic categories set forth in the standards should not be interpreted as being primarily biological or genetic in reference" (np). Moreover, the distinction between social and biological meaning is further confused by the OMB's qualifying statement, that "race and ethnicity may be thought of in terms of social and cultural characteristics as well as ancestry," since ancestry in anthropological interpretation does have a biological (evolutionary-geographic) basis. It is interesting to note that, although the Centers for Disease Control and Prevention (CDC) make practical use of the medical research findings whose clinical studies typically adhere to the OMB identification system, the CDC follows a different, more comprehensive coding schema. The CDC itself reports a seemingly inconsistent opinion, stating that this very practice of using race as shorthand for genetic variation in medical research is problematic (Hahn, 1992; Hahn et al., 1992). These conflicting views are further born out in the CDC's (1993) statement on race in MMWR #42: RR-10: "because most associations between disease and race have no biological basis, race - as a biological concept - is not useful in public health surveillance" (p. 12). This position is later tempered by

a statement of support for collecting and maintaining race data as a means to improve public health. Discrepancy is, therefore, present even within the same governmental system and among federal bodies that are presumably interactive.

Racial categories employed in the closely related fields of biohistorical, forensic and genetic research similarly oscillate between a "preferred" terminology and implied meanings. Language use is heavily dependent upon disciplinary conventions, the personal viewpoint of the practitioner, and the reporting context, thus yielding several variations that need not be mutually exclusive: e.g., Black versus African[-American], Caucasian versus White versus European, Hispanic versus Latino versus [Non-]White. This leaves open the possibility of considerable misinterpretation. For example, Brace has noted (1995) that "Africa of course entails 'black,' but 'black' does not entail African" (p. 72). It is especially telling that he offers this criticism of identification by way of a statement that is itself confused: the geographical continent of Africa does not always entail black. The vocabulary of classification for large scale diversity also shows variability but, in this case, it is better understood as modifications to the popular terminology over time. The typological categories of Caucasoid, Negroid, and Mongoloid have been replaced, for the most part, by the geographic terms of European, African, and Asian. Yet even in the updated editions of the central manual for human osteology, Bass (2005) continues to retain the older typological terminology. Despite these holdovers, early typologies are most commonly cited now only for their value as anachronisms of thought in the earlier 20th century. These changes likely signal a response to the historical patterns of acceptance and rejection of race determination, the shift away from any association with past practices of human "taxonomy," and the reconsideration of group differences in terms of probabilistic-based membership estimation using the framework of populations and ancestry (c.f., Hefner [2009] contra Rhine [1990] on cranial non-metric variation). Confusion is

nevertheless created here when concepts such as ancestry or bio-distance are treated as or criticized for being mere recastings of race under the guise of descriptive statistics and in the service of old questions of racial history, migration and diffusion. This is what Armelagos and Van Gerven (2003) have described as "new wine in old bottles" (p. 57). Similarly, Albanese and Saunders (2006) have pointed out that while France (1988; France & Horn, 2001) has updated the terminology in the subsequent edition of the *Lab Manual and Workbook for Physical Anthropology* by using "ancestry"in the place of "race"(compare p. 30 of the 1st edition and p. 123 of the 4th edition), there is "no change in the underlying concepts, so that determining continental origin has been substituted for color terminology" (p.284).

Negative attitudes towards current efforts in human identification in forensic casework and the fear of an inherent racism within any race-based thinking has led to even further problems of interpretation. Race is too often represented by a variety of pseudonyms which are both rarely defined and erratically used. The subjective inference of authorial intent is required in order to access the real meaning of these variants, to the extent to which such intent can be inferred or relied upon. To interpret Race, RACE, "race", or *race,* the reader must either rely upon inflecting meaning through the textual context of the word or fall-back upon assumed, and potentially faulty, prior knowledge of the personal, ideological or research position of the writer. Even when decoded, the meaning of these variants remains problematic as they are commonly employed in an uncritical routine, driven by some presumed acceptable protocol and not by a true understanding of the implications that come with their use. These problems can even cast doubt on the author's credibility. This is the case when otherwise sound studies on population differences are veiled in an apologetic approach to the study of biologically-defined group differences instead of offering a clear statement on the authors' position *vis à vis* race, ancestry, and the evidence for human population variation as manifested in their data and revealed by their analysis. The

opening and closing statements from Albanese and Saunders (2006) clearly demonstrate the dangers of this approach: "The determination of ancestry population affinities or ethnicity (in the past, referred to as "race") is the most controversial question that a forensic anthropologist must face when assisting in identifying unknown individuals" (p. 282) and "false or misleading information is far worse than lack of information" (p. 309). Although their chapter poses the question, "Is It Possible to Escape Racial Typology in Forensic Identification?," they themselves fall prey to the criticisms implicit within their own arguments. The contradiction here should be clear: they engage in a critical discussion of the use of skeletal estimation methods for individual population-based identification in forensic casework but make only a temporal distinction between race, ancestry and ethnicity (leading to a conflation in terminology). Moreover, they fail to define their meaning of "race" at all. A move to population names only compounds this problem, as can be seen in Feldesman and Fountain's (1996, p.207) publication on stature: "we used a revised data set and a more elaborate research protocol to address questions pertaining to: (1) whether the femur/stature ratios of three quasi-geographic 'races' ('Blacks,' 'Whites,' 'Asians') are statistically significantly different; (2) whether these are statistically (as opposed to biologically) coherent groups; and (3) whether the 'race'-specific ratios are more accurate than the simple generic femur/stature ratio." They also apologetically state that "we recognize [...] that this grouping was crude and did not capture the more substantial store of geographic variation characteristic of modern humans" (1996, p. 211). Although Lieberman et al. (2003) raise the valid point that vocabulary choices and theoretical caveats such as these "suggests that the authors, while accepting the race concept, are aware of the debate over race typologies" and that "they acknowledge the significance of variation in populations" (p.111), it remains highly uncertain whether this type of awareness in any way contributes to resolving the issue of population-specific variation in physical anthropology.

At the other extreme, inconsistency in terminology leaves open the possibility of creating categories that are often hastily defined and whose relationship to biology and society are inadequately supported or explained. This can be seen in a statement made by Rhine (1990) in his contribution to the *Skeletal Attribution of Race*, which was, at one time, a key reference publication on making any such determination: "While the bulk of the sample is Caucasoid, that group can be further subdivided into Anglo (those of predominantly European origin) and Hispanic, a 'neo-race' […] Originally used as a vague 'social race' term, Hispanic is now used in this context as a biological category..." (p. 13). Here, Rhine fails to explain the meaning of these groups and the basis for his subdivisions (their origins and how they received these designations, e.g., self-reported or by phenotypic identification). Also absent are both his justification for asserting that the Hispanic category has transitioned from social to biological race and his definition for a "neo-race" in general and as it relates specifically to the identification of Hispanics. The reader is understandably left confused over what objective criteria are used to define a race. Is Caucasoid the same type or level of race as its subdivision, Anglo? Can races be nested? Where and how can the line be drawn between social and biological race? How can it be known that the designation "Hispanic" has crossed it and, if so, when this transition took place? Are all cases of admixture on their way to becoming, or are already, a part of a neo-race? Does this race-making imply the acceptance of racial purity to some degree? At what point, then, are populations too mixed to be partitioned? It seems that Rhine's work, one of the few classic references for isolating human population variability in the skeleton, offers only a tutorial for "practice" (i.e., instruction on what predefined features should be observed, and how to make assignments using these morphoscopic observations). This offers little help for assessing the utility of these expressed variations and fails to provide a broader understanding of trait patterning. Rhine has eschewed any discussion of how these attributions of race

fit within the larger distribution of human diversity, so that the definition of race goes unstated. Any question of meaning is resolved by presenting the reader with an ambiguous *fait accompli,* and this covert acceptance of race aligns closely with earlier "essentialist" views on human variation. Rhine's work needs only to be considered in light of the definition of race given by Hooton (1926): " A great division of mankind, characterized as a group by the sharing of a certain combination of features, which have been derived from their common descent, and constitute a vague physical background, usually more or less obscured by individual variations, and realized best in a composite picture" (p. 312).

In the face of these difficulties in language and substance, and the ongoing debate over meaning and utility, there are those who feel that there are sufficient common threads among the many explanations of race and ancestry that enable a widely-acceptable and reproducible definition for population-specific variation that is both relevant to the human sciences and useful for the questions of interest to physical anthropology. Among many of those scholars who are deeply rooted in biological theory, the race debate is a *non sequitur* in logic: Crow and Mayr both independently argue that the (correct) definition of race is (or at least should be) common knowledge, even among anthropologists who also deal with the social aspects of the human experience. This is because, as Crow (2002) says, "the concept itself is unambiguous, and I believe that the word has a clear meaning to most people" (p. 85). Mayr (2002) makes an even stronger statement against the non-existence of human races and dismisses the rejection of race in scientific discourse: "Those who subscribe to this opinion are obviously ignorant of modern biology...Recognizing races is only recognizing a biological fact" (p. 90-91). A similar presumed synchrony in contemporary meaning is also assumed by Relethford (2009), who writes that the "most current uses of the term are generally geographically based, such as the idea of geographic race articulated by Garn (1965), which generally subdivides humanity into continental groupings (e.g.,

Africans, Asians, Europeans) or subcontinental groupings (e.g., East Asians, Southeast Asians)" (p. 16). He also maintains that "most classic definitions of race focus on phenotypic traits, such as skin color, craniofacial shape, and hair color and texture, among other features" (p. 16). Keita and colleagues (2004, p. S17) stand out, however, against this "current use" and offer up what they believe to be the common interpretation. They argue that "in anthropology, the meaning of race became formalized for humans and restricted to units based on biological variation in keeping with general zoological practice" (p. S17). They consider that the "correct use" of race is a taxonomic one, where the term dictates "variation below the species level," "natural distinctions," and "differences not susceptible to change." They argue for the adoption of this idea because "it has been formalized." Their point-of-view is in keeping with the earlier definition provided by Mayr (1969), wherein race is "[a]n aggregate of phenotypically similar populations of a species, inhabiting a geographic subdivision of the range of a species, and differing taxonomically from other populations of the species" (p. 41). Hartl and Clark (1997) have chosen to take up the meaning of race from a specifically population genetics perspective; defining races as groups of populations within which the members are more related to each other than they are to individuals of other such groups. Their viewpoint is sufficiently general to allow for the incorporation of external statistical measures for designating groups based on the distribution of some pre-established traits. Consider, for example, the "75 per cent rule," a common convention in animal biology that acts as the "limit of differentiation required for the recognition of subspecies" (Amadon, 1949, p. 2650). Hartl and Clark's definition is in keeping with the criteria offered by Brues (1977): "a division of a species which differs from other divisions by the frequency with which certain hereditary traits appear among its members" (p. 1). It also encompasses Templeton's (1998, p.646) "modern" explanation of race in terms of "a distinct evolutionary lineage within a species" and it also conforms to Dobzhansky's (1970) often cited definition based on gene frequency: "a race is a Mendelian population,

not a single genotype; it consists of individuals who differ genetically among themselves" (p. 268). Populational thinking can also be associated with more essentialist definitions that invoke expectations of local breeding populations, such as Garn (1962) does: "the contemporary approach to race stems from population genetics, where a race is viewed as a breeding population, *neither more nor less*" (p. 6). Keita and colleagues (2004, p. S17) argue however that in-depth discussions of breeding groups are typically avoided and this type of definition is dismissed by moderate thinkers because, as a practical matter, they pose "no serious problem" to making their philosophical or biological arguments. They assume that their colleagues who are in agreement with the biological race concept recognize that that "nature is too variable" to meet such demanding expectations of isolation and discontinuity (Ridley, 2004, p. 349) and that they are likely aware that any criticism using an argument of racial essentialism was already answered by Dobzhansky (in Mead et al., 1968) a half-century ago when he explained that "there is no careful and objective definition of race that would permit delimitation of races as exact, nonoverlapping, discrete entities. Indeed, such criteria do not exist because if they did, we would not have races, we would have distinct species" (p. 165). Nevertheless, the relationship between race and breeding populations has been recently revived in support of the anti-race stance, being critically reinterpreted and frequently extended to a point of confusion. Take, for example, Zack (2002) who wrote that supporters of the existence of races "to this day… assume the following: (1) races are made up of individuals sharing the same essence; (2) each race is sharply discontinuous from all others..." (p. 63). In defining "local populations" as "isolated" units, this type of definition stands outside of the other biological explanations of race by running counter to well-known arguments that races and populations are open-systems of biological exchange (e.g., Washburn, 1963). It recognizes types as implicit realities and in doing so, at best, tacitly accepts and, at worst, explicitly supports, the existence of pure races. As their argument applies to Zack's use of race, Armelagos and coworkers (1982) are right to point out that

a biological definition promoting ideas of permanently bounded, discrete groups serves only to rephrase the typological study of race in the more fashionable terms of population studies. Any such misappropriation of population genetic theory recalls to some degree Coon's (1962) racist ideas of "essence" and independently evolving racial lineages and ignores the meaning behind the change in concept that accompanied the language shift from race to population. It is worth noting that Zack's definition is even more rigid than that presented by the eighteenth-century typologist Blumenbach, which is associated with his classic systematic, scientific attempt at racial taxonomy. Even Blumenbach (1969) recognized (to some limited degree) the reality of biological fluidity when he wrote that "one variety of mankind does so sensibly pass into the other, that you cannot mark out the limits between them" (p.99). Zack is wrong to presume that the common understanding of race would "to this day" be so limited in its definition and, so, operate under her inflated paradigm of essentialist taxonomy. As it has been aptly observed by Andreasen (2005): the evolving conception of race "has been greatly influenced by science" (p. 100; see also Appiah, 1999, p. 269).

It is, therefore, wise to proceed with caution through the literature both for and against the concept of race and to exercise critical judgment when choosing among acceptable point-of-views and definitions. The frequent ambiguous use of race and the poor practice of inconsistent explanations have led to many misinterpretations, and even manipulations. The extent of the misappropriation of race is such that Klein and Takahata (2002) have adopted the controversial opinion that conclusions on race are "reached more on the basis of political and philosophical creeds than on scientific arguments" (p.384). It is important, therefore, to extrapolate the substance within any definition, explanation or exposition on race (to tease out the science from the social component), while at the same time come to an understanding of the historical milieu reflected within the ideas produced, as well as the philosophical

atmosphere out of which they are disseminated. As the descriptions above demonstrate, each of these biologically-based definitions vary in their emphasis and reflect the theoretical origins of their supporters despite the frequent assertion by each author that their definition represents the consensus in thought among their colleagues or within their discipline. It is also apparent, as Edgar (2009, p. 59) has already observed, that these definitions, nevertheless, communicate four identical points, whether or not they are explicitly stated: (1) biologically-based races are better understood in terms of populations, (2) each population shares traits (genes or phenotypic characters) at frequencies that are different from other populations; (3) these group-frequencies are explained by a shared geographic and ancestral history; and (4) the number of traits required to define a population or biological race is not fixed and cannot be predetermined. These salient points, although commonly stated, are not without their own problems. For example, if what Ousley and coworkers (2009) have called "biological distinctiveness" is an acceptable criterion for defining a biological race and no constraints are placed on the number of defining traits used for classification (see Hubbe and Neves [2007] in support of using a large number of traits for craniometric discrimination) accuracy in assignment should increase and the level of subgroup resolution would improve as more traits are considered (see Witherspoon et al. [2007] and Konigsberg, Algee-Hewitt & Steadman [2009] for mathematical explanations). As a result, the number of possible biological races will greatly exceed the number of global partitions originally represented by the traditional three discrete types (e.g., Caucasoid, Negroid, Mongoloid) and supported by the classical race concept in physical anthropology. Even the more refined subdivisions used in discussions of ancestry (e.g., African, European, Asian, Native American, Hispanic) are nevertheless broad geographic groupings which still may underrepresent this diversity (Relethford, 2009). Given these commonalities in the conditions for defining biological race and the problems associated with measures of distinctiveness, it is possible to arrive at a more thoughtful understanding of group-specific

biological variation by first looking away from race as the traditionally formalized concept, with its ambiguity, inconsistency, and closed system essentiality (from Hooton [1931] to Boyd [1950], to Coon [1962], or Garn [1962]), then looking back to the populational thinking of the modern synthesis and the new anthropology (Dobzhansky, 1970; Mayr, 1970; Washburn, 1951, 1963) and finally by looking forward to a nuanced interpretation of among-group differences when classification is of concern (Rosenberg et al., 2002; Bamshad et al., 2003; Jorde & Wooding, 2004). It should not be surprising then that debate lingers over, first, the utility of the many proposed biological concepts of race in explaining human variation (e.g., Sarich and Miele [2004] versus Brace [2005]), and, secondly, the degree to which opinions on biological race have evolved over time within the field of physical anthropology (Lieberman et al., 2003; Cartmill & Brown, 2003). A good definition of biological race may escape the disciplines and literatures of physical and molecular anthropology, human biology and the biomedical sciences because the goal of defining group differences under the constraints of long held expectations of race, "race," Race and RACE (a few bounded, discrete and encompassing categories) is discordant with the emerging patterns of biologically-based variation observed among many geographically and biohistorically different peoples. Returning to Brues' (2009) statement on race which opened this attempt to provide a consistent, realistic and meaningful definition of biological race, the discussion now draws to an end without locating a unified resolution as to its use. There may indeed be "something there" to be observed as Brues has argued, but it is unlikely to be *the* something that satisfies the "common" "correct use" of biological race that has been [mis]understood "to this day." It is appropriate to reiterate Howells' (1995) revision of Livingston's (1962) well-known statement on clines, that "[t]here are no races, only populations" (p. 103) when acknowledging the mounting evidence in support of hierarchically structured human biological variation. Analyses continue to identify geographic patterning in craniometric data (e.g. Howells, 1973, 1989; Vidarsdottir et al., 2002; Roseman & Weaver, 2004;

Pontikos, 2004) and some have recently demonstrated a significant relationship between "geography and genography" (Allocco et al., 2007) using molecular data (Rosenberg et al., 2002; Bamshad et al., 2003; Tang et al., 2005; Xing et al., 2009). Both of these sources of variation are believed to permit high levels of classification accuracy when comparing samples from different parts of the world (Relethford, 2009).

A simple regrouping of these differences in terms of many populations instead of bounded biological races does not bring an end to the discussion of the relationship between biology and race in the social and medical sciences, and especially among physical and forensic anthropologists. A review of the literature reveals that a biological interpretation of race that is based upon the heritability of some shared phenotypic or genotypic traits and operates under the assumption of homogeneity and temporal stability is but one of the ways the *biology* of race is understood among physical anthropologists. Since the assemblage of human biology results from the complex interaction of genetic, environmental, developmental, and cultural processes (Richerson & Boyd, 2005), "the methodological consideration is that if we think of culture as part of human biology, tools for the synthesis of the genetic and cultural aspects of our humanity leap to hand" (Richerson, 2010). Therefore, when addressing population differences, as Edgar writes (2009), "we can best understand biological correlates of folk race taxonomies, the categorizations of race people make about each other in their everyday lives, through an understanding of the interrelated effects of culture and phenotype" (p. 59). By the latter half of the 20th century, two additional race concepts had emerged: "social race" and "bureaucratic race." Albanese and Saunders (2006) provide the following definition: "[s]ocial and bureaucratic race are *socially constructed* concepts for grouping humans that are self-defined by individuals or groups, imposed by certain socioeconomic levels of society on others, or both" (p. 283). They are quick to identify a clear

division between social or bureaucratic race, as "real concepts that have social and economic effects on peoples' lives," and biological race, which is "theoretically based on phenotypic and genotypic variation" and "is not a valid biological concept." For these authors (2006, p. 283) "the conclusions are clear." There is no possibility of negotiation between social and biological race, regardless of the situation or qualification (forensic or medical; ancestry or affinity), because racial groups are not "useful categories for investigating human variation" (p. 283). Edgar, however, introduces a third concept for consideration in the context of biohistorical research: race as "folk taxonomy." In this redefinition, race not only encompasses the culturally-dependent ideas of race (what Ousley et al, 2009 p. 68, consider to be race in its "normal American sense") but also shares features common to several of the anthropological definitions of biological race. She explains (2009) that "while there may be no population component to a folk classification of race, such taxonomies are based on anatomical features in similar combinations shared among people due to common heredity" (p. 59). Edgar's folk taxonomy provides for biohistory what Sauer's (1992, p. 110) "race label" provides for forensic anthropology: the acknowledgement of some level of agreement between social races and skeletal biology, and, from this concordance, the ability to tease out information on individual or group identity from phenotypic variables that is both relevant to scientists and useful to the public. In Sauer's (1992) words: "In ascribing a race name to a set of skeletonized remains, the anthropologist is actually translating information about biological traits to a culturally constructed labeling system that was likely to have been applied to a missing person" (p. 109). Despite Brues' (2009) belief that just providing race in terms of its biological reality is scientific enough and can be kept unfettered with political and ideological weight, it is this practice that has perhaps made the greatest contribution to the ongoing debate over the meaning and utility of race. Arguments against how forensic anthropologists have made use of human biological variation to identify unknown individuals are based predominantly upon three main points of contention:

1. The reality of phenotypic (or genotypic) variation is not consistent with either morphoscopic and metric laboratory methods for identification or the theory that underlies the practice of ancestry estimation and group discrimination.
2. Forensic anthropology has engaged in and has condoned a simple revision of racial terminologies (e.g., the substitution of ancestry for race; the replacement of Caucasoid, Negroid, and Mongoloid with geographic identifiers) without a critical review of the assumptions that are tied to their use.
3. Because of the preceding two factors, the relatively high risk of providing false information in the form of an estimate of ancestry or a determination of race overshadows any value that it may contribute to making an identification for an unknown individual. Providing false or misleading information is taken to be, therefore, far more damaging to forensic casework than an absence of information (Goodman, 1997; Armelagos & Goodman, 1998; Armelagos & Van Gerven, 2003; Albanese & Saunders, 2006).

This last criticism has been even further extended into the realms of professional ethics and general issues of morality (Goodman & Armelagos, 1997). It has been argued that when forensic anthropologists lend credence to the false assumption of a biological basis for race, they perpetuate ideas of "pure" races and, in turn, actively promote both lay and scientific racism (Smay & Armelagos, 2000). While understanding the patterns of human biological variation is essential to revealing population structure, demographic history and the trajectory of modern human evolution, the act of classifying people into populations using phenotypic traits is nevertheless contentious given the ways in which race, as a concept, has been associated with past scientific, political and social agendas. Race is ripe with anxieties over the use of "discrimination" (in both the statistical and social sense of the word) in academia and the fear of an intrinsic and inevitable scientific racism (Brues, 2009). There remains, therefore,

considerable tension between (a) those biological anthropologists who use skeletal variation as a means to access population structure in phenotypic distance studies, to estimate individual parameters of identity when establishing the biological profile, and to meet the practical demands for bureaucratic race made by the medico-legal community (St. Hoyme & Iscan, 1989), and (b) those who take up the position supporting the non-existence and non-utility of race (Wilson & Brown, 1953; Livingston, 1962; Brace, 1964). The latter group attempts to dismiss race, ancestry or population estimates as fictitious information and consider discussions of among-group discrimination as artifacts of personal ideological positions (Brown & Armelagos, 2001). This unresolved opposition emphasizes the importance of reevaluating the data for and against the identification of biologically meaningful patterns of human variation especially in light of the recent, radical innovations in biotechnology and methodology, and the associated advancements in genomics. If Dobzhansky (in Mead et al., 1968) is correct in his now famously voiced opinion on race, "[i]f races did not exist they would have to be invented. Since they do exist they need not be invented, they need be understood" (p. 78), then it is necessary to attempt to resolve three key questions: 1) how can observed differences in human biology (be they broadly phenotypic or genetic, or specifically skeletometric) best be captured and analyzed; 2) do these differences have scientific meaning, statistical significance, and practical utility; and 3) how (if at all) do the concepts of race offer either proxy support for a better understanding of human variation or tangential evidence of some interplay with individual or population biology? It is essential to respond to Sauer's (1992) dilemma, "If races don't exist, why are forensic anthropologists so good at identifying them?" In doing so, it may be possible to offer further clarity on this question as it relates to the theory of human phenotypic diversity, the methods of forensic anthropology, and their value to the medico-legal community.

Race in Early Anthropology

"there is no there there" (Stein, 1937, p. 289)

In her self-reflective statement on place and memory, Stein wrote that there was "no there there"during a return trip to her childhood home in Oakland, CA. It can be read however as a larger commentary on the collective condition: the place of childhood is lost by all, and, in adulthood, that place, or, more likely, some a complex revisionist idea of that place, is held close to the heart. All people retain ideas of what exists or existed but these ideas rarely match with the truth of it. A similar complex relationship can be argued for the concept of race: there was a belief that race existed (*there is a there* that is embodied by some population-specific gestalt), but in time this belief grew to hold little scientific value. This change in scholarly acceptance of the reality of this idea, or idealization, was in part based upon emerging evidence of low-levels of biological variation among peoples of different "origins" (regardless of the identification terminology used, e.g., classes, races or populations), relative to those observed within these same groupings. But this paradigm shift was also, perhaps even more so, encouraged by a mounting backlash against racism, its history in the United States and its manifestation in anthropology both at home and abroad. As Gill (2000) has posited, this reactive position "seems to stem largely from socio-political motivation and not science at all" (np). Newman's (1963) criticism of contemporary identifications of geographic populations using "anti-racial" genetic data demonstrates the coming together of these two sources of rejection: "[they]...harbor many conceptual left-overs from the days of typologist thought in racial anthropology" (p. 192). Muir (1993), voicing an opinion shared by Goodman and Armelagos (1997), has further warned that any form of racialism (positive or negative) opens the door for racists so that by holding onto some scientific truth in biological distinctiveness a social illness is

perpetuated and this illness has intent to do harm. As early as 1952, Levi-Strauss (1952) had already propounded a similar message, when he stated that the "original sin" of anthropology is the [miss]assumption that race is a necessary, core concept for understanding the "production of civilization" (p. 1-3). This tenuous relationship between race, determinism, behavior and human history is one that has been wrestled with by many generations of anthropologists on both levels of scientific proof and professional responsibility. Even when the importance of race is no longer magnified under the lens of cultural prejudice, it is still retained as a core concept even for new method-based explanations of human variation in both past and present human groups (Armelagos & Van Gerven, 2003; Lieberman et al., 1989). Proving the non-existence of races, therefore, has become, and remains for many, just as much a moral, ethical, and political charge as it is a scientific pursuit. Arguments made recently by Smay and Armelagos (2000) serve not only to clarify this position but also to reiterate the deep-rooted emotions that underlie it:

> [v]ery few of the lay public will accept the scientific rejection of race when they "see" the "evidence" with their own eyes every day. Convincing them will not create a martyr, in the sense of Galileo, but it is as much a feat that requires great courage, forbearance, and tenacity. The difficult role has always been assigned to the individuals seeking to change the status quo. But if political and ethical reasons are not enough to encourage members of the physical anthropological community to "fight the good fight," then rigorous standards of scientific accuracy and commitment to evidential support, should be (p. 26).

The emergence, retention and import of this revised position on topics relating to patterns of biological distinctiveness can better be appreciated if the shift from concerns for races as discrete entities is traced

through population-based studies, and finally to the extreme outward rejection of any recognition of biologically-driven differences among human groups. This trajectory of thought on variation (that encompasses races, ancestry or ethnic groups, and populations; efforts in classification, membership estimation, and trait-based quantification; self-ascription, social predetermination, and forensic identification) can also be mapped from its original adoption by specific factions within anthropology before being translated to the discipline at large. This evolution in understanding is best understood in the light of the historical trajectory of research in human biological diversity that has led to the current "mainstream" consensus position, made explicit by the two main representative bodies, the American Anthropological Association, AAA, and the American Association of Physical Anthropologists, AAPA, in their publically available position statements (AAA, 1998; AAPA, 1996). As the organization of humans into populations, ancestral groups, or races has been a fundamental principle in anthropological practice, the debates that race and classification have spurred can be followed along a well-trodden path leading from pursuits in anthropometrics at the turn of the 20th century to innovations in genetics in the 1970s, to a return to skeletometrics and interest in biological distance, to the rise of forensic anthropology in the late 1980s. As the role of [biological] race in describing human biological variation has a lengthy history of debate in anthropology, with its development already well underway in the areas of research that preceded the establishment of the discipline of physical anthropology proper, this project does not attempt a detailed review of the history of race here but refers the reader to the broadly-based discussions offered by Kennedy (1976), Shanklin (1994), Marks (1995), Graves (2001) or Brace (2005) and to others who address more specifically the topic of human biological variation from modern and evolutionary perspectives (e.g., Mielke et al. [2006] and Wolpoff and Caspari [1997] respectively).

Even a short discussion of the history of racialized thought in anthropology cannot go without addressing, in the spirit of Stocking's (1968) historiographic principle, first the early twentieth-century contributions of Hrdlička, at the Smithsonian Institution, Hooton, at Harvard University, and Boas, at Columbia University. As colleagues, all three occupied a similar sphere of interaction in time, geography and professional discourse, with Hrdlička building the Physical Anthropology Division of the National Museum of Natural History and Hooton and Boas inspiring the growth of academic anthropology through their respective university appointments. Yet, each had a unique and profound effect on the modern school of American anthropology in both its direction of scholarship and the beliefs that shaped the discipline (Spencer, 1981; Blakey, 1987, 1997; Baker, 1994; Jantz & Spencer, 1997). They are *the contributors* to what Caspari (2009) has identified as *the key* "three perspectives on race and human variation." Although all three supported very different opinions on the scientific validity of biological race, they, themselves, set the stage for future thought on variation; each contributing, for example, to the first volume of the *American Journal of Physical Anthropology*, AJPA, issued in 1918. Their collective and lasting effect on the advancements of pan-disciplinary research over the last century is palpable in the growth of studies in (1) modern global diversity in craniofacial shape and bodily form, (2) the use of large reference collections, quantitative data sets, and statistics for the comparative study of and discrimination among people and populations, the calculation of biodistance, the assessment of plasticity, and the estimation of ancestry, and (3) the applicability of this information to questions of identification in forensic contexts and its value to the medico-legal community. Their individual impact on the evolution of perspectives on race in the United States can also be felt in the research and pedagogic legacies of their students, so that with even a century having passed since their work in physical anthropology first began, their contributions remain current. Their effects are still regularly witnessed in contemporary views of human skeletal variation and its world-wide distribution by race,

ancestry, population or some other popular proxy (Armelagos et al., 1982; Wolpoff & Caspari, 1997; Caspari, 2003; Armelagos & Van Gervan, 2003; Caspari, 2009).

If the emergence of anthropology as a formalized discipline is defined by Boas' founding of the AAA in 1902 or even if its official recognition is delayed until put in print with the public launch of the AJPA and the inaugural issue in 1918 at Hrdlička's hand, it is apparent that the setting of its broad theoretical foundations and the development of physical anthropology proper came at a politically charged moment in time. This social context had considerable bearing upon the early maturational stages of anthropology as a field and the directions its research would take. Although beliefs in biological determinism, the nature of racial origins (monogenetic and degrading, or polygenetic with many unequal kinds), and the hierarchies of idealized or "primordial" types were implicit within the research foci and designs of anthropologists at the turn of the century, there had not been an explicit emphasis on race as an area that required focused scientific inquiry since the competing schools of the early and mid 1800s (Green, 1959; Stanton, 1960; Gosset, 1963; Harris, 1968; Caspari, 2003; Brace, 2005). Yet, by the time of WWI, a renewed interest and revalidation of the "science of race" emerged in response to encouragements for a "renaissance" in human taxonomy from both governmental and private sources (Kevles, 1985). The formative stages of this early 20th century "racialist movement" in America are commonly associated with the ideas of Grant, a conservationist and eugenist, who, along with his colleagues, Davenport and Laughlin, at the Eugenics Record Office, took up the problem of the "unfit" in contemporary society. Grant explicitly argued for the extermination of "undesirables" and degenerative "race types" and offered strong support for restrictive immigration policies and anti-miscegenation laws. His scholarly output on race, *The Passing of the Great Race* (1916) and *The Conquest of a Continent* (1933), not only provided a social scientific validation for the policies of forced sterilization and

euthanasia that would be adopted by the Nazis, but also established his position as an influential figure in anthropology. Work in anthropometry afforded Grant, among others of like-minds, the sway required to lay the foundations for "acceptable" and even "essential" eugenic research in physical anthropology and emphasize the importance of racial biology to studies of human variation within this field. The ease with which scientific racism was [re]introduced and embraced can perhaps be credited to a continued but underlying support for the earlier and generally accepted tenet that "races were real and races were rankable" (Armelagos & Van Gerven, 2003, p. 55). This "fact" held true regardless of whether it was advanced from within a mono or polygenist perspective (for example, Blumenbach [1795] versus Morton [1844]) and in the presence of the emerging evolutionary and genetic science of Darwin in 1859 and Mendel in 1900. Lewin (1999) offers a simple explanation for how this shift in science was incorporated into these theories of racial determinism and hierarchy: "[i]nequality of races - with blacks on the bottom and whites on the top - was explained away as the natural order of things: before 1859 as the product of God's creation, and after 1859 as the product of natural selection" (p. 3). This applied racial anthropology, in the spirit of eugenics and the ideological movements within Europe, therefore, received considerable academic and external endorsement despite opposing attitudes offered by researchers who saw themselves rooted in scientific principles rather than political agendas. Conflicting views on racial biology were, however, strengthening within the larger anthropological community. Many of the challenges to the emerging deterministic and qualitative concept of race can be attributed chiefly to Boas, who at the time, dominated the philosophic position of the AAA and practiced a culture-based four-field approach to the discipline (Stocking, 1960, 1968). It is during this period that physical anthropology in the United States experienced its first major *crise de conscience,* with the fissioning of anthropological thought on what human differences were indeed bred in bone and mounting tensions between two well-known, opposing advocates: Hrdlička, who supported the older traditions of

European racial anthropology and served as the driving force behind the founding of the *American Journal of Physical Anthropology*, and Boas, whose concern for the consideration of culture alongside biology encouraged skepticism towards the utility of race and the validity of deterministic typology and, in turn, led to a refocusing of the disciplinary approach emphasized in the *American Anthropologist.* It is notable that this tension has yet to be fully resolved, despite a century of debate on the meaning of race as a biological concept and its potential contribution to science and society.

Both Hrdlička's (1907, 1920) early and late contributions to physical anthropology can be classified as traditional research in skeletal biology: they were focused on careful morphological observation, defined groups, and broadly based "anthropological" questions. Operating within this framework, his analyses of New World skeletal material offered support for arguments against a pre-Pleistocene occupation of the Americas; while his study of the Asiatic origins of Native Americans led to the identification of racial traits that were, and often still are, considered to be telling of group membership and patterns of interactions (e.g., linking American Indians and Asians with "shovel-shaped incisors"). It is his practical contributions to the discipline, as administrator, curator and consultant, that have garnered Hrdlička his greatest recognition. Hrdlička (1918) envisioned an applied physical anthropology that was firmly rooted in comparative typology, taking as its focus "the study of racial anatomy, physiology and pathology" (p. 4) for the purpose of understanding human variation; a goal that set it apart from many of its sister fields. At the same time, he was also actively promoting its promise both as a practical science and as a discipline of general public use. Hrdlička made great strides in elevating physical anthropology to a level of observational, comparative research that paralleled the rigor of other areas in [human] biology. As he grew his Division of Physical Anthropology within the Smithsonian, he actively engaged in the large-scale acquisition of human skeletal materials, amassing a collection that

served well for the reference resources needed for his expanding work on skeletal identification and variation. With over 20,000 specimens by some counts, it represented the single largest set of comparative materials of its kind (Ubelaker, 2000). Hrdlička (1918) worked to contextualize the importance of the comparative study of the races of man: "[t]he paramount scientific object of Physical Anthropology is the gradual completion, in collaboration with the anatomist, the physiologist and the chemist, of the study of the normal white man living under normal conditions...Such knowledge of the white race is eventually indispensable for anthropological comparisons"(p. 18). With this new formal realization of the field of anthropology and his increasing collection of comparative data, Hrdlička saw an opportunity (or even a calling) to couple the earlier traditions of racial description with an emerging concern for addressing social issues through racial biology (Hrdlička, 1918; Blakey, 1987). He advocated for the ease of applicability and clear utility of this race-based physical anthropology to the eugenics movement, its ability to offer support to the war effort, and its great potential for actively improving the human species. With this interest in practical service at the fore, he played a formative part in defining the parameters of skeletal biology as both a forensic science (establishing its preliminary goals, methods and research needs) and a social service. He sought to demonstrate its promise as an applied field that could benefit areas outside of the immediate scholarly purview of a purely theoretical physical anthropology and he found a place for the anthropologist as expert witness in legal testimony - a position *vis à vis* casework that remains similarly occupied by forensic practitioners today. Hrdlička offered expert testimony in medico-legal cases and, as "the best informed man in the United States on anthropology," he served as the first consultant from the Smithsonian Institution for the Federal Bureau of Investigation, reporting on at least 37 cases, participating in legal discussions regarding the ancestry of American Indians, and remaining in close correspondence with the then FBI director, J. Edgar Hoover (Ubelaker, 1999, 2000). Despite the change in his professional focus (from once largely clinical or

anatomical and theoretical to applied medico-legal anthropology), the philosophy underlying his analyses remained the same: a pursuit firmly rooted in racialized biology and driven by racial determinism. Hrdlička's vision for the goals of physical anthropology is best left said in his own words:

> [f]inally, the ultimate aim of Physical Anthropology is that it may. . . show the tendencies of the actual and future evolution of man, and aid in its possible regulation and improvement...The growing science of eugenics will essentially become applied anthropology...Progress in this direction stipulates...the necessity of perceiving and formulating the true goals of mankind, physical and intellectual, for the two are inseparable, and then working toward their realization (1918, p. 21).

It is obvious that this applied, racialized physical anthropology of 1918 was rooted in the acceptance of biological determinism and Linnaean ideas of racial hierarchy, thus, making explicit a scientifically sanctioned bond between biological and lay concepts of race and permitting physical anthropology to be used in support of social agendas. In 1926, for example, the AAPA and the National Research Council (NRC) formed a Committee on the Negro, which included Davenport, Hrdlička, and Hooton among its members (Hrdlička, 1927). This committee's findings not only reflected the racist feeling towards Black Americans at the time but their publication in the *AJPA* attempted to "prove that the negro race is phylogenetically a closer approach to primitive man than the white race," encouraging typological determinism under the guise of scientific scholarship and with the sanction of the anthropological community (AAA, 2007). The keen observations made by Cobb as early as 1942 quantitatively express the early dominance of racialized anthropology in the United States. Cobb (1942, p. 126) collected a total of 412 "Negro" studies, of these only 39 were of "foreign origin;" the remaining 373 reports were generated by American institutions. He further noted that "[s]ince the American Journal of Physical

Anthropology was founded in 1918, nearly every important study relating to the Negro has been either published or reviewed in this journal" (1942, p. 118).

This early racialized physical anthropology supported the type of causal, ideology-motivated scholarship that would be later [re]considered as racist propaganda, poorly disguised by a weak overlay of science: to appreciate its import, it merely needs to be considered in light of contemporary and equivalent sorts of racial science proffered by, say, Rushton (1995), or advocated by the theories in "the unified science of man" that are regularly put in print by *The Mankind Quarterly*. Any uneasy peer reaction to Hrdlička's attribution of biological reality to racial categories represents the same fundamental discomfort that is felt among many current anthropologists today, when, albeit modified, ideas of race, biological distinctiveness, qualitative classifications and quantitative methods of discrimination are engaged. The problem that remains in contemporary anthropology is the same fear that any "scientific" argument linking individual biology with social identity necessarily (1) endorses biological determinism, (2) relegates any race-based discourse to describing racial categories and encouraging racial profiling, (3) validates the existence of types through exercises in assignment, and (4) reduces the study of the complex evolution of human diversity to simple racist thinking that is reliant upon an integrated system of biology, morality, and value-based hierarchies. Hrdlička (1918) clearly intended to set the public stage for the future direction of physical anthropology with his aptly titled contributions to the inaugural issue of the AJPA, "Physical anthropology, its scope and aims." His closing thoughts stressed the importance of carrying out anthropometric surveys in the United States comparable to those already conducted at the same national level by other countries (a goal supporting the objectives of the Committee on Anthropology). In his estimation, this data would "serve as an index of progress, stagnation, or deterioration of and within the nations and thus afford indications of vital importance to agencies for

eugenics and for legislation" (1918, p. 23). Already, then, at the very inception of physical anthropology as a separate and meaningful subdiscipline, the challenge of understanding race by scholars and ideologues in the context of the theories of variation and their applied utility to the academic, medico-legal, and political communities was well underway. As Graves (2006) somberly reminds scholars, "[h]istorically, scientists studying human racial variation have varied in political motivation from outright fascists to socialist revolutionaries. No scientist at any part of the spectrum was immune from socialized racial ideology." Race would continue, therefore, to create a philosophic divide among the preeminent researchers as anthropology approached the Second World War and when human taxonomy in the service of eugenics was realized in its most deleterious and degenerate manifestation as the Holocaust (Wolpoff & Caspari, 1997).

Like Hrdlička, Hooton and Boas also undertook detailed studies of the biological basis for phenotypic variation (each addressing, for example, racial affinity in their 1918 AJPA publications) and they also amassed considerable comparative-reference and analytic resources, collecting anthropometric, craniometric and nonmetric data from a variety of sources (from criminals to military personnel, from archaeological skeletal populations to recent immigrants) at rates and in quantities that have not since been matched for skeletal material and measures of bodily-form. Although this trend in research and data collection may initially suggest a parallel trajectory taken by all three, their interests and intentions were indeed divergent. Neither Hooton nor Boas fully adhered to the central tenet of physical anthropology, the study of the "more primitive human races and their subdivisions,"as it was interpreted by Hrdlička (1918, p. 19), and they also pursued goals in scholarship that were equally dissimilar. As they remained at odds throughout their professional careers, their differences in approach encouraged diversity in their projects. They demonstrated not only a great breadth of interest in areas

that are now considered as "classic" physical anthropology but also covered a wealth of topics that extended the scope of their research into the fields of sociology, psychology, statistics, and politics, among others. Hooton (1939), as a classicist *cum* anthropologist, ushered in the biological tradition in American criminology by collecting anthropometric data on nearly 14,000 criminals and 3,000 members of the general population in support of new research in criminal anthropology (Rafter, 2006). Although his project may have been designed in the spirit of Lombroso (1895, 1897), who held much intellectual sway on the interrelated themes of nature, behavior and biology in the late 19th century, Hooton produced a work that was focused on data analysis in addition to offering morphological criteria for the identification of "the criminal" type. He (1939) documented significant differences between criminals and noncriminals in 19 of the 33 measurements taken and reported frequency data on the relationship between race and criminal acts. This quantitative approach suggests perhaps a greater commonality with the interests of Bertillon (possibly transmitted through his successor, Galton), who combined anthropometry with the still primitive methods of multivariate classification, than with the qualitative style of work produced by Lombroso's Italian School of Positivist Criminology (Gillham, 2001; Buikstra & Beck, 2006). Reactionary interest in Hooton's more controversial psycho-physiognomic work, that was both driven by his polygenist and biological determinist views and informed by the eugenics movement that was popular in his formative professional years, has often eclipsed his contribution to the study of human biological variation. Much like the attention given to Hrdlička's contentious, racist, agenda in many historical reconstructions, concerns for documenting periods of scientific racism have also overshadowed discussions of Hooton's place within the history forensic anthropology at the time when modern skeletal identification was moving towards gaining acknowledgment as a scientific pursuit. Of particular interest is Hooton's combined application of qualitative and quantitative methods to questions of race. Their successful marriage is clearly observed in his treatment of the craniometric and

nonmetric trait data obtained from the Pecos Pueblo site (1930). With a large skeletal sample available (over 500 individuals representing multiple cohorts), he was able to evaluate variability in morphology and, subsequently, identify the co-existence of eight "morphological types," doing so, first by gross inspection using a morphoscopic approach to craniofacial characteristics and then, later, by conducting statistical analyses on the available skeletal measurements (Woodbury, 1932, p. 142). This data served as key evidence for the presumptive identification of biological heterogeneity within the Pecos population. His assessment not only provided demographic information for the archaeological site itself (which he coupled with his extensive pathological studies), but served as testament to Hooton's polygenist arguments for the detection of typical racial forms that were group-specific and biologically derived. His results lent support to concepts of racial hierarchy and demonstrated the validity of subracial typing, which he later made explicit in his books, *The American Criminal* (1939) and *Crime and Man* (1939). Hooton divided contemporary humans into three primary races (White, Negroid, and Mongoloid). He then subdivided each of these into secondary races (e.g., separating individuals of Negroid race into the African Negro, Nilotic Negro, or Negrito categories), and he even further refined this taxonomy by defining a third-level racial division that included what he described as "pseudo" morphotypes (i.e., cohesive groupings nested within the larger classes that demonstrated distinctive morphological features). Individualized definitions for these [pseudo-sub] types became possible using his own understanding of biology and concepts of heredity; explaining that, for example, the "Old American" criminal was "a native born White of native parentage-from two or three generations of American nativity." Hooton (1939) also made a causal connection between physiognomy and deviant behavior, claiming that "[c]riminals as a group represent an aggregate of sociologically and biologically inferior individuals" (p.304). He, therefore, promoted the use of multi-tiered biologically-based partitions in order to shed better light on the distribution of deviant behavior within the American

people and, in turn, to use these types with their associated qualities to encourage segregation; a sentiment he reiterated in his later work: "[I]et us cease trying to make the world safe for morons, and endeavor rather to save it from them" (Hooton, 1940, p.230). Race, therefore, showed scientifically supported promise as a discriminatory tool that could be easily and credibly extended to problems in skeletal identification. Hooton's views on purity and hybridity in race and the related concepts of nested races at the subspecific level are presented more fully, with efforts toward cohesion in, clarity for and wider applicability of an essentialist typology, in his seminal work on the human condition, *Up from the Ape* (1931; and in its later 1946 edition).

This enduring belief in discreet human races, morphological types and principles of biological determinism significantly influenced the course Hooton's research took and the historical reception of his work. Nevertheless, Hooton's tenure at Harvard lasted over 40 celebrated years. During his time there, he has been credited as much for the important advances he made in understanding human skeletal variation as for his fundamental developments in data collection and analytical methodologies, which together have left their mark on many of the practices in physical and forensic anthropology still in use today (Shapiro, 1954; Krogman, 1976; Birdsell, 1987; Giles, 1998). Already, Hooton had begun formulating his method and theory of "difference" with his contribution to the inaugural issue of the AJPA, "On Certain Eskimoid Characters in Icelandic Skulls," although he had yet to standardize his interpretation of race. As Caspari (2009, p. 9) has observed, in this 1918 publication, race was used in many different ways and could represent any number of multi-tiered hierarchies: in reference to large continental groupings (e.g., American or European), as subgroups within a continent (e.g., the "European taxonomy of Nordics, Alpines, and the Mediterraneans"), as individuals from many different levels of regional subdivisions that served as reference groups (e.g., Southern Californians or Italians)

and qualitative partitions based on bio-cultural criteria (e.g., "primitive" versus "civilized races"). The further unfolding of these ideas can be followed in his later work on the skeletal materials from Pecos Pueblo. From a timeline of his publications and projects a progression or development in thought can be imagined: starting with his early macroscopic observations, moving into his defined measures and nonmetric traits as racial features, and eventually emerging as a standardized "list" that gained research and pedagogic currency with each use. The history of how Hooton's list became known as the Harvard Blanks can be easily traced, and it is possible to track the ways in which it was passed from mentor to student through time. The popularity of this work is attested to by the fact that it was regularly borrowed by his students when establishing their own laboratory practices, and was even frequently referenced by and shared among members of the "opposing" schools of thought: Montagu (1966), although a student of Boas at Columbia University, included a version of the Harvard Blanks in his handbook for anthropometry. This version has been attributed to Angel, who likely received it, while still a student working under Hooton himself (Brues, 1990). The important role that this "list"assumed in physical anthropology is attested to by the broader-scaled perpetuation of the Hootonian tradition in individual identification. These ideas of group differences and the associated by-race traits have been codified over time as *the* morphological signatures of interest and emerged among the "standards" of good laboratory practice. As such, Hooton's ideas, definitions, and rules of practice have been retained implicitly in many contemporary guidelines for skeletal analysis, both those from the past and those presently used, forming the theoretical structure that supports many efforts in human identification and the attribution of race (see Brues, 1990; Rhine, 1990). Hooton's criteria are also in the process of being retooled by an emerging generation of professional anthropologists in explicit attempts to improve classification and estimation of individual group membership in the context of forensic casework under

the "new"parameters of ancestry estimation and in response to the changing expectations for expert testimony and evidentiary requirements (Hefner, 2009).

While his continued influence pays tribute to his enduring scholarship, Hooton's greatest contribution to physical anthropology may be, in fact, pedagogic. As Shapiro (1954) wrote for Hooton's obituary, which was published in the *American Anthropologist*, "Earnest Hooton possessed a number of exceptional gifts, but none was greater than his genius as a teacher" (p. 1081). He started producing PhDs in 1925 and, over the course of his academic position at Harvard, he alone taught, mentored, and advised 28 graduate students (Shapiro, 1954; Giles, 1997). He personally advanced the careers of more physical anthropologists than any other professor at that same time, including Boas at Columbia University. Caspari (2009, p. 6) has even posited that "[m]ost practicing American physical anthropologists can trace their intellectual ancestry to Hooton at Harvard." Hooton most significant achievement may be, therefore, building a legacy of students who themselves became influential players in the areas of skeletal identification and interpretation of variation as well as successful professional mentors (Spencer 1981). Hooton adopted, and many of his students ultimately retained, a polygenist approach to human variation: they subsequently upheld his interest in the ratification of typological categories and the proved existence of distinct biological races (Brace, 2005). This philosophic heritage is visible, for example, in the studies on group morphology produced by Coon, whose dissertation research addressed the adaptive significance of racial traits and "*typical* racial forms"and whose later publications argued for a hierarchy of types in the spirit of evolutionary determinism (Wolpoff & Caspari, 1997, p. 155), or by Shapiro, whose work was concerned with "hybrid vigor"in "racially mixed"populations and therefore implied ideas of the existence of original, identifiable pure types (Spencer, 1981). Regardless of philosophic intent and despite what seem now to be radical ideas on race, Hooton's research, data

collection, and astute observations have shaped the appreciation of cranial diversity among populations. They not only helped formulate the modern approach to the estimation of ancestry but have also been credited with broadly informing all method and theory in forensic anthropology; either when practically applied in the service of the medico-legal community or more generally in skeletal research and scholarly pursuits (see Hefner et al., 2004; Hefner, 2007). Hooton's legacy is of special interest to this project both as much for the methodologies in data collection and analysis inherited by his students, as for the developments in thought on variation and its relationship to the global distribution of races that his successors engendered.

In reaction to the new "anti" race sentiments that were burgeoning in America after the Holocaust and in dismay over the concomitant change in the intellectual climate, Hooton is reported to have said to his former student, Washburn, at the 1950 Cold Spring Harbor Symposium: "Sherry, I hope I never hear the word 'population' again!" (Marks 2008, p. 23). Interestingly, it was Hooton's other student, Howells (1995), who made the definitive claim that helped to reconceptualize race in anthropology twenty years later: "there are no races, only populations" (p. 103). Howells effectively summarized the critical paradigm shift that followed a post-war crisis in racial studies: old Platonic ideas of race were supplanted by a new generation's focus on geographic variation and the essentialist paradigms of racial biology had shifted towards population-based theory (Krogman, 1943; Kuhn, 1962). This growing concern for and acceptance of the "population" was enabled quite remarkably, even if only in part, by Hooton's legacy. Howells carried forward Hooton's dedication to recording procedures and deep appreciation of the value of comparative data: he ultimately produced a globally representative craniometric data set that has been indispensable to the study of human variation in recent scholarship, especially since it was posted online (c.f. Howells, 1996; http://konig.la.utk.edu/howells.htm). This data

was not only the focus of much of Howells' (1973, 1989, 1995) own research, but it has also served as the key information source for (1) tests of population differences and distances (Relethford, 1994, 2002; Rosemann & Weaver 2004; Konigsberg, Algee-Hewitt & Steadman, 2009), including the Finite Mixture Analyses presented here, (2) the development of discriminant functions for classifying individuals according to their ancestral population and, (3) the concomitant identification software for forensic applications (FORDISC by Ousley and Jantz [1996, 2005]). The Howells craniometric data set remains as an invaluable resource that is both publically available and easily accessible to a broad geographic and disciplinary research community (Howells, 1996).

At the same time that the fledgling field of physical anthropology was intensifying its scholarship through expanded resources, increased data collection and the introduction of new methods, the discipline of modern statistics was undergoing a similar [r]evolution. Around the turn of the 20th century and for the few decades thereafter, increasingly advanced statistical techniques were being developed for the study of variation and the comparison of characters among sizable samples or within large data sets. The nature of the demands of this research (both in terms of data requirements and the promise of its application in areas of interest to the socio-political movements that were dominating thought at that time) inspired an interesting relationship "in science" between the statistician and the quantitatively-minded anthropologist (see Armelagos, Carlson and Van Gerven [1982] for a detailed account of the early statistical foundations of skeletal biology specifically). At first, most of the research produced in these early years of modern statistical history made use of biological variables and took up topics of human anthropometric or morphological variation. Yet, this formal statistical interest in anthropometry was predominantly pragmatic. Biostatistical analyses were conducted with the purely mathematical intentions of testing a particular statistical method or implementing a new procedure

using readily available data and repeatable measures. The process was, therefore, the focus. Understanding the biological aspects of human skeletal morphology and explaining the meaning underlying the observed statistical phenomena fell second among statisticians to the theoretic goal of demonstrating and, hopefully, substantiating the merit of emerging statistical methods using large, multivariate samples and variables that were familiar in the values that they represented. Data sets of substantial size and composed of common measures, such as maximum breadths or lengths, taken on distinct and easily recognized anatomical regions mitigated the basic issues of significance and interpretability for the purely "test" situation. For these reasons, craniofacial and body-shape studies gained significant currency in practical applications, where human metric variation was more important for its contribution to the presentation of the mathematics rather than for its utility in improving an understanding of biological relationships. While the early, pre-WWI statisticians like Lee (1901), Fawcett (1901-1902) and MacDonnell (1901) all published their research on correlation analysis using craniometric variables in this spirit, it was not until Pearson and Davin (1924) took up work on patterns of organic correlations in the human skull that an emerging, combined concern for both statistical and biological significance can clearly be seen. Although their methodology and topics of analysis do demonstrate growth out of the earlier studies (including Pearson's own, first, application of regression analysis to cranial data and questions of heredity in 1896), Pearson and Davins' interest in eliciting biological meaning out of the observed statistical phenomena suggests a new direction in data analysis. Such a new direction is evident in the research that followed by Stoessiger (1927), Collette (1933), Morant (1935), Pearson and Woo (1935) and among others who shared in Pearson's concerns. This new trend may have reflected not only a natural progression in the development of any statistical procedure (from purely theoretical testing to method validation, before its application to problems with real world implications) but also a sensitivity towards the socio-political climate that weighed heavy in the post, pre

and intra-war eras. Their interest in biological data likely reflected a concern for validating historical reconstructions and theories of racial origins in the service of movements concerned with race, ethnicity, and national identity. Stigler (1999) offers an apt characterization of Pearson that permits insight into how society came to drive science in this case: "[h]is interest in eugenics predated his discovery of Galton's statistical methods in the 1890s, but once that discovery was made, all the ingredients were there for an activist social scientist: boundless intellectual energy, an ambition to direct scientifically guided social action, a clear and forceful writing style, and the patience to undertake detailed, painstaking empirical investigations" (p. 13). Pearson's (1926) "coefficient of racial likeness" and the development of distance measures by Fisher (1936) offered the means to support the concept of race and reinforce validity in racial biology (Hunt, 1959; Andersen, 1996). Early researchers who worked under the label of eugenics were not uniform in their intention: in the case of Fisher, he inherited the editorship of the *Annals of Eugenics* from Pearson when he became the Galton Professor of Eugenics at University College in 1933, but also worked closely with Wright and Haldane on what has become the basis of the modern evolutionary or neo-Darwinian synthesis. Regardless of their initial intention, however, these advances in statistics were used to strengthen the dominant typological and taxonomic ideas that were well established in physical anthropology. For example, support for these ideas run heavily through the scholarship of Hrdlička and Hooton. These emerging statistics provided the tools necessary to obtain mathematical support for making distinctions between skeletal groups. The ideological judgments that alone begat racial determinism, systems of hierarchies, and fixed types found new support from this sound (misuse of) biostatistical theory and race, which became firmly rooted in (weak) science.

Of the three dominant figures in physical anthropology at this time, Boas appears to stand alone in his position on race and, in turn, in the meaning that he ascribed to the morphological differences observed among members within the human species. Boas presented what has been called an "unracialized view" of human variation: he was a strong supporter of racial equality and the independence of biology, intelligence, language and culture at the time when the biology of "the normal white man living under ordinary conditions" (Hrdlička, 1918, p. 19) was the standard against which all other peoples were compared (see Boas, 1894; Baker, 1994; Armelagos & Van Gerven, 2003). Boas chose to concentrate his efforts in understanding group differences on (1) statistical tests of the concordance of racial features; (2) inquiries into the common assumptions that worked to support the prevailing concepts of invariant "types"; and (3) critical reviews of the quality of the evidence that had been used by other anthropologists in support of, and, to perpetuate scientific racism (Comas, 1961). Boas was also unique among his colleagues in his interest in quantitative methods (Xie, 1988). He took up a biometric approach early on in his career and, before it became fashionable in anthropology, he questioned the validity of long-held anthropometric practices (noting, for example, the instability of the cranial index in his 1910 and 1912 publications on immigrant populations). He also applied statistical methods to diverse, large sample studies of human variation to less common goals. In doing so, Boas made great strides in both anthropology and biostatistics (Jantz & Spencer, 1997; Camic & Xie, 1994). His work included impressive data collection projects on the subtopics of growth and development, geographic variation, and environmental plasticity, which he pursued to a rigorous understanding of the distributional properties of human variation and the external factors or internal mechanisms that contributed to the differences in phenotypes observed within and across groups. His large-sample studies on growth in school children from Oakland, to Toronto, to Worcester, to Milwaukee and to St. Louis (revisited in 1932, 1933, 1935), the effects of environment and heredity among U.S. immigrants

and their offspring (1910, 1912), the pattern and causes of trait distribution among Swedish men (1918), and the phenotypic results and biological consequences of admixture in "Half-Blood Indians" (1894) stand as testaments to the diversity of his scholarship and his concern for obtaining quantitatively significant results. Over the course of his anthropometric surveys he obtained data from approximately 27,000 individuals representing Native American, Siberian and European ancestries living at the turn of the century (Jantz, 2003) and, for his immigrant study alone, he organized the collection of multivariate data on 17,821 subjects (Boas, 1928). These numbers remains very impressive today, despite contemporary advances in data acquisition, new recording techniques and ease of information exchange, storage and retrieval. The magnitude of the immigrant data set, in either its raw form or its later public iteration as *Materials for the Study of Inheritance in Man* (published in 1928), has gone unmatched even among the most ambitious data collectors and projects. For example, Howells (1989, 1996) and Hanihara (1993, 1996; Hanihara & Ishida, 2001) obtained measurements on 2,504 and 11,757 crania respectively, and the Forensic Anthropology Data Bank (FDB) contained 1731 positively identified individuals at the time of the last website posting of its figures (2005).

Boas, like Pearson, took up an approach to variation that married biostatistics with questions of interest to anthropology. His work focused on the ways in which a particular statistical procedure (or combination thereof) worked in concert with a set of measures representing some aspect(s) of human morphology: his intention being to produce both statistically and biologically meaningful results that could be used to develop arguments of anatomical relationships and derive evolutionary, environmental, or developmental explanations for the observed statistical phenomena. What is especially notable about Boas is his overt rejection of the ideological framework to which the majority of the research conducted by his contemporaries in both anthropology and statistics adhered. It was under

this framework that questionable "biological"explanations were built, using models that were founded upon typological traditions and which were themselves grounded in subjective reconstructions of historical origins and deterministic hierarchies. As Boas (1940), himself, explains, the dominant view among physical anthropologists accepted that humankind was composed of but a few races, invariant types, or "'permanent forms' which have lasted without variation from the beginning of our modern geological period up to the present time"(p. 35). Boas unseated the prevailing assumptions that made up the very foundation of racial typology by posing rhetorically simple questions that demanded complex answers: (1) If averages are obtained from the sum of deviations, how can an average represent the norm? (2) If plasticity can be shown in what are considered racial traits, how can the concept of fixed races be supported? (3) If races are in a constant state of morphological or phenotypic change, how can a consistent number of groups be identified and a fixed system with a permanent taxonomic order be established? These inquiries had profound effects on how the source(s) of variation in human biology were indentified and understood, and, the continued ease with which the existence of invariant biological races could be accepted under *pro forma* arguments in physical anthropology (Armelagos & Van Gerven, 2003).

By the time Boas had taken up his tenure at Columbia University in 1896, he, like his colleagues, was deeply concerned with race; yet, then-unpopular ideas on racial variation had already begun to dominate his thought process and guide the direction of his research. As Stocking (1968, p. 192) explains, Boas "refused to allow the reality of biological variation to be obscured by arbitrary idealized types." Although Boas never denied the possibility that races do exist, his opinions were guided by an otherwise uncommon concern for "process" and the evolution of the individual rather than a preoccupation with the [pre]definition of types (Lewis, 2001). This is a manifestation of what Herskovits

(1943, p. 50) has called his "habit of thinking culturally."It is not surprising then that Boas opposed the validity of racial determinism. He clearly drew heavily upon this culture-based four-field approach to anthropological problems when making the argument that: "historical events appear to have been much more potent in leading races to civilization than their faculty, and it follows that achievements of races do not warrant us to assume that one race is more highly gifted than another" (Boas, 1894, p 303). This argument represents a point of concern that continued throughout Boas' scholarship. It received its greatest attention and achieved perhaps its most significant end in his study on plasticity of head form among the descendants of European immigrants in New York City (Boas 1910, 1912, 1928). Here, he not only demonstrated that bodily form could be altered under the influence of a new (U.S.) environment (i.e., common environmental conditions exerted a homogenizing effect on cranial morphology so that children of immigrants, representing different geographic homelands, converged onto an "American type"), but he further extended his findings to support a broader argument against deterministic racial biology. He posited that when "features of the body change, the whole bodily and mental make-up of the immigrants may change" (1910, p. 7-8). He took his results as evidence *for* the "strong and increasing effect of the American environment"(1910, p. 17) on cranial shape and as equal evidence *against* the fixity of racial types based on assumptions of stable, race-specific morphologies. These views were also retained in his position against racial determinism that was more fully elaborated in his work, *The Mind of Primitive Man* (1911; revised in 1938).

For many years and regardless of changing topics and peoples of interest, from the physical characteristics of indigenous Native Americans of the North-West Pacific coast (1891) to his specific cranial-based studies on immigrant Americans local to his home at Columbia University in the North-East (1910, 1912), Boas continued to develop and refine his "unracialized" position. In the face of heated

criticism from colleagues and the public alike and even despite the dismissal of his conclusions by the Immigration Commission itself (Baker 1998), he nevertheless maintained the constant underlying theme that racial differences cannot be understood by quantitative measures (Caspari 2009). Early on in his career, Boas challenged the common belief that valid racial typologies, human taxonomies, and biocultural hierarchies could be reliably developed from measures of racial phylogeny (e.g., measurements, angles, ratios, traits, and phenotypic observations). In accordance with his greater concern for the effects of environment (and culture) on biology than for the affect of biology (or race) on culture, he questioned, for example, the long-standing preoccupation with the cephalic index; a topic on which he had expressed previous interest (1899, 1903) and returned to in his immigrant study. This index held special scholarly importance among other measures of variation. The index served not only as a reasonable proxy for head form (which was highly valued among racial criteria) but, by the turn of the century, it also already carried with it the weight of tradition among phrenologists, racial biologists, and anthropologists (Gould, 1996; Montagu, 1997; Marks, 2002). Boas himself reflects (1912) that it had "always been considered one of the most stable and permanent characteristics of human races" (p. 5), while, Tanner (1981) highlighted the curious fact that such methodological importance was ascribed to such a simple ratio (head breadth: head length), elevating it to the level of the "central tabernacle of the doctrine"(p. 250) of racial classification. For Boas, the findings of his immigrant study demonstrated, as Gravlee, Bernard and Leonard have reiterated (2003), "not only plasticity of human cranial form but also plasticity of human potential" (p. 127). Boas followed up this study with his 1918 paper that revisited an earlier analysis of anthropometric data collected on Swedish males, first published by Retzius and Furst in 1902. His approach was directed by a now well-established interest in trait distributions. He focused, therefore, on how measures of hair color, eye color, head length and breadth, stature, and the coefficient of correlation between head length and breadth varied across the 25 Swedish provinces that

were sampled. His analysis of this data not only complemented his wider interest in the proper characterization of variation among Europeans and the identification of their causes, but his findings also helped to further undermine the conventional expectations of a "real" relationship between subjectively defined racial and biologically arbitrary psychosomatic types. Boas reported findings for the Swedish data that included gradual changes in trait distributions, geography-based differences in trait variability, heterogeneity in eye color and the lack of trait concordance. Despite the intellectual climate of the times and the common trends in scholarly expectations, Boas (1918) overtly referenced race only once with regards to tests of the concordance and distribution of traits and, even then, he maintained a conservative opinion: "I think it may be possible that a detailed study of distributions based on exhaustive material obtained from small local groups may show us in how far there exist close correlations between the various features. This would help us clear up the racial significance of various anthropometric traits"(p. 422). Even almost a decade after the submission of his initial report on the observed changes in cranial form among immigrants and using a data set that was very different in its parameters and acquisition, Boas continued to offer scientific support that challenged racial biology and discredited its many assumptions. A question of accuracy nevertheless remained at the forefront throughout his professional career and has regained considerable attention in recent years: in the straightforward words of the editors of *American Anthropologist* (2003), "did Boas get it right or wrong?" For example, Gravlee and co-authors (2003) believe that their reanalysis of the immigrant data does "confirm Boas's overarching conclusion regarding the plasticity of cranial form " (p. 326), arguing that environment plays a significant role in cranial shape (i.e., "plasticity") and can drive the direction of morphological change observed in individuals who would otherwise represent fixed racial types. Using this same data, however, Sparks and Jantz (2002, 2003; Sparks, 2001), have consistently come to different conclusions. They determined that because the immigration effect is small relative to the

among-population variation (with 99% being attributed to differences among "ethnic groups" [Sparks and Jantz, 2003, p. 336]), their "reanalysis of Boas' data not only fails to support his [Boas'] contention that cranial plasticity is a primary source of cranial variation but rather supports what morphologists and morphometricians have known for a long time: most of the variation is genetic variation" (Sparks & Jantz, 2002, p. 14637). These same authors (2003, p.333) offer possibly the best characterization of this debate: "these two camps," represented by this recent flurry of analyses and rejoinders, "interpret the changes reported by Boas (1911) as small, occasionally significant and important (Gravlee et. al., 2003), or small, occasionally significant and unimportant (Sparks, 2001; Sparks & Jantz, 2002)." The commonality of both of these views lies in what Relethford (2004) has noted as well: "the different influences on craniometric variation do not erase or obscure other influences" (p. 385). The real question that remains is which of these effects is the most important. As can be seen from Boas' own work and the reanalyses discussed above, any answer will depend upon the design of the individual study, its concerns and questions, and the focus of its tests. Boas' contribution may be less about the accuracy of his findings, and more about the changes in the climate of thought that his work precipitated, as well as the open forum for discussions on phenotypic variation that his work still continues to produce.

Almost exactly one century after he first submitted his preliminary report, *Changes in Bodily Form of Descendants of Immigrants* to the United States Immigration Commission in 1910, Boas' early contributions to the study of race, whether "right" or "wrong," still continue to exert considerable influence upon the state of the discipline, shaping its direction and profoundly affecting its people, research and theories. The value of his corpus of scholarly output, and especially his immigrant data, to current anthropology and its cultural import is palpable in the pages of peer-reviewed journals (Darnell,

2000; Sparks & Jantz, 2002, 2003; Holloway, 2002; Gravelee, Bernard & Leonard 2003; Relethford, 2004). It has also found a place in a public discourse among controversial science-writing that has reached the popular press (Bergstrom, 2002; Wade, 2002; AP, 2002). His research has been revived by trendy stories with interest-catching tag lines, like "Going Head-to-Head Over Boas's Data" (Holden, 2002) in *Science* or "The Measure of America: How a Rebel Anthropologist Waged War on Racism" (Pierpont, 2002) in *The New Yorker.* From a historical point of view and as an aid to understanding the broader trends in the progression of anthropological thought since his research first began, the question if Boas "got it right" is of little consequence. What is essential to acknowledge is the implications of the outcome of his conclusions in terms of public perceptions of race in American society, the trajectory of future scholarship in variation, the disciplinary positions on the biological meaning of race, and the continued value of the race concept in anthropology. Already by 1923, Boas' student, Kroeber, had begun the dissemination of the Boasian view when he wrote that "it is a difficult task to establish any race as either superior or inferior to another, but relatively easy to prove that we entertain a strong prejudice in favor of our own racial superiority"(p. 85).

In the *Mind of Primitive Man* (1911), Boas wrote that "differences between different types of man are, on the whole, small as compared to the range of variation within each type" (p. 94). The similarities between the language on variation he used and the message against race he expounded in this statement and the conclusions that would be drawn almost 60 years later from findings based upon emerging and far more complex genetic data is remarkable. Compare for example the early words of Boas (2011, p. 94) with these of Lewontin (1972):

> [i]t is clear that our perception of relatively large differences between human races and subgroups, as compared to the variation within these groups, is indeed a biased

> perception and that, based on randomly chosen genetic differences, human races and populations are remarkably similar to each other, with the largest part by far of human variation being accounted for by the differences between individuals. Human racial classification is of no social value and is positively destructive of social and human relations. Since such racial classification is now seen to be of virtually no genetic or taxonomic significance either, no justification can be offered for its continuance (p. 397).

In making this connection, it is apparent, as Caspari (2003, 2009, p. 11) has also observed, that Boas' unorthodox rhetoric on race "prestaged" what would become the modern convention, as new scholarship in genetics and the transfer of concern from "types" to "populations" in studies of variation challenged the essentialist tenets that played a central role in the founding of American anthropology. It is especially interesting that it was left up to Boas, as the "institutional and intellectual founder of the contemporary four-field discipline" and preeminent cultural anthropologist, to break the trend in racial studies among physical anthropologists and statisticians using their own quantitative methods (Darnell, 2000, p. 896). It should not be surprising, therefore, that his pedagogic legacy is diverse and included alongside his students pursuing projects in culture and ethnography (among them was Mead, who received her PhD from Columbia University in 1929) are future leaders of the movement against race under the rubric of biological or physical anthropology. Perhaps the best example in this context is Montagu, who after receiving his PhD in 1938, produced a series of works that, in the spirit of Boas' research in variation, openly questioned the validity of race as a biological concept and criticized its supporters. Of these, his best known may be the seminal text on the "new" position against race, *Man's Most Dangerous Myth: The Fallacy of Race* in 1942 (and in 1998 was in its 6th edition), and his follow-up work, *The Concept of Race*, which was first published in 1964. Boas can be credited, therefore, with precipitating the second major shift in the understanding of race in physical anthropology. In Gossett's

(1997) words, Boas was "the man who did more than any other to lay the ghost of racism in scientific disciplines"(p. 450). This changing tide came at the close of WWII, and was an effect of the right conditions, which included: (1) the emergence of the modern synthesis, anthropological genetics and the concomitant populational studies (*Dobzhansky,* 1937; Mayr, 1942; Huxley, 1942), (2) Washburn's (1951) call for and introduction of the "new" physical anthropology, and (3) Montagu's (1950, 1951) drafting and authorship of the first UNESCO Statement on Race, and the wide dissemination of this as "The Race Question." The confluence of all of these factors, people, and events created an environment that enabled a collaborative challenge of the race concept. Some made their contributions inherently, as they were carried along in this direction by the nature of their pursuits, results, and conclusions or with the tide of their colleagues' changing approaches to scholarship. Others, however, acted intentionally, waging attacks on specific ideas, assumptions and even individuals. Montagu's publications were especially controversial and his inflammatory intentions were clear when he included trigger words like "myth" and "fallacy" in the title of his 1942 book. He did not, therefore, simply reject race in his scholarship, but he also openly mocked the science of racial typology, demoting its claims of "truth" to the level of ideology. Montagu's position ultimately emerged as the "expert opinion"shared by the members of the 1950 UNESCO conference, who concluded that "race was less a biological fact, than a social myth." This final statement was quickly reported by The New York Times (July 17, 1950) under the heading: "No scientific basis for race bias found by world panel of experts." As Montagu's arguments were not only actively voiced among his immediate circle of peers but were widely distributed through his published texts and policy work, he publically invalidated the bulk of scholarship that had defined physical anthropology's academic origins and which continued to sanction its (often exclusively) racialized motivations. As Caspari (2003) has remarked, "Boas's followers held some of the most powerful positions within U.S. anthropology and held a dominant voice in the AAA" (p. 66) and, so, it

should be of no surprise that the rejection of race, which began with Boas and was carried forward by his students, would take hold of mid-century scholarship in all of anthropology. For biological anthropology specifically, it is also this Boasian heritage that would inspire the next chapter in its history.

The Synthesis of "New" Ideas

"The anatomy of life, of integrated function, does not know the artificial boundaries which still govern the dissection of a corpse." (Washburn, 1951, p. 130)

The First World War encouraged American xenophobia, fed feelings of patriotism and nationalism, fostered an environment ripe for imposed racial stratification, and, ultimately, supported a tradition of typology that was rooted in a fear of the "other" and depended upon a pseudo-science that preserved and protected the identity of the normal American "self." The aftermath of the Second World War, however, opened up another floodgate for racial ideas through which flowed a very different set of concerns. America's public (including soldiers, families, friends and allies) had felt first-hand the consequence of what Arendt (1963) came to call "the banality of evil." The Nuremberg Trials introduced racism to America as it was exercised in its most horrific form: extermination in the preservation of social order. As Göring (1946) explained in his testimony on the justification of the concentration camps, "[i]t was a question of removing danger." Although the Holocaust epitomized the extreme manifestation of typology, Nazi propaganda shared a racial rhetoric with the work of the American eugenics movement and the efforts in human taxonomy of many early modern anthropologists that was couched in the all too familiar terms of "normalization" (Stein, 1988; Gasman, 2004). The impact of these

changed times on the anthropology community was clear. Washburn's closing remarks in his AAA Presidential address, The Study of Race, delivered at the annual meeting in 1962, reveal the extent of this transformation: "[w]e know the price of discrimination is death, frustration, and hatred. We know that the roots of happiness lie in the biology of the whole species..." (1963, p. 531). Biological and social scientists alike were now acutely aware of the realities of race: how theoretical musings on racial hierarchies could swiftly become the scientific justification necessary to cement ideology in biological fact and sanction otherwise unthinkable acts of discrimination, apartheid, slavery, and genocide. This post-war rejection of scientific racism among academic communities gained momentum in the 1950s with its emergence into the public eye, first in earnest with the case of Brown v. Board of Education (1954) and the followed by the desegregation that resulted from the monumental overturning of the "separate but equal" doctrine of Plessy v. Ferguson (1896). This rejection continued to grow in both scholarly and cultural currency with the mounting support for the U.S. civil rights movement as well as the anti-colonial and independence efforts that were occurring concurrently in Asia and Africa. By the 1960s, concerns for political liberation and indelible human rights were felt world-wide as the spread of freedom movements led to the United Nations General Assembly's resolution on the Declaration on the Granting of Independence to Colonial Countries and Peoples (1960). Out of these trying times grew a "new" physical anthropology. In the 1950s its practitioners first took up an inspired review of the patterns and meaning of human variation in light of the changing attitudes in science and society and the increasing access to previously untapped sources of information. By the early 1960s supporters of this new anthropology had fully adopted an aggressive stance on race. This position against race emerged not only as a natural progression from the type of evolutionary and population-based thinking that the new anthropology promoted, but also in reaction to the essentialist, deterministic claims on the origins, intelligence, and hierarchy of worth that continued to be published as the paradigm of human

variation and which was used to validate public policies promoting racial inequality at a time when civil rights were of immediate concern. The emergence of this new anthropology, therefore, marked a turning point both methodologically and philosophically as it redirected its scholarship towards "anthropological science" (Jackson, 2001, p. 263).

With the publication of his seminar paper, the *New Physical Anthropology* (1951), Washburn issued a call to bring science and the scientific method to bear on the resolution of problems in anthropology. Taking inspiration from the earlier ideas of the agents of the Modern Synthesis, including Dobzhansky (1937), Fisher (1930), Haldane (1932) Huxley (1940, 1942), Mayr (1942), and Wright (1931), Washburn set out to initiate a transition in anthropological thought from a focus on measurement and classification toward a focus on the processes and mechanisms of evolutionary change. He argued for multi and interdisciplinary approaches to teasing out the complex interactions between human behavior, biology, and history and, so, made a call for a practice that integrated both biology and anthropology (Fuentes 2010). In keeping with the anti-establishment opinions that were forming within anthropology's other subfields around this same time, Washburn (1951) admonished physical anthropology for its dependency upon tradition: what he saw as a long-held, uncontested, and non-experimentally verified "group of attitudes and assumptions" (p. 124). In much the same spirit that Willey and Philips (1958) levied their criticism against archaeology, arguing that it "lacks a systematic body of concepts and premises"(p. 1), Washburn (1951, p. 124) also identified an absence of developed theory in physical anthropology. Drawing on this absence, he contested the dominant attitude that underpinned physical anthropological scholarship and rendered the methods and opinions of its practitioners as "static, with emphasis on classification based on types." The significance of Washburn's critique is evidenced not only by its effect on the views of his colleagues in physical anthropology, but

also in its pan-disciplinary contributions. The impact of Washburn's revolutionary ideas on the other subfields is palpable: his argument precipitated the change, for example, in archaeology that led to Binford's (1968) development of processual thought and the emergence of archaeological systematics under the rubric of "New Archaeology." In contemporary anthropological research, it has also served as the inspiration for Lett (1997) and Kuznar's (1997) arguments for "scientific anthropologists" and, particularly, for the integration of science into cultural studies.

Washburn's chief contribution to physical anthropology lies in the theoretical transformation that his ideas inspired. Taking his cue from what he called "the revitalized and revolutionized" study of human evolution, in which the mixture of genetic theory into paleontology and the development of systematics refocused scholarly interest on the processes and mechanisms of evolutionary change (influenced especially by the work of Mayr [1950] and Dobzhansky [1944, 1955]), Washburn argued that physical anthropology must undergo a similar redirection of its fundamental point of view: from technique to process (1951, p. 124). In refocusing physical anthropology's perspective towards biological theory and on studies of population genetics, Washburn revived the "old" physical anthropology as the "new." He saw the irresolvable problems of the "old" physical anthropology as rooted in the core of its methodology, especially due to the development of the conventional physical anthropological methods prior to the acceptance of many key evolutionary ideas and long before the post-war emergence of genetic theory. He argued that the application of measurements, computations, classifications and correlations was a self-limiting trend. The restrictions inherent in the old typological, classificatory methods narrowed the possibilities for the sort of questions that could be posed while they constrained the potential of the results obtained and ensured predictable conclusions that were presupposed by the techniques used. Physical anthropology, therefore, lacked an adequate theoretical foundation that

could elevate it from a technique to a legitimate field of research. Washburn sought to introduce what he called "process"to his "new anthropology." This transformation of static technique into dynamic practice contributed threefold to what he saw as a more holistic and relevant discipline. Firstly, it removed the problematic theories and assumptions that permeated early anthropological work from future inquiries, ending the cycle of misguidance, misinformation, and theoretical stagnation. In Washburn's words, "it is important for the young anthropologist to recognize the wolf even when appearing in a very peculiar kind of sheepskin" (1950, p. 437). Secondly, it recognized the points of merit within the preexisting corpus of anthropological work, encouraged the retention of these ideas, tenets and concepts, and facilitated their improvement within a revised methodological and theoretical framework. Thirdly, it permitted a better, more complete understanding of the trajectory and mechanisms of evolution and the problem of its interpretation in the context of modern human biology. This turn, allowed for the complex questions of human morphogenetic variation to be addressed more efficiently and the results obtained to be guaranteed by experimentally verifiable, biological mechanisms. According to Washburn (1960), biology and, especially, genetics would offer anthropology the tools and the theory to confront old problems but with a freedom from the assumptions and theories of the old techniques.

The differences in approach between what Washburn categorized as these "new" and "old" physical anthropologies are particularly visible in his study of race. Washburn employed the language of the growing field of population genetics (Dobzhansky, 1962) as he made his unequivocal statement that "races must be based on the study of populations"(1963, p. 521). Washburn (1963) argued that there is neither a scientific basis for racial discrimination based upon phenotypic typology, nor for the truthful categorization of breeding populations into the closed systems of "types." He also (1963, p. 523) located

the formation of biological races within the process of evolution (e.g., the areas of migration, drift, selection and mutation), as he argued that "[g]enetics show us that typology must be completely removed from our thinking if we are to progress." Washburn attributed a vital role in establishing the relationship between biology and culture to modern genetics. As he took the understanding of culture as an essential step in the course of teasing out what he saw as anthropologically relevant information from genetic data, his approach drew Dobzhansky's (1962, p. 18) ideas on "a feedback between biological and cultural processes" into anthropology. He thereby also reaffirmed and reemphasized the long-held tenet of anthropology as a discipline: that the individual can and should only be understood as both a biological and a social organism (Washburn 1963, p. 522). Moreover, his work supported a broader practical argument for increased communication between all kinds of anthropologists and the integration of methods and ideas from the subfields of anthropology and other scientific disciplines (Washburn 1951, p. 125). This revised practice and new set of goals brought into obvious question (1) the "old" anthropological concerns with anthropometry, and specifically craniometry, in the establishment of a human taxonomy and in the service of classification and (2) for racial studies, the practice of establishing essentialist types and the parceling of people into smaller and smaller racial packages. This new perspective also highlighted the fundamental problem with the "old" approach: the conventional preoccupation with divisive and classificatory techniques that emphasized unique traits, using them to differentiate between artificially constructed types, ignored the overlaps in the distribution of human variation and, so, both obscured the basic assumption that humankind is a unified species and impeded the study of the evolution of humanity as a single group. To this end, Washburn (1963) stated that "[t]he point here is that racial classification tells us very little. The classification poses problems, it does not solve them"(p. 526). In order to solve these problems, or to properly address the differences in living humans, Washburn (1951) advocated for a scientifically rigorous and consistent

analytical protocol: "1) diagnose the complex, 2) develop methods appropriate to describe variations within it, 3) discover the genetic background of these variations"(p. 129). He included in his 1951 paper the directive that "if traditional physical anthropology was 80% measurement and 20% concerned with heredity, process and anatomy, in the new physical anthropology, the new proportions may be approximately reversed"(p. 129). It is essential to note that Washburn stressed here the importance of the addition of genetics to anthropology, but he did not champion the abandonment of traditional physical anthropological techniques. Instead, he encouraged the smart use of these techniques in concert with modern science in a movement towards a focused research goal in place of the routine application of a technique. Through this transformation in approach, Washburn called for the stalled anthropology of the nineteenth century to assume its rightful position among the fields of twentieth century science as it adopted the study of process, supported by an attention to biological and genetic theory.

While Washburn's introduction of the "new anthropology" in the 1950s did inspire introspection into the traditional methods and theories that long defined the parameters of research in physical anthropology, it was not until the 1960s, with the mainstream change in climate on race and the concerted efforts to confront the misrepresentation of science in the service of discriminatory social policy, that the reality of race and the validity of the type concept in physical anthropology was seriously questioned (and repudiated). The tensions around this rejection of the race concept could now be met in a public forum. Washburn (1983) has explained how, for himself, his new perspective on human variation first emerged in response to the continued proliferation of race-based research by his mentor, Hooton. As described above, Hooton, early on, had gained a respected position at Harvard while clearly making use the "old" anthropological techniques to establish a scientific basis for racial profiling in

support of the eugenics movement. He later continued to carry forward this same typological and values-based tradition even into the mid 1950s, with, for example, his work on the morphological identification of nine unique Irish races, each of which were defined by their own set of biosocial characteristics (Hooton & Dupertuis, 1951, 1955; Hunt, 1981). It should not be surprising that it is at this same time that Washburn made his initial call for change. For his contemporaries, the broad-based paradigmatic shift from the "old" to the "new" anthropology likely spawned from both a combination of Washburn's developing arguments and the discourse on scientific racisms that emerged in reaction to the publication of segregationist literature that used the "science" of anthropology to undermine the civil rights movement in the early 1960s. Most notable of these racist tracks were Coon's *The Origins of Races* (1962), Putnam's *Race and Reason* (1962), Garrett's *The Equalitarian Dogma* (1961), and George's *The Biology of the Race Problem* (1962), which, in the tradition of the biosocial hierarchies (ranking Caucasoids > Mongoloids > Negroids) created by the cranial "science" of Morton (1849) and using misrepresentations of evolutionary concepts and ideas of human subspeciation, presented scientific "evidence" that justified racial inequality. Although Coon kept himself cloistered within academia, never taking a political stand on government or social policies, his passive contribution to scientific racism has been considered as deleterious as the accusations of hidden truths and leftist conspiracies loudly vocalized by contemporary segregationists. The Putnam Letters Committee or the Pioneer Fund, for example, made free use of his literature to bolster their arguments for racialized public policy: this was also a cause to which Coon was sympathetic (Jackson, 2001; Caspari, 2003). Wolpoff and Capsari (1997) contend that Coon's work was indeed overtly racist because of his "clear insensitivity to social issues, born of his attitude toward anthropology that was engendered by his social class and training" (p. 171). Coon, therefore, was guilty of a dangerous disregard for social responsibility when, for example, he presented "proof" for the argument that American Blacks were 200,000 years "less evolved" than their

White neighbors. This evidence was used to justify the history of African-American oppression and to sanction current efforts to uphold laws mandating segregation. In a need to "defend anthropology against what was seen as a very real segregationist threat to the discipline" (Jackson, 2001, p. 250) members of both the anthropological and biological communities united for what Jackson (2001, p. 248) has called "a scientific turf battle of sorts" that sought to ultimately bring about "the demise of the race concept" (Caspari, 2003, p. 66).

Dobzhansky (1961) published a review of Putnam's *Race and Reason*, entitled "A Bogus 'Science' of Race Prejudice" in which, against the "pseudo-science" of racist scholarship, he implored that "[g]eneticists and anthropologists may well give their immediate attention to the danger of misuse of their sciences for propagandistic ends" (p. 190). Willey, as the AAA president in 1961, offered the first official response to Putnam on behalf of the anthropological community, asking for the acceptance of a resolution stating that:

> [t]he American Anthropological Association repudiates statements now appearing in the United States that Negroes are biologically and in innate mental ability inferior to whites, and reaffirms the fact that there is no scientifically established evidence to justify the exclusion of any race from the rights guaranteed by the Constitution of the United States. The basic principles of equality of opportunity and equality before the law are compatible with all that is known about human biology. All races possess the abilities needed to participate fully in the democratic way of life and in modern technological civilization (Press Release on the 60th Annual Meeting of the American Anthropological Association, quoted in Jackson [2001, p. 263]).

This resolution was unanimously passed with a vote of 192-0. One year later, Washburn (1963 [1962]) followed suit and, as the current AAA president, delivered a public address at the AAA's 61st annual meeting that openly targeted the racist rhetoric presented as "science" in Coon's newly published book. These early responses set the stage for later collaborative efforts, such as the 1966 American Association for the Advancement of Science Symposium spear-headed by Mead and Dobzhansky (and its related edited volume, [Mead et al., 1968]). These efforts worked to dispel misconceptions about the scientific basis of race, identify the misrepresentations of real science and the truth of its findings, and resolve the confusion over the patterns of observed human morphogenetic diversity. The work of Livingstone (1962) and Brace (1964) also did much to stress the inadequacy of the race concept in explaining the distribution of this variation, emphasizing the clinal and geographic-based expression of phenotypic and genetic traits and arguing, in turn, that it is "extremely difficult to say where one population ends and another begins"(Brace, 1964, p. 104). The continued influence of the Modern Synthesis and the largely genetic focus of its architects and proponents helped to replace the typological studies of race with a fresh perspective for understanding human biological differences brought about by a new focus on open biological systems, called populations. Practitioners with an interest in genetics and evolutionary mechanisms found themselves at the forefront of anthropological discussions, putting forward the data and ideas that would come to cement the ascendency of the "new" anthropological practices inspired by Washburn. Molecular studies, in particular, came to provide the compelling evidence in support of the argument for the non-existence of races that would drive the research design of future studies in human phenotypic, and especially skeletal, variation. It is at this point, with this "demise of the race concept," the rise of populational thinking, and the emergence of population genetics as a key partner-field to biological anthropology, that the historical period of physical anthropology comes to a close. All subsequent studies have direct, rather than historical, bearing on

how research is conducted today, guiding the questions asked, the methods choosen and the data accessed, and what findings and interpretations are considered to be valid, informative, and contributive to the further advancement of the discipline.

Race at the Advent of Genetics

"The development of mathematical genetics in the first quarter of the twentieth century proved a triumph of reason over invective, effectively uniting Darwin's arguments with Mendel's observations. It provided a powerful heuristic for evolutionary biology and set the stage for the New Synthesis, while generating the foundations of modern population and quantitative genetics." (Waser, 2000, p. 510)

The realization of the New Physical Anthropology in its full form in the 1960s and 1970s and the concomitant rejection of the biological race concept by its proponents owed many of its key developments to the recent availability of more sensitive technologies for obtaining molecular data and the advanced quantitative methods for analysis, which mid-century research in human genetics was only beginning to yield (Cavalli-Sforza, 1964; Edwards & Cavalli-Sforza, 1965; Cavalli-Sforza & Edwards, 1967). Yet the role that progress in mathematical genetics played in the time leading up to this scientific "revolution" and disciplinary "revitalization" cannot be overlooked, especially given the lengthy relationship that existed among the often uncomfortable triad of statistics, race, and physical anthropology. With Huxley's (1938, p. 219) introduction of the concept of the cline (as a "gradation in measurable characters"), phenotypic and genotypic traits were reconceptualized as morphological

measures and allele frequencies distributed over a geographic space. Interpretations of human population structure were no longer constrained by ideas of discrete systems and, therefore, could be reconceived within the framework of continuous models. As a result, the static 19th century concept of race was increasingly seen as both an inadequate and misleading explanation for the patterning of human morphogenetic variation, whose observed complexity continued to increase proportionally with the advancements in data acquisition and analysis. Montague, by 1941, had already challenged the core assumptions of the biologically deterministic understanding of race by invoking Mendelian principles, particularly the argument that traits are not transmitted as complexes of characters. Boyd's publication of *Genetics and the Races of Man* in 1951 first introduced the study of blood groups as a precise and objective approach to the explanation of population relationships. In replacing the historical-descriptive methods of skeletal biology with the complex analyses of population genetics and focusing on the use of traits of "known" inheritance, Boyd is said to have declared osteological research as "passé" (Armelagos and Van Gerven, 2003, p. 57) and his output has been called an "obituary" on bones (Armelagos, Carlson and Van Gerven, 1982, p. 310). In 1953, Wilson and Brown made "the case against the trinomen" rejecting the concept of subspecies in biology using clinal arguments for frogs. Within the formative period of the New Anthropology, Livingstone (1958, 1962) extended this clinal theory to human biology, offering a compelling argument for the unracialized view of human variation using genetic data on the disease prevalence of sickle cell anemia. In indentifying a gradual change in frequency of the sickle cell gene over a broad geographic area (including the Mediterranean, Africa, and South Asia), he demonstrated how continuous clinal variation could be observed within populations and across predefined racial boundaries. From this he concluded, now famously, that "[t]here are no races, there are only clines" (1962, p. 279). Ehrlich and Holm (1964) offered further support to Livingston's theory by revealing that when two or more clines are observed, the patterns of the morphogenetic measures,

including the racial traits of conventional wisdom which had previously enabled race-based differentiation, are discordant. It was demonstrated that, for example, the distribution of melanin can be mapped following a north-south geographic trajectory, while patterned frequencies of the haplotype for beta-S hemoglobin appear to radiate out of and spread through specific geographic regions. As defining traits failed to covary in their geographic distribution, different taxonomic organizations would be expected depending upon what number or combination of traits or measures were used (an idea already introduced by Livingstone in 1962 and later discussed as trait discordance by Wright, [1978, p. 449-50]). The long-established definitions for individual races would not necessarily correspond, therefore, to the biological groupings that were variably produced. For example, a single defining trait, like skin reflectance, could yield a partitioning system that corresponded well to races but, as uncorrelated traits are added into the analysis, the discordance between their individual clinal patterns would make any translation of the final ordering of this combined variation into the language of the conventional system of racial taxonomy increasingly difficult if not impossible (see Newman, 1963). Brace (1964), in turn, advanced a "nonracial" approach to human variation that made use of the concept of clines but, considering his method's susceptibility to multi-trait discordance, he advocated for the evaluation of one trait at a time. Taking up a study on the distributional patterns of many of the traditionally observed phenotypes that had been consistently used to construct racial stereotypes, Brace (1964) argued: (1) that independent evolutionary hypotheses could be developed and tested for the origin and distribution of each trait, (2) that, when clinal maps are overlaid, trait distributions overlap and traditional racial boundaries are obscured, and (3), ultimately, because these clinal patterns are not congruent and do not covary, biological races, as they had been previously conceived within anthropology, simply do not exist. At this same time, a revised approach to the interpretation of variation was being aggressively pursued in light of both the novel forms of morphological expression

that were being increasingly observed in recently discovered fossils and the expanding collection of comparative morphogenetic information on human archaeological, historic, and modern populations. This new information brought together mounting concerns for issues of process, mechanism, and function with budding ideas of contact, gene exchange and admixture in order to develop more holistic and testable explanations for human variation and the trajectory of human evolution. These emergent lines of research therefore engendered the concern that "[t]he historical beginnings of anthropometric interest in the dentition and cranial and facial characteristics implanted a conceptual orientation which is seriously in need of re-evaluation" (Horowitz & Osborne, 1971, p. 189). As more and more discrepancies in the traditional taxonomic rulebook were identified through the comparison of trait distributions and as the once straightforward boundaries of the "old" races that were found to no longer account for the complex and overlapping patterns of variation being discovered for the "new" populations, it was increasingly argued that race contributed only poor and confused explanations for the patterning of world-wide biological diversity observed among and within groups. Race, revisited in the context of Washburn's "manifesto for the modern era" (Armelagos & Van Gerven, 2003, p. 57), became, therefore, obsolete and could no longer be considered as a useful explanatory concept in this circle of "new" anthropologists.

Even after the outward rejection of "racism and hierarchy of races" by the "majority of scientists" in the second half of the 20th century (Lieberman & Reynolds, 1996; Lieberman, 2001, p. 72) and despite the shifts in ideology, philosophy, and methodology among "mainstream" anthropologists by the late 1960's (Caspari, 2003, p. 73), fears that physical anthropology would continue as "merely the handmaiden to human history" nevertheless persisted (Lasker, 1970, p. 2). Even in the evolving research climate of the "new" anthropology in the early 1970s, many physical anthropologists held fast to vestiges of the "old"

practices of typological research. Preeminent scholars, such as Howells, displayed an initial resistance to change by reaffirming a traditional scholarly position: "[m]y purpose is not the study of growth but of taxonomy, of the variation between existing recent populations in the dry skull" (1971, p. 210). Others retained typological research under the guise of race-reconciliation, in which ideas were amalgamated (e.g., Brues, 1972) or the appropriate new terminology was simply substituted for the old (e.g., Boyd, 1950; Garn 1962). Although the New Anthropology had divorced itself from scholarship motivated by racist agendas, considerable weight continued to be given to description, identification and classification even when genetic traits were used in place of typical racial classifiers and when the concept of race was exchanged for that of the population. Armelagos and Van Gerven (2003) have identified a concomitant and widespread failure to embrace the "new" practice of placing findings within a larger theoretical context as the predominant shortcoming in the transition from old to new thinking on variation and its relationship to the concept of race in these early years. Although this is likely true, in part, and some comfort may certainly have remained in keeping to the old ways, perhaps the greatest limitation of the New Anthropology was that it had yet to make use of the multidisciplinary scientific approach it had advocated to provide empirical evidence of the non-existence of races, including the use of rigorous testing, powerful analytical tools and the quantitative morphogenetic data of mathematical genetics. Although the contemporary research that classified individuals by race using typologically derived distributions of either phenotypic or genotypic data was openly criticized for being grounded in outdated ideas, for making use of poor methodology and for (necessarily) yielding highly speculative results that were biased and prone to error, the arguments emerging out of the "new" approach to understanding human biology had yet to provide tangible proof against a structured-view of human variation. Evidence for this was thus far limited to a collective recognition of two fundamental weaknesses in the race-based approach to research design, execution and interpretation. Using the

theory developed around the concepts of clines and populations, researchers in support of the anti-race movement concluded that, clearly: (1) the conventional interpretation of race did not correspond well, if at all, with the natural dispersion of human biological variation and the ways morphogenetic traits were now thought to be distributed among individuals, and (2) the factors used to, first, define and, then, identify racial types actually reflected taxonomic preconceptions (or typological misconceptions) of group differences instead of biological reality. It was not until 1972, however, with Lewontin's publication in the annual review, *Evolutionary Biology*, that the full power of quantitative genetics was first brought to bear on the question of race, so that the largely anecdotal conclusions on the amount of biological variation existing among populations drawn in theory were given the weight of variability measures, probabilities and proportions from established statistical practice. As Powell and Taylor (1978) have argued, there are three questions that can be asked about human race:

> (1) "Are there genetic differences between human races?"
>
> (2) "What is the magnitude of the differences?"
>
> (3) "How does the magnitude of difference between races compare with the genetic diversity within races?" (p. 1139).

It was Lewontin's genetic work on population that provided answers to the questions that potentially held the most explanatory value for and influence on group differentiation practices. In recognition of this, Powell and Taylor (1978) have acknowledged that "[t]he relative importance of classification (if indeed there is any) depends largely on questions 2 and 3, and these are the ones addressed by Lewontin (1972)" (p. 1139).

In his seminal article, "The apportionment of human diversity," Lewontin (1972) made the first methodologically rigorous attempt to test the question of the non-existence of human races by applying

measures of genetic variability within and among modern human populations and determining statistically the extent to which traditional racial categories actually account for the biological variation present within the human species. Using distributions obtained from preexisting studies on select biochemical markers for world-wide samples (including nine blood groups as well as eight serum protein and red blood cell enzyme variants) and applying the Shannon Information Measure (taken as a measure of the frequencies of the alleles at given loci; following Shannon and Weaver [1949]), he evaluated the pattern of diversity in allele frequencies for seven "races"(Caucasian, African, Mongoloid, South Asian Aborigines, Amerinds, Oceanians, Australian Aborigines) and the "populations" that comprised them. His study addressed, therefore, two topics that held both popular interest and were of considerable theoretical concern: (1) the percent variation at two nested sub-levels of population structure and (2) the quality of information retrievable from the available molecular data sources (including genetic system proxies for genotype). Lewontin's (1972) publication offered conclusions that would have a lasting impact on how human variation would be studied even as it established the conceptual parameters that determined how such variation should be perceived. He provided both the specific values that would be cited as irrefutable fact across many fields of inquiry well into the present day (c.f., Block, 1995; Andreasen, 1998; Richardson, 2000; Brown & Armelagos, 2001; Machery & Faucher, 2005; Race, Ethnicity and Genetics Working Group, 2005; Mallon, 2006) as well as the broad theoretical findings on biological diversity that would be corroborated by recent studies using comparable models of population structure but with varying data types and methods (e.g., Michalakis & Excoffier, 1996; Jorde et al., 2000; Romualdi et al., 2002; Hinds et al., 2005; Li et al., 2008). Lewontin observed that the amount of variation among major geographic groups (i.e., partitions corresponding to conventional racial classifications) was remarkably low when considered in relation to the total diversity present within the single human species. He reported that the interracial component of the total genetic

variation accounted for only 6.3 %, the component for populations within races accounted for only 8.3%, and the within-groups or individual differences accounted for the majority of the total genetic variation at 85.4%. From these results, Lewontin (1977) concluded for the study of population sub-structure that "[t]he taxonomic division of the human species into races places a completely disproportionate emphasis on a very small fraction of total human genetic diversity" **(**p. 432).

It is not surprising that in this ongoing period of pan-disciplinary discontent over mandates on best-practice in the study of human variation and the future direction of this scholarship, the introduction of scientific data that supported the philosophy of the New Anthropological cause incited arguments from those resisting change in the traditional values that had been long-shared between anthropology and human biology. Shortly after the release of Lewontin's original findings, criticisms abounded and his results were "met with disbelief" (Brown & Armelagos, 2001, p. 35). The questions of validity that emerged were focused most heavily on the on the concern that "the detectability of the gene differences may be different" between the kinds of gene frequency data analyzed (Nei & Roychoudhury 1974, p. 421). It was therefore posited by those invested in maintaining an anthropological focus on taxonomy that the proper choice of data would yield the correct, expected results: results which sustained the conventional and long-standing view of racial differences by locating the true (high) extent of among-group variation (Brown & Armelagos, 2001). These arguments targeted, specifically, the markers that Lewontin used to derive the percent values, as well as the method of their "averaging." The critique was grounded in other, earlier, analyses which had yielded fluctuating values, including, for example, estimates of average heterozygosity for protein loci in the range of 7%-10% (Harris, 1969; Nei & Roychoudhury, 1972; Harris & Hopkinson, 1972) but a higher average estimate of 16% for blood group loci (Lewontin, 1967). Critics concluded, therefore, that the proportions of variation that were reported

in Lewontin's 1972 publication were unrepresentative of the true levels of heterozygosity because the methodology employed was greatly influenced by the choice and treatment of the data. These arguments were quickly deflated, however, as Lewontin's analyses were soon successfully replicated and convincingly validated by a host of studies which adopted similar approaches but sampled different data, while still yielding comparable results. In 1980, Latter made use of published data on 18 genetic systems (like Lewontin, using blood groups [10], serum proteins [3], and enzymes [5]) to quantify biological diversity at three levels of population structure: six "major human subgroups," the subdivision of these "subgroups" into "regions," and the further division of these "regions" into "populations." For this study, Latter made use of the Shannon Information Measure, as Lewontin did, but also applied Hedrick's (1971) genic similarity measure of the probability of non-identity in genotype for two randomly sampled individuals as well as his own measure of the proportion of genes shared between two individuals sampled at random. After considering three measures of diversity, Latter determined that within-population variation accounted for 83.8% to 87% of the total genetic diversity, while 7.5% to 10.4% was accounted for by "major geographical groups"(comparable to geographic races), and the remaining 5.5% to 6.6% by inter-regional and inter-population differences. Although Latter's research did show some discrepancies in the percentages calculated that were linked to the type of marker assayed (i.e., blood group versus enzyme), these differences were found to be small and, so, did not significantly affect the final interpretation or quality of his results. This insignificance of inter-study differences (and, similarly, the intra-study fluctuations) in estimates was recently restated by Long, Li and Healy (2009) when comparing the percent values for the component of diversity between major geographic groups obtained by Rosenberg et al. (2002) at 4.3%, and Excoffier and Hamilton (2003) at 9.2%, with the gold-standard, that is, 6.3%, estimated by Lewontin (1972). In 1982, Nei and Roychoudhury assessed the genetic variation within and between three world-wide groups, Caucasoids,

Negroids and Mongoloids, which they considered to be representative of "the three major races." Previously, they (1973) had set a precedent for the definition of "macrogeographic race" as the distribution of genetic variation in Old World populations across Africa, Asia, and Europe; partitions which saw continued use in the work of Nei and Livshits (1989). In this later study, Nei and Roychoudhury analyzed data from 23 blood group loci and 62 protein loci and generated estimates of the proportion of the total genetic variation attributable to racial classifications that were in line with the by-system results produced by Latter (1980) and were comparable to the average values provided by Lewontin (1972). They reported estimates of 9% to 11%, and, when partitioned by data type, 8.8% for the protein loci and 10.9% for the blood group loci specifically. Ryman, Chakraborty and Nei (1983) also offered support for the repeatability of these findings, successfully replicating earlier results using two HLA types, nine blood groups, and fourteen proteins. What is interesting about this study is that their model for analysis considered population structure in two ways: (1) using the classic definitions of world-wide races, which followed the traditional biogeographic divides and (2) using algorithmically derived groupings, which were created from dendrograms of genetic relatedness and designed to maximize the differences among the groups. Even when inter-population differences were enhanced using this computational modeling method, only a slight increase in among-group differences was observed (on average, 1.3%). Their study found that, again, individual genetic differences within the populations account for the largest proportion of the total genetic variation, with a mean estimate of 86% and a range from 82.7% (for the blood groups) to 90.3% (for the proteins types). Subsequent analyses by, for example, Dean et al. (1994), Barbujani et al. (1997), Excoffier et al. (1992), and Seielstad et al. (1998) not only followed suit in yielding convincing corroborative results, but also contributed interesting new findings of their own. Dean and colleagues (1994) obtained smaller estimated values of among-group or interracial variation, ranging from 4.4% to 4.75%, when the number of RFLPs

(restriction fragment length polymorphisms) used was increased. In their analysis of 30 microsatellite loci and 79 RFLPs, Barbujani and coworkers (1997) failed to identify more than one locus for which individual differences accounted for less than 50% of the total variation. Finally, the studies by both Excoffier et al. (1992) and Seielstad et al. (1998) demonstrated that within-group variation still accounted for the greatest percentage of diversity (75%-80% and 81% respectively) even when mtDNA data was used. Their findings not only validated the conclusions drawn from the early biochemical (proxy-based) studies and were aligned with the newer results coming out of research at the DNA level, but further discredited the arguments for a scientific basis for race as mtDNA had been hitherto accepted as "the best marker for 'racial origin'" when considering biological data (Brown & Armelagos, 2001, p. 37).

With the growth of interest in applying the theory of population genetics to questions of the magnitude and relative proportions of among-population differences that followed Lewontin's seminal work came a rapid departure from the initial goal of the [in]validation of Lewontin's method and, instead, a move towards other complementary quantitative approaches. It was found, for example, that Lewontin's results could be alternatively duplicated by (1) considering the genetic data in the context of continental population hierarchy and then, (2) using a measure of among-group variation such as F_{ST} (after Wright 1951, 1965, 1969, 1978) or some closely related population structure statistic (Cockerham, 1969; Nei, 1987; Excoffier et al., 1992) to determine the proportion of total genetic variation attributed to differences between these world-wide groups (Long & Kittles, 2003; Jorde & Wooding, 2004). Wright's Fixation Index saw the greatest use (and remains very popular still) as it offered a convenient measure of the degree of subpopulation differentiation, taken as either a decrease in heterozygosity relative to that which would be expected given the assumption of random mating throughout the entire population

(Wright, 1951, 1965) or as a gain of homozygosity in the progression towards fixation (Wright, 1969, 1978). Research has shown that calculations of F_{ST}, measured among samples representing major global populations, consistently produce estimates at 0.15 or less (Long & Kittles 2003). As F_{ST} is constrained in its estimated value within the range of 0.0 (no differentiation) to 1.0 (complete differentiation, where subpopulations are fixed for different alleles), the average of 0.15 (15%) routinely obtained for world-wide inter-population variation suggests, in theory, that the nested hierarchical substructuring observed among human groups is only weakly genetically defined (see Wright [1978] for qualitative guidelines on the interpretation of F_{ST}). Although some single alleles are known to have divergent (higher or lower) values from this average, these have been commonly dismissed as both exceptional and insignificant when the practitioner considers, as the methods of analysis dictate, a large number of alleles or loci collectively (Bowcock et al., 1991, 1994). It is apparent, therefore, why F_{ST} has been seized upon in recent history as an especially useful statistical tool for making a quantitative argument against group partitioning along racial divides. Consider, for example, the partition of Old World populations into the three continental groups (Africa, Asia, and Europe), which correspond to the traditional macrogeographic definition of the "three major races" and have commonly served as the model of population structure for many of the key population genetic studies already discussed. It has been be demonstrated, using F_{ST} and a variety of data sets, that 85% and 90% of the total genetic variation is observed within these broad continental groupings, while only an additional 10%-15% is found among them (Jorde et al., 2000; Watkins et al., 2003; Jorde & Wooding, 2004). A sample of individuals drawn from a single continent should exhibit, therefore, around 90% of the total genetic variation found within the human species and only roughly 10% more variation would be detected if this sample consisted of African, Asian and Europeans (Jorde & Wooding, 2004). Taking up this line of evidence for questions of evolution, Owens and King (1999) have argued that because *any* "population" would contain

approximately 85% of the total human genetic variation and since no major genetic discontinuities across populations have been observed, "the possibility that human history has been characterized by genetically relatively homogeneous groups (‘races’), distinguished by major biological differences, is not consistent with genetic evidence" (p. 451). They have also offered commentary relevant to the debate over the utility of the concept of race to questions of contemporary bio-social substructuring, as they posit that:

> [v]ariation in traits, including skin color, popularly used to identify "races," is likely due to straightforward mechanisms involving limited numbers of genes with very specific physiological effects. Of course prejudice does not require a rational basis, let alone an evolutionary one, but the myth of major genetic differences across "races" is nonetheless worth dismissing with genetic evidence (p. 451).

The advantages of this mounting genetic evidence for the argument against race were apparent even early on in the history of this genes-race debate. Such evidence suggested that (1) the core biological differences detected between predefined populations were "rather small" (Nei & Roychoudhury, 1974, p. 421) and that, therefore, (2) human races, as they were traditionally and typologically defined, are genetically "remarkably similar" (Lewontin 1972, p. 397). These advantages were already obvious to Lewontin in 1972 when he contextualized his own findings:

> [i]t is clear that our perception of relatively large differences between human races and subgroups, as compared to the variation within these groups, is indeed a biased perception and that, based on randomly chosen genetic differences, human races and populations are remarkably similar to each other, with the largest part by far of human variation being accounted for by the differences between individuals. Human racial

> classification is of no social value and is positively destructive of social and human relations. Since such racial classification is now seen to be of virtually no genetic or taxonomic significance either, no justification can be offered for its continuance. (p. 397).

Equally obvious to Lewontin was the potential advantage of for marrying the "social" to the "sciences": this is apparent when he later takes on the topic of human biological diversity in ways that appear to be as much about making a statement on social responsibility as about addressing issues of biostatistical interest. In his 1974 book, *The Genetic Basis of Evolutionary Change*, Lewontin, for example, extended his argument to tackle the related issue of classification, writing that:

> [t]he apparent homogeneity within races as compared to the "obvious"difference between them stems partly from the fact that our consciousness of racial differences is constantly being reinforced socially because racial distinctions serve economic and political ends, and partly because the very characters we use to distinguish races-skin color and texture, hair form, eye, nose, and lip shape-are those to which we are most keenly attuned for the purpose of distinguishing individuals (p. 152).

In his book, Lewontin (1974) also positioned his own work within the broader developmental trajectory of the field of human biology by making the pithy remark that: "[i]ndeed the whole history of the problem of genetic variation is a vivid illustration of the role that deeply embedded ideological assumptions play in determining scientific 'truth' and the direction of scientific inquiry" (p.157). It was through weighty statements such as these, which were both grounded in the testable and repeatable output of statistical genetics and encoded with bio-political commentary on the role of science in promoting social justice, that race in biology was rejected and opinions on the non-existence of human races became instead the accepted biological fact. Lewontin's findings, accompanied by the strength of

corroborative studies that where chronologically extensive and spanned the range of available methodologies, gave the empirical evidence and, with it, the credibility of "scientific" support, necessary for making an irrefutable case against race and effectively championing the anti-race movement within human biology. This rejection of race not only influenced research design in the study of variation but also broadly affected pedagogical practices in human biology. Cartmill (1998) succinctly summarized the trends charted by Littlefield et al. (1982), explaining how "the 1970s had witnessed a sudden and widespread abandonment of racial classification in textbook presentations of human biology" (p. 654). This new philosophy, therefore, set the standards for what would be considered quality information for dissemination while it also defined the "fundamental" theory for use in instruction. Lieberman (2001, p. 72) expanded the boundaries of this phenomenon's reach, identifying a thoroughly widespread abandonment of the race paradigm among "most scientists," across "all disciplines," throughout the latter half of the 20th century. Changing trends in the varying acceptance of race as a biologically meaningful explanatory concept, specifically among anthropologists, and its frequency of use in their predominant pedagogical and research publications have been tracked temporally over the course of several surveys conducted by Lieberman and coworkers (Lieberman & Reynolds, 1978; Littlefield et al., 1982; Lieberman, Stevenson & Reynolds, 1989; Lieberman, Kirk & Littlefield, 2003; Lieberman & Kirk, 2004). These studies found, for example, that, while in 1985, 41% of responding physical anthropologists disagreed with the concept of biological human races (Lieberman, Stevenson & Reynolds, 1989), by 1999, this figure had grown to 69% (Lieberman, Kirk & Littlefield, 2003). This indicates a considerable (25%) increase in the rejection of race over a ten year period. Moreover, the results of their 1999 survey revealed that across all subdisciplines of anthropology, 91% of respondents agreed with the statement that "[h]uman biological variation is best understood in terms of continuous gradations (clines) not races."

This strong rejection of race suggests that a "mainstream" consensus among anthropologists was indeed established. This should not be surprising given that it was around that same time that both the AAA (1998) and AAPA (1996) issued their official (revised) positions on race, which sought to formalize opinions that represented "generally the contemporary thinking and scholarly positions of a majority of anthropologists"(AAA, 1998). Only a few points need to be drawn from the "AAPA Statement on Biological Aspects of Race" in order to see how the disciplinary change in the status of race in anthropology was well underway. In this statement, race was formally rejected as a useful explanation of human variation in both theory, with the codification of an *anti*-race concept for understanding human evolution and the distribution of diversity, and in practice, with the abandonment of classification-based approaches using measures of biological substratification. As the text states:

> [t]here is great genetic diversity within all human populations. Pure races, in the sense of genetically homogenous populations, do not exist in the human species today, nor is there any evidence that they have ever existed in the past... For centuries, scholars have sought to comprehend patterns in nature by classifying living things. The only living species in the human family, *Homo sapiens*, has become a highly diversified global array of populations. The geographic pattern of genetic variation within this array is complex, and presents no major discontinuity. Humanity cannot be classified into discrete geographic categories with absolute boundaries. Furthermore, the complexities of human history make it difficult to determine the position of certain groups in classifications. (AAPA, 1996, np).

The content of the AAA's and AAPA's published position statements on race, combined with their own observations on the opinions of their colleagues have led Lieberman, Kirk and Littlefield (2003, p. 112) to

conclude that "the paradigm of race has approached the point where its survival is in doubt. Even those who continue to use the concept have serious doubts about its utility." The confluence of the ideas driving the New Synthesis and the New Anthropology with the scientific findings of population genetic research, which convincingly argued for weak population substructuring and the non-existence of biologically defined races, resulted in what Brown and Armelagos (2001, p. 34) have called a 30 year "assault on the race concept." This movement began in earnest the 1970s and continues still today within what many would consider "mainstream" anthropological thought, practice and pedagogy (Caspari, 2003, p. 73). It is essential to realize that for physical anthropology, in particular, this movement towards the rejection of race also had (and still bears) considerable effect upon the validity of any population based studies. If individuals in one group are shown to be more similar in general to individuals from another group than to other members of their own group, then the definitions given to "populations" also fail in explaining the true patterning of human variation. Populations can be said to be just as biologically meaningless and obfuscatory to the study of the mechanisms of evolution as races. In Brace's (1964) words, populations, represent "the last bastion of the 'type concept'" (p. 314). It is often argued, therefore, that any biological information captured by race, or the related concepts of ethnicity or ancestry, is necessarily confounded by personal ideologies and flawed methodologies and that, consequently, there is no possibility for an unbiased explanation of human variation. This critique of racial biology has moved well beyond an "assault on the race concept" to an inquiry into the validity of any study that uses morphological variation as a way to explore relationships among human groups. Considerable tension has developed between those biological anthropologists who use skeletal markers of variation as a means to access past population structure in phenotypic distance studies and estimate individual ancestry in forensic contexts, and those who dismiss any discussions of group differentiation as artifacts of biased ideological positions. This continued and expanding conflict stresses the

importance of reevaluating the question of identifying biologically meaningful patterns of human variation. Any understanding of the contemporary place of "race," Race, RACE, *race* or race, as well as the importance of its proxies, related sub-groups, and linked terminology (e.g., populations, ancestries, ethnicities), within current physical anthropology requires four central components: (1) an explanation of how population substructuring has been treated throughout the recent period of post-racial scholarship and how it is currently viewed in research and practice, (2) a reassessment of the evidence that supported the rejection of race in skeletal studies and, specifically, in craniometrics throughout this 30 year period of assault, (3) a clarification of why and how ideas on race and the concept of ancestry have been maintained within the subfield of forensic anthropology and how they have emerged as key factors in the biomedical sciences, and, finally, (4) a discussion of how very recent advancements in genetic research, in the forms of genome wide association studies (GWAS) and ancestry informative markers (AIMs), are challenging the long-standing view that races, populations, and ancestry or ethnic groups are not biological realities. Such research suggests the possibility that these divisions may be useful in understanding the patterning of human variation, to individual identification, and to making practical contributions in clinical, forensic, and legal contexts. The questions, controversies and trends that are important to the current anthropology of human skeletal and genetic variation are addressed in the discussion that follows in Chapter 2.

Chapter 2: Human Variation

Race in the Genomic Age

"If my theory of relativity is proven successful, Germany will claim me as a German and France will declare that I am a citizen of the world. Should my theory prove untrue, France will say that I am a German and Germany will declare that I am a Jew." (Einstein, 1922, np)

In his Paris address to the French Philosophical Society at the Sorbonne, Einstein (1922) not only made a self-reflexive statement on the uncomfortable negotiation of his identity, speaking to the concerns over race that were felt ubiquitously throughout the troubled political times of northern Europe in the period following the Treaty of Versailles and preceding WWII, but also brought to the fore the complex system of social hierarchy that dictates the conditions of individual membership within some formally named or predefined groups, like nationalities, populations, races and ethnicities. There are two points that can be made here in light of Einstein's remark, both of which are especially important to an anthropological discussion of human biological differences: the first point addresses the criteria for group membership and the second deals with the promise of genetics in the explanation of human variation.

(1.) As it has already been shown for the meaning of race, the definitions given to race as a concept or to races as individualized partitions nested within the human species are inconsistent. Although human variation has been shaped by the mechanisms of evolution and the long-term processes of population history and even though this history can carry quantitative information on shared biological variation or "ancestry" (Kittles & Weiss, 2003), since race nevertheless remains elusive, the criteria that define

variation-based categories are malleable, membership in one group does not preclude membership in several, and the transition among or in and out of many groups may be fluid. For this reason, the standards invoked for allocation of membership or self-identification are most often situationally dependent so that a thoughtful understanding of any perceived patterns of association requires the external observer to consider the context of the classification carefully, with a foreknowledge of the statistics, conditions, and agendas that underlie its making. Moreover, as can be recalled from the earlier discussion of clines, the core principles of this theory further confound any straightforward assignment based upon a direct linkage between an observed trait complex and a predefined group. For human variation, it is argued that the distribution[s] of the biological traits that have been conventionally taken to be characteristic of distinct groups do not conform to the type of patterning indicative of strong and meaningfully organized separations among these groups. They, in turn, fail to support the world-wide existence of discrete morphogenetic entities: the basic condition necessary to sustain the traditional concept of biological race. These traits are rarely, if ever, uniquely constrained to one group but are instead distributed gradationally across groups. For example, traits may be concentrated geographically within one region but are dispersed gradually over neighboring areas without sharp discontinuities and are shared, albeit in different and often lower frequencies, by other contact groups (Barbujani, 2005). When human variation is taken to be clinally distributed (with patterns of global gradients of human diversity) or when the biological evidence suggests character overlaps within and among groups, the observed patterns of morphology are often explained as either reflecting localized gene flow and isolation by distance (IBD), as successive founder effects or as drift resulting from migration events throughout human evolutionary history, and, to a lesser extent, selection in response to environmental conditions (Cavalli-Sforza, Menozzi & Piazza, 1994; Relethford, 2004, 2010;

Roseman, 2004; Serre & Pääbo, 2004; Prugnolle, Manica & Balloux, 2005; Manica, Prugnolle & Balloux, 2005; Prugnolle et al., 2005; Ramachandran et al., 2005; Betti et al., 2010; Jantz & Logan, 2010).

(2.) Thinking retrospectively on Einstein's remark and placing it in its milieu of early modern science and politics, it becomes apparent why Whitmarsh and Jones (2010) make the analogy that, "[g]enetics has been promised to be, for the twenty-first century, what physics was for the twentieth," so that the rise of genetics, like the rise of physics,"captured the public imagination and offers solutions to an astonishing range of problems" (p. 2). Because of this, expectations and predictions for success are lofty. It has been said that "[g]enetics will decipher the mysteries of evolution, improving our knowledge on the history of life on earth and of the human race" (Whitmarsh & Jones, 2010, p. 2) and even that the "new genetics" will become the "apotheosis of modern rationality" (Lindee, Goodman & Heath, 2003, p. 4) so that a thing of nature "will be known in such a way that it can be changed" (Rabinow, 1996, p. 93). Most importantly, a high scientific and public profile has emerged alongside the developments in the field of genetics. These profiles, when taken singularly and combined, have had what many scholars and news makers consider to be both the single greatest and most unanticipated impact on race. Their influence is palpable not only in ideas information value but also in terms of the immediacy and breadth of interest shown by specialists and lay people alike (Koenig, Lee & Richardson, 2008). The potential pan-disciplinary effects on understanding race can only be imaged, when the abundance of new, globally accessible genetic information is considered. Recent international efforts to decode the information embedded in the human genome and the free availability of molecular resources are initiating enlightened conversations on human diversity and engaging researchers world-wide. This can be seen, for example, in the discourse surrounding the National Center for Biotechnology Information (NCBI) Human Genome Resources public website. The profound implications of genomic research were

already recognized at the inception of the Human Genome Project (HGP), when the National Research Council (NRC) made the following statement in their 1988 report: "[w]hatever its scientific merits, a concerted effort to map and sequence the human genome would have profound social significance. Human beings are fascinated with the reasons we are what we are, both for what those reasons tell us about ourselves and for the insights they give us into those around us" (p. 100). The National Human Genome Research Institute's (NHGRI) Ethical, Legal and Social Implications (ELSI) Research Program was founded soon thereafter, in 1990, as an integral part of the HGP (NHGRI-ELSI, 2010). Its charge was (and still is) to address the new biocultural issues that were emerging concurrently with the advancements made by the HGP by identifying the immediate areas of concern and the "grand challenges" (Collins et al., 2003) facing future genomic research. Of particular note is the inclusion of Grand Challenge III-2: "**t**he impact of genomics on concepts of race, ethnicity, kinship and individual and group identity**"** (NHGRI-ELSI, 2010). With the completion of the HPG at the beginning of the new millennium (International Human Genome Sequencing Consortium, 2001) considerable focus has been placed on the findings that stress human similarity and unity while promoting the "constructions of scientific legitimacy," which, in turn, have reopened the realities of race for discussion (McCann-Mortimer, Augoustinos & Lecouteur, 2004). This new genetic data has become the "sounder" source of evidence for many scholars and is even regarded as the definitive source of scientific proof that all human beings are remarkably similar (Marks 1994). The knowledge that differences among individuals amount to only 0.1% at the DNA level and that no genetic features are unique to any one population, race, or ethnicity, has been adopted as *the* confirmation of the conclusion that races are social classifications (Foster and Sharp 2003). In his White House speech announcing the completion of the "working draft" of the Human Genome Project, Venter himself (2000) confidently declared to the watching American nation, and by extension to the world, that "the concept of race has no genetic or scientific basis" (np). The HGP's

potential for enormous public impact is obvious when the optimistic remarks delivered by President Clinton (2000) at that same time are recalled:

> [a]fter all, I believe one of the great truths to emerge from this triumphant expedition inside the human genome is that in genetic terms, all human beings, regardless of race, are more than 99.9 percent the same. What that means is that modern science has confirmed what we first learned from ancient fates. The most important fact of life on this Earth is our common humanity (np).

In general discourse, these sentiments expressed against the biological concept of race remained equally as strong even when reconsidered in hindsight. One year after the initial excitement of the project's completion, Collins and Mansoura (2001) wrote in reflection upon the outcome of the HGP that "those who wish to draw precise racial boundaries around certain groups will not be able to use science as a legitimate justification" (p. 221). In actual practice, however, the success of the HGP renewed and intensified interest in identifying inter-individual and inter-population, -race or -ethnicity differences (Foster and Sharp 2003). This resulted in the fast-paced production of genetic resources (e.g., GenBank, OMIM, HapMap databases) that were intentionally designed to aid in the study of human biological variation in allele or trait frequencies (Benson et al., 2008; McKusick, 2007; Thorisson et al., 2005) and to address questions concerning genetic variants associated with, for example, disease susceptibility (Collins & McKusick, 2001), environmental response (Olden & Gutherie, 2001), and drug metabolism (Nebert & Menon, 2001). This practice has had a notable effect which extends beyond the bounds of individualized areas of scholarship to influence many broader perceptions of human variation. It has been observed that there is a tendency to collapse social and biological definitions of populations and to overlook the nuanced study-specific language that qualifies a discussion of biosocial group separation and informs an understanding of aggregations of like-individuals and the distance measures

that quantify the difference between them (Hubbard & Wald 1993; Nelkin, 2001; Condit et al., 2004; Dubriwny, Bates & Bevin, 2004). Similarly, genetic information is commonly afforded greater authority than any social information (Nelkin & Lindee, 1995). Genetic data linked, therefore, to concepts of *a priori* partitioned groups (e.g., races, populations and ethnicities) can encourage "the reductionary reconfiguration of those categories along simplified biological lines" (Foster & Sharp 2003, p. 845). Group-targeted research and journalistic misrepresentations of molecular findings work, however unintentionally, to reify race-based thinking for the human biological sciences and may ultimately initiate a new era of racialization (Kahn, 2005; Lynch & Condit, 2006; Condit, 2007). This awkward relationship between discourse and practice in both the academic and public arenas, coupled with the high level of difficulty in interpreting the many diverse and unstable social categories and the complexity of genetic data, may have led to an increased frequency in the use of racial terms in scientific publications. This is a phenomenon recently documented by Gissis (2008):

> Based upon my findings I argue that the category of race only seemingly disappeared from scientific discourse after World War II and has had a fluctuating yet continuous use during the time span from 1946 to 2003, and has even become more pronounced from the early 1970s on (p. 438).

A similar problem of disconnect has been observed in anthropology. It has been shown that scholarly publications, opinion pieces, pedagogical discourse, and professional rhetoric all present inconsistent views on the interpretation of human variation (Lieberman, Stevenson & Reynolds, 1989; Armelagos & Goodman, 1998; Wienker, 2001; Lieberman, Kirk & Littlefield, 2003; Lieberman and Kirk, 2004; Lieberman et al., 2005; Hart & Ashmore, 2006; Štrkalj, 2007), especially as it pertains to human identification methods in forensic casework (Sauer, 1992; Kennedy, 1995; Goodman, 1997; Smay & Armelagos, 2000; Brown & Armelagos, 2001; Armelagos & Van Gerven, 2003; Williams, Belcher &

Armelagos, 2005; Albanese & Saunders 2006; Gordon 2009). It has been posited by some that this inconsistency may mask a wider chasm in thought that has been obscured by concerns for racialized science and scientific racism, so that current positions on variation were made (and are kept) mainstream less by compelling arguments against the utility of measured human variation and more by the politics and peer pressure of the times (Klein & Takahata, 2002). Among the physical anthropological community, strong and telling opinions on this issue have emerged accordingly. Gill (2000), for example, has offered the charged remarks that:

> [a]t the beginning of the twenty-first century, even as a majority of biological anthropologists favor the reality of the race perspective, not one introductory textbook of physical anthropology even presents that perspective as a possibility. In a case as flagrant as this, we are not dealing with science but rather with blatant, politically motivated censorship (np).

Similarly, Harpending (2000) has provided commentary on the dichotomy of this issue felt among "most professionals familiar with the literature on human differences and most natural scientists," as he argues that the non-existence of races is commonly understood as an "absurd fantasy," but nevertheless "[a] poll about views of race would be like a poll about Marxism in East Germany in 1980. Everyone would lie" (p. 85).

A multi-field discipline, like anthropology, is not only comprised of internal factions with divergent lines of study and espousing often equally disparate core philosophies, but also draws in, and is understandably influenced by, external collaborators who contribute to realizing multi-focused research goals. It should not be surprising, therefore, that the rush of new genetic information from the 1970s onward only helped to widen the theoretical divide over the conceptualization of human biological

variation and to fuel the debate over the scientific validity of partitioning human variation into biologically defined (e.g., phenotypic-trait based) groups. Arguments rooted in the often obvious, but always subjective, patterns of difference that were observed mophoscopically and morphometrically were increasingly weakened by the empirical gene-based knowledge that emerged with Lewontin (1972). These challanges continued to mount with the introduction of new markers and methods, and culminated with the HGP findings indicating that, genetically speaking, humans are 99.9% the same. Bolstered by this evidence, an aggressive stance against race has taken hold of the debate and has remained the dominating force in recent decades (c.f., Lieberman, Reynolds & Stevenson, 1989; Appiah, 1990; Hull, 1998; Brown & Armelagos, 2001; Haslanger, 2005). This is often presented as both the mainstream and the only rational position:

> [c]ontemporary genetic studies of human populations have revealed that there are no alleles distinctive of this race or of that, and, although a few researchers like J. Philippe Rushton—'ogre naturalists,' as Ian Hacking aptly dubs them—continue to seek such simple genetic differences, there is a widespread consensus among anthropologists that races are not "biologically real" (Kitcher, 2007, p. 293–4).

A definitive list of the flaws in the concept of race and, similarly, the problems with the identification of distinct biological groups using phenotypic and/or genotypic data has been formalized over this time. Moreover, this list has been repeatedly invoked in current studies on the emergence and dispersal of modern humans and the evolution of current patterns of variation (Lieberman & Jackson, 1995; Keita & Kittles, 1997; Templeton, 1998; Long & Kittles, 2003; Jorde & Wooding, 2004):

1) Common group-specific characteristics are non-concordant, and, so, if partitions are defined by traits that have been shown to follow independent geographic distributions, the difficulty of

making a meaningful group identification increases proportionally with the addition of each discordant trait.

2) There is no consensus on when to formally define (i.e., provide biosocial names for) groups that may be better thought of as aggregates of local populations.

3) Meaningful among-group boundaries are difficult, if impossible, to obtain for phenotypically intermediate groups and admixed individuals.

4) Allocation of membership is equally challenging for individuals who trace their ancestry to two or more named groups (e.g., races, populations, ethnicities) or self-identify with a composite group composed of socially similar but geographically, historically, and genetically different subsets (e.g., Hispanic, Latino).

5) The association of many human groups with distinct evolutionary lineages is problematic.

6) The variation within groups has been repeatedly shown to greatly exceed the variation between groups.

7) F_{st} measured among samples of world-wide groups has been found to be consistently low (with estimates 15% or less when computed as an average over many alleles or loci) indicating that subpopulation differentiation among humans is weak.

8) Only 0.1% of human DNA is known to vary among individuals.

This dismissal of meaningful human biological differences using genetic "proof"of "sameness" has also contibuted to the extension of the argument against the "truth" in human population substructuring beyond the typological tradition of race. This argument has sought to not only reject the idea of race , but also to generalize the terms of the debate to include group-specific biology. This is a decision which, in turn, has necessarily confounded the definitions of race, ancestry, population, or group (either as a meaningful and useful biological reality or an empirical, yet purely social construct of no scientific

consequence). Researchers advancing this position have therefore come to include any studies that make use of morphogenetic variation as a way to explore relationships among human groups under their rubric of rejection. This remains true even when the goal is not simply taxonomic (e.g., tracing the trajectory of human evolution, accessing past population structure in biological distance studies, or estimating individual ancestry in forensic casework). Out of this recent period of scholarship many critical studies that seek to support or address the decline in use of concepts of unique biological and *identifiable* groups as scientifically accurate means for explaining the variation observed within and between members of the single human species have therefore emerged. It is for this reason that Brown and Armelagos (2001) have aptly explained the history of physical anthropology over last 30 years within the framework of an "assault on the race concept" (p. 34) As this period is long and its literary output is considerable, it is necessary to highlight only the salient points of this debate below.

Race and Human Biological Variation

"Anyone who continues to believe in race as a physical attribute of individuals, despite the now commonplace disclaimers of biologists and geneticists, might as well also believe that Santa Claus, the Easter Bunny and the tooth fairy are real, and that the earth stands still while the sun moves." (Fields, 1990, p. 96).

For almost a half century, influential quantitative studies of modern human diversity have supported the argument made by Livingstone in 1962 that "[t]here are no races, there are only clines." These, in turn, have reiterated the common conclusions drawn about the patterns and processes of human biological

variation: (1) that world-wide populations are not bounded biological entities, (2) but that they, instead, diverge along clinal or isolation-by-distance patterns and move towards neutral variation, and (3) that traditional racial traits may have some geographic meaning but they do not correlate well with biological variation among groups. As Wade (2004) explains, the implication for the human biological sciences generally is that "[m]ost scholars argue that 'race' has no relation to human nature. A minority contend that it does" (p. 157) For physical anthropology specifically, opinion polls on this "great race debate" (Gill 2000, np) have yielded similar results (Lieberman, Reynolds & Stevenson, 1989; Lieberman et al., 2003; Cartmill & Brown, 2003). Position statements and research topics show that the majority of current "mainstream" anthropologists appear to agree with the non-existence of races. They have integrated this theory into their own approach to human variation using phenotypic traits and have made use of this information as the supportive evidence for arguments on the arbitrary nature of racial classifications (AAA, 1999; AAPA, 1999). While Wright (1978, p. 439) reiterates that in every day interactions humans may be quite adept at classifying individuals into different groups based upon some biosocial understanding of the relationship between external phenotype and identity (considering how morphological characteristics, like bone structure, skin color, hair texture, and soft tissue features are suggestive of group membership), there is a well-established viewpoint (c.f., the "independent variation argument" or "the fact of discordance") accepted among philosophers (Appiah 1996, p. 99; Andreasen 2004, p.428, 2005, p.94, 2008, p.484; Machery & Faucher 2005, p.1209; Glasgow 2009, p.88) and anthropologists alike (Livingstone 1962; Goodman 1997, p.24; Brown & Armelagos 2001, p.34) that these phenotypic categories lack objective validity. Highlighting this methodological lacuna, Diamond (1994) clearly explains that, "[t]here are many different, equally valid procedures for defining races, and those different procedures yield very different classifications… Faced with such differing classifications, many anthropologists today conclude that one cannot recognize any human races at all" (p. 84). There

are, in short, two different kinds of "knowing" race: one that permits an understanding and identification of race, and another that does not. A perceived gap exists, therefore, between what can be conceptualized intuitively based upon current social constructions of human diversity and ideas of the self and others (i.e., taking race as "a cultural category which can become an embodied part of the human experience") and what can be tested and corroborated empirically using best practices and following accepted standards of scientific rigor (Wade 2004, p. 157).

For a brief clarification of the procedural problems most commonly cited in physical anthropological literature, it is possible to turn to the many recent reviews of the "race debate," whose summary work aligns well with the mainstream positions against the reality and scientific utility of race or group discrimination presented in the key organizational statements (AAA, 1994, 1999; AAPA, 1996 [1999]; ASA, 2003; UNESCO, 1964; Mueller-Wille, 2005; Brattain, 2007). Although these studies are often rooted in genetic theory, reminding their readers of the findings in support of low level of among group variation, their approach is frequently most strongly focused upon the arbitrary, subjective and even user-directed nature of skeletal biological classifications. In the opinion of Brown and Armelagos (2001), for example, any differentiation between "races"is dependent upon the particular, yet random, subset of morphological traits employed in the analysis. It is the use of a predefined and subjectively selected suite of traits that, they argue, confounds any attempt at classifications. The research design of any racially based study, and specifically the selection of a preferred configuration of racial signatures, is, therefore, not only arbitrary but severely biased as the choice of materials and methods is closely tied to, and is essentially reflective of, the analyst’s own cultural milieu, understanding of bio-social "norms," research concerns, and personal, philosophic and/or ideological agenda(s). It extends logically for these authors, and,in genera,l for many advocates against the study of by-group biological differences, that

there is no possibility, by the very operational nature of the currently employed and commonly accepted procedures (i.e., the well-known morphoscopic and morphometric methods), for an unbiased explanation of human variation. Embedded within these arguments are three common methodological criticisms: (1) most traits are non-concordant, (2) classification accuracy is inversely proportional to the number of traits considered in the analysis, and (3) traits must be *a priori* selected for their known association with predefined groups.

In accordance with clinal theory, a perceived lack of correlation among "racial" traits used for population discrimination is taken to confound any morphogenetic attempt at classification. Brown and Armelagos (2001) explain this issue of non-concordance by arguing that as genetic traits are continually undergoing evolution following different trajectories of change (i.e., at varied rates and magnitudes and in divergent directions), achieving a concordance between these traits is a biological impossibility. It holds, for many researchers that, since phenotypic traits are uncorrelated or discordant, the success of their classification depends upon the number of traits used. It has been long argued that as the number of traits employed incrementally increases, the success of racial classification decreases. To wit, Livingstone had observed already in 1962 that

> [i]f one genetic character is used, it is possible to divide a species into subspecies according to the variation in this character. If two characters are used, it may still be possible, but there will be some ‘problem populations,' which, if you are an anthropologist, will be labeled composite or mixed. As the number of characters increases it becomes more nearly impossible to determine what the 'actual races’ really are (p. 279).

As a result, it is widely believed that any consideration given to additional (known-to-be) "informative" traits does not, in fact, enhance the analysis and improve its outcome; instead, the many-trait approach is considered detrimental to the accuracy of group allocation. Multi-trait analyses are, therefore, "noise"ridden and their classification results are, in turn, uninterpretable, so much so that "the problems in racial classification become insurmountable" (Brown and Armelagos, 2001, p. 34). Because successful classification is limited by the requisite small, if not singular, subset of traits, the quality of the classification is compromised by the imposition of an inevitable methodological constraint. This is a concern which had been already expressed by Lewontin (1974) in his discussion of genetics and race:

> [t]he apparent homogeneity within races **as** compared to the "obvious"difference between them stems partly from the fact that our consciousness of racial differences is constantly being reinforced socially because racial distinctions serve economic and political ends, and partly because the very characters we use to distinguish races-skin color and texture, hair form, eye, nose, and lip shape-are those to which we are most keenly attuned for the purpose of distinguishing individuals (p. 152).

A similar argument is made for craniofacial data. Traits must be typically prescreened for their "known" ability to signify membership so that "in practice, racial classifiers have to select the genetic traits and morphological features that support their preconceived notions of race"(Brown and Armelagos, 2001, p. 35). As a result, there is "an astounding mismatch between actual cranial variation and the variation modeled by racial constructs." (Armelagos and Van Gerven, 2003, p. 61). This subjectivity in procedure has lead these and other authors to conclude that races are biologically meaningless cultural artifacts because "the races we get are little more than a function of the trait or traits we use" (Armelagos & Van Gerven, 2003, p.61) and, so, classification results represent nothing more than "forced allocations" (Keita, 2007). Given these concerns, concepts of meaningful group differences are frequently discredited

following arguments on their own weaknesses in theory and method, and interpretation and use. They are, for example, taken to be (1) perpetuated by the racialized research strategies of anthropological practitioners, misrepresented under the guise of human biological variation (as "populations" or "ancestries"), (2) validated by confused and often biased findings of group differences, and (3) used to support grossly naive conclusions about the mechanisms of evolution, the distributional patterns of human diversity across space and through time, and on the usefulness of classification and group discrimination, ancestry estimation and the morphogenetic assessment of membership in bioarchaeological contexts and in forensic casework or medico-legal identification (e.g., Goodman, 1997; Goodman & Armelagos, 1997; Armelagos & Goodman, 1998; Smay; & Armelagos**,** 2000; Brace, 2000; Brown & Armelagos, 2001; Armelagos & Van Gerven, 2003; Williams, Belcher & Armelagos, 2005; Albanese & Saunders, 2006; Hubbe & Neves, 2007).

As these criticisms on human variation in physical anthropology have transcended the specific use of "race"to one that also includes "ancestry,""ethnicity," "population," or for that matter, any comparable pre-defined aggregation of individuals or peoples that has been imbued with some unique biological meaning, the rejection of race has emerged as a larger issue that marries the concerns for procedure with issue of perception. This assault on race not only questions the methodology of any approach that attempts to trace the magnitude and direction of human variation as observed within and between modern human groups but also takes on the complex relationships between professional and personal agendas, ideologies and philosophies. Any proxy for morphogenetic distinctiveness or grouping methodology is problematic because it is unavoidably bound to the project of human classification. The acknowledgement of the existence of hierarchical substructure within the unified human species and any attempt to classify an individual into one of these nested groups based upon some biologically

defined criteria of membership are contentious to many as they recall at least in their concept, if not in intent, the typological tradition that enabled human taxonomy in the service of racial profiling, biological determinism, and discrimination. Although there is tacit recognition that "racial thinking is not necessarily synonymous with *racist* thinking," there is an explicit emphasis on the unavoidable dangers in supporting a taxonomic or systematic-inspired approach to human population diversity in anthropological discourse, what Keita and Kittles (1997) have called the "racial" point of view. Critics argue that all typological schemas embrace equally the same fundamental tenet: types and their defining traits can be accurately associated with unique geographical origins or "homelands." This theoretical position supports an outright rejection of any form of parallel microevolution and invokes arguments for gene flow, colonization and admixture among groups to explain the presence of trait complexes in areas other than in their accepted place of origin, as well as any anomalous departures from the classic phenotypic expressions, e.g., the continental morphotypes of Caucasoid, Mongoloid, and Negroid. Every categorical approach for many researchers, therefore, perpetuates notions of authentic, continentally bound morphologies and effectively validates the racist question of the existence of human subspecies (Klein, 1989) or secondary races (Coon 1962, 1965). Within this anti-typological framework, discrete human groups are understood as conforming to Platonic forms: they are unnatural creations based upon an observed combination of "important"discriminating characteristics, they exist only as these artificial agglomerations, and members are defined and bound to these units by *a priori* assumptions of morphological similarity and difference (Goodman, 1997; Keita & Kittles, 1997). Given these parameters, classification is "anti-evolutionary to its core" (Goodman, 1997, p.21). The fear that any typologically-based research design is necessarily colored by racial stereotypes has continued to fuel criticisms of classification based studies that are concerned with modern human population affiliation (i.e., ancestry or admixture estimation) using morphogenetic, but especially craniometric,

data. These critique persist despite recent improvements in sample sizes and reference collection diversity (e.g., the Forensic Anthropology Data Bank or the W.M. Bass Skeletal Collection) or advances in data acquisition and analytical procedures (e.g., DNA extraction and genotyping, geometric morphometrics, and computationally intensive algorithms). It is necessary, therefore, to review independently the key papers on the distribution of human biological variation as it applies to questions of detecting morphogenetic differences between groups for physical and, specifically, forensic anthropology. In doing so, it is essential to particularly consider variation in craniometrics and the related topics of world-wide craniofacial diversity and ancestry estimation.

Human Craniometric Variation

"Some people think that the philosophy a scientist accepts is not of very much importance; his job is to observe the phenomena. This is a gross oversimplification and it involves the subsidiary hypothesis that all scientists are fully equipped with serendipity. A sensible philosophy controlled by a relevant set of concepts saves so much research time that it can nearly act as a substitute for genius; it may be that this is what we mean by genius. Thus equipped we avoid the pseudo-problems or, more correctly, the problems that are real in adjacent fields but pseudo in our own. A scientist can have no more valuable skill than the ability to see whether the problem he is investigating exists and whether the concepts he is using are applicable." (Pirie, 1952, p. 280)

The statistical genetic approaches developed out of Lewontin's seminal work and adopted by corroborative studies using similar genetic system proxies and methodology as well as those formative

studies that made use of different markers and statistics offered consistent, easily interpreted solutions to the otherwise vexing questions of the magnitude and the meaning of genetic differences among and within humans groups. This can be seen, for example, in the work of Nei and Livshits (1989), Excoffier et al. (1992), and Barbujani et al. (1997), who respectively used protein polymorphisms, blood groups, and DNA. These novel findings not only ratified the ongoing revolution in the quantitative interpretation of human variation, but also provided scientific support for the rejection of race. Ultimately, they aided in discrediting discriminatory tools as neither reliable nor helpful approaches to understanding structured human differences and demonstrated the impossibility of any positive use of the information gleaned from their application. The consequences of these genetic studies to the race debate in skeletal biology were, at least for the dominating circle of "new" anthropologists, conceptually clear. Reiterating Lewontin's (1972) conclusions, they empirically showed that the biological differences identified between individuals align poorly with predefined group (e.g., racial or ethnic) boundaries. The fate of race in physical anthropology was, therefore, initially genetically determined. By extension, projects in classification that relied upon quantitative skeletal data (like craniometrics or scored traits) but which referenced predefined groups were deemed equally ill-conceived as population genetic research continued to convincingly and simply show that variation within human groups exceeds variation among human groups, regardless of the type and number of system(s) surveyed. These systems include, for example, mitochondrial (Merriwether et al., 1991), Y-chromosome (Hammer et al., 2001) and various kinds of autosomal polymorphisms (Stoneking et al., 1997) as well as combined systems (Jorde et al., 2000). As Lewontin (1972) had already explained, "[h]uman racial classification is of no social value and is positively destructive of social and human relations. Since such racial classification is now seen to be of virtually no genetic or taxonomic significance either, no justification can be offered for its continuance" (p. 397). Despite arguments that the human cranium has been of "exceptional fascination"

among skeletal researchers and its preferential use is said to be tied to "historical trends within the field of physical anthropology itself" (Kemkes-Grottenthaler, 2002, p. 52; Henschen, 1966), these robust methods of population genetic theory that were revolutionizing molecular research were slow to be applied to questions of human craniofacial diversity. This held true whether they were applied as a further test of Lewontin's findings in the context of skeletal biology, a reply to the subsequent studies of neutral genetic variation that have replicated his original results, or as a focused investigation into the craniometric patterning long observed through other means of analysis by employing a new approach to explaining the perceived similarities within groups and the differences among them.

The skeletal morphology of the craniofacial complex has been studied extensively to evaluate relationships among human groups in a variety of cases that consider both fossil and extant populations. Craniometric traits in particular have been used to address questions about three key topics:

(1) the origins of modern humans in Africa, the subsequent migration out of Africa, and incidents of admixture with preexisting populations at points throughout this dispersal (e.g., Howells, 1989; Relethford & Harpending, 1994; Harvati & Weaver, 2006; Manica et al., 2007; Perez et al., 2007; von Cramon-Taubadel & Lycett, 2008; Betti et al., 2009).

(2) the effects of evolutionary forces on human phenotypic development, including the interactions among genetic drift, gene flow and natural selection (Relethford, 2004, 2007, 2010; Roseman, 2004; Roseman & Weaver, 2007; Weaver et al., 2007), and the relationship between morphological and molecular diversity (Relethford, 1994, 2002; Roseman & Weaver, 2004; Roseman & Weaver, 2007; Smith, 2009; von Cramon-Taubadel, 2009; Strauss & Hubbe, 2010).

(3) the reconstruction of population history, the assessment of ancestry and the interpretation of race from biosocial, clinical and forensic perspectives (Howells, 1973, 1989, 1995; Relethford,

2001, 2004, 2009; Harvati & Weaver, 2006; Martínez-Abadías et al., 2006; von Cramon-Taubadel, 2009; Ousley, Jantz & Freid, 2009; Konigsberg, Algee-Hewitt & Steadman, 2009; Jantz & Logan, 2010).

The use of cranial shape in population studies may draw some of its popularity from the fact that the skull is often the best-preserved part of a recovered skeleton (Walker, Johnsonn & Lambert, 1988) and that it has its roots of interest in the popular pseudo-science of craniology (Kemkes-Grottenthaler, 2002). These factors resulted in the widespread collection of crania beginning in the nineteenth century and, by consequence, the development of all types of assessment and estimation methods that focused first on this available material: for example the study of suture closure and dental wear patterns for age estimation in the mid-1800s (c.f., Masset, 1989; Rose & Ungar, 1998; Spencer, 1997). The widespread use of cranial data, however, is more likely attributable to its "clear" and "strong" geographic patterning (Ousley, Jantz & Freid, 2009; Strauss & Hubbe, 2010) and the ease of which its structure can be explored by the straightforward application of multivariate statistics to standard measurements and commonly collected traits (c.f., Howells, 1973, 1989, 1995; Hanihara, 1996, 2008; Hanihara et al., 2003). In forensic anthropological practice, its use is also definitive as the most reliable source of ancestry estimation (Bass 1995). As Steadman and Andersen (2009) have instructed: "[f]or cases in which ancestry might be discriminatory and therefore assist with personal identification, forensic anthropologists first look to the cranium" (p. 19). Even given the history of interest in detecting population-specific phenotypic traits and the focus on morphological variation in the human cranium for the elucidation of past and present biological and demographic trends, it was several years after the publication Lewontin's findings that population genetic theory was rigorously applied to craniofacial data in order to better understand the differentiation processes and the distributional properties of quantitative cranial traits. The popularization of a growing corpus of genetic evidence that attested to a lack of intergroup

differentiation brought about, at first, arguments that contrasted the global patterns of variants in classic genetic and DNA markers with those revealed using phenotypic characters, highlighting the apparent contradictions between them (Disotell 2000). Within these arguments, reference was commonly made to the idea of obvious, quantifiable differences in geographic regions, which was produced by, argued for, and reproduced in contemporary analyses of craniometric data (Nei & Roychoudhury,1982; Stringer & Andrews, 1988; Wright, 1992; Howells, 1995; Gill, 1998; Hennessey & Stringer, 2002). It was not until the mid-1990s that a quantitative genetic, model-bound methodology was first applied to craniofacial data. Relethford used a multivariate-pooled variance approach (**R**-matrix method) based on an equal and additive effects quantitative genetics model (i.e., the Relethford-Blangero method) to address world-wide craniometric diversity and offer empirical information on the relationship of phenotypic (morphological) to neutral genetic variation. He did so through a direct comparison of the level of among-region differentiation obtained from craniometric traits with the amount that now had been commonly observed for genetic markers (Relethford, 1994, 2001,2002; Relethford & Blangero, 1990; Relethford & Harpending, 1994; Relethford, Crawford & Blangero, 1997).

From a combined anthropological and population genetic perspective, Relethford, offered a much needed, updated discussion of modern human craniofacial diversity, first in 1994, with his pioneering study of global morphologies, in which he considered only craniometric data and, again, in his later expanded analyses, in which he included additional phenotypic and genetic systems (2001, 2002, 2004, 2009; Relethford & Harpending, 1994). In order to account for the magnitude of craniofacial diversity that had been documented among and within global populations, Relethford has made regular use of the world-wide craniometric data set originally collected by W.W. Howells (1996), which appeared particularly suited to his research goals for several reasons. This data set was sampled for regional

morphological differences across far distances (Relethford, 2009), it was believed to represent "distinct biological populations" (Relethford, 2010, p. 106), and it had been already used extensively to evaluate population-based patterns of morphological differences using metrically derived cranial variables (Howells, 1973, 1989). His methodology followed established practice in population genetics literature to express the degree of variation among groups relative to panmixia (and under certain assumptions) using the standard measure, F_{ST} (c.f., Excoffier, 2001; Beerli & Felsenstein, 1999; Long & Kittles, 2004). This index was derived from Wright's (1951) theory of allelic fixation and his concept of *F-statistics, which have proven effective in quantifying* the extent of reduction of heterozygosity in a subdivided population at differentiated hierarchical levels of a population structure (Hartl & Clark, 1997; see Wright, 1969 for G_{ST}). *These measures* can be used to assess the degree of differentiation within a population among groups of demes (F_{SG}), within groups among demes (F_{GT}), and within a population among demes (F_{ST}). They do so by returning calculated values in a range from 0 (indicating no differentiation between the overall population and its subpopulations) to a theoretical maximum of 1 (although in actuality the observed fixation index is typically much less than 1, even for highly differentiated populations). For human populations and when using diverse autosomal, and presumably neutral, markers (e.g., blood groups, protein polymorphisms, RFLPs, STRPs and SNPs), the classic values obtained for interregional human variation consistently range from 0.10-0.16, and, so, suggest low to moderate group differentiation (Tishkoff & Verrelli**,** 2003; Pasternak, 2005; see Rosenberg, 2002 for even smaller estimates [0.03-0.05]). Relethford estimated an analog of Wright's (1951) F_{ST} for quantitative phenotypic traits. In doing so, he facilitated the application of certain population genetic concepts to cranial morphology and improved the investigation of the structure of craniofacial variation on a global scale by its comparison with the results obtained from neutral genetic markers (e.g., Lewontin, 1972; Barbujani et al., 1997; Bowcock et al., 1991; Rosenberg et al., 2002; c.f., Roseman,

2004; Roseman & Weaver, 2004, 2007). He adopted F_{ST} as a measure of among-group craniometric diversity under the assumption that these phenotypic characters vary among regions and among individuals within regions according to an equal and additive effects model (Relethford & Blangero, 1990; see also Spitze, 1993; Whitlock, 1999; Merilä & Crnokrak, 2001). He also worked under the assumption that variation in human cranial shape is, on average, moderately to highly heritable within local populations (c.f., Konisgberg & Ousley, 1995; Mousseau & Roff, 1987; Lynch & Walsh, 1998; Raposo do Amaral et al., 1989; Sparks & Jantz, 2002; Roseman, 2004; Sherwood et al., 2008). Despite fluctuations identified for some traits (Carson, 2006; Martínez-Abadías et al., 2009), this assumption still holds and it has been even argued that interregional differences are already well established early in ontogeny (Vioarsdóttir, O'Higgins & Stringer, 2002). Taking the average narrow sense heritability, h^2, of cranial measurements as ≈ 0.55 (after Devor, 1987), Relethford (1994) retuned F_{ST} values of 0.112 for his three region and 0.114 for his six region partition of the world-wide populations represented by the Howells data set. In fact, he found (1994, 2002) that estimates of F_{ST} for all 57 traits were constrained to the range of 0.11-0.14, depending upon the number of geographic samples used. From these results, he concluded that craniometric variation among major geographic groupings is low and, importantly, it is consistent in magnitude with the expectations for the amount of genetic variation explained by differences among human groups (i.e., values in the range of 10-15%) when using either classic or DNA markers (c.f., Nei & Roychoudhury, 1974, 1982; Latter, 1980; Nei & Livshits, 1989; Jorde et al., 2000). From this data, Relethford has argued that the correspondence between both phenotypically and genetically derived F_{ST} estimates suggests that morphological and molecular data are in agreement both qualitatively and quantitatively. He rejects, therefore, the earlier presumptions made by, for example, Nei and Roychoudhury (1982) or Stringer and Andrews (1988) that (1) morphological traits display a significantly larger degree of among group variation within the human species than do genetic traits and

that (2), for this reason, racial differences are highly conspicuous. His findings suggest instead that global craniometric variation is in agreement with the signatures expected under a neutral model of genetic drift balanced by gene flow. His results not only corroborated Lewontin's now classic conclusions by considering yet another biological system, but were also used by many other researchers to validate the view that concepts of race and even populations are poor approximations for the true patterns of craniofacial diversity. In part, his findings offered support for the opinion that craniometric variation among major geographic regions is very limited: a position that was even adopted (albeit tentatively) by Howells (1989), himself, following his comparative study of modern and archaic craniofacial forms. These findings also gave new weight to the "antagonist" perspectives on the biological significance and anthropological meaning underlying the distribution of human variation and on the "science" of its substructuring into predefined groups (e.g., populations, races, ethnicities). These results were therefore frequently used as justification for the popularized, and often highly charged, view that "[t]here is no such thing as a biological entity that warrants the term 'race'" (Brace, 2000) or "[t]he population is the last bastion of the typologist" (Brace in Armelagos & Van Gerven, 2003, p. 60).

In his later studies on global diversity in skin reflectance, Relethford (2002, 2009) expanded his analysis to other patterns of human phenotypic variation, testing first the relationship between craniofacial shape and measures of skin reflectance. It is important to note that in doing so, he explicitly considered the two sets of data that have long received considerable attention, both scholarly and public, in discussions of group distinction and socio-biological affiliation and which have experienced a lengthy history of use as racial proxies. Relethford's findings suggest a clear discordance in F_{ST} values: assuming a heritability of 0.66 (after Williams-Blangero & Blangero, 1992), F_{ST} is estimated for skin color at 0.88 for a five region partition (Relethford, 2002), while the range of F_{ST} values typically obtained for genetic and

craniometric data is restricted to 0.10-0.15 (Barbujani et al., 1997, Relethford, 2002). He also reports that skin reflectance presents an opposite pattern to genetic or craniometric markers in the apportionment of its diversity. For skin color, approximately 88% of total variation exists among regions, 3% is among local populations within regions, and 9% is present within in local populations, while, for craniometric variation, approximately 13% of total diversity occurs among regions, 6% among local populations within regions and 81% within local populations (Relethford, 2002). Relethford (2002, 2009) has consistently interpreted these results in terms of selection pressures, adopting a position that is in line with other contemporary analyses. He considers the pattern observed for pigmentation to have been shaped by past natural selection for tolerance to ultraviolet (UV) radiation, where skin color is represents the global geographical distribution of local groups (i.e., "light"skin reflects north/south geography, while "dark"skin suggests origins near the Equator). Other studies have also identified a strong relationship between levels of pigmentation and latitude and between levels of pigmentation and exposure UV radiation (Roberts & Kahlon, 1976; Relethford, 1997; Jablonski & Chaplin, 2000). Skin color shows, for Relethford (2009, p. 17), "a classic pattern of clinal variation." He concludes, therefore, that, despite the high F_{ST} values, "the elevated degree of differentiation in human skin color by itself does not provide support for a view of discrete and differentiated human races" (Relethford, 2009, p. 17). For the distribution of craniometric variation, he offers an explanation in terms of on average selective neutrality, positing that interregionally differing selection pressures from the recent past have played only a limited role in producing overall patterns of human craniometric diversity. These conclusions are of particular note as they are in disagreement with several earlier studies that had identified significant effects of selection on some aspects of human craniofacial architecture. Findings of correlations between climatic variables (like temperature and humidity) and cranial features (e.g., the nasal region especially) and documented regional differences in specific traits and/or overall craniofacial morphology

that may be tied to factors of "environment" (like altitude and lifestyle) had previously demonstrated that cranial variation does not always behave in a manner that is geographically uniform (Carlson & Van Gerven, 1977; Guglielmino-Matessi et al., 1979; Carey & Steegmann, 1981; Beals, Smith & Dodd, 1983; Rothhammer & Silva, 1990; Franciscus & Long, 1991; Gill, 1998; Steele & Bramblett, 1988).

As Rosemann and Weaver (2007) have already observed, "Relethford (1994) did leave open the possibility that some aspects of craniometric variation have been subject to interregionally differing selection pressures in the recent past" (p. 257). It was expected that he (2004, 2009, 2010) continue to evaluate the effects of geographically limited gene flow and genetic drift on the retention of population structure and to engage these topics in popular discussions on human origins, on the influence of migration and environment, and on the methods of identification of population-specific departures from neutrality as well as their importance to the interpretation of global craniometric variation. In his 2004 publication on isolation by distance (IBD), he tested the correspondence among several available systems (genetic and phenotypic) for the study of variation under a model which assumed decreasing gene flow with increasing geographical distances. Using red blood cells, polymorphisms, microsatellite DNA markers and craniometric traits, he assessed the levels of biological variation between global populations and found that a common pattern of global gene flow, mediated by geographic distance, exists across all four of these major, but different, data types. He argues that the correspondence between these systems and their consistent association with geography suggests, again, that cranial morphometrics can be reliably employed in the study of population structure and history. For questions of human origins, Relethford (1994, 2004) posited that this agreement between genetic and geographic distance and the agreement between craniometric and genetic variation in locating the highest level of diversity in sub-Saharan Africa attests to both an evolutionary history that involved the diffusion of

Anatomically Modern Humans (AMH) out of Africa and a recent single origin for modern humans. This is a theory that has been recently supported with morphogenetic data by Manica et al. (2007), who have shown a decrease in morphological variability relative to geographic distance from Africa and a correspondance with the pattern observed for neutral genetic markers (c.f., Cavalli-Sforza, Menozzi & Piazza, 1994; Serre & Pääbo, 2004; Prugnolle, Manica & Balloux, 2005; Manica, Prugnolle & Balloux, 2005; Ramachandran et al., 2005; Liu et al., 2006; Lintz et al., 2007). For questions of population structure, Relethford (2004) used craniometrics to explore the possible impact of migration on the patterns of cranial variation by evaluating cranial plasticity, natural selection and increased gene flow for the Boas immigrant data (a topic debated previously by Sparks & Jantz [2002] and Gravlee, Bernard & Leonard [2003], and recently revisited by Jantz & Logan [2010]). Here, he argues that the major forces driving the observed craniometric patterns are gene flow and genetic affinity and, in turn, he draws the important conclusion that developmental plasticity and climate adaptation do not erase or obscure underlying patterns of population structure and history. Relethford's study not only offered external support for the position taken by Sparks and Jantz (2002) in their earlier analysis of the Boas data but also worked, in concert with other approaches that sought to understand the influence of selection and stochastic evolutionary processes on craniofacial shape (e.g., Cheverud, 1982; Ackermann & Cheverud, 2004). This, in turn, set the stage for the studies on the genetic background of cranial morphology that would soon follow (e.g., Carson, 2006; Martínez-Abadías et al., 2009; Carson & Algee-Hewitt, 2009). These findings have been recently corroborated and extended, for example, by Martínez-Abadías and colleagues (2006) who, in showing that the admixed offspring of two different parental populations exhibit, generally, an intermediate cranial morphology, have demonstrated that these phenotypic (cranial) traits are, in some large part, under the control of an additive polygenic system.

In 2009 and 2010, Relethford revaluated Howells' craniometric data set to address re-emerging questions of the skeletal retention of population structure and history in light of advances in molecular research and the controversies that surround many current approaches to skeletal analysis. In this work, Relethford gave special attention to concerns for (1) the importance of potential deviations of global variation from a neutral genetic model given recent findings (e.g., Roseman, 2004; Roseman & Weaver, 2004) that evidence supporting the neutral evolutionary argument only holds when cranial morphology is evaluated as a whole and across broad geographic space, and for (2) the debate over the usefulness of the biological race concept and, by extension, group partitioning using measured differences in craniofacial morphology. In his earlier work on craniometric diversity, neutrality and IBD, Relethford (1994, 2002, 2004) has suggested that, if natural selection did play a role in shaping global diversity in craniofacial morphology, its effects were not sufficiently pronounced to influence a pooled analysis of craniometric variation. In his more recent work, he (2010) analyzed residuals of craniometric distances from a geographically-based neutral model of population structure in order partition these effects and explore other population-specific influences. Using a Procrustes rotation method on scaled configurations of the phenotypic and geographic distance matrices obtained for all pairs of populations, he identified three outlier populations: the Buriat, Greenland Inuit, and Peruvian samples. He attributes the deviations observed for the Buriat and Greenland Inuit to cold adaptation, while, for the Peruvian sample, he offers an explanation of population movement, where this specific population deviation reflects the history of the (evolutionarily) recent migration into South America. In order to better understand the import of these findings, as Procrustes rotation can be affected by large outliers, he conducted a second analysis, which repeated the process but with the Buriat, Greenland Inuit, and Peruvian samples removed. As the plot of residuals failed to reveal any additional outliers and the fit between geography and phenotypic variation improved in this revised analysis, he concludes that the

three populations of concern are highly anomalous as they are responding to extreme conditions of or extreme changes in their environment. From these results, Relethford (2010) argues three critical points: (1) the absence of additional outliers, after the exclusion of the three pre-identified populations, makes "the case for essentially neutral variation in overall craniometric distances is even clearer," (2) "selection will sometimes act to *obscure, but not erase* the signature of population history and structure," and (3) craniometric traits are useful to population studies despite some acknowledged effect of selection on different parts of the cranium and on particular populations existing under unique circumstances (p. 109-110). Comparable conclusions are offered by Jantz and Logan's (2010) recent reanalysis of the Boas immigrant data, which found that population specific factors (e.g., cranial modification among the Hebrew children and biological stress leading to impaired growth among the Sicilians), were not adaptive responses to the selective pressures of a new "homogeneous American environment" and contributed to the observed change in cranial morphology among the immigrants' U.S. born offspring (p. 702).

Relethford's 2009 paper on global patterns of phenotypic variation can be used in concert to with these three key conclusions to clarify their contribution to the identification of the underlying patterns of biological affinities among world-wide populations, to better interpret the significance of these findings, and to relate these to the renewed debate over race and group discrimination using skeletal traits. Relethford's (1994, 2002, 2004, 2009, 2010) studies have demonstrated consistently that there is an association between craniometric variation and geography, and many now agree that a general acceptance has been reached on the assumption that underlies this relationship: populations that are located farther apart geographically will tend to shows fewer similarities than those populations who are positioned more closely together. And, Relethford is not alone in making the extended argument

that because this association has been retained throughout recent evolutionary history in the face of selective forces, craniometric variation can be taken as informative of both human origins and modern human population structure. Relethford is, not unexpectedly then, a major proponent of the argument that the study of craniometric variation can thereby enable a "reasonable reconstruction of population history" (2009, p. 17). But, what is interesting is that he provides qualifications for his results on this observed biogeographic patterning, clarifying their interpretation by stating that "there are no abrupt breaks in the relationship between phenotypic and genetic distance" (2009, p. 19). Most importantly, from this observation, he asserts that when group partitioning is the goal, "decisions for subdivision into clusters (or races) are going to be subjective" (2009, p. 19). Relethford draws this conclusion from his reconsideration of the geographically structured, global variation represented in the Howells data set in light of the issue of classification: considering, specifically, its role within the theory and methods of forensic anthropology. This is a concern which is of particular importance to this larger dissertation project. Relethford begins his reappraisal of the Howells data set by conducting a preliminary test to determine if the craniometric variables collected, when subjected to discriminant function analysis (DFA), could accurately classify the individual crania represented by Howells' sample into their correct, predefined geographic groups. Using all 57 measurements and six geographic regions, he obtained very impressive rates of correct classification (97% for males and 96% for females) even when the jackknife method is implemented (96% for males, 94% for females). Taken alone, these results can suggest that human craniometric variation is highly structured and that the partitioning of global craniofacial shape (as represented by Howells's samples) into six geographic regions is an appropriate and, perhaps, even meaningful solution. Moreover, if these results are reviewed in the spirit of the use of DF-based analyses in the context of forensic anthropological casework, they may be seen as containing essential information that can contribute reliably to the individual's biological profile and assist the medico-legal

community in making identifications for "unknown" skeletal remains. Nevertheless, Relethford (2009) appears skeptical of these results when he writes that "[t]o argue against this subdivision in the face of such high accuracy often seems counterintuitive. . . at first" (p. 19). For Relethford, the critique of classification using craniometrics lies in the arbitrary nature of determining the right number of group partitions. As he discovered, when the geographic regions used by Howells are defined differently, so that the sample is divided into, for example, seven or eight groups instead of his original six, classification accuracy remained not only high but very close to the results obtained for the six group solutions (96%-97%). All solutions appear equally "correct." Faced with this dilemma, he (2009) posed the largely rhetorical question: "[i]f defining different numbers of geographic races all produce essentially the same rate of accuracy in classification, then what justification exists for preferring one number of races to another?" (p. 20). In lieu of providing a reply, Relethford chooses to offer an explanation for the observed phenomenon, proposing that it can be accounted for in terms of both sampling strategies and research design. He argues that if one defines clusters in such a way that geographical distances among clusters typically exceed geographic distance within clusters, and if one uses enough traits in the analysis, then there will be a high degree of accuracy in classification. He further contends that "There is no justification based on classification accuracy for preferring five, six, seven, or eight groups. From the perspective of racial classification, if we take these different groups as equivalent to races, then we see that the number of races, and the geographic cutoffs used to define them, are subjective decisions" (p. 19). Relethford (2009) concludes, therefore, that "Geographic race is a crude way of trying to describe the geographic structuring of human biological variation" (p. 19).

For many who rely on F_{ST}, clinal theory, and arguments of statistical and methodological weaknesses for making their case against the identification of human population substructing, biologically-based

classification and the estimation of ancestry and its practical translation into biosocial terminology, Relethford's findings offer quantitative support for the non-existence of definable human [biosocial] groups. His observations of (1) relatively low values for among-group human variation using craniometric traits, (2) the correspondence between morphometric data and neutral genetic markers (both in terms of the magnitudes of difference and meaning for group differentiation), (3) the lack of concordance between phenotypic traits (e.g., craniometrics and skin reflectance) and (4) the overlap in trait distributions and questionable cluster solutions all aid in this argument. By extension, they validate, for many, the rejection of identification or classification methods that rely on morphological signatures of population membership and the dismissal of scholarship that seeks to detect meaningful group differences using craniofacial traits for tracing the trajectory of human history or for practical application to, especially, forensic identification in the service of the medico-legal community. Take for example, the following rhetorical statement made by Brace and Hunt (1990) on the selective neutrality of cranial morphology, in which these opinions operate inherently: "[w]hat possible difference in selective value might there be in having a cranial contour that is ovoid as opposed to pentagonal – or any other of the myriad of named shapes? Why on Earth would it be better to have high rounded orbits under some circumstances and low rectangular ones under others?"(p. 334). If observed differences in craniofacial morphology are meaningless, as this statement would imply, and the "population" based subdivision of human craniometric variation for the purpose of classification is ambiguous and subjective because, as Relethford (2009) has demonstrated, any preselected number of groups is both equally "correct" and "works" equally "well," then, the resultant substructuring of craniometric variation and, in turn, the individually assigned group memberships, can be dismissed as arbitrary or as biased reflections of the practitioner's unique choice of samples, variables, and *a priori* identifiers (see Brown & Armelagos, 2001; Armelagos & Van Gerven, 2003). This perspective understandably has considerable implications for the

subdiscipline of forensic anthropology and the basis for its theory, method and research. While repercussions may be felt, specifically, on the global validity of the forensic practitioner's approach to the question of ancestry, the effects of these conclusions extend to the estimation of other essential parameters of the biological profile because "knowing" ancestry is frequently a precondition for the most effective estimation of other demographic variables (e.g., age, sex, and stature) that are key to aiding the medico-legal community in making an identification for unknown, mostly skeletal and often fragmentary, human remains. The invalidation of group discrimination, therefore, undermines the very foundation of forensic anthropological practice while it devalues, and may even inhibit, the contributions that a thoughtful skeletal analysis can make to assisting law enforcement investigators in excluding cases, finding a presumptive match among open cases, and eventually establishing a positive identification by isolating the unique biological traits that characterize *an* individual as *the* individual. The tension between the biological anthropologists who use craniometric variation as a means to access population structure and estimate individual ancestry in medico-legal contexts, and those who dismiss any discussions of group differentiation as artifacts of personal ideological positions suggests the importance of reevaluating the criticisms of identifying biologically meaningful patterns of human variation as they apply to forensic anthropological practice.

Classification and Quantitative Methods

"Anthropology has been involved in making racial distinctions on the basis of biological characteristics through much of its history. This is still an important pursuit in forensic anthropology

at least, although a significant goal of this latter research appears to be the substantiation of law-enforcement folk taxonomies."(MacEachern, 2000: 359)

The place of race within the scientific discourse on human variation and the related operationalization of ancestry groups, ethnicities, or populations in biomedical [etiologic and pharmaceutical] or forensic [identification] practices are problems that have vexed anthropologists and geneticists alike for the last half century. Duster (2006) aptly explains that "[t]he unenviable task has been to try to walk a tightrope – to figure out a way to effectively deploy in research the concept of race (or population groups designated by race) without endowing "race"with a false sense of genetic essentialism" (p. 3). Race or any concept like it, that acts either directly as its surrogate or as a biosocial, biogeographic or biohistorical proxy for morphogenetic distinctiveness (Burchard et al., 2003; Duster, 2003), is problematic because it is unavoidably bound to the project of human classification (Risch et al., 2002). Classification is considered, among symbolic anthropologists, to be innate: located at the very fundamental level of the human psychosocial makeup and even essential to the human experience. The natural world, as Douglas (1986) has mused, is often appropriated as a rhetorical resource for use in social categorization: "[t]here needs to be an analogy by which the formal structure of a crucial set of social relations is found in the physical world, or in the supernatural world, or in eternity, or anywhere..." (p. 48). As a result, social structures are retained and perpetuated by tacit logics and assumptions based upon visible referents, so that, "when the analogy is applied back and forth from one set of social relations to another, and from these back to nature, its recurring formal structure becomes easily recognized and endowed with a self-validating truth" (Douglas, 1986, p. 48). Race, and concepts like it, act together, therefore, as basic and necessary organizing principles. Researchers concerned with the social production of scientific (evolutionary, biomedical, forensic) knowledge, and especially those

interested specifically in the molecularization and institutionalization of concepts of biological identity, take this as cause to argue that racial, ethnic, and ancestral categories have maintained an influential role in current research on the distribution of human diversity. Fullwiley (2008) explains that "[d]espite the fact that the U.S. Office of Management and Budget (OMB) specifies that American racial descriptors 'do not conform to any biological, anthropological, or genetic criteria,' many scientists come to understand race as a 'physical world' referent for sorting human genetic variation" (p. 149). This begs the question: if classification is natural, normal and even necessary part of the human experience, and biological categories have found a place within other valid scientific discourse, why does classification in anthropology, in the form of ancestry estimation or population partitioning, generate such strong and negative reactions, even and especially when it is mobilized in the public service of identification (i.e., in the contexts of human rights work, death investigations, and missing persons cases)? This natural response to patterned differences among groups, when put to either theoretical or practical use, appears for physical anthropology to be, however, complicated by the effects of pan-disciplinary nurturing. The acknowledgement of the existence of hierarchical substructure within the unified human species and any attempt to allocate an individual into one of these nested groups based upon some biologically defined criteria of membership remain contentious. For many, they recall in their concept, albeit not in intent, the typological tradition that mobilized human taxonomy in the service of racial profiling, biological determinism, and oppression. As MacEachern's statement above reveals, there is, for some, a *de facto* association between the contemporary approach to the estimation of ancestry and the racialized biology of early anthropology. This can, by extension, include a fear that encoded into any form of classification is the socially dangerous and racist ideas that drive discrimination, sanction movements toward inequality, and potentially culminate in eugenics, genocide, apartheid, and slavery

(Gill, 2000). Goodman (1997), for example, has argued from this position in his criticisms of forensic identification methods and the applied perspective of their practitioners:

> I do not for a moment think that knowing race is a myth eliminates racism. But as long as well-meaning investigators continue to use the concept of race without clearly defining it, they reify race as biology. In doing, they mislead the public and encourage racist notions. According to the American sociologist Douglas E. Muir, those who continue to see race in biology but mean no harm by it are nothing more than 'kind racists.' By continuing to legitimize race, they inadvertently aid the 'mean racists' who wish to do no harm. Far too many scientists, unfortunately, still belong to both categories (p. 25).

This discomfort surrounding discussions of grouping concepts and methods of classification for "making racial distinctions" and the negativity cloaking this "pursuit" in or "goal" of forensic anthropology (MacEachern, 2000, p. 359) is not limited to theoretical discourse on human variation and anti-race rhetoric in mainstream anthropology (Lieberman & Kirk, 2004). This view has also emerged within the forensic anthropological community itself, so that even some of its members would agree that "[t]he determination of ancestry population affinities or ethnicity (in the past, referred to as 'race') is the most controversial question that a forensic anthropologist must face when assisting in identifying unknown individuals" (Albanese & Saunders, 2006, p. 282). In scholarly, sub-disciplinary dialogue, the title of a recent best-practices chapter by Albanese and Saunders (2006), from which the quotation above was drawn, stands as a clear signal of this concern: *"Is It Possible to Escape Racial Typology in Forensic Identification?*" In response to this rhetorical question, the authors provide arguments against the use of "'race'-determination methods" (2006, p. 281) while simultaneously conflating the clearly

social interpretation of race with the practice of estimating ancestry in personal identification contexts. They write that

> [t]he recent trend in the forensic literature has been to use the term 'ancestry' instead of 'race,' with no change in the underlying concepts, so that determining continental origin has been substituted for color terminology. Regardless of the terminology used, the underlying assumption *in forensic applications* is the same: using morphological, metric, or a combination of data, it is possible to assign an unknown individual into one of a limited number of continental or racial groups (usually two to six groups) (2006, p. 282).

Moreover, with the help of mass-media attention given to identification in "forensics" and the popularization of skeletal analysis specifically, forensic anthropology has found that it is not immune to many aspects of the "CSI effect" (Tyler, 2006; Algee-Hewitt, Wilson & Jantz, 2007). Examples of media depictions of forensic science occur on television series such as "CSI" (Donahue & Fink, 2002), "Law and Order" (McCreary & de Segonzac, 2001) and "Bones" (Hanson, 2006-2011), in non-fiction books (Bass & Jefferson, 2003, 2007; Roach, 2003) and suspense novels (Jefferson Bass series, 2006-2010; Cornwell, 1994). The public awareness of the relationship between biology and "race"has also been abetted by the commercialization of genetic ancestry testing (Bolnick et al., 2007; Harmon, 2008; Healy, 2008; Salkin, 2008; Weiss, 2008; Roberts, 2010). The growing interest in ancestry estimation and race identification from the skeleton has spread out, therefore, to inform lay understandings of "casework" and "best-practice." Too often there is a disconnect between what can be done, what is done and how it is done and the information that is retrievable from forensic anthropological procedures in reality. These public [mis]perceptions are apparent in a recent *Explainer* article featured in the daily online magazine, *Slate*. In his piece entitled, "Alas, Poor Yorick ... or Is It Othello?," web-journalist, Brian Palmer (2011), set out

to answer the question**:** "Can you tell a person's race from his or her skull?" Crediting *Ann Bunch* (The *College at Brockport*) *and John Williams* (*Western Carolina University*), *Palmer* writes in explanation:

> [r]acial classification is an inexact science, if that's even the right word for it. Forensic anthropologists never make definitive ancestry pronouncements. They say a bone is 'consistent with' European ancestry or 'likely' of Asian ancestry. And practitioners say it takes years of experience to achieve mastery, since you have to see piles and piles of disembodied mandibles to be able to recognize the sometimes subtle differences among them...The practice of inferring race from bones is also somewhat controversial. While today's forensic anthropologists don't like to talk about it, the discipline has its roots in the pseudoscientific 19th-century practice of using skull measurements to prove Caucasian intellectual superiority...There are people who think even the modern techniques are bunk (2011, np).

Despite the "bunk," within two days of its posting on Tuesday, January 4, 2011, at 4:59 PM ET, this article had already been "Liked" on Facebook by 74 readers and had generated 27 comments, one of which speaks to the ease with which misperceptions can be formulated and misrepresentations can be perpetuated, even within academia and among those "educated" in the field. User-ID, Clinton Thomas, wrote:

> I took a class on physical anthropology in college where we did exactly what is discussed in this article. One portion of our final exam was to classify the age, race and gender of five human skulls. It was only a 200-level class, so the people who were doing it weren't seasoned CSI-types, but normal college students with no experience. I got an A. It's really not that hard once you know what to look for. People of Asian descent have shovel-shaped incisors ... Native Americans have

> distinctive cheekbones ... males have more robust skulls than females ... each fissure of the skull's plates fuses at a certain age, so age can be determined within 3-5 years based on which fissures are closed and which are not (2011, np).

These examples get at the larger problem with the discussion of biologically based classification in forensic anthropology: in conflating race with ancestry and reducing the assessment procedures to a set of black-boxed steps or list of traits, there is a lack of understanding of the concepts and an oversimplification of the biological theory and statistics that underlie good estimation methods (i.e., what is being estimated, how the estimating is being derived, and in what ways the results are either scientific or subjective). In consequence, this enables both a tacit tendency to assume that a direct linkage exists on some level between biogeographic ancestry and "bureaucratic" race, and a self-seeking tendency to use these presumptions to advance opinions against forensic methods and theories, which are argued "to substantiate typological thinking and to reinforce the type concept itself" (Williams, Belcher & Armelagos, 2005, p. 344). Together, these trends can perpetuate larger philosophical positions on the apportionment of human variation: "[t]he assignment of skeletal racial origin is based principally upon stereotypical features found most frequently in the most geographically distant populations...it fails to identify populations that originate elsewhere and misrepresents fundamental patterns of human biological diversity" (Williams, Belcher & Armelagos, 2005, p. 345). These subdisciplinary problems may be in part due to (1) the type of criteria that has been traditionally employed (e.g., Hooton's "Harvard List" [Brues, 1990]; see also Rhine [1990, 1993] and Gill [1998] for skeletal attributes of race and biogeographic affinity), (2) the perspective of the publications that have characteristically dominated the discussions on the topic (St. Hoyme & Iscan, 1989; Brues, 1990; Rhine, 1990, 1993; Hinkes, 1993; Gill 1998), (3) the systems into which human variation has been artificially

structured (i.e., as a tripartite system of major racial stocks or types), and (4) the ways in which ancestry assessment has been represented in beginner textbooks and, in turn, taught to impressionable entry-level students with limited, biased, or no experience (Bass, 2005; Byers, 2005; Klepinger, 2006; see also Hefner et al., 2007 and Wheat, 2009). For example, trait lists and morphoscopic methods require little theoretical fore-knowledge, are easily observed and learned in isolation, and are satisfying for their tactile approach and emphasis on personal [subjective] observation. Simultaneously, the retention of terminology like "Caucasoid," "Negroid," and "Mongoloid" allows for the simple partitioning of human variation and, in turn, enables an easy correspondence between skeletal characteristics and the continental types that they are traditionally said to define (see Hefner [2007, 2009] for problems with and new quantitative approaches to non-metric data). It is important to recall a key point made by the online-commenter, Clinton Thomas, on his experience of learning how to classify skulls by race: "It's really not that hard once you know what to look for" (2011, np). It is also worthwhile to note an interesting parallel in opinion between this "normal college student" and some experts. For example, consider Gill's (2000) personal reflection on both his own and his colleagues' professional identification practices:

> I have been able to prove to myself over the years, in actual legal cases, that I am *more* accurate at assessing race from skeletal remains than from looking at living people standing before me. So those of us in forensic anthropology know that the skeleton reflects race, whether "real" or not, just as well if not better than superficial soft tissue does. The idea that race is "only skin deep" is simply not true, as any experienced forensic anthropologist will affirm (np).

Pan-disciplinary statements such as these, oversimplify human identification practices and overlook the scientific theory that underlies the robust forensic anthropological methods, and may overemphasize

the reliability of subjective experience and education. When their impact is considered, it is not unbelievable that, even when well-established statistical procedures are thoughtfully applied to standardized measurements, arguments against classification-based studies have garnered considerable scholarly attention (Goodman, 1997; Brown & Armelagos, 2001; Armelagos & Van Gerven 2003). Recent social and technical developments have also had a cumulative effect on classification studies and the discipline. These include the new demands placed upon expert testimony and evidentiary requirements in the wake of the Daubert ruling (1993), the renewed talk of race and ancestry in biomedical and genetics literature alongside developing technologies for genomic research (reviewed in Fujimura, Duster & Rajagopalan, 2008), the emergence of advanced morphometric equipment (e.g., digitizers, laser and CT scanners) and virtual anthropology techniques for large sample and multi-dimensional data collection, visualization, manipulation and analysis (Slice, 2005; Weber & Bookstein, 2011), the common-place access to fast high powered computers that now facilitate the implementation of complex statistical procedures and the quick execution of iterative algorithms, and the related dissemination of software (e.g., CRANID, FORDISC, 3D-ID) specific to forensic anthropological identification (Wright, 1992; Ousley & Jantz, 2005; Slice & Ross, 2009). In the wake of these developments it is not surprising that the field is witnessing a high yield of publications that set out to test, explain, reject or accept new and prevailing methodologies in ancestry estimation and classification and to finally respond adequately to the question first posed by Sauer (1992) almost 20 years ago: "if races don't exist, why are forensic anthropologists so good at identifying them?" (c.f., Ousley & Jantz, 2002; Freid et al., 2005; Williams, Belcher & Armelagos, 2005; Steadman, Adams & Konigsberg, 2006; Hubbe & Neves, 2007; Keita, 2007; Spradley et al., 2008; Hefner, 2009; Ousley, Jantz & Freid, 2009; Steadman & Konigsberg, 2009; Konigsberg, Algee-Hewitt & Steadman, 2009).

There are good reasons to be offered for adopting a multivariate metric approach to the quantification of human craniofacial variation (e.g., precision and reliability of measurements, the suitability of multivariate statistics to continuous and correlated data, the conservative behavior of continuous variation, and the moderate to high heritability of cranial architecture). Similarly, craniometric data has been shown to estimate ancestry with considerable accuracy when consideration is given to the context of recovery and when appropriate reference populations and priors are chosen (Howells 1966, 1969, 1972, 1973, 1989, 1995; Van Vark & Howells, 1984; Van Vark, 1985; Sjøvold, 1984; Devor, 1987; Jantz & Ousley, 2005; Carson, 2006; Pietrusewsky, 2008; Spradley et al., 2008; Konigsberg, Algee-Hewitt & Steadman, 2009). Nevertheless, critical opinions of forensic identification practices that make use of classification methods for the estimation of ancestry are highly charged and continue to be popularized alike in academic literature and in science-writing for public consumption. For example, Goodman (1997) has challenged identification validity by highlighting four "problem" cases. From these selected instances, he has determined that, for all ancestry estimation procedures employed in forensic anthropological assessments, the identifications made are not only inaccurate but random. Paraphrasing Boas (1940, p. 173), he has later argued with Armelagos (1998) that his assertions are accurate because "what is key to racial classifications depends largely on the previous experience of the observer, not upon the morphological value for the observed traits" (p. 359). Together, Armelagos and Goodman (1998) have also emphasized the consequences of these perceived methodological weaknesses and potential errors in allocation by offering the sobering caution that "the use of race in forensic research has probably led to countless misidentifications" (p. 370). Finally, in light of these concerns, Albanese and Saunders (2006) have presented best-practice recommendations that deemphasize ancestry estimation on the basis that "False or misleading information is far worse than a lack of information. The relatively high risk of false information may outweigh the value of determining 'race' may possibly have

for the positive identification of an unknown individual" (p. 281). It was Giles and Elliott's (1962) application of multivariate statistics to a sample of measured skulls, representing both modern (but turn-of-the century) American Blacks and Whites from Missouri and Ohio and prehistoric Native Americans recovered from the Indian Knoll site in Kentucky, that (1) first suggested the potential of discriminant function analysis (DFA) for the identification of "race" in medico-legal contexts (i.e., obtaining 80%-90% accuracy using a set of eight craniometric variables and specimens drawn from the same skeletal populations), (2) opened discussions of procedural concerns (Snow et al., 1978), and (3) initiated the skepticism which later brought about the rejection of their approach (Goodman, 1997). Armelagos and Van Gerven (2003), most notably, contend that "when applied to independent samples of known composition (the true measure of its success), the method is less than 20 percent accurate (Goodman 1997) -- a figure that hardly inspires confidence in forensic anthropology's ability to race a skull..." (p. 60). The dispute over the merits of statistical approaches to ancestry estimation did not, however, really begin in earnest until the first commercial release of the computer software, FORDISC 1.0 (Jantz & Ousley, 1992). These same critics have argued that through this software, a "seriously flawed" method of "skull diagnosis," with a demonstrated "poor performance," has been "sold." This debate has continued to mount, gaining in interest and, for some, in traction, with both the release of versions 2.0 (Ousley & Jantz, 1996) and 3.0 (Jantz & Ousley, 2005) as well as with the growing number of studies that seek to test the program's utility.

Physical anthropologists have long a history of using [anthropo- or skeleto-] metric data for the quantification of human skeletal variation: one that dates back well before the formalization of the discipline as a field of scientific study. Forensic anthropologists, in particular, have been aware for well over a half-century of the merits of applying discriminant function analysis (DFA) to craniometrics when

the estimation of sex and ancestry is required for medico-legal casework, in the context of "making" a presumptive identification of "unknown" human remains: this awareness can be witnessed in the landmark studies by Giles and Elliot (1962, 1963) and Giles (1964) (c.f., Steel, 1962; Hanihara et al., 1964; Birkby, 1966). Despite the methodological value of producing classification using an objective statistical procedure using standard measurements, many respected practitioners have continued to construct individual biological profiles using craniofacial estimates of ancestry arrived at through visual assessment (i.e., the anthroscopic or morphoscopic method) based upon differing experiential knowledge of the patterns of human cranial variation obtained from their personal observation of reference collections and their exposure to a unique combination of skeletal cases. For example, Stewart (1979) has advised that "Sometimes it is only the anthropologist's experience that tells him there is an undefinable 'something' about the skeleton that suggests one race over another"(p. 231), a sentiment that Hinkes (1993) has recently reiterated, stating that "A skilled eye and practiced judgment are the most valuable tools for a forensic anthropologist. Increasing experience will allow him to make sense out of small bits of evidence" (p. 52). For some, this emphasis on subjective assessment may be in part due to the computational demands and the prerequisite familiarity with the appropriate statistical theory that an application of a multivariate method, like DFA, to craniometric data would require of the practitioner. As remains are often fragmentary, varying in unpredictable ways between cases, sample-specific formulae need to be developed for each analysis (Ubelaker, 1998). It was not until Jantz and Ousley released the first version of their user-friendly software, FORDISC: Computerized Forensic Discriminant Functions, in 1992, that DFA was made practically available to the anthropological community at large and, in turn, fully accessible for use in forensic casework, regardless of each user's own level of competence in the application of quantitative methods.

Developed in association with the Forensic Anthropology Center (FAC) at the University of Tennessee, Knoxville, the current release, FORDISC 3, is defined by its authors as "interactive computer program, running under Windows, for classifying adults by race and sex using any combination of standard measurements" (FORDISC webpage 2011). FORDISC, in all three of its releases, draws upon the statistical theory that underlies the multivariate classification procedures known collectively as discriminant [function] analysis (c.f. McLachlan, 1992; Huberty, 1994). Since classification is a fundamental part of data organization in most fields (Keogh, 2005) and the family of discriminant function analysis, specifically, has experienced a lengthy history of development and application (e.g., Fisher, 1936), it is not surprising that this family of procedures is often described as "one of the simplest and most widely used classification methods" (Fielding, 2007, p. 100). Discriminant analysis (DA or DFA), as a tool for predictive modeling in its most general form, is concerned with the relationship between a categorical variable (a directly observable "group" or an otherwise known but latent "class") and a set of interrelated measured variables that characterize each observation or entity within the complete sample data set (McLachlan, 1992). If practitioners have prior information about a sample of cases, obtained from several predictors, they can investigate if and how well this data can be used to define groups (that is, *discriminate* between them). When classification accuracy is found to be high for the initial reference sample, this training data can be used to then predict the group to which a target "unknown" individual belongs. In an unsupervised learning based approach, an initial knowledge of correct groups status is neither algorithmically necessary nor theoretically desirable for parameter estimation, clustering and classification, so that the grouping variable is not predefined and the number of true groups does not guide the partitioning (e.g., in the context of finite mixture analysis or when adopting a model-based clustering approach, of which DFA can be seen as a specially constrained adaptation). The application of supervised classification techniques, like DFA, however, supposes that

the data in use consists of a finite number (g) of distinct groups or categories or classes or populations (denoted as $G_1,...,G_g$), whose existence are *a priori* known (McLachlan, 1992; McLachlan & Peel, 2000; Fraley & Raftery, 2002). These categorical identifiers are fundamental to the initial execution of the DF algorithm when analyzing training data, and are equally integral to the classification procedure that subsequently "identifies" the affiliation of the individual test or target cases. DFA strictly presumes that the individual of interest shares membership with one of these predefined groups, G_i, and that this membership is exclusive of all other groups, i.e., the categorical variable, z, denotes a singular affiliation, so that z=i indicates that the target belongs to group G_i (i=1,...,g). The p-dimension vector $x = (x_1,...,x_p)'$ contains the measured values on p available features of the case in question. While it can simply be said that discriminant analysis is, therefore, concerned with the relationship between the group identifier, z, and the feature vector, x, in actuality, there exists within this rubric a series of problems that lie along the range of its statistical methodology: from questions of inference to questions of decision. In the context of in forensic anthropology, where DA is used for identification purposes, work is done within the framework of decision, under which the intention is to estimate z using the information contained in x. Specifically, for the estimation of ancestry, the affiliation (a documented identifier, like race or ancestry) of the individual of interest or target (the unidentified, unclassified forensic case) is unknown. The goal of the analysis is, therefore, to facilitate the estimation of ancestry, contribute information to the biological profile, and, in turn, assist in making a presumptive identification. The practitioner uses DFA to perform an outright allocation of the unknown individual into one of the g predefined groups, which have been chosen for their relevance to the target specimen[s] and for which individual affiliations are *a priori* "known" (e.g., for a "forensic" prior, consider the modern, diverse and well-documented reference populations making-up the Forensic Anthropology Data Bank; see Ousley & Jantz [1997] and Shirley [2009]). These allocations, or *predictions* of membership, are based upon the unique

craniometric values that have been obtained from the target individual's remains, in accordance with some standard data collection protocol and following specific measurement definitions (Martin, 1956; Howells, 1973; Moore-Jansen, Ousley & Jantz, 1994). In FORDISC (Jantz & Ousley, 2005), the linear discriminant function (LDF) is used to convert the craniometrics into discriminant function (DF) scores using a linear combination of the original values that maximizes the separation, that is, the morphological differences between the preselected groups. The DF score of the unknown individuals can then be compared with the mean score calculated for each reference group and used to assign by-case membership with, or to "classify" these "unknowns" into, the group with the closest average score and with which they, therefore, share the highest probability of affiliation. If the analysis dictates the use of many, or at least more than two groups ($g>2$), it is possible to construct more than one discriminant function and, so, calculate multiple DF scores. In this case, a canonical variates analysis (CVA) is performed: group differences are determined from multiple axes and, although the allocation rule remains conceptually similar, the assignment of the unknown individual to its best-fit group it is now based upon the overall distance to each group centroid size (i.e., the group mean score in multi-dimensions) across all axes.

As with any investigative procedure, obtaining an optimal solution is heavily dependent upon the conditions of the analysis. Any statistical output obtained through the implementation of DFA is without meaning if due consideration has not been given to the factors that may weaken (e.g., the presence of outliers) or bias the analysis (e.g., sampling strategies, like resubstitution), or to those which violate key assumptions (e.g., the ratio of the sample size to the number of variables, distributional normality, variance-covariance matrix homogeneity, and the relationship in composition between the training data and the target's population-at-large) or invalidate the results (e.g., overfitting and inflated apparent

error rates) (Johnson & Wichern, 1992; Huberty, 1994; Freid et al., 2005; Konigsberg, Algee-Hewitt & Steadman, 2009). When the DFA expectations are met or are disregarded at the practitioner's discretion and their consequences on the solutions produced are acknowledged and accounted for in the design of the analysis, evaluation measures and validation techniques can provide information on the quality and the success of the training sample's partitions or the targets' classifications. Zigler and Phillips (1961) have argued that there are three main criteria when assessing the results of a classification approach: homogeneity (the similarity of individuals in the groups), reliability (consistency or agreement in who should be allocated to a particular group), and validity (how well group membership informs the analyst about their characteristics). The results of DF procedures executed in FORDISC can be easily assessed with these concerns in mind by using the raw information output and following the program-specific recommendations provided in the software's documentation. To determine how well the measurements selected discriminate between the chosen reference samples and the individual, FORDISC permits the user to first (1) review, in general, the classification matrix for confounding factors (like an inadequate comparative sample size or highly imbalanced subsample size ratios), (2) judge classification accuracy from the estimated error rates or cross-validated [LOOCV] percent-correct classification rates, (3) assess indicators of weak versus strong data structure (observing the natural log of the determinant or the p-value for the Kullback test of VCVM homogeneity) and (4) review the Mahalanobis distance (D^2) matrix and the significance of the derived distances (Jantz & Ousley, 2005). When the user is sufficiently confident in the strength of the classification, FORDISC offers standard probability-based measures that enable the user to then evaluate the quality of the individual "target" classification. Allocations of an unknown test or target individual to the "correct" group can be quickly assessed using two key indicators of reliability, which can be also used to develop decision thresholds that define "acceptable" classifications: (1) the posterior probabilities of group membership and (2) the typicality probabilities

(Jantz & Ousley, 2005). Although these tools are made easily accessible in FORDISC and these concepts may be familiar as useful membership statistics for other clustering or classification methods such as mixture modeling (McLachlan & Basford, 1988; McLachlan & Peel, 2000) or variants of discriminant function analysis (McLachlan, 1992), it is essential that the user understands their appropriate interpretation in the context of DFA and this software. While, for example, both posterior and typicality values are probabilities, each is concerned with different aspects of the analysis (posterior probabilities speak to the group allocations, while typicalities take into account the Mahalanobis distance or D^2) and therefore conveys different information. In this case, the posterior probabilities (PP) are generated under the assumption that the unknown individual belongs to one of the chosen reference groups (Tatsuoka, 1971) and, so, these values, which are here only *estimates* of membership, may vary considerably for the same target individual given which groups are included by the user in any given run of the program. Nevertheless, the authors offer guidelines for interpretation in the FORDISC documentation. As posterior probabilities represent the prior probabilities of group membership (i.e., those estimates obtained *before* any target metrics are considered, and treated by FORDISC as equal across the reference groups selected by the user) that have been modified with respect to the craniometric information representing the unique morphology of the "unknown" individual, they can vary between 0.00 and 1.00. In the ideal situation of a strong classification, however, the input case data will alter these prior expectations *a posteriori*, in such a way that all but one value produced are very small; so that the larger the posterior probability, the greater the confidence in allocation. Jantz and Ousley (2005) provide an explanation of meaningful posterior probability thresholds: "[a] high posterior probability (>0.9) indicates that the individual is more similar to that group than to any other and hence is more likely to be correct than one with a low posterior probability (<0.7). If posterior probabilities are spread out more or less evenly among the chosen groups, less confidence can be placed in the

classification..."(np). Although frequently appropriate, the uncritical use of this type of high-low decision criteria can mislead the practitioner into accepting classifications that are, in fact, weak in terms of their larger statistical foundation and often nonsensical when the context of recovery (i.e. the specimen-specific population-at-large), the population history and the microevolutionary processes are considered. For example, a posterior probability approaching 1.00 can result when a target individual has very unusual measurements (relative to the specimen's "true" origin) and this particular configuration of craniometrics can only be [possibly] found in the reference population to which it has been assigned the closest membership (via FORDISC's DFA). In order to check for misallocations that may be seen as artifacts of the analysis or complications arising from an individual's unique biological expression, Jantz and Ousley (2005, np) recommend that the user also inspect the typicality probabilities (TP) in three forms (Typ F, Typ Chi and Typ R, each of which have their own properties) for a more nuanced picture of the individual's allocation. These values can suggest measurement or data entry error, skeletal distortion (e.g., pathological abnormalities or boney alteration by postmortem taphonomic processes), or, perhaps most importantly, that the individual's true affiliation is not represented by the reference groups included within FORDISC's database of comparative populations. Jantz and Ousley (2005) propose that "in general, typicality probabilities above 0.05 can be ignored" and "values less than roughly 0.01 could be considered atypical, indicating that only 1% of that population, 1 in 100, would be expected to have such a differing morphology"(np). In cases when all typicalities are found to be low, therefore, the associated posterior probabilities and the assigned membership should be viewed with caution and, in turn, the reliability of the "identification" results questioned. As a typicality probability indicates where a given specimen falls within the variation expressed in a chosen reference group, the three typicality statistics calculated by FORDISC can offer the user valuable information on the likelihood that the target individual belongs to the group into which it has been

classified (e.g., Albrecht, 1992). As typicality probabilities are tied to the predetermined DF rules and the conditions of the analysis defined by the user and often constrained by the target specimen itself, their information value is not without external complications. For example, low typicalities can be, in part, user-induced when the chosen groups are very similar and too few predictor variables have been considered or high typicality probabilities can result when the target individual is craniometrically similar to the centroids of several of the preselected populations, but not very close to any one of these reference groups (yielding, therefore, low accompanying posterior probabilities). While most DF analyses produce results that fall in between these cases of extreme and discordant values (posterior and typicality probabilities of 1 or 0, which are in opposition to each other), and validation techniques and confirmatory statistics such as these are well-known to be useful and conceptually satisfying given their probabilistic approach in the "estimation" of membership, a knowledgeable evaluation of the interplay between the posterior and typicality probabilities remains a crucial analytical step in making a responsible metric assessment of individual ancestry (Freid et al., 2005). Moreover, the meaning of these analyses are reliable and their interpretation is informative only when they are properly understood and deployed in the service of trustworthy data, well-structured analyses and sensible tests of "unknowns" given the acknowledged limitations of the DF method and the FORDISC software. These are concerns that Freid and coworkers (2005) have already clearly illustrated with their test-classifications of a fantasy "alien" cast, a male gorilla skull and a soccer ball, for which they have reported that "The three 'unknowns' in this case show distinct East Asian/ Pacific Island affinities with the alien classifying as Guam male, the gorilla as Moriori, and the soccer ball as Buriat. These are 'strong' classifications of .966 to 1.000, yet all show typicalities approaching zero. These results indicate that soccer balls, aliens and gorillas are all extremely atypical of the reference groups in FORDISC and should not be classified by the program."

On the one hand, it is the true that for many exploratory cluster, pattern recognition, and classification techniques there may exist as many optimal solutions as there are ways to structure the investigation. This is a problem that has been recently highlighted by Relethford (2009) for the global partitioning of human craniometric variation into well-structured subsets when the identification of distinct populations (or "race" equivalents) is of concern. In this case, decisions made by the analyst regarding, for example, initial group partitioning or the prior choice of "true" cluster number, heavily influence what "correct" results are produced, how they are interpreted as "best fitting" and if, from these, the analysis is considered "successful." On the other hand, the classification of unknown individuals into ancestry groups by the application of discriminant function analysis to a suite of standard cranial measurements imposes restrictions upon the solutions produced for each case under evaluation. As discriminant function analysis requires that known groupings of some quantitative measures are used to infer the structure of the training data as a whole and categorize other target individuals, FORDISC's results are constrained by the predefined groups (G_i of the DFA) that have been embedded within the software and whose craniometrics compose FORDISC's reference database. For research in human craniometric variation, the "skeletal series" that typically make up each reference population and is "encountered most frequently in skeletal biology" represents "aggregates of individuals found in a specified area at a given time or some subgrouping below the species level" or "a collection of skulls housed in museums documented as having been collected from a specific village, island, or region" (Pietrusewsky 2008, p. 497). In general terms, these *a priori* "known" populations are all essential to the successful execution of FORDISC's "comparative" classification algorithm. For specific "target" analyses, a thoughtfully chosen subset of these (those sex-specific populations to which the practitioner believes, based on combined experience and deductive logic, that their "unknown" case is most likely to belong)

is essential to the good execution of an individual classification given the unique demands of the case under investigation. FORDISC 3 includes 12 reference populations selected from the University of Tennessee's Forensic Anthropology Data Bank or FDB (Ousley & Jantz, 1997). The FDB contains skeletal and demographic information for almost 3000 forensic cases representing multiple (including admixed) groups; a considerable number of which have been positively identified and have sex and race reported or known (Forensic Anthropology Data Bank website, 2011). The FDB sample included within FORDISC is composed mostly of modern individuals of both sexes and is associated with several biogeographic ancestries or self-reported races, including American White, American Black, American Indian Hispanic, Guatemalan, and Asian groups (Jantz & Ousley, 2005). As the FDB's contribution to the FORDISC database represents much of the current diversity of United States, it is, therefore, particularly relevant for use in the classification of medico-legal cases that have been recovered in North American contexts and/or are believed to have been drawn from this population-at-large (Ubelaker, 1998). FORDISC 3 also incorporates the full set of cranial measurements for a mixed sex sample of 2524 individuals from 28 pre-defined populations, compiled by Howells from 1965 to 1980. These data now comprise the Howells World-Wide Craniometric Data Set (Howells 1973, 1989, 1996), which has been used extensively for studies testing the patterned distribution of human craniometric diversity and issues of race (most recently, see Relethford [2009, 2010]or Konigsberg, Algee-Hewitt and Steadman [2009]). As the Howells populations demonstrate clear biogeographic separation (i.e., sampled for their far interregional geographic distances, to maximize comparative value) provide global coverage (from the Americas, Europe, Africa, Asia, to the Australia/Pacific Islands), and represent considerable temporal depth (from modern times to 600 BCE), their inclusion in FORDISC significantly broadens the range of unknown crania that can be allocated to one of the available reference groups with some statistically interpretable level of certainty and, in turn, whose group membership (in terms of individual ancestry or

biosocial race) can be estimated or identified with measurable success. Speaking cautiously to the fact that the Howells cranial measurements were taken on "museum collections of archaeologically recovered remains from around the world," Ubelaker, Ross and Graver (2002, np) have remarked that "although it would be surprising if an unknown forensic case originated from any of the specific Howells samples, the option could offer some potentially useful insight into the likely ancestral origins of the unknown individual." It is important to note, as Jantz and Ousley (2005) themselves have warned, that the inclusion of the FDB or Howells' groups within the FORDISC database does not imply that the full complement of samples will be available for use for every program run, even when a specific subset of groups is especially relevant to a given analysis. The availability of the comparative data is dependent upon the correspondence between the measurements collected on the individual of interest (the "unknown") and the unique configuration of values obtained for each reference crania, which may differ between individuals within and among groups due to data collection protocol: the complement of predictor variables considered (Howells, 1989; Hanihara, 1993, 1996; Hanihara & Ishida, 2001; FDB forensic data forms, 2011) and the degree of preservation or fragmentation of the remains (Waldron, 1987, 1994; Walker Johnson & Lambert, 1988). Sample sizes of the comparative groups may be significantly reduced by the variables selected (potentially necessitating the exclusion of key "probable" groups) and, as a result, the information value of the analysis may be limited. These constraints are also coupled with both theoretical and practical restrictions on the number of variables and groups assayed in the target analysis. Jantz and Ousley (2005) advise that a well-chosen subset of variables (suggesting 10 variables as a "reasonable minimum" and a maximum that is dependent upon sample size) that are the combined with few, but, highly relevant reference populations will yield the optimal classification results. While these expectations are accepted among many members of the larger statistical community (as they apply to human biological and non-human/non-biological data alike), there are

equally reliable studies that present conflicting results. Some have found that the use of a large number of groups leads to increased misclassification rates (Breckenridge, 2000), but others have shown that cluster or classification-based analyses run on data composed of very many true groups yield, in fact, better classifications (Kuiper and Fisher, 1975; Milligan, Soon & Sokol, 1983). While arguments for adhering to criteria that treat the problem of variable inclusion as an issue tied to sample size and tend to restrict the acceptable variable number following generalized "rules of thumb" (e.g., the number of variables must be at least three or four times the sample size) are commonly cited for their success and precedent (Huberty, 1994; Tabachnick & Fidell, 2001), studies that test increases in the number of variables used have demonstrated improved prediction and high rates of correct classification (Breckenridge, 2000; Hubbe & Neves, 2007; Lubke & Muthen, 2007) and have allowed for more permissive criteria to be created, i.e., the size of the smallest group needs only to be greater than the number of predictor variables (Tabachnick & Fidell, 2001). Like in most situations of applied data analysis, it can be seen that here too exists the practical problem of model multiplicity. In explanation, Faraway (2002) provides some thoughtful commentary:

> [a]nalysts have personal preferences in their choices of methodology, use software with varying capabilities and will interpret the same graphical display differently. In comparing the competing analyses of two statisticians, it may sometimes be possible to determine that one analysis is clearly superior. However, in most cases, particularly when the analysts are experienced and professionally trained, a universally acceptable judgment of superiority will not be possible (p.134-135).

This issue of uncertainty permeates all aspects of the discussion on the quantification of craniometrics variation presented here and it will be addressed to differing degrees of satisfaction by this project. As this project moves forward in a discussion of what Faraway (2002, ch.11) calls "statistical strategy" it is

important that it also retains the emphasis that he has placed upon "experience" and 'training." As "data analysis is not an automatic process," the better the choices made, the higher the success of the model, and the greater the robusticity of the results (Faraway, 2002, p. 134).

Forensic Anthropology and FORDISC

"But our results were all to hell, and I said to Cullumbine one day; 'You know, we really need to have a statistician look at these data because they are very variable.' And he said, 'Yes, I know, but we can't get a statistician; there isn't one available. What do you know about it?'" (Box in DeGroot, 1987, p. 239)

Although FORDISC offers a comprehensive package of statistical tools (i.e., from a simplified user-interface, to the integration of a stepwise procedure for variable selection to the automatic output of several validation statistics) and comparative samples (e.g., "world-wide" and "forensic" reference populations of known sex and ancestry/race) for the easy implementation of DFA for all levels of researchers, it is, as the authors have noted repeatedly, an "interactive computer program" (Freid et al., 2005, p.103; FORDISC webpage, 2011). The conditions of its proper use requires that "the researcher guides the analysis," with external, user-specific, consideration given to these factors as they relate uniquely to the case under evaluation (Freid et al., 2005 p.103). The forensic practitioner must, therefore, rely upon her breadth of experience and foreknowledge of the conditions of the remains and the context of their recovery to independently guide FORDISC's procedures (Freid et al., 2005). She must choose the best-available (typically sex-specific) subset of the FDB and Howells populations against

which to compare the target case (limiting the number of groups into which the "unknown" can be classified) and negotiate a balance among the number predictor variables, the amount of variation these are believed to explain, and their potential effect upon quantities of per-sample reference specimens available for comparison and, ultimately, membership allocation.

It is this "interactive" design, coupled with the requirements for *a priori* defined, program-specific, reference groups and *a posteriori* assessments of target memberships that are integral to FORDISC's discriminant function approach, that have been brought up as the points of potential critique. These concerns have been highlighted in discussions of (1) the general application of constrained clustering and classification methods to craniometric data, (2) the utility of FORDISC to the broad question of finding structure in human craniometric variation and its focused application to problems of human identification in the context of medico-legal casework and biodistance studies, and (3) the validity of any allocation tools that are used to estimate membership or corroborate biogeographic ancestry, biosocial or bureaucratic race, or other subpopulation proxies (like ethnicity or nationality) and, which, may eventually lead to assignments of "identity" (assistive or positive; historical or archaeological) based upon some measures of craniofacial similarity. FORDISC has received considerable attention in print from studies that seek to test its utility in isolated identification contexts and to evaluate general, or broad evolutionary, relationships between specific populations of interest and relevant target specimens using well-defined [paleo]demographic questions and knowledge of group histories. Moreover, FORDISC's methodology has been the central focus of discussion for projects that adopt a theory, philosophy, and even ideology driven approach, to test its accuracy in making correct classifications of cranial remains within the larger framework of appraising its value to "mainstream anthropology." Within this critical milieu, FORDISC has been assessed as a tool for understanding

human biological variation from the preconditioned perspective that population differences are neither scientifically quantifiable nor biologically meaningful and races [or forensic estimates of ancestry] do not exist.

FORDISC has been used in paleoanthropological, bioarchaeological and biodistance studies to identify patterns of relatedness and population-specific craniometric trends. These analyses operate under the assumption that "populations that share recent common ancestry and or exchange a large number of migrants should resemble one another more than geographically isolated and distantly related populations" (Roseman, 2004, p. 12824). In these cases, similarity to one of the software-specific reference populations does not necessarily imply a straightforward classification in terms of "specific identity or origin" (Leach et al., 2009, p. 549). The goal is, therefore, often exploratory, where the user takes advantage of the fact that FORDISC must classify all crania inputted "into the group with which they have the closest affinity" (Spradley, Ousley & Jantz, 2008, p. 199). These studies have achieved varying degrees of success and demonstrate the potential strengths and weaknesses of this approach.

For questions of archaeological significance, Mangold and coworkers (1993) used FORDISC to validate the identification of a single set of remains from for a well-preserved Albee Phase burial as those belonging to a Native American female. Using two groups and 21 traits, they found that the software "strongly aligns the specimen with Amerindian females rather than males (n=73; posterior probability = 1.000; typicality probability = 0.04)," which corroborated the determination made from the observation of discrete cranial and pelvic traits (1993, np). In 1998, Maish and Fukuzawa reported on their use of FORDISC to classify 59 complete and fragmentary crania representing male and female adults from two identified Ontario Iroquoian sites. Using the forensic reference groups and a selection of the most

relevant populations from Howells' series, these authors concluded that "FORDISC 2.0 is not a reliable package for determining the racial affinity of Ontario Iroquoian crania" (1998, np). In a recent study with similar goals but different geographic concerns, Hughes and coworkers (2010) investigated the ability of DFA to differentiate between Californian prehistoric/historic Native American remains from modern forensic cases using craniometric data that is comparable, in part, to the data included in the current version of FORDISC (e.g., the Santa Cruz Island [Chumash] and Guatemalan [Mayan] samples). They also evaluated the classification accuracy obtained from a morphoscopic analysis using seven variables. Their DF test using only nine craniometric variables and two test samples produced a group classification rate of 87% correct, while they found that morphoscopic data predicted membership with a reduced accuracy of 77%. These results support, therefore, the argument that that it is indeed possible to differentiate between these chosen populations with considerable accuracy using craniofacial data and that morphometrics yield the best results. These authors (2010) contended in their summation that "such a method contributes to the efficient and empirical determination of temporal and geographic affinity, allowing for the repartriation of Native American remains to their tribes, as well as the accurate analysis of forensically significant remains" (np).

Lovvorn and colleagues (1999) have demonstrated how FORDISC can be applied to questions of microevolution and paleopopulation history in the Americas with their analysis of a middle archaic burial from Sidney, Nebraska. Their results indicated that the Sidney male is not strongly affiliated with any one of the 10 preselected archaeological reference groups, but shares the greatest best-fit membership, first, with the Inugsuk Eskimo from Greenland (with a posterior probability of 0.374 and typicality of 0.487) and, second, with the Ainu of Japan (with a posterior probability of 0.18 and typicality of 0.328). Lovvorn et al. (1999) explain these weak classification results in terms of the limited data extractable

from the fragmentary remains (i.e., the use of too few craniometric variables as only 6 measurements were available due to missing data for the orbits and midface) and as indicative of an "intermediate morphology not well represented by any recent population" (p. 539). While they "consider both factors likely" (p. 539), they nevertheless rely, in part, upon these results to draw conclusions on the broad question of the peopling of the Americas. They propose that the Sidney male presents a "blend of Amerindian and earlier protomongoloid traits" (p. 527), which they believe to be consistent with the "hypothesis that Plains Amerindians descended from the earliest wave of Paleoindians who crossed the Bering Straits" (p. 527). A similar analysis of the Sidney cranium using FORDISC was conducted by Jantz and Owsley (2001) who, on the one hand, also demonstrated that "it is atypical of any historic Plains tribe" but, on the other hand, obtained an alternative best-fit assignment. Their different [and less constrained] choice of comparative populations produced a different classification; in this case it was allocated into the Tasmanian group. Bulbeck's (2005) work on the larger project, *The Contribution of South Asia to the Peopling of Australasia* (Bulbeck et al., 2003), has similarly demonstrated the applicability of FORDISC to reconstruction of the evolutionary trajectory of modern humans at different time-depths and geographic locations. Tests of Australian fossil crania against preselected subsets of the Howells populations using FORDISC produced results of affiliation that empirically corroborate the "widely recognised" belief that "Australian skulls at the Pleistocene/Holocene boundary differ metrically from recent Australian Aborigines" (Bulbeck 2005, np; c.f., Brown, 1989). Although straightforward identifications are inappropriate given the limited populations represented in the FORDISC database, the classification of the Sahulland fossil crania as Tasmanian and Tolai (and not as Australian), for example, support the argument that "during the late Pleistocene, Tasmania and New Britain (via New Guinea) were connected to mainland Australia, and accordingly all their inhabitants constituted a single interbreeding population" (Bulbeck 2005, np). These FORDISC-based findings here too shed important

light upon the history of population relationships and offer external evidence for theories of group interaction in the distant past that have come to shape current craniometrics trends.

Williams, Erickson and Laird have demonstrated the utility of FORDISC to questions of population homogeneity for the purpose of archaeo-historical reconstruction of site histories and demographic profiles. In 2000, they reported on their application of FORDISC to a mass burial of partial and complete skeletal remains of 30 individuals recovered from the 18th Century mortuary of the Little Dutch Church, in Halifax, Nova Scotia. While it was originally assumed that these skeletons represented a single cemetery population of German colonists, the results from their FORDISC analysis indicated that this mass grave also contained the remains of "non-Europeans,"which agreed with the findings from additional dental and biometric studies. With a similar goal of teasing out within-sample structure, Leach et al. (2009) recently employed FORDISC "to investigate and quantify the diversity noted during the anthroposcopic assessment" of two Romano-British archaeological populations: the Railway and Trenholm Drive cemetery groups in York. While they doubt that the true population origins can be accounted for by the reference samples included in FORDISC and acknowledge even the impossibility of defining a "typical Roman" reference sample due to considerable urban fluidity and migration, they believe that the craniometric variation captured within FORDISC's comparative data set should reveal broadly based trends in affiliation. As Keita (2007) has also suggested, "FORDISC can be used as a tool for comparative morphometrics anatomy, treated as an exploratory instrument, whose database contains a range of craniometric patterns" (Leach et al., 2009, p. 557). The results of their craniometric analyses indicate that 11%-12% of the samples were likely of African descent, which agrees well with external evidence attesting to an urban, mixed military and civilian population in York, comprised of

varied ancestral heritages and geographic origins and including "strong affinities to populations from North Africa"(Leach et al., 2009, p. 559).

From a historico-forensic perspective, Stojanowski and Duncan (2008, 2009) applied FORDISC and independent LDF analysis to assess the population affinity of a single partial cranium, "The Fort King George Skull" (FKG-121), and, in turn, speak to its positive identification as one of the "Georgia Martyrs" killed in Spain during the Guale revolt of 1597. Using the European American, African American, Native American and Hispanic samples available in FORDISC 2.0, the specimen was classified as European American but with a low typicality probability of 0.072, which the authors attribute to the modern composition of the FDB. An additional comparison of FKG-121 against the Howells populations failed to offer improved results; suggesting an Egyptian affiliation with a slight increase in typicality (0.106). These authors reported, however, that, when conducting LDF analyses independent of the program's constraints, the accurate classification of FKG-121 was achieved when archaeological and historical data was used to gain access to the occupational history of the recovery site and to reconstruct the demographic profile of the population-at-large with which this individual most likely shared membership. They (2009, p. 275) conclude that their study of FKG-121 "highlights the unique challenges of historical forensics cases relative to those of traditional jurisprudence, as well as the potential for using historiography to overcome those challenges in future analyses." Seidemann, Stowjanowski and Rich (2009) have attempted a similar historico-forensic analysis to provide identification information on a single human cranium, Skull 33-238, seized from an eBay sale by the State of Louisiana. Making use of sex-specific reference samples from the FDB and Howells data sets, they conducted independent DF analyses to assess population affinity for the "eBay cranium," which had been presumptively identified as a female of Native American origin, a probable disinterment, and a case of archaeological (and

NAGPRA) concern. Classification results using the FDB samples allocated Skull 33-238 into the Native American female group with a posterior probability of 0.659 but a low typicality probability of 0.09. The DF analysis using the Howells populations offered comparable results, assigning Skull 33-238 to the Arikara group with an equally high posterior probability of 0.681. Although the associated typicality value remains low at 0.263 in this second attempt, its improvement when using the Howells data suggests a better fit with Plains populations from the colonial period than with the modern FDB individuals, many of whom represent Native Americans from the southwestern U.S. (see Jantz & Owsley [2001] for a discussion of the morphometric variation in early North American crania and the history of population relationships that this craniometric data suggests). The improvement in typicality with the increase in sample age suggests that the failure to achieve optimal classification may be related to the limited temporal depth represented in the comparative populations, coupled with the conditions of the remains (the presence of cranial modification). As this study demonstrates, time is a potential factor in confounding identification when the antiquity of the remains is not known or is not adequately accounted for by the reference samples available for use in the analysis. The authors conclude that "caution is warranted and these results should be interpreted appropriately" (2009, p. 1250). Wright, Williams and Crass (2003, p. 93) have expressed similar concerns given the low typicality value (0.051) obtained for their FORDISC assessment of a 19th Century burial on Sapelo Island, Georgia, by which the "race" of the "Sapelo Island Man" was identified as European (posterior probability of 0.992). As the authors state, "the results of FORDISC 2.0 should be taken with caution because this program was developed using modern skeletal remains and its accuracy in classifying remains from the historic period is unknown" (p. 93).

Despite the acknowledged complexities in the application of FORDISC to studies involving questions of considerable time-depth and "skeletal populations" for which appropriate samples are presently lacking in the software's reference database, it is the use of FORDISC in the service of its primary design as an assistive tool for forensic anthropologists that has been the single greatest source of critique. As an aid to human identification for medico-legal purposes, FORDISC's greatest analytical strength lies in the inclusion of comparative cases obtained from the Forensic Anthropology Data Bank (FDB) that represent both the modern American population-at-large and account for the diversity observed in much of the contemporary casework conducted in North America (Ousley & Jantz, 1997; Ubelaker, 1998; Jantz & Ousley, 2006). These contributed "forensic" samples have enabled the widespread use of the FORDISC software, both at home and abroad, as a valuable resource when making individual identifications in collaboration with the law enforcement community and for use in multi-team projects related to the documentation of human right violations or in mass disaster recovery. The important role that FORDISC plays is made clear, for example, in a press-report posted on the Arlington National Cemetery's website (2000), which speaks to its use in recovery and identification operations by the Army Central Identification Laboratory (CILHI):

> [t]he lab tapped into the smarts of a forensic computer program developed at the University of Tennessee, called ForDisc, which automates the process of matching skeletal remains, Byrd said. ForDisc is based on an extensive skeletal database that comprises samples of racial and body types found throughout the population, Byrd said, and allows scientists from CILHI to quickly determine the probability of whether a femur of a certain length matches a tibia of a certain length, for example. Recent new methodology extends that capability to bone fragments as well. This is a key piece of software, because the CILHI scientists work "blind" when they begin analysis

> of skeletal remains, with no prior knowledge of the physical characteristics or even the number of individuals involved in an incident, according to a command briefing. It's also more useful than DNA in cases where the number of individuals involved raises the possibility that the same base pair sequence will show up in more than one set of remains, Byrd said (np).

The impact of FORDISC on the forensic anthropology community is also obvious from the ubiquity with which it has begun to appear in laboratory manuals, case studies and textbooks (Burns, 1999; Fairgrieve, 1999; Byers, 2002; Steadman, 2009), and the regular demand for training workshops on its use by forensic practitioners (Jantz & Ousley, 2008). While FORDISC's history of use is short, it is notably varied and intensive. Its history of publication is, however, relatively limited. Positive or explanatory reports are few in number (Freid et al., 2005; Ousley, Jantz & Freid, 2009; Konigsberg, Algee-Hewitt& Steadman, 2009), the interpretations given to the available FORDISC "tests" are many and conflicted (Uberlaker, Ross & Graver, 2002; Hubbe & Neves, 2007; Keita, 2007), and the most critical of these have tended to gain the greatest scholarly and professional attention (e.g., Kosiba, 2000; Belcher, Williams & Armelagos, 2002; Leathers, Edwards & Armelagos, 2002; Williams, Belcher & Armelagos, 2005; Wienker, 2005; Elliott & Collard, 2009).

In 1994, Groh carried out a test of the first version of FORDISC using 63 adult crania of forensic significance obtained from a skeletal collection of 414 cases in Dade Country, Florida. She also made craniometric comparisons between the Black and White individuals of both sexes from her test group and craniofacial data obtained from samples representing 19th century American Black and White populations, including the Terry and Hamann-Todd anatomical collections. While Groh (1994) found that FORDISC achieved a classification accuracy rate of 70%, she also detected significant changes in cranial

morphology over time, which she describes as an increase in basion-bregma height in White males, White females and Black females. Given these observations, she concluded that, while FORDISC was moderately successful in its ancestry-sex assignments, the identification of secular change in the cranial architecture of some groups has "serious implications in the identification of unknown individuals" and "physical anthropologists must update their criteria to determine race and sex in accordance with these variations" (1994, p. 0368). These relatively recent changes in cranial morphology, especially in the last 150 years, have been more thoroughly addressed for American White and Black groups in recent studies that make use of both traditional craniometrics and the potentially more revealing geometric morphometric methods (Jantz & Meadows Jantz, 2000; Jantz, 2001; Wescott & Jantz, 2005). Findings have shown collectively that (1) crania of both groups have become higher, longer and narrower, with a notable inferior shift of the cranial base, especially at the landmark basion, that (2) secular change is proximately related to a decrease in cranial breadth and an increase in cranial capacity and (3) it may be best explained as the result of novel environmental conditions that offer improvement in nutrition, health and disease control during infancy (Wescott & Jantz, 2005). The acknowledgment of the confounding factor of cranial change highlights the importance of choosing appropriate reference samples for DFA comparisons and the potential weaknesses of FORDISC in classifying "modern" individuals for whom a birth cohort, and thus their contemporaries within the FDB samples, may not be easily determined (this a problem that becomes magnified in the negotiation of relationships between archaeological and present-day specimens).

More recently, Ubelaker, Ross and Graver (2002) have taken up this issue of reference sample incompatibility and the problems that it may present for FORDISC's widespread applicability. As they (2002) explain,

> FORDISC 2.0 has emerged as a powerful tool in the forensic analysis that is routinely employed in most North American forensic laboratories. Because the Forensic Data Bank was developed largely from North American forensic cases, its obvious use would be in applications to North American cases... With the increase in today's world travel, a representative of almost any population in the world could end up being a forensic case in almost any place in the world (np).

Ubelaker, Ross and Graver (2002) test FORDISC using an archaeological skeletal population from Spain that is both temporally and spatially unrepresented by the Howells data set and for which the use of current FDB correlates (e.g., Hispanics, White Americans) is presumably ill-conceived. Now under curation in Madrid, this collection originated from a large secondary ossuary recovered from the Church of Santa Maria (in Wamba, northwestern Spain), and has been "likely" dated to the 16th and 17th centuries. Although the authors note a considerable degree of morphological heterogeneity within this burial group from their visual assessments and despite the fact that neither sex nor unique ancestries are documented for each individual, these remains are believed to represent a relatively balanced sample of local individuals associated with the Church, who are therefore best identified collectively as "Spanish." Given these conditions and the reasonably large sample size of 95 measured individuals (58 presumptive males and 37 presumptive females), the authors contend that this collection can be used with success to examine the behavior of FORDISC when the program is applied to a group of target specimens that is geographically and temporally cohesive, morphologically diverse and whose population designation is not directly represented by the Howells or FDB reference samples in the software's (version 2.0) comparative database. As sex determinations were made from the assessment of gross morphological indicators and prior "sex-only" tests of FORDISC showed the difficulty in corroborating the morphoscopic estimation of sex using the craniometric data alone (with overall small

size being the likely confounding factor in distinguishing male from female crania), the authors conducted two sets of tests for each sample comparison to account for this issue of sex. Their analyses obtained the following results using the FDB: when sex was disregarded, FORDISC classified the crania across seven race-groups. Among these groups, 44% were classified as White, 35% as Black, 9% as Hispanic, 4% as Japanese, 4% as American Indian, and the remaining (3) individuals as Chinese and Vietnamese. Although the dominant group switched from White to Black when sex was considered, the distribution of the outstanding individuals remained similar. The analysis using the Howells populations produced a similar pattern of classification: the test specimens were most readily assigned membership with the Egyptian group and the classifications produced for the remaining crania were widely dispersed over the "other" (21) groups. A greater disparity was observed in the male-female assignments when using the Howells database, however: 58% of the test sample was classified into European or North African groups when sex estimates were discarded, while a close 55% was allocated into these same groups when the sex information, obtained by morphoscopic assessment, was used. Regardless whether sex was considered known or unknown, the single most common classification consistently remained Howells' Egyptian population. For both analyses, the typicality values associated with the individual allocations offered limited insight, as they were split between the European or North African groups and the other FDB and Howells groups assayed (for example, when sex was "known," European or North African groups were considered typical in 63% of the classifications, and 51% percent of the targets with typicality values greater than 0.05 were classified as European or North African). The variety in the "racial classification" (i.e., with almost always, and often more than, half of the male and female samples being assigned to many "other" groups) and the frequent allocation of the Spanish crania into groups with no interpretable [bio]geographic or strong [bio]historical relationships led Ubelaker, Ross and Graver (2002) to argue that "the samples clearly were inadequate to elucidate the specific geographical

origin of the overall Spanish sample...it is more likely that the Spanish sample simply is quite different from samples in the existing databases" (np). These problematic findings and the overall "variety of results" demonstrate the potential limitations of FORDISC's utility outside of the context of North American casework and, so, lead Ubelaker, Ross and Graver (2002, np) to conclude that "caution is called for when applying FORDISC 2.0, as well as other similar discriminant function approaches, to samples that are not well represented in the relevant databases." They (2002) offer a recommendation of FORDISC with reservation: they identify the need to "improve an already useful forensic tool and make it more applicable to international forensic cases" and advise that " such approaches should complement, not displace, overall evaluation of remains" (np).

In 2005, Wienker reported on a series of tests of FORDISC using craniometric data from Cuba and India, representing modern cases of [mostly] known genetic ancestry. In the discussion of his research design, he indicates that his intentions are (1) to evaluate the program's ability to classify the individual crania into their correct racial category (or category of closest affinity in the case of the Indian sample) and (2) to test its ability to provide consistent classifications when the target samples (e.g., the Indian crania) are not directly represented by the reference group in the FORDISC database. He offers this work, therefore, as a "modern" extension of the tests already conducted on archaeological samples. It is of particular interest that Wienker (2005) also writes that "the purpose of this report is to demonstrate the premise [that] human races do not exist except as social constructs, convenient mental typologies of common agreement" (np). It can be seen from this perspective how classification can be inextricably tied to racial taxonomy in the minds of those who simultaneously engage in the anti-race debate and reject the meaningful patterning of human craniometric variation. This reveals how, by extension, any acceptance of FORDISC's analytical utility can be construed to offer scientific justification of the race

concept, just as the identification of FORDISC flaws can be perceived as giving validation to the arguments that support the non-existence of biological races and which contest the skeletal estimation of biosocial group membership that is core to the practice of forensic anthropology. Winker (2005) himself best summarizes his methodological approach in support of his thesis:"...the method of demonstration is the result of the application of FORDISC analysis..." (np).

Wienker begins his analysis by addressing the now well-known issue of choosing the comparative (training or control) data necessary for executing any DFA. In order to determine the best subset of reference populations representative of the target sample under investigation, Wienker adopted a selection procedure based on the posterior probabilities obtained for each run, which he applied consistently to each sample:

> In each analysis, the same methodology was used. Initially each skull was analyzed using all 11 demographic categories. Categories that yielded posterior probabilities (the probability that the unknown skull falls into the specific demographic category) of less than .05 were discarded and the analysis was run against the remaining categories until all demographic categories yielded posterior probabilities of at least .05 (2005, np).

Using the groups preselected in this way, Wienker proceeded to analyze four different target samples separately: a series of 35 White Cuban males, 16 crania originating from the Indian subcontinent, a Havana cemetery population of 47 Chinese Cuban males, and a sample of 50 Cuban forensic cases obtained from the Office of the Havana Medical Examiner. Wienker's test of FORDISC on the White Cuban sample produced a classification accuracy of 56% (with 16 of the 35 individuals identified as "White" and 4 identified as "Hispanic"), with the highest posterior probabilities spread across the White

(0.58), Hispanic (0.46), Black (0.45) and Japanese (0.53) groups but with the highest typicalities associated with Amerindian (0.99) and Chinese (0.825). He does, however, obtain improved classification accuracy for the Havana ME sample, with these forensic cases being correctly classified into their most relevant group at a rate of 78%. While Wienker finds neither of these percent accuracy values sufficiently high or meaningful and takes these as evidence of the poor performance of FORDISC, in reality his results should be neither surprising nor dismissed. They may, in fact, be very useful in offering insight into how the FORDISC software deals with cases that are not heavily represented in the U.S. population-at-large and not well-represented in its comparative database. Moreover, these results may aid in educating FORDISC users on the difficulty in making good connections between the available reference samples, the context of the remains, and the often ambiguous [self-]identifiers that are used to characterize certain admixed individuals or encapsulate multiple ethnic/geographic/ ancestral groups. The inconsistencies and inaccuracies in the classification of the "Cuban" crania can be most profitably explained, therefore, in terms of the complexity of the term "Hispanic," as a social identity and as a biological category. Many studies have clearly demonstrated the problems with obtaining or defining a "good"classification accuracy for "Hispanic" skeletons (c.f., Ubelaker, Ross & Graver, 2002; Slice & Ross, 2004; Ross at al., 2004; Ross, Slice & Pachar, 2005; Spradley & Jantz, 2005; Spradley & Algee-Hewitt, 2007; Spradley et al., 2008). Difficulties arise as any identification of "Hispanic"may include within it many populations, the exact combination of which may change with geographic location and even vary within in the same country or region (e.g., Hispanics of the American Southwest versus the Southeast). The collective term "Hispanic"may therefore characterize individuals and encapsulate groups with quite different ancestral or genetic histories (e.g., Hispanics from Mexico, the Caribbean, or of White European-Spanish descent). Jantz and Ousley have frequently acknowledged the potential problems with the Hispanic group included in the FORDISC software in their public discussions and workshops and,

to this end, they provide the following cautionary statement in FORDISC 3's documentation: "This group has been the most problematic as far as 'race' is concerned, and includes individuals born in the U.S., Mexico and Central America. There are no individuals born in the Caribbean area, who more often classify as Black (Slice & Ross 1994). Most of these come from New Mexico and some of their identifications are based on context" (2005, np). When Wienker's results are reinterpreted and their cultural and bio-geographic context considered and with the composition (and small sample size of only 88 males) of the Hispanic group in FORDISC known, then his two sets of Cuban samples should be treated as target cases for which there are no identical reference populations available, even in the current, updated version of the software. It should therefore be expected, as his findings show, that each "Cuban" individual would be allocated to the group with which he has the greatest ancestral affinity and that this membership would be made with limited statistical confidence, in the absence of a perfect target to training sample match. Similar arguments can be made for the nature of the classifications obtained for the crania representing the Indian subcontinent. In the absence of a comparative "Indian" population in FORDISC, Wienker focuses his test on the program's ability to consistently classify this agglomerate of target individuals into the available populations that are closest geographically to India, e.g., as Vietnamese or Chinese. He reports that FORDISC failed to classify the Indian sample into these geographically relevant categories and instead produced moderately high posterior and typicality probabilities (suggesting "good" allocations) for a variety of groups: White, Black, Amerindian, and Japanese males and females. Although Wienker (2005, np) does explain these poor classifications in terms of "the fact that the peoples of India are an extraordinary genetic hodge podge...who have substantial hereditary influences from East Asia, Southwest Asia, and Europe," he fails to recognize that, given the absence of a comparative sample (and the complicated links between biology and race in the presence of the Indian caste system and religious/social taboos and the many

subpopulations that these cultural factors engender and support, e.g., Brahmans versus Dravidians), the goal of his analysis should be to tease out information on ancestry from the allocations made by following bio-distance concepts and knowledge of population affinity. Instead he employs his findings to support an earlier study conducted by Kosiba (2000), which also simply stated that the FORDISC program is unable to produce consistent classifications. Finally, in contrast to the other samples tested, Wienker (2005, np) reports that FORDISC did "do a pretty good job of classifying the Chinese skulls as Asian." He is quick to clarify, however, that "it could not accurately identify even half as actually Chinese." This discrepancy between a correct, but general classification as "Asian" at a rate of 83% and the more tenuous allocation of Wienker's specimens to FORDISC's Chinese group specifically (49% classification accuracy) may speak more to the incompatibility of the target and reference samples than to FORDISC's ability to successfully assign membership to Chinese individuals when the software is employed in modern forensic contexts. As FORDISC's Chinese sample is a subset of the FDB groups and, so, is composed of contemporary individuals (specifically, Hong Kong University cadaveric specimens [Jantz & Ousley, 2005]) and since Wienker's cemetery population of 47 "Chinese Cuban male skulls," although presumed to be "Chinese in ancestry," are "probably not completely modern...some may date to more than 100 years ago, if not somewhat longer" (2005, np), his straightforward comparison of two samples representing two or more different time-depths fails to consider the potential problems of craniofacial change over time and the issue of antiquity, especially when dealing with possibly century-old cemetery remains which may be better considered archaeologically significant than exemplary of current forensic casework. Despite these oversights, Wienker (2005) concludes, for his four analyses combined, that "[t]he results indicate that the program does not classify almost all of the skulls into the correct racial category, nor does it classify some skulls consistently. In some cases, it is poor in correctly

classifying skulls with regard to traditional racial categories. The primary reason for the inaccuracy and inconsistency is that human races are a social construct, not a biological reality" (np).

Of the research projects that have set out to evaluate the FORDISC platform (from its user-interface, to its algorithmic underpinnings, to its control data sets), perhaps the best known are three complementary studies that have taken ancient Nubian crania as the subjects of their analyses. While there is no exact population match in FORDISC's comparative database, the Nubian samples are expected to share biological affinity with "their nearest neighbor," Howells' ancient Egyptians, or with other African populations (Williams & Armelagos, 2007, p. 285). These projects are, therefore, heuristic in their design; basing their analysis of "FORDISC 's practical use" in forensic anthropology on the principles of relatedness that are common to biodistance studies and to trends in bioarchaeology (Williams, Belcher & Armelagos, 2005, p. 341; Buikstra, Frankenberg & Konigsberg, 1990; Larsen 1997, 2002; Pietrusewsky, 2008). Motivated by the perceived problem that "practitioners of this program are using a tool that has not been systematically tested for validity," these projects collectively seek to realize three goals, which are both methodological and theoretical in nature (Williams, Belcher & Armelagos, 2005, p. 341). Their intention is to not only (1) test the software's classification accuracy but also to (2) critically engage the theoretical and methodological principles that underlie its design and implementation and (3) challenge the conceptual foundations that support the quantification of human biological variation and enable the estimation of the unique parameters of an individual's identity (Belcher, Williams & Armelagos, 2002; Leathers, Edwards & Armelagos, 2002; Williams, Belcher & Armelagos, 2005). While these studies may represent their author's initial tests of the FORDISC software, they are built upon a foundation of arguments that conflate the study of race, ancestry, and the patterns of human cranial variation with [mis]assumptions, bias and information value: arguments

to which they themselves have contributed. Consider that, in the most recent of these reports, Williams, Belcher, and Armelagos (2005) begin their discussion of FORDISC with the opinion that

> Forensic anthropology has been much more reluctant to divorce itself from the premodern partitioning of human biological variation into races (Smay & Armelagos 2000), despite the fact that human biological variation in genetic markers (Lewontin 1972, Stoneking 1993) and cranial morphology (Relethford, 1994) is quantitatively greater within than between major geographic regions or races (p. 340-31).

Their arguments also align well with earlier critiques of the use of discriminant functions for the metric estimation of ancestry and the identification of race. For example, in a review publication that seeks to represent the state of the field of skeletal biology, Armelagos and Van Gerven (2003) foreground findings that argue against the utility of DFA, and, in doing so, call into question this important part of forensic anthropological practice: "Goodman (1997) demonstrated that the 85-90 percent accuracy claimed by forensic anthropologists is seriously misleading. High levels of accuracy can be achieved only when the skulls meet extremely limiting criteria" (p. 60).

In 2002, Belcher, Williams and Armelagos assessed FORDISC's value as a tool for the determination of sex and ancestry using data collected on an archaeological population representing Meroitic Nubians. For a sample of 47 crania they report inconsistent membership results for both the attribution of sex and the identification of biological affinity. Based upon the poor classifications obtained, they not only dismissed the program as "flawed" (c.f., Kosiba, 2000) but also extended these findings to a larger commentary on disciplinary problems, through which they argue against "the utility of any forensic application that attempts to constrain worldwide human cranial variability" (2002, p. 42). In a second study in 2002, Leathers, Edwards and Armelagos tested FORDISC on a larger series of post-Meroitic

Nubian crania. For a sample of 89 individuals, they also achieved low accuracy for their expected best-fit classifications (e.g., Egyptian) and, even when the more general biogeographic allocations were accepted and combined, only 57% of the test sample was classified as African. From this limited success, Belcher, Williams and Armelagos (2002) concluded that FORDISC-based classifications are "not morphologically or biologically accurate" (p. 99). The conclusions drawn from these preliminary tests were seized upon in subsequent studies by other researchers with similar interests, who shared their philosophic perspectives on human variation. These findings not only gave further credence to arguments against the use of DFA, but also provided the scientific results necessary to issue empirical statements on its execution in FORDISC and its implications for the ongoing debate over the quantification of ancestry and race in human identification. Armelagos and Van Gerven (2003), for example, argued shortly thereafter that "Poor performance has not disabused forensic anthropologists from selling the method. Fordisc 2.0 (Ousley & Jantz 1996) is a computer program designed to diagnose any skull into one of Howell's geographic populations. The program, however, is seriously flawed..." (p. 60). These Nubian studies also gave support to the closely-related arguments that warn against using population-based approaches in anthropology in light of its lengthy and troublesome tradition of human taxonomy. Williams, Belcher, and Armelagos (2005) have, for example, used these findings to tie this history to the present discussion of FORDISC: "Forensic anthropology has often been called upon to substantiate typological thinking and to reinforce the type concept itself"(p. 344). Armelagos and Van Gerven (2003) have also brought their interpretation and research direction to a broader focus: "[w]hat we see with the African test is the result of an astounding mismatch between actual cranial variation and the variation modeled by racial constructs.... Sadly, the response has been directed more toward fixing the program than fixing the approach" (p. 60-61).

Certainly the best known of these three projects, is the last, in which Williams, Belcher and Armelagos (2005) retest the efficacy of FORDISC as a "computer program designed to standardize the attribution of population affinity for forensic analysis and to provide a user-friendly method for ascribing 'social race' to unknowns" (p. 341). While their initial goal is procedural, attempting, again, to classify a sample of Meriotic Period Lower Nubian crania into the population with which they demonstrate the greatest phenetic similarity, these authors also expand the focus of their analysis to address what they feel are the larger theoretical problems that are driving the high rates of misclassification obtained by this and the previous (2002) analyses using this particular population of interest: (1) regardless of the combinations of samples and the reference databanks (e.g., Howells or FDB) used, FORDISC is a poor predictor of biological affinity or direct membership because its analytical design is dependent upon statistically contrived or culturally mediated, but not necessarily biologically defined, groups and, so, it cannot account for the heterogeneity that one would expect to observe in a single natural population, (2) FORDISC is an equally poor predictor of biological affinity as its comparative groups represent artificial constructions of modern design and the complex biological history, and, in turn, the diversity in morphological expression and relationships of real populations cannot be fully accounted for by these static agglomerates or "types," and (3) FORDISC is "fundamentally" and "conceptually flawed" because human biological variation is not organized into a pattern of closed units or true biological entities (like species) that human taxonomic systems, like race or ancestry, imply and discrete classification approaches require (Williams, Belcher & Armelagos, 2005,p. 344; Williams & Armelagos, 2007, p. 287). This study not only received wide readership, with its publication in *Current Anthropology*, but, most importantly, elicited for the first time a public reply from the architects and experienced users of FORDISC (Freid et al., 2005) which initiated a volleyed discussion in print over the criteria that should

define the best practices for its use (e.g., Naar, Hilgenberg & Armelagos 2006; Hubbe & Neves, 2007; Williams & Armelagos, 2007; Keita, 2007; Williams, Belcher & Armelagos, 2007; Elliott & Collard, 2009).

Treating their sample as "unknown individuals,"Williams and colleagues (2005) considered 12 cranial measurements on 42 Meroitic Period Lower Nubian crania to test the classification accuracy of FORDISC 2.0. To evaluate the full potential of the software's comparative database, they included, as their reference groups, the complete set of Howells populations, as well as the modern American races from the Forensic Anthropology Data Bank. Not unlike the preceding analyses, the overall classification results obtained were poor and, under closer scrutiny, the majority of the individual allocations that were produced failed to conform to even general biogeographic or historical expectations. In the absence of an exact match between the control samples and their test cases, Williams, Belcher, and Armelagos (2005) conducted independent tests of similarity between the Nubian and Howells' African data to determine the degree of sample comparability: if their Nubian and Howells' Egyptian crania did not differ significantly and, still, poor classification results were produced, the problem should be seen to lie with the program itself and not with the design of their study (i.e. their use of an ancient sample and limited variables). They also expected that any methodological errors could be investigated by reviewing the typicality probabilities produced for each classification, which they believed would appear "significantly different from the populations ascribed by the program" (Williams, Belcher & Armelagos, 2005, p. 343). As they obtained insignificant results for their *t*-tests of group-differences, they hypothesized that, when the Howells reference groups were used, the Nubian specimens should cluster together in the single Egyptian category, their nearest geographic neighbor. When also considering their data in more general terms and leaving open the possibility for complications arising from some unaccounted training-target mismatch, they posited that the Nubians should be, at least, consistently

allocated to other continental African populations, such as the Teita of Kenya or the Dogon of Mali, notwithstanding the issue of different depths of time. When considering the comparative forensic, FDB, data, they assumed that all allocations made would result in a limited identification as either as Black or White. In actual practice, however, FORDISC "correctly"assigned only 8 of the crania to their predetermined best-fit population, Late Predynastic Egyptians. The remaining individuals were classified variably across the Howells reference groups, with their allocations not only extending beyond continental Africa but covering a wide geographical range that included North and Central Europe, Easter Island, the Andaman Islands, Japan, Taiwan, South Africa, Australia, and North America. They noted (2005) that for these cases the "typicality statistics were neither uniformly high nor low" (p. 342). Similarly, when assignments were made using the FDB reference data, the Nubians appeared to share membership with all but one of the available biosocial groups, (including Black, White, Hispanic, Chinese, Japanese and Native American but not Vietnamese) but typicality values identified 14 specimens as significantly different from the group to which they were allocated. They also reported that, for 11 of their test specimens, memberships varied when changes were made to the individual determinations of sex, suggesting that knowing sex identifications, or estimating it reliably, is a basic, and potentially difficult, requirement for the successful application of FORDISC.

Williams and coworkers (2005) have attributed their inaccuracy in classification (by geography, affinity or race) to three key factors, all of which extend beyond the constraints of their Nubian sample and the bring into question the validity of any analysis of human variation using craniometrics: (1) that the polymorphic nature of cranial variation transcends geographic and population boundaries, (2) that world-wide human variation is distributed in clinal pattern, and (3) that certain cultural and environmental effects on craniofacial shape are always unknowable. When also considering the

program's specific design, they argue that the control populations built-into the FORDISC software fail to capture enough of the morphological variation distributed globally to ensure the reliability of this type of analysis and instill confidence in the classifications it produces. They contend that because the limited reference samples available have been constructed from socio-cultural definitions of group membership (race, ethnicity, geopolitical identity) and not some scientific measure, FORDISC's predefined "cranial populations are suspect as real biological units" and are poor "ideological" representations of true craniometric diversity. When these are used to satisfy the computational requirements of discriminant analysis, and even when reasonably well-matched groups are considered, any classification results produced using FORDISC are always already without meaning, regardless of any semblance of statistical success from the analysis. It should not be surprising then that FORDISC has been subsequently censured for its requisite use of priors and for participating in the systems that predefine and populate these categories. Armelagos and Van Gerven (2003), for example, have spoken directly to this issue in their interpretation of the application of FORDISC to the Nubian crania: "[t]he program forced a solution based on *a priori* racial criteria presumed (as all racial schemes do) to delimit patterns of real human variation" (p. 60).

In 2005, Freid et al. challenged the findings of these Nubian studies, and, in doing so, also responded to the growing concerns for the variability among independent test results (e.g., high misallocation rates, low consistency, and questionable information value) by demonstrating how "FORDISC, like any good tool, performs best when applied to the task for which it was developed, and the user will only achieve maximum results with a full understanding of the strengths and limitations of the program"(2005, np). They argue that the "incorrect" classifications that have spawned these criticisms are "often due to inappropriate reference samples and failure to properly evaluate the typicality and posterior

probabilities provided by the program" and take up their own analysis of Nubian data to illustrate these points (Freid et al., 2005, np). Using a large combined sample of 144 adult males and 93 adult females from the western necropolis and the walled cemetery at Mirgissa as their test cases and the Howells populations as their comparative reference groups, they obtained explicable results, and even reasonably good classification accuracy considering the acknowledged problem of training and target sample incompatibility. They report that, for the 144 male specimens, 56% were classified as Egyptian, 18% were allocated among three African groups (the Bushman, Dogon and Zulu), and the remaining classifications were split between the East Asian/Pacific Island groups and Europeans. Of the 93 females, FORDISC identified 66% as Egyptian, 16% as African, 6% as European and the remaining 12% were distributed among the Circum-Pacific groups. From this data, it can be seen that, in fact, the overall classification patterns obtained satisfy the expectations of memberships proposed by Williams and colleagues (2005) for their skeletal sample, given the temporal-geographic relationships among the Nubian and Howells groups: 75% of the males and 82% of the females were assigned to either the Egyptian or one of three African groups, Egypt appeared as the first or second smallest Mahalanobis D^2 distance in over 76% of the male and 86% of the female specimens, high posterior probabilities were reserved for the majority of the crania that were classified as Egyptian and low posterior probabilities were consistently associated with individuals classified into the other reference groups. Freid et al. (2005) that these findings, when properly interpreted, demonstrate how "classifications outside of the expected reference group with these low values should not be considered 'incorrect'" (np). Instead, these authors believe that they do not demonstrate a failure on the part of the algorithm or software to correctly identify the biological affinity of ancient Nubian crania but instead attest to an issue of the user's poor methodology: "[i]n order for FORDISC to perform effectively, it must be applied in regions or contexts where reference samples adequately represent cranial variation" (Freid et al., 2005, np). In a

more recent reply, Ousley, Jantz & Freid (2009) have added that Williams and colleagues (2005) based their conclusions on group relationships by compiling individual classifications, when, in fact, they should have considered the group centroids. Moreover, they suspect that "based on the individual results of Williams et al., their Nubian sample centroid is most likely closest to Egypt" (ibid, p. 73).

Several discussion pieces have also emerged in response to the third Nubian study (2005), to which Williams and coworkers have replied in turn (2007). These reanalyses have further commented on the problematic classification results, attempting to identify additional, potential sources of error and offer recommendations for improved tests. In reaction to the methodological explanations provided by Freid et al. (2005), Naar, Hilgenberg and Armelagos (2006) revisited the issue of the availability and choice reference sample. Using the Egyptian population included within the reference database as their test group, they found that FORDISC still had difficulty in producing reliable results when the criteria recommended by Freid and coworkers (2005) was followed (i.e., a typicality of greater than 0.1[2] and a posterior probability greater than 0.5). For 111 Egyptian cases, they found that only 55 (49.5%) were correctly identified, 16 (14.4%) were classified as Egyptian but did not meet the threshold for the typicality and posterior probabilities, 14 were classified into other groups with significant probability and typicality values, and 26 (23.4%) were also misallocated but with insignificant probabilities. They used these findings to refute the earlier claims that, across the Nubian studies, their authors had collectively "failed to apply the methodology correctly" and that, because the Nubian samples were not adequately represented by FORDISC's predefined reference groups, they "should not have expected successful

[2] Both Freid et al. (2005) and Naar, Hilgenberg and Armelagos (2006) refer to a typicality threshold of greater than 0.1; Jantz and Ousley (2005), however, recommend that individual classifications with values less than 0.01 should be considered atypical.

classification" (Williams & Armelagos, 2007, p. 286). FORDISC's inability successfully classify its own training sample was interpreted as reinforcement for the prior criticisms that the program itself is flawed and that craniometrics variation lacks the structure to support taxonomic research that is based upon cultural identifiers. These findings reiterated the conclusions that Williams, Belcher and Armelagos had already made in 2005: "While we acknowledge that many scientists are fully aware of the problems inherent in Fordisc 2.0 and in race determination generally, our experiment provides a point of view that is rarely examined within forensic anthropology, that is, that human cranial populations are suspect as real biological units equivalent to the cultures, languages, geographic regions, and social histories they are intended to represent" (p. 344).

In 2007, Hubbe and Neves provided the first printed rejoinder to the publication by Williams et al (2005) in Current *Anthropology*. While they agreed with Freid et al's (2005) critique that the problem with this Nubian study was one of poor methodology, Hubbe and Neves (2007, p. 285) argued that "the design of their experiment is flawed" and proposed, specifically, that its weakness lies in the number of variables employed. They contend that, while Williams, Belcher, and Armelagos selected 12 variables that "have been shown to have diagnostic value"(2005, p. 342), these "12 variables are far from enough to classify a skull on the basis of discriminant functions" (2007, p. 285). Hubbe and Neves support their critique by using the Howells populations to test for changes in classification accuracy when four different suites of craniometric variables are used: the full complement of the 57 measurements available in the Howells data set, a reduced subset of 11 variables, 22 facial variables, and 21 neurocranial variables. They reported a difference of over 30% between the classifications made using the complete data set (57 variables) and those using the smaller set of 11 measurements. Remarkably good classifications are obtained when all-available measurements were used, with the mean correct classifications

consistently exceeding 90%. They also call attention to the fact that no differences were detected in accuracy when either facial or neurocranial metrics were used, which suggests that correct allocation rates are based upon the number of variables not their anatomical location and also works to dispel claims that classification success may be tied to the fact that the cranium is "developmentally very sensitive to the conditions of growth"(Williams, Belcher & Armelagos, 2005, p. 344). In a final summation, they argued that their [revised] analysis offers clear methodological explanation for the Nubian findings under debate and demonstrates "how misleading a morphological study based on a small number of variables can be" (2007, p. 285). These findings of high classification accuracy also have broader implications for the theoretical foundations upon which FORDISC (and any quantitative approach to the estimation of ancestry) rest, as Hubbe and Neves conclude that "[m]ost of our craniometric variance is unquestionably found within local populations, although the small amount of among-populations variance can be very informative about our evolutionary history" (2007, p. 285). Williams and Armelagos (2007) have responded to Hubbe and Neves with an argument based upon three perceived differences in intention: (1) they identify the incomparability of their methodologies, attesting that, while they adopted a "heuristic" approach, Hubbe and Neves undertook a "classification experiment" (2007, p.287); (2) they highlight the incompatibility of their philosophies, arguing that "Hubbe and Neves attempt to polarize biological anthropologists into one group (theirs) that sees population affinity as the most important signal and another (presumably, that belonging to Williams and Armelagos) that views population affinity as largely inconsequential" (2007, p. 287); and (3), ultimately, they identify an opposition in their focus, explaining that, while Hubbe and Neves "inadvertently reiterate the essentialism of the race concept by suggesting that a greater number of variables is all that is needed to classify individuals," their "results were intended to show is that Fordisc is a poor predictor of population affinity using either the Howells data set or the Forensic Data Base

because the heterogeneity of a single population is not captured by the analysis of craniofacial traits that the program employs" (2007, p.287).

In a second discussion piece in *Current Anthropology*, Keita (2007) offered further remarks on Williams and coworker's (2005) study, following-up on comments made in Hubbe and Neves' (2007) critique and focusing on the role of genetics in shaping cranial morphology and the potential for population signatures that may enable the detection of among- group differences using classification techniques. To this, Williams, Belcher and Armelagos (2007, p. 427) offered a reply in which they stress again the question of intention. In terms of their approach, they write that "[o]ur results simply confirm the heterogeneity of Meroitic Nubia and the lack of fit between conceptual models in osteological research and actual patterns of human biological variation." Moreover, in regards to their philosophy, they reiterate their concerns for practice, identifying the danger of condoning the uncritical use of FORDISC, given its noted "conceptual" and "fundamental" flaws: "by using the program (which often yields incorrect classifications) practitioners may be inadvertently reinforcing contemporary epistemologies that greatly misinterpret human biological diversity" (2007, p.426). While, in the opinion of the original authors "Keita now criticizes us for a paper that we did not write. Specifically, he critiques our discussion of noninherited factors, namely, the conditions of growth, which we cite as contributing to craniofacial variation" (Williams, Belcher & Armelagos, 2007, p. 426), Keita's comments do bring up an essential failing that overshadows their entire study. Keita (2007) reminds us that "[t]he goal of forensics is the actual *identification* of the unknown with a recognized social group known to law enforcement and not the discovery of relative similarity (phenetic affinity)"(p. 425). From this it can be surmised that any attempt to test the utility of FORDISC out of its intended context, using populations that are not represented in the reference database and for which the correct assessment of biological affinity and

not the identification of a potentially "true" membership is the best-case result, are unsatisfying conceptually if not equally as weak statistically.

Despite the popularity of this classification debate and while many cogent arguments have been made in explanation of the FORDISC software, its application and utility, the issue of its "systematic testing for validity" (Williams, Belcher & Armelagos, 2005, p. 341) has continued to remain, nevertheless, unsettled for many anthropologists: forensic, bioarchaeological and cultural alike. Recently, Elliott and Collard (2009) have argued that resolving the debate over FORDISC's merit as a reliable (sex and ancestry) identification tool is a "pressing need," especially "given the importance in bioarchaology and forensic anthropology of determining the ancestry of unidentified skeletal specimens and the frequency with FORDISC is used" (p. 849). To this end, they report results obtained through the distillation of a larger project that ties together the various lines of evidence that have been popularized by these earlier, critical tests of FORDISC's utility and that have been used to highlight the program's flaws (see Elliott 2008). They evaluated these critiques by speaking to what they, and others, have identified as the "main points of contention"(2009, p. 849): (1) the absence of an appropriate reference population, (2) the choice of the number of variables, (3) the effect of "knowing" sex on target classifications. They address these potential sources of error by offering expanded method tests, but remain within the framework of the well-known theoretical concerns, i.e., the perceived weaknesses of the software's statistical foundations, design and user-specific implementation. In their study, they employed a larger sample of test specimens, designed their analyses in such a way that enabled specimens to be classified with and without the inclusion of their true population of origin as a reference group within in the comparative database, and they controlled for the potential effects of variable selection and sex by repeating classifications under different conditions and with opposing configurations (e.g., adjusting the

number of non-overlapping variables considered and varying the assumption that the target's sex is known). They also took on the underlying issue of classification validation and investigate the relationship between the allocation of the individual target specimen and their associated posterior and typicality probabilities, following the recommended sectioning points for these values (PP>0.50 and TP>0.01) provided in the FORDISC user's manual (Jantz & Ousley, 2005) and the more conservative limits (PP>0.80 and TP>0.01) communicated to forensic practitioners in a recent software workshop (c.f., Jantz & Ousley, 2007). As many of the arguments against FORDISC are rooted in the problem of the appropriate choice of test subjects for evaluating the software's classification accuracy, Elliott and Collard attempt to circumvent this procedural issue and simulate "favorable conditions" (2009, p. 851) by drawing their target specimens directly from the same Howells data set that FORDISC uses to generate its discriminant functions. Their prediction of membership is based upon samples that serve, simultaneously, as training and target specimens. They assembled their test sample from the Berg, Hokkaido Japanese, Santa Cruz, Tasmanian and Zulu populations, and for the execution of "source-population-excluded" analyses, they chose the Norse, Kyushu, Yauyos, mainland Australian Aborigines, and Teita as the "most closely related populations" to each of their test groups, respectively (p. 851). "Correct" classifications occur, therefore, when the test specimens are allocated to their predetermined, best-fit group. The choice to evaluate the quality of by-case predictions of membership by making use of a subset of the same individuals who have been employed as the DF training data for the initial partitioning of the reference groups is not only methodologically suspect (Fielding 2007), it is also worthy of special note. Elliott and Collard (2009) present arguments that discredit the use of FORDISC on account of sample acquisition biases associated with Howells' craniometric data set that influence the representation of the true distribution of variation by group and sex: the non-random selection of specimens that were considered "typical"of their identified group, the often extreme reduction of

sample size on account of the exclusion of morphologically unusual crania (e.g., only 50 out of the almost 1800 available ancient Egyptian crania were selected), the exaggeration of between-sex differences resulting from the reliance upon morphological estimates of sex for most groups and the exclusion of crania for which sex could not be confidently determined, and the artificial reduction in the overlap among the different populations (and their sex-specific subgroupings). Given these concerns of information quality and notwithstanding the issue of methodology, the choice of Elliot and Collard to reuse a selection of Howells' data as the best-available test-set should affect the reliability of their results, a point to which they themselves concede but only in furthering their argument against FORDISC: "the analyses reported here probably overestimate FORDISC's ability to attribute ancestry" and that "the corollary of this is that the success rate of FORDISC in the analyses in which sex was specified may have been artificially high" (Elliott and Collard, 2009, p. 851).

The findings from these directed tests confirm for Elliott and Collard (2009) that "the use of FORDISC for ancestry determination is problematic" (p. 850) and its "utility in research and medico-legal contexts is limited" (p. 849). It is notable that when their analyses were conducted under certain conditions, they obtained classification accuracies that are comparable in success to the rates reported by Freid et al. (2005). When all 56 variables were used and the "unknown" specimen's population of origin was included as one of the comparative reference samples, they produced reasonably good classifications, i.e., correctly allocating more than 70% of the test cases to their true population of origin. The results of all of their other analyses were, however, quite inferior: classifying less than 40% of the target specimens into their correct group. After close inspection of the posterior and typicality probabilities associated with the correct and incorrect classifications, Elliott and Collard (2009) also found that there were always some misallocated individuals whose PP and TP values exceeded the recommended

sectioning points. When these acceptable limits were recalculated, incorporating information on the individual assignments and true memberships, they redefined the acceptable PP and TP values, as 0.991 and 0.952, respectively. These new highly constrained values have considerable implications for best-practice procedures. They recommend that memberships should only be accepted if the posterior and typicality probabilities meet the criteria: PP>0.991 and TP>0.952. Under these new guides and when only considering the classifications representing their "best" analysis, 99% of Elliott and Collard's identifications fall under the category of "failed attributions" (2009, p. 851). When interpreted collectively, Elliott and Collard believe that their findings offer decisive evidence disproving the validity of FORDISC as a useful forensic tool and, in turn, resolve longstanding questions over its ability to produce reliable and accurate results and enable responsible and meaningful identifications. Elliott and Collard (2009) make the decisive statement that "even in favorable conditions - when the focal specimen's source population is present in the reference sample, the focal specimen is nearly complete and its sex is known - FORDISC has no more than 1 per cent chance of success" (p. 851). If correct, these findings have considerable ramifications on human identification protocol and methodological best-practices, as these authors note "many of the ancestry determinations that have already been obtained using FORDISC are likely to be unreliable" (p. 851). It can be argued that conclusions such as these extend well beyond the academic exercises or theoretical debates for which tests of FORDISC are of subdisciplinary interest. While the exact nature of their implications cannot be immediately known for past casework, it can be imagined how, now, in this, the post-Daubert era, they may easily lead to significant practical disciplinary consequences. They may affect the contributions of forensic anthropologists to the medico-legal community, given the current context of newly emerging directives for the forensic sciences (e.g., SoFS), and, specifically, the ongoing development of best-practices for forensic anthropology (e.g., SWGANTH). They may also have a broader impact given the currently

evolving judicial conditions, where standards for the admissibility of evidence and acceptability of testimony are in flux (c.f., NRC report 2009, Daubert v. Merrell Dow Pharmaceuticals, 1993).

Engaging the FORDISC Debate

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." (Box & Draper, 1987, p. 74)

While the test subjects and sample compositions may change and conflicting results may be reported, these common critiques of FORDISC, not only test its utility as "user-friendly" software but, in evaluating its abilities and strengths, they also question its theoretical foundations, both disciplinary and statistical. These analyses take aim at the application of multivariate statistics to craniometric data for the purpose of group discrimination, the estimation of individual affiliations, and, in the context of forensic casework, the eventual translation of these estimates of ancestry to the more practically interpretable categories of biosocial race for medico-legal investigations. The constituent arguments that together formulate the basis of these critical studies remain consistent: (1) classifications based upon craniofacial data are subjective constructs that depend upon the preselected traits and reference samples, (2) user-guided analyses are always already biased by the theoretical position of the forensic practitioner on race, ancestry and human variation, (3) all individual allocations produced using "fundamentally" and "conceptually flawed" tools will unavoidably fail to represent biological or statistically meaningful realities, and (4) the information gleaned from these faulty results may adversely affect the direction of the investigation and potentially undermine the success of collaborative efforts towards making a

positive identification (Goodman, 1997; Armelagos & Van Gerven, 2003; Williams, Belcher & Armelagos, 2005). To this list, respondents have also added the broader, and often irresolvable, argument that the resolution of the classification may be compromised by the interplay of external and internal factors in the design of the analysis and the interpretation of the outcome. Problems may be unwittingly induced by, for example, a limited or biased understanding of the patterns of population relationships that may yield a biohistory that is unique to the specimen's origins, the quantity and quality of information that may extracted from the specimen, the unknowable effects of individual environments (from lifeways to pathology) on craniofacial morphology , and the inability to distinguish between shared-primitive and shared-derived similarities in cranial shape (Williams, Belcher & Armelagos, 2005; Keita, 2007; Elliott & Collard, 2009). Many critics embrace these concerns and continue today to share the view voiced already by Goodman (1997) at the time of FORDISC's emergence: "At best, in other words, racial identifications are depressingly inaccurate. At worst, they are completely haphazard"(p. 22). Despite a decade of methodological refinements and revisions on the utility of race, many "mainstream" scholars have integrated this opinion into their current characterization of ancestry estimation in forensic anthropological practice, so that "[f]orensic anthropologists have been criticized by other anthropologists for upholding the race concept whenever they estimate ancestry of a skeleton" (Steadman & Andersen 2009, p. 19; c.f., Albanese & Saunders, 2006).

In a recent practical discussion of the theory and application of personal identification methods, Steadman (2009, p. 4) counsels that "Forensic anthropologists must remember that their reports, testimony and conduct have immediate and long terms ramifications for defendants, families, and the judicial process as a whole." Any responsible practitioner, must adopt, therefore, "a strict adherence to scientific principles and professional protocols" (Steadman, 2009, p. 4). This is especially true, now, at

the time when federal guidelines, like Daubert, are becoming increasingly rigorous: for example, by ascribing new meaning to "admissibility" and dictating the use of "generally accepted" methods (Saks and Koehler, 2005, p. 892). The implications of these updated rulings on evidence and testimony may be considerable for the standard identification techniques which are commonly taught, widely used, and highly valued in the forensic anthropological community (Wiersema, Love & Naul, 2009). In accordance with this "coming paradigm shift" in the science of human identification (Saks & Koehler 2005), there is merit in Elliott and Collard's claim (2009, p. 849) that there is a "pressing need" to resolve the debate over FORDISC, as either a flawed program or a useful, but often misused, tool (Freid et al., 2005). The prevailing critical approaches to the [in]validation of FORDISC and the assessment of the forensic estimation of identity, however, should not be uncritically accepted. These are too often indiscriminately adopted and too easily used "to determine which of these claims is correct." Their design makes for weak and conflicted tests of the software and, more importantly, fails to adequately address the fundamental premise regarding the structured organization of human variation that is, in fact, driving all criticisms of FORDISC. The information value is limited in not only Elliott and Collard's (2009) recent reassessment but also in those earlier tests that their results appear to corroborate. It is also necessary to take issue with the freedom with which their findings are used to comment on the body of past ancestry estimations that may have contributed to the making of an identification and even to courtroom testimony. While it must be understood that the interchange between proof and disproof is an essential negotiative process in the many steps that make up the scientific method, caution must be heeded when the ramifications of this theoretical exercise have real-world consequences and statements that question the findings of medico-legal importance must be grounded in the best scientific arguments and facts. It is these casual predictions of "countless misidentifications" (Armelagos & Goodman, 1998,p. 370) and suspicions of "unreliable" past identifications (Elliott and Collard 2009, p.

851) that necessitate a reevaluation of the entire statistical approach to the "testing" of ancestry estimation through the [in]"validation" of FORDISC, in a way that (1) disassociates this evaluation from the assessment of the software and from any ideological stance on race and classification that may direct research design, and (2) instead, repositions the question of the contribution of metric-based classification techniques to the estimation of ancestry within a broader, yet more powerful, framework of statistical methodology. To demonstrate the problems with these previous efforts that have attempted to "systematically test validity" by working within the confines of FORDISC and the system of its DFA, four interrelated points are introduced here for discussion: the issues of predefined reference populations, the incorporation of prior demographic information, the choice of test data and the selection of variables.

As it is described above, discriminant function analysis requires an *a priori* selection of reference groups. These predefined "known" samples serve as a training set upon which to build the initial model. When individual classification is the goal, the conditions of the prediction algorithm require that comparisons are made between this training set and the "unknown" target data (McLachlan, 1992). FORDISC's statistical underpinnings dictate that the unidentified specimen will be classified with respect to the available comparative, control populations, so that a best-fit classification will always be produced regardless if it is, in fact, the correct (sex, ancestry/race) identity for the unknown test case (Freid et al., 2005). The "success" of FORDISC may be affected by several factors, but the topics of the most frequent discussion are (1) the lack of representation of the test case's true population of origin in the software's database as well as (2) the unique "atypical" aspects of the individual's morphology that may position it at either of the extreme tails of normal variation, relative to distributions of the control groups. Jantz and Ousley (2005) explain how the quality of this classification can be increased by the choice of

appropriate reference samples and the correct interpretation of the probability statistics output. They, with Freid et al. (2005), have used these recommendations to formulate a cohesive methodological argument with which they have responded to the critics of FORDISC. Even further improvements, however, can be made to DFA-based analyses if the practitioner takes into consideration the demographic profile of the population-at-large from which the "unknown" individual is believed to have been drawn (i.e., based on external indicators, like recovery location, personal effects, preliminary morphological assessments). Recently, Konigsberg, Algee-Hewitt & Steadman (2009, p. 78) have demonstrated the potential of this refinement to FORDISC's approach "from a highly empirical perspective," showing that "an informed prior based on the context of a forensic case will provide a better estimate of ancestry from craniometric data than an uninformed prior." This information can be incorporated into the discriminant function analysis by altering the prior probabilities of group membership so that they represent the demographic composition of the test case's recovery context. The posterior probabilities of group membership can be calculated given these informed, prior expectations and, in turn, the unknown individual's allocation can be more relevantly made.

In the context of "systematically" testing FORDISC's "validity" as "a powerful tool in forensic analysis" (Ubelaker, Ross & Graver, 2002, np; c.f., Williams, Belcher & Armelagos, 2007, p. 426), the Forensic Data Bank stands as an acceptable informative prior for modern casework (c.f., Konigsberg, Algee-Hewitt & Steadman, 2009; Langley-Shirley & Jantz, 2010). The FDB's morphometric data offers, therefore, the most relevant choice of reference groups when evaluating the performance of FORDISC's built-in [DFA] procedure in "diagnosing race" (Williams, Belcher & Armelagos, 2007, p. 421) for modern medico-legal casework. It is, in fact, this application which even the archaeological studies (e.g., the Nubian publications) set out to test. It can also be argued that reasonable prior expectations for race can be

obtained for any given modern American context by using the most recent U.S. census data for casework with (presumed) current dates of death or by consulting the most appropriate past survey when estimates of the post-mortem interval are considerable (e.g., reevaluation of a "cold case" or the discovery of "old" remains). While it is true that census definitions of race are not without issues of interpretation and representation and they have suffered historically from the problems of consistency and parsimony in terminology (AAA 1997), Konigsberg, Algee-Hewitt & Steadman (2009) have shown how useful frequencies of self-reported "races" for particular locales (state, states, region) can be extracted from this demographic data, and how these can be productively used to characterize the population-at-large for the geographic area of interest. From this perspective, it may not be simply enough to follow the prescribed FORDISC routine, i.e., choosing best-fit reference groups, conducting DFA using equal priors, and formulating conclusions based on post-hoc reviews of the classification and probability output. While a careful assessment of the individual posterior and typicality probabilities of membership is an indispensable step for determining if the case under investigation belongs to one of the many *a priori* groupings within the Howells and FDB data sets or shares membership with all or some of these groups, the inclusion of well-chosen prior data is an important consideration for conducting more meaningful DF analyses. As Konigsberg, Algee-Hewitt & Steadman (2009) have clearly shown by example, when forming posterior probabilities, prior information can be invaluable to both the estimation of membership and the presentation of the allocation made as valid evidence of individual affinity, ancestry or race. They write: "we know, for example, that Mr. Johnson's remains were found in eastern Iowa in 1999. Knowing this, is it reasonable to assume that the remains are as likely to be from a descendant of 26th–30th Dynasty Egyptians or Easter Islanders (both samples are contained within the Howells' data) as from relatively recent Americans?" (Konigsberg, Algee-Hewitt & Steadman, 2009, p. 79). If understanding and correcting for context is a critical step in an informed

discriminant function analysis, then perhaps the better argument to be made against FORDISC is not one that focuses singularly on the reference sample problem but, instead, addresses the user's inability to independently modify the classification procedure to account for changing demographic expectations: ideally, users should be able to oscillate between the software's constraint of equal priors and other different values of informative priors that are case-specific. Alternatively, perhaps a better defense can be mounted against the typical criticisms of FORDISC by extending the rebuttals given by its authors for the common instances of "incorrect" classifications to include a larger common-sense argument. Instead of focusing their reply on the (valid but narrowly focused) methodological explanation, that "these disputed results are often due to inappropriate reference samples and failure to properly evaluate the typicality and posterior probabilities provided by the program," they would also do well to aggressively question the logic of testing this forensic tool outside of "the task for which it was developed" and for which it has been shown to "perform best" (Freid et al., 2005, np).

An effective response to the comments made by the authors of the Nubian and Spanish studies on FORDISC's weak classification results for their unrepresented groups can be mounted by addressing the bigger issue of [un]informative or [un]equal priors. While the authors of these test projects do demonstrate statistically that "the samples clearly were inadequate to elucidate the specific geographical origin" of their test crania (Ubelaker, Ross & Graver, 2002, np), it can be argued that these "significant" findings may not be meaningful or helpful to their ultimate research goal. Although Williams et al. (2005) imply from the title of their study that they are concerned with "forensic [mis]classification," their "empirical" test of "validity," using the ancient Nubian crania in actual practice, poorly approximates the real-world conditions (e.g., sample origins, recovery context, available data) of a forensic case and the related procedures of making a reliable, assistive identification. Their

decision to evaluate the FORDISC software using (1) irrelevant test subjects and (2) an uninformative prior (i.e., for the Nubian data, using a prior probability of 1/28, that is, one over total the number of groups in their study or the full Howells data set) is highly questionable. It is suspected here that, in both the Nubian and Spanish studies, the test crania would be classified differently, and with improved results, if the value of the priors were reduced for any comparative groups that represent areas that are geographically removed from Europe and Northern Africa, i.e., in keeping with a reasonable, and still broadly-based, expectation for their population-at-large or origin. For the misclassification of the Nubian samples as Easter Islanders when using equal priors across the 28 comparative groups, Konigsberg, Algee-Hewitt & Steadman (2009, p. 84) point out that "knowing that the 42 crania in Williams et al.'s study were those of archaeologically obtained Nubians, the prior probability that they were Easter Islanders seems vanishingly small."

When reflecting specifically on the aims of these test studies, it is important to logically ask: what is the likelihood that ancient Nubian or archaeological Spanish remains would be recovered in the United States, identified as both an individual of "unknown" identity and a case of medico-legal significance, and merit a forensic anthropological assessment in an American laboratory? Moreover, what experienced, well-trained practitioner would not realistically challenge the appropriateness of implementing either "test" procedure, using the Nubian or Spanish data, given her practical knowledge of local casework, i.e., in terms of the typical recovery conditions, state of the remains, and observable signatures of modern skeletal morphology? In reply, it is possible to look to Konigsberg, Algee-Hewitt and Steadman (2009, p. 84), who have already advised that "[w]e consider forensic anthropologists as being implicit Bayesians because they often do bring prior information to their cases, though this information is typically implicit, unstated, and not quantified." If the intention of these critical studies is

to evaluate FORDISC's ability to aid in the personal identifications made by North American forensic anthropologists (the practitioners to whom their criticisms of "tool use" are directed and the circle of developers and supporters of the philosophies, theories and methodologies that they reject), then these analyses should be conducted on test cases that represent the typical modern American male or female (e.g., the individual with the highest probability of being a forensic case, conditional upon the current demographic parameters of the United States). If these studies are interested in the global applicability of FORDISC to all types of international casework, as, for example, Ubelaker, Ross & Graver (2002) propose, then the test case should represent, what Tomanio and colleagues (2011) have called, the "most typical human." When consideration is given to the global population projections for 2011 (given a planet of seven billion people), this consensus individual is defined as a 28-year-old, Han-Chinese male (Tomanio et al., 2011). Neither the Nubian nor the Spanish test samples adhere to these local or global real-world criteria. It is also important to note that these "typical" identifications are subject to variability. They are dependent upon population-specific growth rates and potential fluctuations in sex proportions (attributable to, for example, rates of natural population increase, net migration, birth and death, etc.). Consider, for example, the projected change in ranks and designations on the global scale: since growth trends suggest that by 2030 China will be replaced by India as the top ranked nation in population density, an individual of Indian biogeographic ancestry and biosocial race will assume the position of the new "most typical human"; moreover, as the present world-wide male to female ratio is very close (1.01:1), there is a possibility for quick, short exchanges to occur in the designation of the more populous sex (Tomanio et al., 2011). What these fluctuating statistics indicate is that understanding both the conditions of the "unknown" individual and the effects of methodological decisions on the interpretability and significance of the resulting solutions is critical to any analyses that consider real-world, present-day questions and data. As Konigsberg, Algee-Hewitt and Steadman (2009,

p.84) reaffirm for these attempts at meaningful classification, "[r]egarding the use of race in estimation and evidentiary problems, this, like many problems in forensic anthropology is very context specific."

In sum, it is worth noting that reference sample problem identified by FORDISC's critics is not limited to the program or the predefined groups populating its software package or even to DFA alone, but has implications for many classification based studies regardless of the field or research agenda. It is well known that the application of many comparative or "clustering" analyses to markers, traits, individuals, and populations typically demands the pre-allocation of individual specimens into user-defined groups and that ascertainment methods often involve a process of selection that deviates from random sampling. Even classification results undertaken by the most vigilant "blind"researcher are unlikely to be devoid of some unrecognized or pre-acknowledged biases. What is important to realize is that this problem can be mitigated, if not fully resolved when these biases are used to informative ends: not only by guiding the choice of the reference populations but also by modifying the prior probabilities to reflect what is "known" about the case given the collective experience of the user and the extrinsic (geographic location, population-at-large) and intrinsic (craniofacial morphology, personal identifiers) conditions of the target individual. Contrary to the common belief, neither FORDISC nor its underlying discriminant function algorithm "force[...] a solution based on *a priori* racial criteria" as Armelagos and Van Gerven contend (2003, p. 60). While there are predetermined "constructs" in the form of the comparative skeletal populations that enable FORDISC's classifications, the predefinition of these groups in discriminant function analysis does not mean that their incorporation within the program affords the procedure some independent ability to generate a statistically significant or biologically interpretable outcome for the classification of a test case if in fact the true strength of craniofacial similarity (and, so, membership) between the unknown "test" individual and the reference group(s) is weak. The FORDISC

program uses DFA to make best guesses from the data available and defines these solutions probabilistically, using posterior probabilities of group membership, which are themselves qualified by typicality values. As the potential for a good allocation is conditional upon both the compatibility of the reference samples available and some prior expectations of membership, a thoughtful choice of informative priors may improve the statistical strength of these guesses. More meaningful posterior probabilities can be calculated and, in consequence, the accuracy and reliability of the classifications may be increased. DFA using informative priors should produce, in theory, more robust results, with better classification accuracy or interpretability (for problem specimens especially) than those generated by FORDISC 3.0, under its current equal prior limitation.

While this revised approach to the implementation of DFA does offer the possibility of improving classification accuracy, there will always remain potentially irresolvable problems for any analysis that relies upon external "informative" data (whether it be in the form of prior demographic information or comparative metrics) that can be influenced by the individual assumptions of the forensic anthropologist. The quality of a particular analysis may be unknowingly compromised by circumstances that are unique to the case in question. These conditions can misguide the practitioner into making logical but erroneous prior assumptions, which may ultimately impose deleterious effects on the analysis, thereby inhibiting correct classifications and obscuring valuable membership information. Confounded results may stem from the contributions of one or many sources of error:

(1) In the current high mobility environment with increasing fluidity across borders at all levels (Hall, Smith & Marciszewska, 2006; NEXUS program 2011), the site of the recovery of the remains may be different from the site of the individual's origin due to antemortem travel or intentional (e.g., trophies, mass graves and dump sites) and natural (e.g., animal or fluvial) postmortem transport

(Taylor et al., 1984; Sledzik & Ousley, 1991; Nawrocki et al., 1997; Haglund, 1997; Verano, 2003; Baraybar & Gasior, 2005; Harrison, 2006).

(2) As demographic priors obtained from general census data, or even from sources as forensic-specific as the FDB, document self-reported race information, the incorporation of this data into any prior-based analyses dictates the use of "bureaucratic" or "social race" and the acceptance of a presumed correlation between skeletally defined ancestry and membership in self-identified biosocial categories (Sauer, 1992; Albanese & Saunders, 2006; *Ousley,* Jantz *& Freid,* 2009).

(3) The demographic make-up of a skeletal collection of "forensic cases" may not represent a random sample of the population-at-large (e.g., as quantified by census records or homicide statistics; c.f., Wilson & Algee-Hewitt 2009) and the degree of correspondence may vary by region (c.f., Marks [1995] for cases from Tennessee).

(4) Skeletal or virtual repositories of positively identified cases may be the best source of prior data, but case-by-case documentation is often poor, access can be limited or restricted, and large scale databases (like the FDB) or "contemporary" collections (e.g., William M. Bass Donated and Forensic Skeletal Collections, McCormick Collection, Maxwell Museum Documented Collection) are rare and are frequently geographically constrained, temporally specific, and demographically biased (Ousley & Jantz, 1997; Wilson, Algee-Hewitt & Jantz, 2007; Meadows Jantz & Jantz, 2008; Komar & Grivas, 2008).

While some of these sources of error can be avoided or ameliorated by improved sampling strategies and robust statistics, others may remain unknowable or irresolvable simply because any study concerned with real-world, complex biological data is fraught with many, and often correlated, obstacles. Without a reasonably easy and convincing solution, these factors may go unaccounted for in

the course of research design, the analysis and the processing of the case. It is common to face, therefore, the conundrum of working with what is available; this may involve, in Churchill's words (1950), deciding if "[p]erhaps it is better to be irresponsible and right, than to be responsible and wrong." This dilemma is well known in forensic anthropology, as it permeates all aspects of the biological profile and the methods used in the estimation of its parameters of identity, i.e., from sex to ancestry to age to stature estimation. Klepinger (2006, 58), for example, has offered up her opinion on "best" practice for age estimation when indicators are limited: "I am inclined to disregard cranial suture closure when other markers can be used. There are, however, occasions when the cranial sutures are all there is , and then the best option to employ one of Nawrocki's (1998) regression formulae for various combinations of endocranial, ectocranial, and palatine suture closure." For ancestry estimation, however, Albanese and Saunders (2006) have argued an opposing view, contending that "False or misleading information is far worse than a lack of information. The relatively high risk of false information may outweigh the value that determining 'race' may possibly have for the positive identification of an unknown individual" (p. 281). Moreover, Steadman and Konigsberg (2009) and Konigsberg, Algee-Hewitt and Steadman (2009) have proposed a shift in focus from estimation in isolation (i.e., the question of method and result accuracy) to one that also includes the issue of evidentiary significance (i.e., the question of the findings' contribution, as one or one of many points of similarity, to making a statically informed identification). The most frequent (and often the most immediately profitable) reply to the question of deciding between a potentially poor estimate or no estimate is, however, one that is framed in terms of observer experience and the specifics of the case. With this approach, the decision to estimate (at all, with or without priors, by [not] applying an accepted "best fix") is largely philosophical, situational, and variable.

If the choice is deliberately made to assume equal (or uninformative) priors, the analyst may argue in justification that, for example, the risk of using inappropriate prior expectations for a population-at-large to which the unknown case may or may not belong outweighs the benefit that could be gained by contextualizing the classification, i.e., a potential rise in allocation accuracy. While this approach may remove some ambiguity in the research design, it does not correct for all areas of uncertainty in the statistical procedures that require the comparison of the unknown specimen against known reference populations. Perhaps one of the more commonly overlooked problems is the researcher's inability to account for initial misidentifications in the control groups themselves - an issue that has, only recently, been thoroughly reinvestigated for DFA (McLachlan 1992, Holden & Kelley 2010). Supervised classification techniques, like DFA, that allocate individuals into non-overlapping groups, must assume that the training data used to build the decision model are "perfectly classified" when they are reused to estimate parameters and make single-group membership assignments, e.g., for hard clustering of target unknowns (Holden & Kelley, 2010, p. 36). This statistical classification is, however, "never 100% accurate" in practice (Holden & Kelley 2010, p. 37). When class is latent, misclassifications are understandably common and unavoidable. Most practitioners would agree that in this context true class can never be known for every individual training case with certainty. Latent class analyses are easily conceived conceptually, therefore, as probabilistic, with a focus on estimation (c.f., Clogg, 1981, 1995; McCutcheon, 1987; Hagenaars & McCutcheon, 2002). As a special case of model-based clustering for multivariate discrete data, it assumes that each observation comes from one of a number of classes, groups or subpopulations, and models each with its own probability distribution (Lazarsfeld, 1950; Lazarsfeld & Henry, 1968; Wolfe, 1963; McLachlan & Peel, 2000; Fraley & Raftery, 2002). When class is directly observable and discussions involve the *a priori* definition and *a posteriori* use of known groups for data training and by-case classifying, coming to an understanding of certainty, or a lack thereof, in

the face of knowable identifiers (like observable morphologies, or self-declared and documented statuses) is more problematic. There is frequently an overall confidence placed in the truth of the individual *a priori* group identifiers, which subsequently begets a second-level confidence in the composition and characterization of the control groups themselves. This naturally engenders the facile belief that these constructed groups are each accurate representations of a given population in its traits and true distribution, thereby establishing as these reference groups and "perfect" baselines against which all comparisons are made. Group compositions are accepted as "real" and "unbiased" simply because their members were *a priori* labeled (i.e., external to the statistical procedure and before implementing the discrimination/classification algorithm), despite the variability in the degree of objectivity with which their personal group identifier was assigned. When these assumptions are reviewed critically and in the context of DFA, two key questions emerge. When the research design employs a method that requires the use of predefined "known" reference groups and expects that these categorizations have been made without initial mistakes in "observation" or "identification," in what ways are the resulting analyses vulnerable to error? What effect would misclassifications in the training data set have on the discriminant function procedure and the analyst's ability to obtain both statistically accurate and biologically meaningful classifications for the target specimens? As Holden and Kelley (2010) recall that "statistical classification methods can only be as good as the predictors used" (p. 37), it is easy to understand the ways in which significant error can be introduced even at the most fundamental stages. Problems at the level of data acquisition and preprocessing, for example, may be derived from simple mistakes in recording procedures or may be linked to more complex issues relating to poorly stated definitions, weakly delineated objectives, biases in ascertainment, and unresolved inter and intra-observer variation during collection and documentation (Boss, 1954; Katz & McSweeney, 1979). Once obtained, the condition of the data, the preprocessing choices made, and its distributional

characteristics (e.g., sample size, ratio of group sizes, effect size, degree of overlap, covariance structure, lack of sphericity, presence of outliers and inliers) may also influence classification accuracy by interacting with initially misclassified data proportions over the course of the discriminant function analysis - from model building to classification (Kuiper & Fisher, 1975; Blashfield, 1976; Edelbrock, 1979; Harrell & Lee, 1985; Van Ness & Yang, 1998; deCraen et al., 2006; Finch & Schneider, 2006; Holden & Kelley, 2010) .

In response to the explosion of many computationally intensive clustering and classification techniques in the last few decades, a variety of simulation studies have been conducted to compare competing and complementary statistical methods over a range of misclassification conditions, using different data types and models. Their results suggest that the performance of these techniques varies considerably between methods and that the nature of their responses can even be situationally specific, producing nuanced within-method discrepancies in success. In the context of hypothesis testing is the example of chi-square analysis, for which misclassifications of the initial data categories have been shown to have little effect on the validity of the significance test, although the overall power of the test may be reduced (Boss, 1954; Assakul & Proctor, 1967; Katz & McSweeney, 1979). Or, in the context of clustering, unsupervised classification methods can be considered (e.g., the mixture-modeling techniques essential to machine learning, A.I., and advanced pattern recognition), as their algorithmic design requires limited or no *a priori* information on true membership. As these methods seek to uncover the number of groups that best represent the structure of the data without identifiers or external reference populations, their behavior remains consistently stable despite the introduction of initially misclassified cases into the overall "mixture" of data (Holden & Kelley, 2010). Methods, whose execution is dependent upon a control population or reference groups, and for whom some

"knowledge" about the "real" identities of their members is presumed, are, however, not so stable. For discriminant function analysis, the presence of misclassifications in the training data set has be found to have a demonstrable, significant effect on classification accuracy when partitioning target data sets or making by-case allocations (Lachenbruch, 1966, 1974, 1979; McLachlan, 1972; Chhikara & McKeon, 1984; Grayson, 1987; Holden & Kelley, 2010). The percentage of correctly classified cases is known to be systematically underestimated. Early studies of supervised classification methods, and especially DFA, have also detected a patterned difference in the procedure's behavior when a distinction is made between the two types of misclassifications resulting from either a random or nonrandom source of error (Lachenbruch, 1966, 1974, 1979; McLachlan, 1972; Chhikara & McKeon, 1984). While it has been argued that random misclassifications of DFA training data have a greater influence on classification accuracy than do nonrandom misclassifications, recent reevaluations of this hypothesis, conducted without the same limitations placed "on the scope of the variables investigated" (e.g., ratio of group sizes), have come to the opposite conclusion (Holden & Kelley 2010, p. 38). Holden and Kelley (2010, p. 41) have not only established that, in fact, the error introduced by the presence of nonrandom data points (i.e., those training cases with differing probabilities of being misclassified, depending upon their relative position in the group distribution) has the greatest impact on the ability of discriminant function analysis to produce optimal classifications. They have also convincingly argued that these nonrandom, problem misclassifications tend to be the "borderline" cases (i.e., those which lie closer to the overlap of two or more distributions) and not the "outliers" (i.e., cases lying at the distributional tails) as they had been traditionally perceived and to which DFA is highly sensitive (Kuiper & Fisher, 1975; Edelbrock, 1979; Van Ness & Yang, 1998). When DFA is broken down into its linear (LDA) and quadratic (QDA) forms and these procedures are compared for their responses to misclassified data (with random and nonrandom types combined), Holden and Kelley (2010) have concluded that LDA is less affected than

QDA. Corroborating Lachenbruch's (1979) original findings, they have also demonstrated that the sample covariance matrices diverge systematically from the population parameters as the number of misclassified cases grows. Since FORDISC employs only LDA, these results may be helpful in providing practical justification for choosing such a simple variance-covariance model (equal for all groups) beyond the argument that LDF is "the most commonly used DFA" (Ousley & Jantz 2005; c.f., Konigsberg, Algee-Hewitt and Steadman, 2009, for QDA when estimating race) - but it is only in part reassuring to learn that LDA is the better performing of the two potentially weak approaches.

While this new knowledge that DFA is, in practice, robust in the presence of random misclassifications in the training data and the reconfirmation that LDA, specifically, is less vulnerable to both types of data error are encouraging findings, their importance is greatly overshadowed by the fact that nonrandom misclassified data "poses such a threat to classification accuracy using DFA" (Holden & Kelley, 2010, p. 51). This discovery is troubling for two reasons: (1) nonrandom misclassifications tend to be "borderline" in nature and (2) uncorrected nonrandom data errors are "more likely to occur in practice," are "more realistic in many situations," and are expected to dominate errors in human "diagnostic" contexts especially (Holden & Kelley, 2010, p. 41, 51). The trend towards higher rates of misclassification in these marginal cases has significant implications for human identification when the "hard" classification approach of DFA for target allocations and the often "fuzzy" qualities of craniofacial data and group identifiers are considered. Ancestry or race signatures often overlap, individuals may hold multiple memberships, and many groups may be mixed in their composition (i.e., representing several populations with differing degrees of contribution). The "initial" preallocation of a marginal case into, say, Group A may therefore be just as valid as a preallocation into Group B , C, or D. Socio-cultural factors and the degree of parental contributions in the context of admixture may, however, direct these

borderline classifications towards the identification with a particular group in a nonrandom fashion in one sample subset and in a different but equally valid direction in another subset (which may be composed of similar individuals who are linked in some other way). It may be argued that there is a problem of compatibility between the "hard" classification approach that requires rigid assignment into one single group and the reality of inter-population fluidity, where one individual may meet at any one time several aspects of the criteria that define multiple groups. The information and practical value of any subsequent target classifications that rely upon these complex and confused partitions of the training data for their allocation may be seriously reduced. For studies using group qualifiers that fall under the rubric of demography (for which there are many documented, acceptable, and mutually-inclusive identifiers of race ancestry, ethnicity, nationality, etc) and draw upon comparative skeletal populations that are, on the one hand, composed of disparate samples but, on the other hand, are defined by a single best-fit or catchall term for the "collection" as a whole (which likely adheres to some naming convention and not the reality of their biodistances or relatedness), the greatest confounding factor is the negotiated link between a personal interpretation of membership terminology (when self-identifying or during the ascription of status to others) and the biological variables that enable individual assessment and group quantification. This is an issue that is not only present in these initial stage of research design and model building, but is one that will remain omnipresent in all subsequent analytical steps. This problem holds particular importance for any forensic anthropologist who works with recently admixed groups, whose parental population contributions are variable and often geographically structured and whose complex self-identity is often poorly captured (grossly simplified or irregularly specified) by the bureaucratic terms for and definitions of race and ethnicity that permeate personal identification documents and vital statistics reports (e.g., equal employment opportunity self-disclosure forms, driver's license and state ID applications, medical records, and census data). These complications

for casework have arisen in earnest in more recent years, growing out of rapid transformations to the global demographic structure and the unprecedented "modern" human experience, and, many experts predict that this issue will only expand, and its importance intensify, with time. In a 2009 press statement made by the U.S. Census Bureau Newsroom (Release cb09-75 2009), it was estimated that nearly half of children under age 5 represent minorities. Both geographic and social mobility have increased interpopulation contact and the patterns and rates of admixture; moreover, new political, economic and social climates will likely continue to relax previous biocultural barriers. These factors are making for, what some researchers have called, "a shrinking world," in which "mixing populations are scrambling genetic signals" (The Genographic Project, 2011, np) and with which two phenomena are frequently identified: (1) "distinct peoples, languages, and cultures are quickly vanishing into a 21st century global melting pot" (The Genographic Project, 2011, np), while, at the same time, (2) census data is revealing that changes are occurring in the diversity of groups represented, that the numbers of minority individuals making up these groups are shifting, and that complex biosocial groups accompanied by new self-identification rules and terminology are steadily emerging (Feldman, Lewontin & King, 2003; U.S. Census Bureau Newsroom Release, cb09-75 2009). To illustrate the impact of the current state of demographic flux on American forensic anthropology, it is only necessary to consider the problem of making reliable, quantitative identifications for contemporary skeletal remains of Hispanic origin. Hispanics now rank as the single largest minority population in the United States (U.S. Census Bureau NC-EST2009-03). According to a recent report by the Pew Hispanic Center (Passel & Cohn, 2011), the number of Hispanics counted in the 2010 Census surpassed the growth expectations previously released by the Census Bureau: ongoing data analysis for 33 states suggests that Hispanics represent the majority (58%) of the population expansion recorded in the last decade and the current combined total of 38.7 million Hispanics across these states includes roughly 590,000 individuals (or

1.5%) more than accounted for by the Bureau's projections. This discrepancy in estimation has been attributed to the unprecedented growth in the number of unauthorized immigrants. The implications of this immigrant influx to forensic anthropology and its impact on the nature of identification casework cannot be overstated. Border Safety Initiative (BSI) reports on the "target-zone" have shown that, between 1998 and 2005, the number of migrant border-crossing deaths recorded annually almost doubled, from 266 in 1998 to 472 in 2005 (U.S. Government Accountability Office GAO-06-770).[3] For the Arizona-Mexico border alone, the non-profit non-governmental human rights organization, Coalición de Derechos Humanos (2010, 2011), has estimated that, in fact, over 5,000 migrant deaths have occurred *since* border control policies were implemented in the 1990s and provides on their website a detailed list of 253 (170 males, 32 females) recovered remains for the period between October 2009 to September 2010 alone. This external documentation is ongoing. The Coalición de Derechos Humanos maintains an regularly updated count online (59 individuals, from October 1, 2010 to February 28, 2011) and similar statistics, with information on personal identities, can be obtained through the Arizona Daily Star's searchable Border Death Database, which provides web-access to existing cases for the last decade (Arizona Daily Star website 2010).

Despite the growing numbers of Hispanics, there is a notable lack of morphogenetic information that adequately captures the range of their biological variation. As a group, they represent individuals of many geographical origins with complex sociobiological histories. The U.S. Census Bureau includes within their largely linguistic (Spanish-speaking) definition of "Hispanic," individuals originating from

[3] The methodology established by the BSI in 1998 defines border-crossing deaths as "those occurring in furtherance of an illegal entry" and its target zone as "an area consisting of 45 counties on or near the southwest border with Mexico...that includes 9 of the 20 Border Patrol sectors responsible for enforcing U.S. borders and securing official ports of entry." (U.S. Government Accountability Office GAO-06-770, p. 1)

Mexico, Puerto Rico, Cuba, South or Central America, or other Hispanic／Latino origins (Ramirez & de la Cruz, 2002). For skeletal data in particular, it is this diversity that complicates both the study of Hispanic skeletons and the codification of solid identification standards on both practical and theoretical levels. Sample sizes that adequately represent all of the contributing groups are largely impossible to obtain, and so efforts in recording and in the interpretation of the distributions of biological traits that characterize these groups individually and, in turn, "Hispanics" generally, are significantly constrained. While the high number of border-crossing deaths of migrants from Mexico into the Southern United States does ensure that many Hispanic cases pass through the offices of the county Medical Examiner and Anthropologist, the personal data that accompanies a modern reference skeleton is unreliably available and positive identifications may be but a small percentage of a yearly caseload. There is, as the U.S. Government Accountability Office (GAO-06-770) has noted, an "uncertainty that arises in those cases involving bodies discovered in the desert or other remote areas" (p. 26), so that

> [i]n some of these instances both Border Patrol agents and local medical examiners must use their professional judgment in determining whether circumstantial evidence is sufficient to classify a decedent as a migrant who died while in furtherance of an illegal entry. Both Border Patrol officials and local medical examiners with whom we spoke reported relying on such evidence as the type of clothing worn by the decedent, whether or not the person was carrying water jugs (as evidence that the person intended to travel some distance on foot), as well as any personal documents or identification that might indicate country of origin. Border Patrol officials and others also reported that, in many cases where the decedent had no identification or only skeletal remains were found, they may conclude that the decedent was a migrant attempting illegal entry because the

remains were found in a remote area that was a known migrant-crossing corridor (p. 26-27).

While there are no large scale comparative skeletal collections of Hispanics that are analogous to, for example, the Terry or Hamann-Todd collections for American Blacks and Whites available to date (Usher 2002), the Forensic Data Bank, maintained through the Forensic Anthropology Center at the University of Tennessee, represents the largest skeletal database for modern Hispanics. Although it contains both metric data and demographic information for over 150 individuals, of which 88 South West Hispanics and 83 Guatemalans are included in Fordisc 3 (Ousley & Jantz, 2005, np), it is, however, limited in its representation of the many substructured groups that make up this composite population. The FDB, as well as the William M. Bass Donated and Forensic Skeletal Collections, contain samples mostly from Mexico and the American South West and with but a few examples of other biogeographic groups that would better represent the casework of forensic anthropologists in the South and North East, including for example Cubans, Puerto Ricans, and Dominicans. These imperfect resources and the vexing issues of biological diversity and self-identity are clear from the relatively small Hispanic sample in FORDISC 3.0 (only 88 males) and it is relevant to recall Ousley and Jantz's (2005) comments on the limited representation of Hispanic populations in the reference database (excluding, e.g., Black Caribbean Hispanics). Despite these constraints, research into Hispanic skeletal variation is ongoing. What has been found most recently for craniofacial data is, that (1) "discriminant function analyses indicate that American White criteria provide poor estimations," (2) "ancestry estimation of Hispanic crania is difficult," and (3) that weak classification success, when applying DFA to limited craniometrics samples, may be "because Hispanic individuals can have differing degrees of Native American, European, and African ancestry" (Spradley et al., 2008, p. 21). In a study focusing on the metric identification of Hispanic remains, Spradley and colleagues (2008) reported that when a LDF analysis was conducted

using four control groups (Hispanics from the FDB, Guatemalan Mayans, American Blacks and American Whites), only 45% of the Hispanic training sample was correctly allocated to their *a priori* known group. This low level of accuracy may reflect the morphogenetically intermediate position of the FDB Hispanic between American Whites and the Guatemalan (or, more broadly defined, Native American) samples. When this model was used to classify an external Hispanic test set, only 10 of the 21 cases were correctly identified and, after inspecting the associated posterior probabilities (0.5 on average), the authors concluded "that they are just as likely to belong to another group" (Spradley et al., 2008, p. 27). These findings clearly demonstrate the potential difficulty in obtaining good discrimination when (1) membership characteristics of the original reference sample are "borderline," (2) overlap between groups is considerable and complex for both the training and target samples, and (3) test cases appear to be "intermediate" between two or more parental populations due to a history of hybridization and admixture. The changing demographic trends in the U.S. towards a steady growth in Hispanic minority groups and the realities of high numbers of migrant deaths and border-crossing fatalities attest to the immediate and paramount importance of developing better population-specific standards for identification that account for the unique group characteristics of the composite, diverse Hispanic peoples, as well as best-practice statistical methodologies that enable the reliable, accurate estimation of individual biological parameters.

The methodological weakness in discriminant function analysis for the assessment of individuals of Hispanic origins is not limited to making forensic identifications for groups that meet these "problem" criteria. It is important to realize that the original group organization has implications for assigning meaning to any classification results produced using discriminant function analysis, even when the most reliable data set is used. Relethford (2009) has recently demonstrated how a change to the partitions of

the Howells data set, from the original 18 populations to other equally correct but different groupings (collapsing or pooling the populations by regions), can produce results that call into question the ability to make convincing statements on the reliability of the analysis. Relethford's study does not provide new information on classification procedures, as it has been long known in the statistical literature that the number of predefined clusters, the ways in which the data is originally organized, and the choice of partitioning algorithms (especially when dealing with model-free clustering and classification methods) can yield different solutions, each with equally high degrees of accuracy and statistical significance. What his study does do is contextualize this ambiguity within the framework of an anthropological problem. His findings beg the question: "[i]f defining different numbers of geographic races all produce essentially the same rate of accuracy in classification, then what justification exists for preferring one number of races to another?" (Relethford, 2009, p. 20). Since Relethford believes that "there is no justification based on classification accuracy for preferring five, six, seven, or eight groups," he has argued that "from the perspective of racial classification, if we take these different groups as equivalent to races, then we see that the number of races, and the geographic cutoffs used to define them, are subjective decisions" (p. 20). While this statement is true, the clarity of its impact is reduced by his methodological choice of using discriminant function analysis, i.e., the same approach that has been historically shown to give statistical credence to craniometric classification. While his goal is to evaluate key assumptions about group differentiation, as they are associated with "race" (biogeographic) classification, the nature of his research design makes it difficult to extract his test of "race" from a simultaneous test of discriminant function analysis. This potential point of confusion suggests that in the interest of independent evaluation, another statistical clustering or classification approach is warranted.

Closely tied to the methodological considerations associated with choosing predefined populations and incorporating prior expectations when model building is the selection of the most appropriate target sample and the most informative configuration of variables. Good decisions in these areas are essential to the success of the subsequent classification and validation procedures. While the circumstances of any given case will dictate the options available for the best variable selection or population matching, there is less constraint on these options when the purpose of the analysis is to conduct validation tests of discriminant function techniques and, more specifically, assess the utility of the FORDISC software. The same holds true for the evaluation of broad concepts, like population substructure (and biosocial race or biogeographic ancestry), using large-scale data sets, are the focus of concern. The ideal first step towards this goal is to establish optimal classification accuracy under controlled conditions. This baseline should provide a standard against which similar analyses can be evaluated for consistency and enable effective comparisons under differing conditions, for example when later steps are taken to refine the assessment by introducing potential sources of error (e.g., initial misclassifications, missing data, many or few variables, predictors that correspond to different aspects craniofacial shape, etc.). In this context of experimentation, test cases and the craniometrics that represent them should be thoughtfully chosen so that they not only adhere to robust statistical expectations but best capture the skeletal variation present within and among the groups of interest. In the above discussions of the many "archaeological" and "forensic" tests of FORDISC for the estimation of "affinity" or "race," the theoretical and practical problems that may arise from the indiscriminate use of target cases that are badly suited to the method of analysis or larger project goal have already been isolated. These specimens may be, for example, poorly related to the control samples in terms of time, place and sphere of biohistorical interaction, or, their population history may be complex and inadequately understood, e.g., as hybrids of previously hybridized groups, social categories composed of variably admixed subunits, populations with unknown

origins and disputed trajectories. It also needs to be stressed that the use of test cases with incompatible population characteristics, like sample age (ancient or modern), geographic origins (local, global), unique features or group specific anomalies (genetic or pathological abnormalities, environmental effects, intentional cranial modifications), sample significance (archaeological, historical, medico-legal), is an equally ill-conceived approach. It is to be expected that, even in the face of results that may meet some acceptable level of accuracy or even achieve statistical significance, any classifications made using poorly chosen cases will offer little insight into the utility of a given allocation tool for in the context for which it was devised. From the nature of the discriminant function algorithm, any project seeking to evaluate FORDISC will obtain interpretable output. What is important is whether this output has meaning. Moreover, the findings from this assessment need to be interrogated to determine if they satisfy the project's goal. If, for example, the intention is to conduct a reliable method test, then the good or bad classification of contextually irrelevant samples, like archaeological specimens from ancient Egypt, offers arguably little or no practical information on the value of FORDISC in circumstances related to modern American forensic identification. With this said, it is worth returning to the debate over the tests of FORDISC using the "unknown" Nubian crania.

In reaction to the inappropriate choice of using ancient Nubians as "forensic" test cases, several of the published responses chose to evaluate FORDISC's classification accuracy using, what can be understood as a relaxed approach to validation by re-substitution. Naar, Hilgenberg, and Armelagos (2006), Hubbe and Neves (2007) and Elliott and Collard (2009) all selectively sampled from Howells' craniometric data set in accordance with their individual project intentions and, in turn, generated classifications and "tested" FORDISC accuracy using a subset of the same data that trained the program's discriminant functions and developed the model-classifiers. While their decision to re-substitute may circumvent the

issue of identifying a reasonable independent sample of target unknowns, it does so at the cost of statistical rigor. Their analyses, therefore, suffer from a different, but equally serious, failing in methodology. Although the convenience and simplicity of re-substitution often, initially, recommends this approach and while there is some precedent for its use in the statistical literature, it is most often considered overly optimistic, and, so, is thought to be a poor predictor of accuracy. Its ability to provide realistic estimates of the classifier's utility is limited by the *post hoc* nature of its implementation; it is typically taken to be satisfactory only when data sets are very large and when they have been subjected to formalized procedures or algorithmic rules (see Efron and Tibshirani [1997] for random sampling with resubstitution and Kohavi [1995] for an evaluation of cross-validation and the bootstrap). Global cross-validation, for example, has gained in popularity as a valid large-sample data mining technique when used with discriminant-based decision or classification tree analysis: in this case the full analysis is replicated a specified number of times while holding-out a subset of the training sample (equal to 1 over the specified number of times), and each hold-out subset is subsequently used as a test sample for the cross-validation of the classification tree in question (Breiman et al., 1984). Lin and Pourhmadi (1998) have demonstrated, however, that, while its use for the purpose of model-seeking and improvement of model-fit may yield small prediction errors for the training data, the optimization gained in this way will result in some considerable loss of practical generalization. The extent of this effect on classification accuracy cannot be *a priori* known, may not be easily estimated, and is, therefore, likely hard to control or correct for *a posteriori*. In the context of most traditional classification methods, and especially for the relatively small sample DFA that has populated the conventional quantitative trait literature in anthropology, this approach is understandably highly problematic. It is not uncommon for the final model to over-fit the training data and thereby produce methodologically biased, overly optimistic results (Chatfield, 1994; McLachlan, 1992). Artificially inflated estimates are of even greater concern

when the classifier has been specifically designed to deal with the unique qualities of the original data set. Models are commonly shaped in this way by the routine application of some selection technique, like the implementation of a stepwise selection of predictor variables - a procedure that is typically incorporated into standard discriminant function analyses (and statistical software packages) and, as such, has been included within FORDISC's toolkit (Ousley & Jantz, 2005). In a cautious statement on the potential for problems in an approach that is dependent upon preprocessing, extensive design, and/or automated assessment, Faraway explains (2002) that

> [t]here is a danger of doing too much analysis. The more transformations and permutations of leaving out influential points you do, the better fitting model you will find. Torture the data long enough, and sooner or later it will confess. Remember that fitting the data well is no guarantee of good predictive performance or that the model is a good representation of the underlying population (p. 134).

In this way, it can be seen how extremely good classification results can be obtained when the training and target samples are held the same, but it is very unlikely that these allocation rates are realistic, and, therefore, useful reflections of the model's actual predictive accuracy. In order to independently evaluate how well the DF model (based upon the functions computed from the original "training set") performs in practical applications (making classifications for randomly obtained "unknowns"), it is advisable to classify only "new" test cases, i.e., target specimens that were not used to estimate the model's parameters (the classification functions) and, in the best scenario, were collected by prospective sampling (Fielding, 2007). In an ideal "independent" analysis, the test set is not employed until after the model has been developed and validated. In order to segregate the final assessment from the "tuning" of the performance of the classifier, a more rigorous test using three data sets (e.g.,

training, validation, test samples) is often encouraged (Fielding, 2007). It is expected that the accuracy achieved by using external test data in this way will be much lower than any estimates obtained by reusing the original training data as the test sample; it will be conservative but it will also be more robust in its assessment of the classifier's future performance and ultimate predictive value (Henery, 1994). The re-classification of training cases is typically reserved as a diagnostic tool for the detection of outliers, for the identification of areas where a given classification function appears inadequate, or in exceptional circumstances when exact allocation rates are not the focus of the analysis. The tests of FORDISC that employ this suspect approach should therefore be reevaluated under the critical assumption that any findings reported may be founded upon problematic estimates of accuracy. Conclusions for or against the utility of FORDISC to forensic casework may reflect some statistical artifacts instead of the software's true ability to classify unknowns effectively and reliably under the conditions defined by the user (e.g., choice of reference populations, test samples and variables).

While Hubbe and Neves (2007, p. 285) do not overtly address previous decisions to use archaeological, even ancient, samples to test FORDISC, they do become indirect participants in this discussion when they themselves make a questionable choice of target sample when evaluating the relationship between the number of predictors used in a discriminant function analysis and its classification success. In their reply to the Nubian studies, they argue specifically that Williams and coworkers (2005) used too few (12) variables, despite their presumed "diagnostic value" (Williams et al., p. 342). Hubbe and Neves posit that these predictors are "far from enough to classify a skull on the basis of discriminant functions", and, that it is for this reason alone that the "design of their experiment is flawed" (2007, p. 285). They (2007, p. 285) set out "to demonstrate this" using a total of 18 populations, which they drew directly from the Howells data set and selectively sampled for their geographic representation (i.e., choosing three groups

from each "region"). They employed 57 variables (the maximum number shared by all of the Howells groups), which they used *in toto* and in various subset configurations. They retained the original population designations (and partitions) for use as the identifier of true group membership in *their post-hoc* assessments of classification accuracy, assuming that "a skull was considered as correctly classified only when it was associated with its original populations" (2007, p. 285). The results of their analyses were consistent and clear: in the tests using the full set, classification accuracies always exceeded 90%. These allocation rates were recently mirrored by Elliott and Collard (2009), who also reused the Howells data set as their target sample: when sex was presumed known and the source-population was included in the reference samples selected, FORDISC correctly classified 85.5% of the test cases when 56 variables were used and only 27%, 22% and 5% when 10 variable subsets (representing the neurocranium, face and basicranium) were used (c.f, Elliott, 2008). These findings appear to offer considerable support for the utility of FORDISC as a responsible tool within the anthropologist's identification kit: even two of its critics demonstrate that it can work very well. These conditions of high allocation rates also give important insight into best-practice for study design and into the practical performance of FORDISC with respect to variable selection. Results such as these suggest, in theory, that the FORDISC user should employ the maximum number of variables available for the case or allowed in the program. They also indicate, in practice, that the success of the classification will be heavily influenced by the practitioner's ability to obtain these measurements on cranial specimens of highly unpredictable states of preservation and frequently wrought with missing data. What is critical about these findings is that they go against some of the more common statistical protocols for variable inclusion, e.g., the common-place use of case to predictor ratios of 3:1 or 5:1 and selection algorithms to reduce predictor dimensionality (Huberty, 1994; Fielding, 2007). They also appear to be at odds with the recommendations made for FORDISC. From a statistical perspective, Ousley and Jantz (2005, np) advocate for a step-wise selected

procedure, as it "often removes 'noisy' variables that do not contribute to discrimination." They also caution the user against selecting too many variables, in terms of the practical constraints of the FORDISC database: as missing data is not imputed for the reference groups, choosing all or a large number of predictors will limit the number of reference cases that can be used in the DF analysis. Their advice implies, therefore, that it is preferable to sacrifice variable number for comparative sample size. These views, when read in light of Hubbe and Neves' analysis, call into question FORDISC's capacity to reveal memberships for fragmented specimens that may be lacking key landmarks or morphoscopic features and for whom "ancestry" may not be otherwise observable (i.e., absence or poor representation of the facial region). Williams and Armelagos (2007, p. 286) have written in recent justification of their choice to limit their analysis to 12 predictors that "there is no stipulated number of variables...simply because forensic evidence is often fragmentary." While opinions appear to be split on variable use, regardless of position on the utility of FORDISC, it is difficult to easily accept Hubbe and Neves' conclusions. What is problematic is the matter of teasing out the effect of their decision to re-sample the Howells data: it is essential to determine if the percent values reported in these studies are inflated or truthful representations of the program's potential for success. The fact that Hubbe and Neves (2007) did find a 30% difference in classification accuracy between the classifications based on 57 variables and those using only 11 when the test samples were kept constant does suggest that the number of variables is an influential factor in a FORDISC analysis. These finding are, however, complicated by evidence of an apparent inconsistency in these re-substitution based studies: Naar, Hilgenberg and Armelagos (2006) reclassified Howells' full Egyptian reference population and, despite their reuse of individuals included in FORDISC's control database and the concomitant expectation for inflated estimates of success, their classification accuracy fell midway between the highs and lows obtained by Hubbe and Neves (2007) and Elliott and Collard (2009): of the 111 specimens tested, 55

(49.5%) were correctly identified. On the one hand, this reduction in accuracy may be attributed to the limits imposed on the number of predictors used; on the other hand, it is necessary to question their choice of probability thresholds for identifying a correct allocation (e.g., here, based on a typicality greater than 0.1 instead of "below 0.05 (5%), or certainly 0.01 (1%) for a group," as Ousley and Jantz [2005] have recommended). If all the specimens who were allocated into the Egyptian group are simply considered to be correct, 40 out of 111 specimens were misclassified. From this revised perspective, their percent correct value (64%) is not as egregiously low as the extremely poor classification rates obtained when "too few" variables (10 or 11) were used by the studies that followed.

Accepting these findings with confidence given the discrepancies associated with both sides of the argument is difficult: is it necessary to argue for a change in either best-practice protocol for test design or variable selection to ensure optimal FORDISC results or is it necessary to ultimately reject FORDISC on the grounds that these factors are of little consequence? This decision is further complicated when consideration is given to the fact that equally good, if not better, classification rates have been obtained by others, using appreciably fewer variables than the full complement of 57 or more collected by Howells and often even less than the 24 cranial measurements included on the Forensic Data Bank case recording forms. For example, Brues (1990) obtained a 98% accuracy in assignments by racial group using 17 globally distributed populations and 34 measurements; Ousley, Jantz and Freid (2009) reported classification rates of 80% when using only 2 variables, 95% with 7 variables and 97% with 19 variables for a sample of American Whites and Blacks; and Konigsberg and coworkers (2009) demonstrated that, when prior population expectations are considered, posterior probabilities of membership can reach 99% when only 13 standard craniometric variables are used. It is also necessary to contend with the serious statistical and philosophic reservations for adopting a more complicated analysis that includes

many variables of questionable value. Relethford (2009) has made the point that the nature of the sample and the basic principles of geographic distance, when combined with many variables, may inflate classification results to such an extent that they are not realistic (in terms of procedure and protocol for future applications) or useful (in terms of unique information value):

> I suggest that if one defines clusters in such a way that geographical distances among clusters typically exceed geographic distance within clusters, and if one uses enough traits in the analysis, then there will be a high degree of accuracy in classification. These results again stem from the simple fact that people living farther apart in the world tend to look more different combined with a large number of traits...Given enough markers, one can easily assign individuals to geographically widespread groupings such as Western Europe, East Asia, and sub-Saharan Africa. These analyses are simply another way of expressing the correlation between genetics and geography (p. 20).

In the context of discriminatory power, it is especially important to avoid the inclusion of many redundant or irrelevant predictors. It is only necessary to recall Watanabe's (1960, 1985) "ugly duckling" theorem to understand the critical importance of weighted membership. The retention of variables that contribute little or no new information on group membership may drive an analysis towards the artificial homogenization of the otherwise distinct cases. This, in turn, shrinks the naturally occurring distances between groups, making all (different) groups appear (more) similar, and degrading the classifier's performance. There are flaws, in principle, with the studies that test FORDISC using a maximum-variable approach , particularly in regards to the statement made by Elliott (2008, p. 18) for method justification: "Following Hubbe and Neves (2007), more variables should provide better discrimination power. At worst, adding more variables would simply fail to improve discrimination and result in a plateau effect."

It can be argued that the most statically satisfying and practically beneficial model is one that adheres to the heuristic of parsimony: "*Frustra fit per plura quod potest fieri per pauciora*" (Occam, Summa Totius Logicae, i. 12). After Mac Nally's (2000) apt explanation, the reasons why building a simple predictive model is preferable can be summarized using but four points: (1) over-fitting lessens model utility, as modeling "noise" reduces the classifier's generalization to other independent data sets and under varying test conditions, (2) the configuration with the fewest but most effective predictors is generally best able to secure the good future performance of any given model, when considering different sources and forms of data and changes to circumstances of its collection and analysis, (3) simple models most often give more insight into (or more clearly discernable information on) causality, and (4) from a very practical resource perspective, fewer predictors require less expenditures in terms of time, labor, and cost of resources required during data collection and for future archiving and analysis. It is wise to recall that there is always the fundamental need to balance the classifier's bias against its variance. As Burnham and Anderson (2004) have argued, there is an inevitable conflict between the desire to build a parsimonious model, one that avoids the inclusion of noise and the issue of over-fitting, and the critical requirement for this simple model to not only adequately but most effectively represent the complexity of the data. As Fielding (2007) advises, "the selection of variables always involves compromises but the selections should be driven primarily by domain-specific knowledge" (p. 87). It is crucial to attempt, therefore, to mitigate this issue of all-variable inclusion with optimal variable selection or balanced reduction, by arguing that a predictor-subset, consisting of relatively few variables, can be justified based on projections of prediction accuracy using the design and results of comparable studies or selection techniques that are themselves methodologically robust and have been shown to be successful for similar data and analogous questions. Moreover, while engaging in heady debates over statistical logic, it is important not to overlook the pragmatic matter of application. It must be realized that the

approaches that advocate for many variables show a disregard for practical utility: predictor choices are often externally dictated (and *de facto* validated) by the unavoidable, physical limitations of real-world cases, which often demand analyses of fragmentary, poorly preserved skeletal remains and the common practice of sharing data acquired by other observers, which may have been collected following different standards or procedures, and to specific project ends. With that said, it is also necessary to question the logic of studies that argue for the choice of test samples or methodological approaches based on parsimony (e.g., the limited number of test populations, selectively sampled for geographic representation, coverage of diversity, and known relationships), while also choosing to include, and ultimately advocate for, the maximum number of variables common to all groups. This is a decision which introduces the highly debated and poorly resolved question of missing value imputation, by-variable exclusion, or by-case deletion for multivariate biological data (Rubin, 1976; Little & Schluchter, 1985; Little, 1988; Little & Rubin, 2002). It is important to remain skeptical, therefore, of the findings of any projects that set out to critically "test" FORDISC's methodology or utility, while engaging, themselves, in questionable statistical practices.

While these exchanges over reference group(s), test sample(s) and variable selection between members of the opposing sides of the FORDISC debate appear to introduce more problems than they resolve (bringing to light additional sources of error and poor methodologies in these "corrective" studies), these discussions do make an important contribution to an understanding of the patterned distribution of human variation. To return to the basic beliefs put forth for the non-existence of race and the ways in which these arguments have contributed to making a case for the impossibility of biologically meaningful group discrimination, one of the key issues taken up by the "unracialized" view of contemporary human variation is the number of traits used in classification and their "nonconcordance"

(Armelagos & Van Gerven, 2003). In the early evolution of populational thinking, Livingstone (1962) was among the first to initiate the discussion of what has become the "independent variation argument" or "the fact of discordance" (Appiah, 1996, p. 99; Andreasen, 2004, p. 428, 2005, p. 94, 2008, p. 484; Machery and Faucher, 2005, p. 1209; Glasgow, 2009, p. 88) and his statement on race is now a classic quotation that has been put to frequent use in support of arguments against classification:

> [i]f one genetic character is used, it is possible to divide a species into subspecies according to the variation in this character. If two characters are used, it may still be possible, but there will be some "problem populations," which, if you are an anthropologist, will be labeled composite or mixed. As the number of characters increases, it becomes more nearly impossible to determine what the "actual races really are" (p. 279).

For phenotypic variation in physical anthropology specifically, this discussion of the concordance of traits reached full force among the "new" anthropologists and its impact is clear in the writings of scholars, like Diamond (1994), Goodman (1997), and Brace (2005). Brown and Armelagos (2001), for example, have seized upon this passage for use in their arguments against phenotypic classification, ancestry estimation and biological race, explaining that

> [t]he success of racial classification depends on the number of traits used in ordering the races. A single trait such as skin color will result in a classification system that is easily determined. Add another trait and classification becomes a more difficult task, and there usually are groups that cannot be classified. As you increase the number of traits, the problems in racial classification become insurmountable (p. 34).

Moreover, Armelagos and Van Gerven (2003) have extended this view to the debate over FORDISC, arguing that "we have known for decades, so-called racial traits are nonconcordant, and the races, we

get are little more than a function of the trait or traits we use. Sadly, the response has been directed more toward fixing the program than fixing the approach" (p. 61). While these statements on traits, their relationships and their numbers, are strong in their anti-race rhetoric, the recent deliberations over variable selection have exposed their statistical weakness. They have degraded, therefore in turn, the authority of this old line of reasoning on both levels of the ongoing classification debate: in the context of theory (race and human taxonomy) and practice (FORDISC and ancestry estimation). The high degrees of accuracy reported by those studies that used a moderate number of predictors (Brues, 1990; Ousley, Jantz & Freid, 2009; Konigsberg, Algee-Hewitt & Steadman, 2009) and, especially, the equally high classification success achieved by the many-variable approach tested by researchers representing both sides of the discussion (Hubbe & Neves, 2007; Elliott & Collard, 2009) provide empirical evidence that the classificatory power of the function does not decrease with an increase in the number of variables in the discriminant function analysis. These conclusions refute, therefore, the now long-standing belief that traits are nonconcordant and offer mathematical proof that classification is *not* "insurmountable" (c.f., Konigsberg, Algee-Hewitt & Steadman, 2009). In doing so, they not only invalidate many of the methodological criticisms that have been specifically directed towards the forensic estimation of ancestry and the medico-legal contributions of FORDISC but, in speaking to this issue of nonconcordance, they also aid more broadly in destabilizing many of the classic arguments that have linked quantitative trait analysis to "human taxonomy" and have been mobilized against ancestry estimation in biological anthropology.

This return to the larger question of detecting meaningful, population-specific patterns in contemporary craniometrics reveals that the FORDISC debate is, more often than not, a discursive site onto which unresolved questions over the organization of human biological variation have been transferred under

the guise of testing FORDIC's practical utility. What can be learned from these critical exchanges is that a good test of FORDISC is one that also evaluates the validity of ancestry estimation (and, by extension, the use of race in medico-legal and biomedical research) as it relates to classification and patterned human variation. These two subjects are inextricable and so a good approach is one that marries the program-specific (individual, goal-oriented) concerns with the more profound, and more theoretical, question of detecting world-wide population substructuring using craniofacial shape. In addressing classification in one context, an equal contribution is made to an understanding of it in the other. In terms of practical methodology, the problems that have been identified for the FORDISC "procedure" can be drawn upon in order to facilitate a movement forward from these criticisms. From this perspective, the most profitable first approach involves the vigilant selection of highly relevant test samples, the use of robust variables that capture the greatest amount of variation in the craniofacial skeleton while upholding the rule of parsimony, and the application of a method that is free of the problems associated with *a priori* defined reference samples. Through an analysis that is not entangled in the preconditions of discriminant function analysis, the first steps can be made towards addressing population partitioning without the weight of the contemporary arguments that surround it, as well as the historical concerns over "taxonomy" that dominated the transition from the "old" to the "new" anthropology. The matters of context, prior expectations, and the evidentiary significance of ancestry estimates and biosocial race can, then, be elaborated upon with respect to an established or "known" baseline. Recent tests of group separation or "clustering" using craniometric data have emerged following these methodological lines. Ousley, Jantz and Freid (2009), for example, have made use of a variety of group differentiation and individual membership identification procedures, including discriminant function analysis, K-nearest neighbor, principal component analysis and K-means clustering, to evaluate differences in resolution by method, assess morphological relationships between

biosocial race groups (American Blacks and Whites) of interest to forensic anthropologists in the United States and explore the potential for biogeographic classification given patterns of world-wide variation. Their K-means cluster analysis obtained good rates of allocation and cluster separation when the analysis was focused on a limited number of groups (only 2) with good natural separation and with different population histories that are *a priori* known. For the American sample, they report that "92% of cluster one members were white males, and 80% of cluster two members were black males; 89% of black males were placed into cluster two and 86% of white males were placed into cluster one […]Between-cluster variation was 11% of the total variation"(2009, p. 71). For the Howells data, they found that two major continental differences were detectable; 81% of the Africans were placed in cluster one and 91% of the Europeans were allocated to cluster two. These analyses produced classifications that were better than chance, and indeed very high in their percent-success; moreover, they were comparable in their values to the rates obtained by complementary tests, i.e., most notably those using DFA and stepwise selected variables. These analyses suggest, as the author's explain (2009), that "there is a strong concordance between social race and biological differences" for the American Blacks and Whites and that continental groups "can be separated from each other very effectively using craniometrics" (p. 72), regardless which of their methods were applied. While their tests of multiple methods at varying levels of population structure (within country or continent, between continent) do help to resolve the methodological lacuna that has already been located in Relethford's (2009) recent reassessment of the relationship between racial units and classification accuracy, their tests of subpopulation classification (e.g., Howells' North and South Japanese) instead contribute to Relethford's (2009, p. 20) problem of choosing the proper " geographic cut-offs" for defining races. Ousley, Jantz and Freid (2009,) do acknowledge the "subjectivity" in this: "[i]t would seem that the number of biological

races may be limited only by the number of samples, contradicting the classic view that there are only a few discrete biological races" (p. 72).

In light of growing interest in the importance of considering the "population-at-large" when engaging in ancestry estimation and craniometric classification, there is increasing evidence attesting to the potential weaknesses in studies whose research goals demand the consideration of context but whose research design fails to account for this factor (Konigsberg, Algee-Hewitt & Steadman, 2009). When the goal is individual identification, these concerns over subpopulation discrimination could potentially be resolved for future analyses by the introduction of prior information, which would refocus and refine the analysis and transform the many potentially "correct" divisions into classification results that hold the most useful information on assignments. When the goal is the identification of world-wide population structure, the problem is considerably more complicated as context is no longer a reasonable factor, as the nature of this macro-scale question requires an assessment of the data overall. While Ousley, Jantz and Freid (2009) offer many thoughtful comments on the debate over meaningful, world-wide patterns and they contribute valuable empirical data to the discussion of population partitioning, their approach remains constrained by the need to provide external membership information, either in the form of predefined group affiliations when building their discriminant functions (e.g., a six-way analysis of Howells' European and sub-Saharan African samples, including the Berg, Norse, Zalavar, Dogon, Teita, and Zulu) or the predetermined two-cluster solution (Europeans versus Africans) when generating their K-means solution. Their findings do detect global craniometric structuring at the most fundamental level (and reflecting the most basic geographic, north-south, differences and evolutionary divergences) but, again, they do so while leaving open the possibility for a critical reiteration of the arguments over "forced solutions" and without addressing a truly global sample: they do not extend

their search for structure beyond the simple and obvious two-way, likely clinal, divide. A recent study by Konigsberg, Algee-Hewitt and Steadman (2009) has also taken up the question of cranial-based classification, but for forensic anthropology specifically and from the perspective of evidentiary significance, when obtaining morphometric estimates of sex and race. Konigsberg, Algee-Hewitt and Steadman (2009) applied a finite mixture form of cluster analysis, as implemented by the contributed R package, MCLUST version 3 (Fraley and Raftery 2006), to Howells' series and the Forensic Data Bank's race-based groups and obtained statistically significant results. As two distinct groups were discernable even after removing the effects of size, they concluded that craniometric variation is structured by populations. It is important to note, however, that the focus of this analysis was not to test mixture modeling on craniometric data or to elucidate patterns of world-wide diversity. Konigsberg and coworkers employed mixture analysis for exploratory purposes and to limited ends: the identification of internal data structure by revealing the presence of inherent or "natural" group partitions, in the justification of subsequent, regularized discriminant function analyses using informative priors (demographic data from the U.S. Census 2000). While these authors made use of the more procedurally conservative and, so, statistically robust form of cluster analysis, by choosing a parametric "model-based" clustering method, they conducted the least directed analysis because finite mixture methods do not require the predefinition of the cluster number by the analyst. In avoiding the issue of preselecting the number of components into which the observations will be resolved, the analysis is run based on the adherence to information criteria, or statistical tests and values computed from the data structure itself. Using this approach, their preliminary analysis is free, therefore, of the need for external input which may inevitably reflect, on some level, the observer's subjective ideas about the data's group composition. While Konigsberg, Algee-Hewitt and Steadman (2009) limited the application of the finite mixture method to a single implementation of the procedure, the full-procedure approach has been

informally recommended by Pontikos (2004) for globally mixed data. As one of many contributions to a personal web-blog, Pontikos examined the full Howells data set (28 populations, 57 traits) for the optimal number of clusters and found evidence of fourteen groups with geographic, social, and anthropological meaning. While his study does address biogeographic diversity and race by speaking to the global patterning of craniometric variation and, in revealing the best groupings within the mixture, it has obtained interesting results that support biogeographic races, any confidence in drawing reliable conclusions from Pontikos' analysis is, unfortunately, low on account of several questionable statistical, programming, and philosophic choices made, and for which he offers no explanation or an unsatisfactory justification. Nevertheless, this and the more recent studies, although few, attest to a growing interest in applying clustering algorithms to craniometric data as an alternative to the traditional discriminant function methods. There has yet to be, however, a project that takes full advantage of the benefits offered by robust clustering approaches and critically evaluates their merits, conditions, and limitations for application in quantitative, biological anthropology. There is also a need for a formal analysis of mixture methods that can speak thoroughly to their promise for exploring craniometric data, and, specifically, to the issue of resolving group morphologies as they apply to world-wide population studies and the individual identification context of forensic anthropology.

Detecting Structured Variation: A Mixture-Clustering Approach

"Euclid taught me that without assumptions there is no proof. Therefore, in any argument, examine the assumptions." (Bell in Eves, 1988, p. 165, TT 57)

Throughout its century of subdisciplinary development, biological anthropology has witnessed both a multitude of interpretations offered for the distribution of human phenotypic and, most recently, genetic diversity and an equal number of explanations for how micro-level processes and macro-evolutionary mechanisms have come to shape the patterns of variation that have been observed within and among alternately defined human groups. Despite these many perspectives on the sources, character, and the social versus scientific implications of this diversity, a recent symposium aimed at "reconciling race" and determining how "biological anthropologists view human variation" revealed a remarkable consensus (Edgar & Hunley, 2009, p. 2). While encouraging open communication on these many divergent views of human biological variation and on the data and methods employed by different researchers and anthropological communities in their formulation and justification, the results of this symposium suggest that there is but one "fundamental difference" (Edgar & Hunley 2009, p. 2). The participants disagreed about "the precise nature of the geographic patterning of human biological variation" (Edgar & Hunley, 2009, p. 2) and, so, their opinions diverged along two major theoretical axes:

(1) the distribution of world-wide human biological variation is continuous, i.e., gradated and overlapping

(2) human biological variation is structured, i.e., by geography, in nested hierarchies.

From a review of the publications that followed this meeting (*AJPA* 2009, vol. 139), it can been seen that, while all participants referenced in some way the fact that humans are genetically homogenous, with much of the genetic variation being shared across the world's populations (i.e., 99.99% genetically alike), differences in opinions emerged when this concensus was linked to the problem of population subdivision and classification. Some contributors supported the predominant views that have emerged from the seminal work of Livingstone and Lewontin. They offered the conventional, corresponding arguments that human biological variation is clinally distributed, that the amount of neutral genetic

variation detectable among groups is low, and that both neutral genetic and phenotypic variation conforms to a pattern of isolation by distance. From this perspective, they reiterated how the traditional conclusions on variation can be logically drawn: if there are no geographic discontinuities, there can be no meaningful division of the human species into groups; so that, biologically speaking, races, or any of its biosocial-cultural proxies, do not exist. Other participants chose, however, to highlight the recent developments in molecular biology and population genetics that have called into question the theoretical and empirical foundations that have heretofore upheld the predominant, conservative, and for many, the only, view on human variation. These participants advocated instead that world-wide diversity is structured according to geography and that the population-specific patterns that represent this organization are interpretable. These patterns are the result of a complex sequence of events that reflects both unique group histories as well as the collective human experience (e.g., mate-exchange, migration, population splits, bottlenecks and founder effects) and, so, are informative of the development of both populations and persons. With this "new" position comes a critical implication, as Jorde and Wooding (2004) have recently explained: "[g]enetic variation is geographically structured, as expected from the partial isolation of human populations during much of their history. Because traditional concepts of race are in turn correlated with geography, it is inaccurate, as some have argued, to state that race is 'biologically meaningless'" (p. S32).

It is through this latter view that biological significance can be awarded to the morphological differences observed among groups, and these "idiosyncrasies" can be taken as signatures of population history that are retained, organized, and informative or meaningful when detected (Marks 2008, p. 29). It is the extension of this theoretical premise to skeletal data that gives credence to craniometric-based classifications and the quantitative estimation of ancestry in forensic anthropology. As Bamshad and

colleagues (2004) have stated, "[s]o, although there might be little variation among groups, it is highly structured and therefore useful for distinguishing groups and allocating individuals into groups" (p. 604). The assumption that craniometric diversity is meaningfully geographically structured and that this structure enables classification underlies the methodological workings of group discrimination procedures, as in the FORDISC software. Although both explanatory discussions and critical tests of the common identification approaches emphasize the "unknown" case as the unit of analysis of singular importance (the target cranium for the discriminant function analysis and the focus of the classification), the estimation of individual ancestry is inextricably tied to an understanding of the factors that define human groups and the distribution of craniometric variation among them. Determining an individual's membership requires the program, procedure or practitioner to decide if the individual in question inherited the observed and metrically quantified craniofacial morphology from a particular group or combination of groups. On its most basic level, this cannot be accomplished until or unless morphologically distinctive groups are reliably isolated and these partitions are accepted as statistically significant and biologically relevant. A profitable preliminary stage in any discussion of craniometric classification (whether it be made for or against estimation) would be, therefore, the unbiased demonstration that human craniometric variation is or is not highly structured. As the prior theoretical substantiation of this structuring is essential to the practitioner's interpretation of human identifications as "successful" and revealing of "accurate" information on personal identity (e.g., as biosocial race or population affinity) that reflects biological reality and not user-ideology, it can be argued that a more appropriate, and overlooked, first step in the evaluation of craniometric classification would be to conduct a test of its conceptual foundations. The fundamental question that has yet to be investigated is *if human craniometric variation is structured along these biogeographic lines, is the strength of this structure sufficient to permit its scientific detection with some measure of statistical certainty?*

To respond appropriately, this analysis would require a different methodological approach than those employed by the predominant tests. The question of the existence of meaningful, quantifiable and scientifically detectable population divisions would benefit greatly by the redefinition of the problem of group discrimination as one akin to latent class identification, where all identifiers are truly unknown or intentionally withheld. To test for world-wide data "structure" or "stratification" in a data set of "mixed" population craniometrics, specifically, there is promise in the application of finite mixture analysis or model-based clustering. This type of "*a priori*-free" analysis can be best achieved by applying an unsupervised clustering approach that is grounded in the statistically robust, algorithmically stable and parametric clustering procedures of finite mixture analysis, which is both readily testable and mathematically satisfying on account of its probabilistic, "model-based" approach. Using this form of mixture-clustering, any given data set can be evaluated for the strength of its internal structure by teasing out latent classes, so that any hidden groups found are identified in the absence of external expectations for the clusters that the data set may contain and the individual allocations that may result. In a practical craniometric application, finite mixture analysis can be used to determine if, in the absence of *a priori* group-identifiers (e.g., individual race, ethnicity, ancestry, geography details), it is possible to allocate these now "unknown" individuals with high probability into clusters in which all members share more common ancestry or biological affinity than individuals assigned to other clusters. Moreover, this analysis can be designed to operate at the two levels of resolution which have been, thus far, the focus of major theoretical concern: (1) the level of global populations, by exploring among group variation as biogeographically organized clusters within a state of panmixia, and (2) at the level the local population, by testing for the presence of subpopulation structure, expressed as ethnic or biosocially defined clusters within the local group and formed by cultural, geographic, and/or temporal barriers. By

expanding the research question to include variation within a single population, the data necessary to explore the inconsistencies in among-group estimates already identified by Lewontin (1974) can be accessed:

> [a]lthough there is variation between loci in their relative contributions, the average values show that 85 percent of human genetic diversity is *within* national populations and only 7.5 percent between nations within races and 7.5 percent between major races. The relative division between the last two categories depends in part on how national groups are assigned to races and how many "races"are constructed. Are Hindi-speaking peoples of India Caucasoids or a separate group? Do they belong together with Vedic speakers? Where do the Turks or the Lapps belong? There are many possible ways of dividing groups, depending upon linguistic, morphological, or genetic criteria (p. 155).

It will then be possible to better respond to arguments made on the arbitrary or even subjective nature of the many "correct" cluster solutions, which Cavalli-Sforza, Menozzi and Piazza (1994, p. 19) have already established: "[h]uman races are still extremely unstable entities in the bands of modern taxonomists…As one goes down the scale of the taxonomic hierarchy toward the lower and lower partitions, the boundaries between clusters become even less clear…There is great genetic variation in all populations, even in small ones" (for craniometrics, see, most recently, Relethford, 2009 or Ousley, Jantz & Freid, 2009). By ascertaining if there is meaningful (statistically and biohistorically interpretable) structure at clearly definable, finite strata (e.g., continent, country, nation, state or village), it can be determined if "binning" by biological trait(s) is indeed like "slicing soup," so that "you can cut wherever you want, but the soup stays mixed" (Nature Biotechnology Editorial, 2002, p. 637).

Collectively, mixture-clustering approaches, regardless of distribution or model parameterization, promise robust parametric and probabilistic identification procedures for detecting the morphogenetic boundaries between populations that are not complicated by the problem of predefining (for model creation) and selecting (for the target analyses and individual classifications) suitable reference groups that adequately reflect global craniometric diversity or whose contributions represent the demographic conditions of the population-at-large from which a particular tests case is believed to have been drawn. Model-based clustering methods have several well-noted advantages over supervised clustering and classification techniques, three of which are especially relevant to this discussion: (1) sample data need not be entered in predefined groups as these methods are not dependent upon a comparative training data set and the "correct" prior knowledge of group memberships, which are fundamental aspects of discriminant function analysis, (2) the actual numbers of phenotypic clusters need not be known or specified before the algorithm is executed, as is the case for K-means clustering (for which the number of *k* groups must be predetermined in order to partition the N-dimensional population into *k* sets) (MacQueen, 1967; Hartigan & Wong, 1979; McLachlan, 1992; McLachlan & Peel, 2000), and (3) mixture methods have been shown to work well for small and large sample sizes with few and many variables, accommodate or reliably estimate missing values, and to easily adapt to the incorporation of extension algorithms, including robust variable selection routines (McLachlan & Krishnan, 1997; McLachlan & Peel, 2000; McLachlan, Do & Ambroise, 2004; Raftery & Dean, 2006; Dean & Raftery, 2010). In the very fundamentals of their design, therefore, mixture analyses are free from the common criticisms levied against many of the other competing approaches: that clustering and classification techniques are unreliably exploratory, often requiring extensive user-guidance, are known to defy rigorous validation, and are only confirmative of suspected or previously determined structure. Gower (1972), for example, has spoken to their use in "anthropological classifications" specifically, for situations in which "the

populations are fairly close together and have to be distinguished by varying characters whose values overlap the populations"(p. 20). Moreover, he has further written that "[i]n this situation I am rather dubious of the efficacy of most of the methods discussed in this section, unless there is an enormous amount of data available for groups represented by populations means are to be clustered" (p. 20). Mixture methods are equally unaffected by the concerns engendered by the user or program induced biases of DFA and FORDISC, which may be introduced by the internal limitations of the software, individual choices in reference sample or research design, and by *a priori* knowledge of, for example, population, racial, or ethnic affiliations, individual genetic origins, sampling locations, geographic distributions, phenotypic information, and surrogates or racial proxies. When comparing classification accuracy between finite mixture and discriminant function analysis (either LDA or QDA) under identical, controlled conditions, mixture modeling has been shown to produce equally good if not better results, especially when group sizes are equal but the expected effect size is low (Holden & Kelly 2010). Moreover, when misclassifications are nonrandom and group sizes are discrepant, the direction of its misclassification is towards the smaller group and so its bias is unlike the pattern already well-documented for DFA (Breckenridge, 2000; Lei & Koehly, 2003; Holden & Kelly, 2010). These findings have considerable implications for the practical use of mixture-modeling over a discriminant function approach in the contexts of diagnostic identification, where it is more desirable to err on the side of misclassifying a case within the smaller (or "minority") group than to overlook individuals with subtle, but potentially biosocially significant, features and misplace these cases within the larger more "typical" group, for which individual probabilities of membership may be less informative or corroborative of a personal identification (c.f., Hallahan et al., 2007).

Mixture-modeling has been recently employed within a similar framework to great success in innovative molecular analyses. As a result, these studies have challenged the traditional understanding of among-group variation and have revealed the promise of classification using biosocial and biogeographic parameters. Most remarkably, they have demonstrated that genetic differentiation does exist along racial lines, when defined in terms of major geographic or continental regions (c.f., Stephens et al., 2001; Rosenberg et al., 2002; Risch et al., 2002; Bamshad et al., 2003; Tang et al., 2005). This new genomic work has been coupled with recent critical reevaluations of the hypothesis that supported many of the earlier, traditional views on the distribution of human biological variation and, together, these findings have now generated an explosion of research on genetic ancestry and a renewed interest in the importance of race to the biological sciences. Most importantly, they have has called into question the assumption that "the concept of race has no genetic or scientific basis" (Venter, 2000, np). In drawing upon these new molecular findings and through emerging data analysis tools like the intensive, iterative model-base clustering methods, there is an opportunity for the skeletal biologist to also engage in this renewed ancestry-race debate. Until the relatively recent growth in anthropological genetics, access to population relationships at the macrogeographic level and the identification of biosocial affinity or the estimation of ancestry for the individual was made possible by the morphoscopic and metric assessment of the craniofacial morphology. Populations were sampled and studied in order to (1) identify features that were expressed in select predefined groups with high frequency and which appeared in low frequency in all other groups, (2) establish the distribution of these quantifiable traits (commonly using homologous landmarks and their distances or non-metric scored features) for each available group and to (3) isolate which of the observed traits overlapped only minimally with other groups. The concept of human craniofacial variation that emerged from these seminal studies depends upon the idea of a unique, consistent and reliably identifiable craniofacial gestalt that exists as some biological reality for

each predefined population (e.g., continental, ancestral, or subgroups formed from admixture events). Moreover, it is built upon a theoretical interpretation of a gene-to-phenotype connection, which has seen very limited testing.

Using the criteria above, the development of protocol for interpreting human skeletal variation has followed a logical trajectory: patterned features or measurements observed in each sample aggregate were codified as group-specific morphologies and phenotypic trait analyses were operationalized under the assumption that the successful clustering or partitioning of individuals based upon shared or dissimilar craniofacial measures reflected the actual underlying pattern of genetic relatedness. The acceptance of this approach allows for the easily accessible and quantifiable morphotype to be taken as a convenient proxy for the genotype, so that complex links or lacunae between groups can be discovered. If the well-developed arguments within the larger biological community on the relationship between genes and morphology in a wide variety of species are followed and the new information on race and identity that is rapidly developing out of the increasing knowledge on the human genome is accepted, it becomes both feasible and highly valuable to test the skeletal relationships between human groups using methodological approaches similar to the model-based clustering methods that are achieving very successful outcomes for genetic studies. While the ability of finite mixture analysis to detect the morphological boundaries between and within populations remains to be fully explored in craniometric studies, extensive testing of this alternative model-based, *a priori*-free method may help to resolve the tension between those biological anthropologists who use craniometric variation as a means to access past population structure in phenotypic distance studies and to estimate individual ancestry in forensic contexts, and those who dismiss any discussions of group differentiation as artifacts of personal ideological positions. In both the explanation of this approach and the justification for its use with

craniometric data, it is important to address the changing trends in genetic theory, the emerging "mixture" based research, the meaning of ancestry and the contemporary redefinition of race, as well as their impact upon the anthropological conception of human skeletal variation and the theory that underlies the protocol for forensic identification.

Statistical Fallacies: Revisting Classification

"...the best causes tend to attract to their support the worst arguments, which seems to be equally true in the intellectual and in the moral sense."(Fisher, 1956, p. 31)

Since 1972, there has been much written about Lewontin's seminal study on the apportionment of human biological diversity. The credence that his findings have given to arguments made on the non-existence of races is well known and extensively penned; yet, the consequences that they brought upon contemporary investigations into race, are, perhaps, still best articulated in the opinions given by Lewontin himself. At the genesis this debate, he wrote: "I shall try, in this article, to display Professor Jensen's argument, to show how the structure of his argument is designed to make his point and to reveal what appear to be deeply embedded assumptions derived from a particular world view, leading him to erroneous conclusions" (1970, p. 2). It is interesting to note that shortly after Lewontin issued this invective against Jensen for his racialized science (albeit, rightly in this case, as it pertained to hierarchies of intelligence), Lewontin also began to deploy his quantitative, population genetic knowledge as irrefutable, empirical evidence in support of his own ideological position on race. He

engaged the question of human categorization alongside the proponents of race with equal, but oppositely, biased rhetoric: "[i]ndeed the whole history of the problem of genetic variation is a vivid illustration of the role that deeply embedded ideological assumptions play in determining scientific 'truth' and the direction of scientific inquiry" (1974, p. 157). In Edwards' opinion, Lewontin deliberately "used his analysis of variation to mount an unjustified assault on classification, which he deplored for social reasons" (2003, p.801). It is noteworthy how quickly his findings factored into his remarks on classification and ran through the overriding themes in his 1974 book, *The Genetic Basis of Evolutionary Change*: "[t]he taxonomic division of the human species into races places a completely disproportionate emphasis on a very small fraction of the total of human diversity. That scientists as well as nonscientists nevertheless continue to emphasize these genetically minor differences and find new 'scientific' justifications for doing so is an indication of the power of socioeconomically based ideology over the supposed objectivity of knowledge" (174, p. 156). It is, therefore, impossible to overstate the role that the interpretations given to his findings and their short but influential history in the anti-race debate have played in extending the discussion on the distribution of human variation to the issue of classification, among many scholars in many disciplines and subfields, and with respect to many research foci (c.f., Cooper, Kaufman & Ward, 2003; Haga & Venter, 2003; Rotimi, 2007). Cavalli-Sforza, despite his common identification as the "father" of "genetic geography"(Leslie 1999, np), agreed from very early on that "[r]aces are, in fact, generally very far from pure and, as a result, any classification of races is arbitrary, imperfect, and difficult" (Bodmer & Cavalli-Sforza, 1976, p. 559). Even throughout his work on the Human Genome Diversity Project (HGDP), for which he expressed concerns over the possibility that it might be "feeding 'scientific racism'" (2005, p.333), he has, nevertheless, held fast to this seemingly contradictory position. Only recently, he wrote that "[f]rom an ethical point of view, studies of human population genetics and evolution have generated the strongest proof that there is no

scientific basis for racism, with the demonstration that human genetic diversity between populations is small, and perhaps entirely the result of climatic adaptation and random drift" (2005, p. 340). For "mainstream" physical anthropology, similar anti-race views persist, dominating its core concepts, both in the literature and in their basic instruction (Jurmain et al., 2005; Boyd & Silk, 2009; Fuentes, 2007). Brown and Armelagos (2001, p. 35) have upheld Lewontin's position and continue to advocate for the rejection of race even now, despite the new era of population-based research in molecular anthropology. They continue to argue, as he did, that "racial classification is [...] of virtually no genetic or taxonomic significance" (1972, p. 397). Many of these statements against race, population differences, and classification have drawn additional weight from the fact that subsequent studies of neutral genetic variation have successfully replicated Lewontin's results and, in doing so, have reinforced the validity of the consistently "low" estimates (between 7% and 10%) of among-group or inter-racial variation (c.f., Nei & Roychoudhury, 1982; Barbujani et al., 1997; Jorde et al., 2000; Rosenberg et al., 2002). What has typically gone unacknowledged, when invoking these average values and citing these corroborative analyses as empirical evidence against intergroup biological differences, is that this proportion of variation is, in fact, statistically significant. Moreover, when consideration is given to scale effects, even Wright (1978) himself believed that an F_{ST} value of 0.15 corresponds to a moderate to great level of differentiation (Long & Kittles, 2003, Hunley, Healy & Long 2009). The universal cut-off at or around 6.3% for inter-group taxonomic significance appears to be have been arbitrarily defined and uncritically accepted.

It has become, then, a matter of scientific consensus that, because this among-group genetic variation is many times less than the variation observed within human groups, it is simply too "low" or "small" to be meaningful. The importance of Lewontin's estimates of variation lies for many, therefore, in their

demonstrative effect and not in the scientific value of their statistical significance: "popular articles" make use of these now standard values to "play down the genetical differences among human populations" and they are "often stated, usually without any reference" (Edwards, 2003, p. 798). These mathematical facts alone do not adequately support an argument for the "superficial nature of biology at the classificatory level" (Blackith & Reyment, 1971, p. 5) nor do they justify, in turn, the preclusion of human classification by population. This "significant" oversight can be seen as but one aspect of a much deeper rooted problem related to the use of Wright's Fixation Index, F_{ST}. Recent reevaluations of the F_{ST} statistic have revealed that, while it is of value to matters of evolutionary history, it is a weak contributor to questions regarding the existence, distribution or number of human groups (Jorde et al., 2000). Mountain and Risch (2004, p. S51) have noted that "a low FST estimate implies little about the degree to which genes contribute to between-group differences." Findings such as this suggest, therefore, that "F_{ST} should not be confused with classification" (Jorde & Wooding, 2004, p. 30), despite its ubiquitous presence in discussions of human morphogenetic variation and its history of use in arguments against race, population [sub]structure, and biological affinity. While this problem of use and interpretation can be traced back to the founding arguments against "human taxonomy" that were built upon Lewontin's results, its thorough mathematical explanation has only recently been widely disseminated (Edwards, 2003). The ability of many subsequent studies to reproduce with similar accuracy Lewontin's estimates (with only subtle variation occurring on account of the type of markers used) attests that there is no mistake in Lewontin's statistical analysis; as Ramachandran et al. (2010, p. 596) have asserted, "it is an undeniable mathematic fact that the amount of genetic variation observed within groups is much larger than the differences among groups." These authors have gone on to say, however, that "this [fact] does not means that genetic data do not contain discernible information regarding genetic ancestry" (2010, p. 596). They argue instead that, "minute differences in allele

frequencies across loci when compounded across the whole of the genome actually contain a great deal of information regarding ancestry" (2010, p. 596). "Lewontin's fallacy," as coined by Edwards (2003), lies, therefore, in the assumption that, because the genetic differences between groups (for a random sampling of a limited number of loci) are believed to be swamped by the much greater within-population differences, Lewontin's low estimate of racial variation is relevant to classification. What Lewontin failed to consider, however, is the aggregated effect of among-group differences in allele frequencies for a range of loci; the impact of this univariate perspective on how he viewed and evaluated the biology of race is evident in the question that he posed already 15 years after his original analysis: "How much difference in the frequencies of A, B, AB, and O blood groups does one require before deciding that it is large enough to declare two local populations are in separate 'races'?" (1987, p. 200). The problem with Lewontin's approach is, as Edwards (2003, p. 799-800) explains, the fact that "each additional locus contributes equally to the within population and between-population sums of squares, whose proportions therefore remain unchanged but, at the same time, it contributes information about classification which is cumulative over loci because their gene frequencies are correlated." An easily interpretable analogy using biased coin tosses demonstrates both the validity of Edward's criticism and its straightforward testability (Sesardic 2010). Begin with the assumption that there are two coins labeled H and T, for heads and tails respectively; subsequently, introduce the prior information that coin H is biased towards heads, $p(H)=0.6$, and coin T is biased towards tails, $p(T)=0.4$. If one of these two coins is randomly drawn to toss, the probability of determining which of the coins was selected, based upon the outcome of the flip, is little better than chance (60% success rate). Working only from the results of this one toss, therefore, the resulting information value is very low. If, however, the number of tosses is increased using this preselected coin and the prior probability of heads or tails is held constant, there will be a proportional improvement in the observer's predictive ability. Using

binomial probability theory and Bayes theorem, it can be determined that after 11 tosses the predictive success rate would increase to 75%, after 101 tosses to 97.9%, and after 1001 tosses, any acceptable level of statistic certainty would be met and exceeded, with misidentifications of the correct coin occurring in only 1 out of 10 billion trials. From this simple example, it can be seen how, while the outcome of any single toss is poorly predictable, the findings obtained using joint probabilities can yield very reliable classification or identification results. This test can be extended to genetic frequencies by replacing the coin tosses with biallelic loc: in doing so, it becomes apparent conceptually that, as Edwards has shown mathematically, "the 'taxonomic significance' of genetic data in fact often arises from correlations amongst the different loci, for it is these that may contain the information which enables a stable classification to be uncovered" (2003, p. 799). The probability of misclassification decreases, therefore, as the number of gene loci increases, and the proportion of within-group variation remains consistent (at approximately 84%), while, the probability of misclassification becomes more and more trivial. Classification at the levels of [sub]populations can be supported, therefore, without the detection of a large average amount of variation among these groups, on a locus-by-locus analysis. In sum, this is so because: (1) F_{ST} considers each locus in isolation, (2) it disregards the correlation among loci, (3) the number of correlations increases with the inclusion of more and more loci, and (4) the more correlations present in the data set, the more information which is relevant to classification is overlooked by F_{ST}.

Given the clarity with which this theoretical shortcoming can be identified and explained, it should not be surprising to learn that this is in fact an "old statistical fallacy" (Edwards 2003, p. 799); what is surprising is the little attention that it has received in the relatively recent history of scholarship on genes and race. This problem was well known to both Pearson (1926) and Fisher (1925) in the very early

years of mathematical population genetics (along with the founding contributions of Wright, Haldane, and Malécot) and it even held a place in the literature at the time of Lewontin's seminal publication. Gower, for example, remarked, in 1972, that, when assessing population affinities, "the human mind distinguishes between different groups *because* there are correlated characters within the postulated groups" (1972, p. 3). Soon after, Mitton (1977, 1978) demonstrated that when a measure of genetic distance, which incorporates information from several loci simultaneously, is applied to blood group data, "substantial" differences between races can be identified. Chakraborty (1978), Lewontin (1978), and Powell and Taylor (1978) did indeed comment on Mitton's findings and his use of multilocus genotypic identity; however, they were quick to point out "the fallacy of Mitton's statistical measure" but did not address the fact that Mitton's focus was on human classification alone and not on the calculation of average values of genetic distances (Nei 1981, p. 88). It is interesting to speculate that if Mitton's work had not been so thoroughly dismissed or if Edwards (2003, p. 801) had identified this fallacy when his "article could, and perhaps should, have been written, soon after 1974," scholars would not have been so slow to revisit race and reevaluate classification's role as a valid scientific tool. Nevertheless, the identification of this problem has contributed to the renewed debate over biological race. It has appeared in recent discussions of craniometrics, ancestry estimation and legal evidence in physical anthropology (c.f., Konigsberg, Algee-Hewitt & Steadman [2009], who provide detailed mathematical proofs), while, more generally, it has inspired definitive statements on the potential for yielding meaningful classifications through clustering techniques and the use of race to positive ends for forensic and biomedical sciences in the new genomic age. Bamshad and coworkers (2004), for example, have written recently that "although there might be little variation among groups, it is highly structured and therefore useful for distinguishing groups and allocating individuals into groups" (p.604). Witherspoon et al. (2007) have equally noted that "modest differences between populations can

nonetheless enable accurate classification" (p. 352). Ramachandran and colleagues (2010) have even argued that "given currently technology...it is feasible to accurately identify individuals from populations that differ by as little as 1% in F_{ST} if enough markers are genotyped" (p. 596).

While these reevaluations of Lewontin's findings have encouraged the disassociation of F_{ST} from classification, they have not addressed the potential problems with the measure itself. Only recently, Long and Kittles (2003) have made a contribution to the broader discussion of F_{ST} that is worthy of particular note. While a full account of the differences in the patterns of genetic variation that exist between human and other species has not yet been given, available genomic data suggests that there are aspects of a shared biological variation that are distinctively human. Despite the large global distribution of the species, humans are known overall to be less genetically diverse than other mammals (Li & Sadler, 1991; Kaessman et al., 2001). This is also true when human data is compared to its closest primate relative, with whom human's share considerable biological relatedness: chimpanzee species constrained to central and western Africa alone have, for example, higher levels of diversity than humans (Ebersberger et al., 2002; Yu et al., 2003; Fisher et al., 2004; The Chimpanzee Sequencing and Analysis Consortium, 2005; Marques-Bonet, Ryder & Eichler, 2009). The distribution of genetic variants among and within human populations have also been shown to differ from other mammalian species, for which greater distances have been detected between groups, even though the evolutionary history of humanity has altered allele frequencies and has left genetic signals in certain (geographic) populations (Templeton, 1998; Jorde, Bamshad & Rogers 1998; Jorde et al., 2000; Nordborg & Travare, 2002; Bamshad & Wooding, 2003; Kittles & Weiss 2003). As F_{ST} has played an important role in formulating many of the predominant opinions about human diversity and population relationships, any concerns over its reliability (in the consistency of the estimates produced, its utility under particular

conditions, and its appropriate application) will have significant implications for (1) how anthropologists have come to understand the patterns of variation within and between human groups, (2) the ability to speak confidently about the always already obscure concepts of race, ancestry and any other partitions by biological affinity based upon the statistical theory behind F_{ST}, and (3) an easy acceptance of the empirical evidence that was gained from F_{ST} as, now, a classic population genetic measure. It is for this reason that it is necessary to address seriously the critical shortcomings that Long and Kittles (2003) have identified in the use of F_{ST} and the quite radical conclusion that they have drawn: "[t]here is little meaning to simple partitions of human genetic variation on a worldwide scale, and the broad acceptance of *FST* as a valid measure has prevented a deeper understanding of human variation" (p. 450).

Wright's population structure statistic, F_{ST}, is well known for its explicit definition in the context of a population model, for which each subgroup is taken to be an independent replicate of the evolutionary process (Wright 1951, 1969; Cockerham, 1969; Weir & Cockerham, 1984; Weir, 1996). When F_{ST} has been used to measure the variation among samples of world-wide human populations, a value of 15% (or less) has been consistently obtained for both genetic data and, select, phenotypic traits. Findings of standard F_{ST} estimates using craniometrics (10%-15%) have greatly influenced how the global distribution of cranial morphology is viewed by contributing support to theories of neutrality and isolation by distance models (Relethford, 1994, 2002). The recent work by Long and Kittles (2003) has, however, convincingly shown that this typically "low" estimate ($0.10 \geq F_{ST} \leq 0.15$) is not as statistically robust or as biologically informative as commonly believed. They demonstrate that this, now, conventional and "ubiquitous" estimate, around which ideas of race and classification have been shaped, is the product of a statistical artifact. They argue, in turn, that it not only poorly represents

biological reality, but that it also reflects significant biases. Accordingly, Long and Kittles (2003, p. 450) contend that "estimates of *FST* will fail dramatically to identify important differentiation among groups," because of casual violations of the key but, often hidden, assumptions that underlie the calculation of F_{ST} and the interpretation of its values. Estimates of variation, regardless of size, are meaningful only if it can be proven that: (1) there is evolutionary independence between populations existing within the same geographic region, (2) there is evolutionary independence among groups of populations existing in different geographic regions, and that (3) the levels of variation are equivalent in all local populations and in all geographic regions. Long and Kittles argue convincingly that human groups do not meet this criteria. Furthermore, they demonstrate that any estimate of F_{ST} depends heavily upon allele frequencies so that, in actual practice, its estimates cannot vary freely from 0.0 to 1.0 as was otherwise presupposed. Long and Kittles reveal, perhaps most dramatically, the implications of these problems to questions of magnitude in population differences by conducting a standard analysis using a large data set of dinucleotide repeat allele frequencies at eight loci estimated for eight local human populations drawn from five continents and a sample of chimpanzees (c.f., Deka et al. 1995). When they considered only the human populations, the estimated value returned was in keeping with the general expectations for analyses of among-group variation, i.e., $F_{ST} = 0.119$. When, however, the chimpanzee sample was introduced into their test, they obtained remarkable results: the inclusion of the non-human outgroup *only slightly* elevated the previously estimated value, from 0.119 to 0.183. Moreover, this second estimate neither differed significantly from the result obtained for the human data alone nor from the standard interpopulation value of around 0.15. From a statistical point of view, these findings attest to the constraint imposed upon the estimate's true ability to assume any value within the 0.0 to 1.0 range. In the larger context of interpretability, the impact of these results is considerable and less easily determined. Given the sizeable difference in the levels of diversity already known to exist between

human and chimpanzees and, yet, the insignificant change in F_{ST} values between the tests of the data with chimpanzees excluded and included, it can be imagined that, when used uncritically, F_{ST} can both undermine analyses of group distinctiveness and mislead investigations that seek to identify meaningful patterns within the global distribution of human variation. Put simply, if F_{ST} cannot expose the inter-generic variation between *Pan* and *Homo*, what useful information could it possibly reveal about the relatively subtle, intra-specific differences among human populations or [bio-]social/geographic races, for whom their defining qualities are highly nuanced and their histories of interaction and evolution highly complex? To this point, Long and Kittles (2003) have argued that "the belief that human genetic diversity on a global scale can be reduced to simple statistical partitions has limited our understanding of diversity and thwarted training in biological anthropology" (p. 466). The simple explanations too commonly given to this difficult problem need only be considered. Take for example the rather extreme account of "race" given in the popular textbook by Boyd and Silk (2009), which is now in its 5th edition:

> [g]eneticists computed the amount of variation in these characters within each local group, among groups within each race, and among the races. They found that there is much more genetic variation *within* local groups than there is *among* local groups or among races themselves. Differences within local groups account for about 85% of all the variation in the human species. To put this another way, suppose a malevolent extraterrestrial wiped out the entire human species except for one local group, which it preserved in an extraterrestrial zoo. The alien could pick any local group at random—the Efe, the Inuit, the citizens of Ames, Iowa, or the people of Patagonia—and then wipe out the rest of the humans on the planet. This group would still contain on average 85% of the genetic variation that exists in the entire human species (p.143).

Long and Kittles (2003) argue that, in fact, this would not necessarily be the case because the outcome of any "eradication" would differ depending on which population is maintained. For example, they show that if the Sokoto group is retained *no* gene diversity would be eliminated (F_{ST} = 0.000), but, if the Sokoto are replaced with Papua New Guineans, *one third* of the world-wide genetic variation would be lost. If genotype is indeed equivalent to phenotype as previous skeletal studies have assumed, the enormous effect that this reevaluation of F_{ST} should have on the biological and anthropological understanding of race becomes apparent. Long and Kittles' argument, first, calls into question the way in which meaning is used and assigned to common and even *the* canonical measure(s) of biological diversity; they aid, therefore, on a theoretical level, in identifying potential weaknesses in the methods and statistical bases for the quantification of variation in anthropological genetics. Perhaps more tangibly, at the level of current trends in practice, their argument helps to resolve the apparent inconsistencies in results reported recently for large scale analyses and genome wide association studies. Long and Kittles' compelling demonstration of the constraints placed upon the actual estimatable values of F_{ST} may clarify, for example, the findings that global F_{ST} estimates are lower for microsatellite polymorphisms than for biallelic Alu insertion polymorphisms (Bamshad et al., 2003) and explain the negative correlation between F_{ST} and microsatellite heterozygosity (Bowcock et al., 1994). In terms of the question of race and classification, their argument can also be employed in the service of a better understanding of how recent world-wide studies interested in group-specific diversity can obtain high values for within-population variation (even accounting for as much as 93%-95% of the total genetic variation), when standard F_{ST} analyses are conducted, while simultaneously revealing highly structured inter-group variation that has enabled the geographic clustering of these same samples into preexisting biosocial populations, i.e., races, ethnicities, self-identified groups (c.f., Rosenberg et al., 2002). Although they focus on the problems of F_{ST}, they put forth an argument that complements those

made regarding "Lewontin's fallacy." Long and Kittles demonstrate in a different way that the conventional measures of inter and intrapopulation variation are incompatible with classification approaches. F_{ST} cannot contribute to the identification of population structure for classification because the information that it provides is both dubious in its significance and, most simply, irrelevant to the questions that classification-based studies seek to answer. In this context, Long and Kittles' (2003) analysis provides convincing support for the renewed interest in *scientifically* distinguishing groups and allocating individuals into these groups using biological traits; and, so, their discussion also serves as notice for the need to reevaluate craniofacial classification in a way that current investigations into the patterned (cluster) distribution of morphometric variation are divorced from tests of inter-population diversity using F_{ST}.

These critical reevaluations of Lewontin's findings and the use of F_{ST} as the "classic" measure of population variation are not alone in their identification of problems in mainstream scholarship: other conceptual issues have recently called into question the validity of the predominant anti-race views and are undermining the foundations upon which the broader arguments against human biological classification rest. While traditional discussions of the apportionment of morphogenetic diversity have maintained an important role in shaping the common perspective on variation, the trajectory of current research reveals an equally strong interest in accounting for the geographic patterning of neutral genetic variation (Handley et al., 2007; Hunley & Edgar, 2009). Although mitochondrial DNA and nuclear loci have been studied extensively using many different types of markers and the results reported have varied to some degree with the specifics of each analysis (Tishkoff & Verrelli, 2003), these tests have revealed two consistent trends in findings using modern molecular data: (1) greater genetic diversity and less linkage disequilibrium occurs in African populations than in non-African populations (Tishkoff &

Williams, 2002; Kittles & Weiss, 2003), and (2) allele frequencies undergo gradual change over geographic space without "sharp" geographic discontinuities in traits (Barbujani, 2005). The first trend has obvious importance for coming to an understanding of the evolutionary and demographic history of the human species through genetic approaches (c.f., Cavalli-Sforza & Feldman, 2003; Garrigan & Hammer, 2006; DeGiorgio, Jakobsson & Rosenberg, 2009). Molecular evidence suggests, for many, that anatomically modern humans (AMH) originated in some single region of northeast Africa (ca. 100-200 kya) and that this emergent group expanded within Africa and subsequently out of Africa into Eurasia (ca. 100 kya), and then Australia (by 45 mya) before further dispersals into the New World (ca. 20-30 mya) and a final migration (less than 10 mya) into more remote locations with low accessibility, e.g., the Pacific islands (Jobling, Hurles & Tyler-Smith, 2004; Weiss & Long, 2009). If this progenitor population evolved in Africa before undergoing a global dispersion, the associated temporal depth would have allowed for more genetic differences to accumulate in African groups through mutation over time (Tishkoff et al., 2000; Stringer, 2002; Tishkoff & Gonder, 2006). This greater age could also account for the lower levels of linkage disequilibrium observed among Africans, as recombination leads to its decrease over time (Pritchard & Przeworski, 2001; Kittles & Weiss, 2003). Furthermore, differences between African and non-African populations could be attributed to a possible genetic bottleneck that preceded the major wave of migration, which suggests that only a subset of the variation present in the original, ancestral group was carried out of Africa by a small number of migrants at the time of this dispersal; explaining how non-African populations are found to be less genetically diverse and to exhibit higher levels of linkage disequilibrium (Bamshad et al., 2004; Tishkoff & Kidd, 2004). These in and out-of-Africa patterns have led some to posit, therefore, that non-Africans (the dispersing or colonizing populations) can be interpreted genetically as a subset of Africans (the ancestral group) and, on account of this perceived stratification, to refute, in turn, any explanations of human variation in terms of

identifiable, equal but different, biological units (Quitana-Murci et al., 1999; Underhill et al., 2000; Brown & Armelagos, 2001; Kidd et al., 2004; Barbujani, 2005).

When the conversation is directed towards contemporary human variation, it is the second trend that has received the most attention in debates over the geographic patterning of modern human diversity and the appropriateness of population partitioning using biological traits. The majority of these studies have interpreted the positive correlation between genetic and geographic distances in the spirit of Livingstone's (1962) early but influential theories of character discordance and clinal variation: (1) change in allele frequencies are gradated over geography, with overlap, but no concordance, between trait distributions, (2) populations are closest biologically to their neighbors, and (3) genetic similarity is inversely correlated with geographic distance (Serre &Pääbo, 2004; Prugnolle, Manica & Balloux, 2005; Manica, Prugnolle & Balloux, 2004). These observations have been typically explained using models of isolation by distance (IBD) and localized gene flow (Cavalli-Sforza, Menozzi, & Piazza 1994; Relethford, 2004), where spatial restrictions on mating behaviors and, in turn, on gene exchange produce a pattern of decreased genetic similarity with increased geographic distance (Wright, 1943; Malécot, 1969). The implications of this framework for population stratification should be immediately obvious: Seer and Pääbo (2004) analyzed data from individuals who were evenly sampled across geographic space and concluded that the pattern of global genetic variation is best described by clines and not by clusters of races; Relethford (2002)applied these arguments to the global distribution of cranial morphology, writing that "the apportionment of global genetic diversity estimated from craniometric data mirrors the pattern observed in several studies of genetic markers and DNA polymorphisms in showing that the majority of diversity exists within local populations [...] the results from both genetic and craniometric analyses are consistent with neutral traits under an isolation by distance model" (p. 396-7). Despite the

support of the now long scholarly history on clines and the appeal of the IBD model for its simplicity and tractability, there exists within these arguments a considerably large but often overlooked lacuna that extends back to their foundations and, for some, "call into question their validity" (Hunley, Healy & Long, 2009, p. 35). Lewontin's theory may be better characterized as an axiom than a proven scientific fact; it lacked irrefutable evidence and validation and was perpetuated based upon its "universal" acceptance alone. In his published debate on the non-existence of races, Livingstone (1962) not only failed to provide any empirical evidence in support of his position but he also failed to consider, as Dobzhansky had in their exchange, that, while discordant gene patterns are common throughout human populations, cultural and environmental barriers serve as important impediments to gene flow and, as such, can produce large discontinuities and concordant allele frequencies among globally distributed groups. Although accumulated data on the effects of geographic, climatic and historical factors on human interaction, reproduction and migration do not support archetypical notions of isolated, discreet small-scale populations, an improved understanding of the sources and effects of these partial barriers has increasingly shown that populations or "social categories cannot be completely dismissed" (Foster & Sharp, 2003, p. 846). On account of historical processes (like bottlenecks or founder effects), similar SNPs, ancestral or conserved haplotypes, variants associated with disease (in their incidence, severity, progression and treatment), as well as environmental response genes may be shared among members of one population in frequencies that differ from those patterns identified in other populations (Collins & McKusick 2001; LaVeist, 2002). These ideas are not new and have been inscribed within many of the common techniques used to detect genetic characteristics. It can be recalled that linkage disequilibrium analysis was developed for use on identifiable populations with reasonably unique demographic profiles, with the purpose of searching out complex disease genes, while Haplotype mapping relies upon the assumption that ancestral signatures are common across populations with different social histories

(Jorde, 2000; Goldstein & Weale, 2001; Daly et al., 2001; Weiss & Clark, 2002; Kittles & Weiss, 2003; Tishkoff & Kidd, 2004; Pritchard & Przeworski, 2001). Additional explanations have also emerged that present viable alternatives to cline-based interpretations given the the observed patterns of variation. Hunley, Healy and Long (2009, p. 35) have proposed that "the opposing clines envisioned by Livingstone might form in a purely phyletic process through random character loss," while Ramachandran and colleagues (2005) have argued that the strong global correlation between genetic and geographic data can be better understood as the result of select evolutionary processes, specifically, nested serial population fissions, migration, and isolation. This latter model is worthy of particular note because, while isolation by distance is incompatible with theories that supports the identification of biogeographically distinct populations or biosocial groups, "serial fissions" do allow for the possibility of meaningful taxonomic units and, by extension, a renewed discussion of the identification of populations, races or ethnicities. Several innovative, genetic studies on the inference of individual ancestry and the detection of population [sub]structure have initiated this discussion, while simultaneously challenging the fundamental beliefs that have shaped scholarly knowledge of how human variation is distributed.

Genetic Clustering: STRUCTURE and Race

"It is time to devote some political will, experimental innovation and financial resources to address crucial questions about patterns of genetic variation, geographical ancestry and disease. At the same time, geneticists need to facilitate the incorporation of this new information into complicated

frameworks of knowledge that we already have and attitudes that manifest as public perceptions of race" (Bamshad et al., 2004, p. 608).

As has already been seen, it is well known that 85%-90% of neutral genetic variation in the human species is attributable to differences between individuals within populations, while the remaining 10%–15% of the total variation is distributed among groups (Lewontin, 1972; Barbujani et al., 1997; Jorde et al., 2000). In practical interpretation, this latter estimate represents genetic differentiation between "predefined populations," which may take the form of biosocial races, ethnicities, biogeographic nationalities or continental ancestries (Rosenberg et al., 2002; Bamshad et al., 2003). Although arguably modest in size, the latter between-population variance component has been shown to conform to geographic structuring (Bamshad et al., 2004; Witherspoon et al., 2007) and to wield considerable influence over the manifestations of individual or group-specific biology: e.g., affecting the average differences in physical traits, disease risk, and treatment outcome among populations (Gonzalez et al., 1999; Flanagan et al., 2000; Thio et al., 2002). In order to evaluate of the effect of this variation, in both its magnitude and direction, on characteristics such as these, assumptions are typically made about the genetic structure of the study sample (the number of [sub]populations) and about which individuals are assigned to each of these internal groups (the classifications and membership patterns). In the absence of appropriate, or even limited but corroborative, genetic data, these inferences on the nature and extent of human population structure are typically drawn from information that is itself based upon some level of assumption (Foster & Sharp, 2002). Group definitions and individual categorizations that enable these analyses have been traditionally approached using predetermined labels and self or imposed identities, which have been established not by measure of known biological similarity but on

the basis of culture or geography (Rosenberg et al., 2002). Still today, phenotypic and societal factors, such as skin color, race, and the multitude of malleable social descriptors, continue to be taken as easily accessible, and presumably reliable, proxies for the true genetic relationships that characterize the neutral variation observed among the samples under review (Cooper, 1994; LaVeist, 1997; Williams, 1997; Aspinall, 1998). As this proxy-based approach most often leads to the comparison of phenotypes between cohorts that are described using the language of race (e.g., test samples of "African Americans,""Whites,""Hispanics"), its adoption continues to be met with frequent opposition and the findings it yields are often faced with outright rejection. It should not be surprising that issues over a perceived insertion of racial biology into the field of genetics are at the fore of current discussions (Shields et al., 2005; Nguyen et al., 2006; Kittles et al., 2007), as the race-free study of human "sameness" is, for many, a key focus of contemporary genomics (Venter, 2000). The validity of this system of classification-by-proxy in genetic research has been, therefore, hotly debated and highly criticized, especially for its use in the fast-developing areas of the biomedical sciences and for its translation into the contexts of clinical and forensic practice (c.f., Ossorio & Duster, 2005; Braun et al., 2007; Lundy, 2007; Clarke et al., 2009; Kayser & de Knijff, 2011). As a result, any related methodology has been equally censured for "its weak conceptual underpinnings and its strong assumptions about underlying biology" (Bamshad et al., 2003, p. 578; c.f., Lee et al., 2001; Wilson et al., 2001; Foster & Sharp, 2002). There is the belief that with the rapid-paced emergence of large-scale and continually growing repositories of freely accessible genetic resources, many of the concerns expressed over the appropriate use or general utility of surrogate data for individual genotypes and the reliability of proxy-based analyses can be formally addressed through large-scale empirical testing of this new data, given its multi-locus representation of globally distributed populations (Mountain & Cavalli-Sforza, 1997; Rannala & Mountain, 1997; Shriver et al., 1997). Findings from preliminary studies have not, however,

successfully realized this goal. Instead, these studies have led to the formulation of a host of unsatisfying and often contradictory conclusions. Despite these disparate and often disputed results, the presence of conflicting interpretations, often offered for the same data sets, have, nevertheless, implied for many that there exists a simple answer to the complex questions surrounding population structure, stratification, and classification: that reliably strong inferences supporting structured populations cannot be justifiably made from such weak signals and that proxies inaccurately reflect true relationships (Wilson et al., 2001; Risch et al., 2002; Romualdi et al., 2002). Understandably, in the wake of these statements, the divide over the nature of human population structure has only widened. Moreover, the arguments made against the existence of either a valid or methodologically useful relationship between groups defined by proxy and genetically inferred clusters have gained further traction among scholars in the broader human biological community and within the fields of medical and anthropological genetics (Witzig, 1996; Goodman, 2000; Schwartz, 2001).

Despite the strengthening of these mainstream views, recent innovative genomic approaches to quantifying among-group genetic diversity have been, nevertheless, put forth in the interest of better appreciating the relationships between individual ancestry, race and population structure. These also have been made in challenge to the popular arguments that equate the categorization of humans in proxy-based genetics with "racial profiling," the "medicalization of race," and, most disparagingly, with the "scientific legitimization of a flawed social construct" (Witzig, 1996; Schwartz, 2001). These studies are, therefore, among the greatest current, and, if not, the most controversial, contributions to the ongoing conflicts over the validity of the role of biosocial and biogeographic labels in understanding human genetic variation. They have ushered in a new era in populational thinking that is helping to reshape the trajectory of the present "genomics age" (Smith, 2005). By introducing new techniques for

alternatively designed analyses and new ways to make use of the growing repositories of molecular data, this innovative research is reinvigorating the scholarly discourse on structured variation. In turn, it is also reenabling classification as a (1) viable scientific tool when considering broadly based patterns of similarity and diversity and as a (2) valid scientific goal when focusing on individual membership and affinity, even when these concepts are considered in terms of race, ancestry and ethnicity. Although these studies provide, in part, results that conform to conventional expectations for the (small) among-population variance components and the (low) frequency of private alleles, they also demonstrate that genetic differentiation can be detected among human [sub]groups with high mean accuracy, especially when labels are based upon continental ancestry or biohistorical place-of-origin. Consider, for example, Rosenberg and colleagues (2002), whose analysis "identified six main genetic clusters, five of which correspond to major geographic regions, and subclusters that often correspond to individual populations" (p. 2381). Bamshad et al. (2003) have undertaken similar work that has "demonstrated that, for a collection of heterogeneous samples from sub-Saharan Africa, East Asia, and Europe, the genetic data accurately predicted assignment to clusters that corresponded to major continents" (p. 584). The implications of these findings are weighty, because, as Bolnick (2008) has remarked, the biogeographic groups employed in these new studies and which have been successfully linked to genetically inferred clusters "are essentially races" (p. 73). As Rosenberg et al.'s (2002) five "major geographic regions" represent America, Africa, Eurasia, East Asia, Oceania and Bamshad et al.'s (2003) three identified clusters correspond to Africa, Europe and Asia, their findings have been, quite convincingly, used to indicate how race-based partitioning by means of biogeographic proxies (like the broad labels of continental ancestry, but also, even, ideas of biohistorical origin, including "ethnicity," in the case of admixed individuals) is not without some degree biological significance (Burchard et al., 2003; Mountain & Risch, 2004). Similar attention has also been given to several other studies which

have generated corroborating results even when adopting different methodological approaches (c.f., Wade, 2002). Frudakis and colleagues (2003) developed a linear classification method to incorporate highly informative SNPs into a classifier model. They found that individuals of European, African, and Asian descent could be identified, respectively, with 99%, 98%, and 100% accuracy. Shriver et al. (2004) employed, instead, allele sharing distance and neighbor-joining tree making methods to reconstruct individual relationships among African-American, European-American, Chinese and Japanese samples. They produced trees with genetic clusters that corresponded to these four predefined populations, with roots branching consistently from the African-American clade. Studies such as these, along with more recent, refined work on Ancestry Informative Markers (AIMs), Biogeographical Ancestry Analyses (BGA), and the definition of Personal Genetic Histories (PGHs), have not only been cited as support for genomic substructure and as evidence of an estimatable individual biological identity (serving as the basis for personal genomics in the public sphere and commercial arena), but have also been used to reinforce the larger, theoretical and philosophic, premise that there exists a scientifically meaningful relationship between individual genetic histories and the modern concepts of human races (Shriver and Kittles, 2004).

Although several approaches have been successful at accessing genetic structure at varying levels of population stratification, studies, like those first presented by Rosenberg et al. (2002) and Bamshad et al. (2003), are of particular interest for the methodological schema through which their associations between the patterns of genetic differentiation and biogeographic groups are produced. This specific approach to the inference of population structure merits special emphasis for two reasons: (1) for its influence on the course of analytical procedures in population genetics and (2) for its statistical methodology and computational procedures, which define and enable this type of analysis. Firstly, this

approach has steadily gained in popularity, especially since its introduction for use on human genetic data, within the framework of the user-interface, STRUCTURE (Pritchard, Stephens & Donnelly, 2000), its companion software (e.g., Rosenberg, 2004; Jakobsson & Rosenberg, 2007) and subsequent extensions (e.g., Falush, Stephens & Pritchard, 2007; Hubisz et al., 2009). The impact of this method on anthropological genetics is palpable in the number of publications in which it, and variations on it, have been featured as the main mode of analysis or the main topic of discussion (for human data alone, see Rosenberg et al., 2002; Bamshad et al., 2003; Falush, Stephens & Pritchard, 2003; Falush et al., 2003; Hoggart et al., 2004; McKeigue, 2005; Tang et al., 2005, 2006; Zhu et al., 2006; Montana & Hoggart, 2007; Bolnick, 2008; Weiss & Long, 2009). The reach of its effects are further revealed by the fact that even experts in the field of human population genetics consider this "Bayesian K-populations cluster analysis" approach as the "the tool of choice for analyzing the large-scale human population genetic data that are now available" (Weiss & Long 2009, p. 704). Given the conceptual difficulty of deriving explicit hypotheses from the complex evolutionary processes that, it is believed, have actively shaped the patterns of differences observed among human groups, as well as the practical expertise needed to design effective statistical and computational tools in order to test these hypotheses, it is not surprising that current scholarship bears witness to a growing "tendency for researchers to rely on readily available user-friendly computer packages" (Weiss & Long, 2009, p. 704). Practical demands are, therefore, making software packages that execute this mixture-clustering approach desirable and useful commodities to many researchers, who share in a widespread interest to better understand the global distribution of human genetic variation from a population perspective and are keen to use novel data types and apply cutting-edge technology to their unique research questions (especially given the current ease of access to world-wide samples and free-ware tools), but who are neither statisticians nor

computer programmers themselves (for example software, see Pritchard, Stephens & Donnelly, 2000; Corander et al., 2004; Chen, Forbes & François, 2006).

Secondly, genetic clustering algorithms "have become an important tool in a number of fields," but have received considerable attention in their application to human data for their statistical design, assumptions and output (Hubisz et al., 2009, p. 1322). Discussions of this form of computational genetics have been driven especially by the fact that this approach is known to produce consistent (albeit controversial) results, both across marker types and chosen populations, when similar study designs are followed and replicated. Despite the calculation of low levels of among-group variation (and F_{ST} values), it has been shown that individuals sampled world-wide can be organized into distinctive clusters, with high rates of success when a large number of genetic markers are used (e.g., ≥ 60 for a mean classification accuracy of 90% and ≥100 loci to achieve 99%-100% correct assignments) and with measurable degrees of statistical certainty (see Bamshad et al., 2003). These "genetically inferred" populations often correspond to the politico-geographic boundaries that enable the inference of biogeographic ancestry and reflect affiliations with commonly used biosocial proxies, including self-identified racial, ethnic and ancestry groups (Risch et al., 2002). When the statistical procedures that are common to these so called "Bayesian K-populations cluster," "STRUCTURE," or "*structure* -like" analyses are considered (Weiss and Long, 2009; Pritchard, Stephens & Donnelly, 2000), the underlying methodology is of particular interest for its relationship to the finite mixture and model-based clustering methods that have been already proposed here for resolving world-wide craniometric data into structured groups. These procedures have been formalized and put to the test under a variety of conditions, including different model parameterizations, distributional assumptions, and with the use of simulation (c.f., Dawson & Belkhir, 2001; Anderson & Thompson, 2002; Evanno, Regnaut & Goudet,

2005; Pella & Masuda, 2006; Chen, 2007; Huelsenbeck & Andolfatto, 2007). They have been applied with success to the problems of detecting latent, hidden or cryptic population structure within a [pan]mixed data set, defining this structure in terms of distinct populations, allocating individuals to these populations, identifying admixed individuals or migrants from members of their parental populations, studying zones of hybridity and inferring instances of historical population interaction and the production of admixed groups, and estimating population-specific allele frequencies in the presence of high levels of migration and admixture (e.g., Pritchard & Rosenberg, 1999; Rosenberg et al., 2001, 2002; Falush et al., 2003; Albert et al., 2006; Lecis et al., 2006; Ostrowski et al., 2006; Falush, Stephens & Pritchard, 2007; Hubisz et al., 2009). What is key to extract from these analyses is the fact that they actively engage statistically robust model-based clustering algorithms, they can execute their procedures in the absence of any personal or population identifiers (i.e., assuming individuals are "unknown"), and can account for the unique properties of each data set by adopting a "kind of admixture approach to human variation," whose probabilistic methodology allows for flexibility in membership assignments (Weiss & Long, 2009, p. 704). These model-based clustering methods offer for genetics, therefore, a proven means to explore concepts of race, ancestry, and group-specific characteristics. They promise to give researchers the ability to evaluate *scientifically*, with statistically satisfying and intelligible measures of certainty, their relationships to the patterned distribution of human biological variation, while employing a methodologically sound alternative to the classification of morphogenetic data requiring *a priori* known, predefined populations. In light of these advantages, and the obvious procedural precedent that they provide, it is worthwhile to briefly address the general features of these types of analyses (i.e., the commonalities in approach regardless of the specific program used for implementation), and, given the popularity of STRUCTURE software, to also include an explanation of the theoretical conditions and algorithmic underpinnings of this specific package.

For the sake of the following explanation, it is assumed that it is both appropriate and useful to group a collection of individuals, whose biological information makes up a reasonably sized sample of genetic data, into distinct and already defined populations. It is worthwhile to consider two situations of common interest in population genetic applications for which this clustering and classification would be desirable, if not completely necessary: (1) the researcher begins with a set of chosen samples, and through these individuals, seeks to investigate the distributional properties of certain traits, to evaluate the patterns of among-group relationships and, ultimately, to better understand the properties of human "populations" or (2) the researcher begins with some prior knowledge about the existence of [predefined] populations and, using estimated allele frequencies in each population at unlinked loci, seeks to allocate individuals of unknown membership into the groups with which they have the greatest likelihood of affiliation. The former scenario represents the utility of population structure in the case of global diversity studies, where the population is the unit of interest and tracing the evolutionary relationships among modern populations through time and across space is the goal (e.g., Cavalli-Sforza, Menozzi & Piazza, 1994). The latter example demonstrates, alternatively, the need to account for the presence of "nuisance" and often "cryptic" population structure, which can lead to practical problems of considerable import. Take for instance, case-control studies, where erroneous associations ("false positives") between a candidate gene (the "disease marker") and a phenotype can result in deleterious research, health, and treatment effects (c.f., Cardon & Palmer, 2003; Thomas, Haile & Duggan, 2005). Similar problems exist in the context of forensic identification by DNA, in which the degree of population substructure is an essential piece of information for estimating the probability of a false match using Bayesian methods (c.f., Foreman, Smith & Evett, 1997; Roeder et al., 1998). Regardless of the fundamental differences in goals or the potentially infinite variety of circumstances that these goals may

beget, the first procedural step in either of these situations remains the same: the analysis should begin with the identification of population [sub]structure. Although it is common to see this identification made through the use of proxies (which often include self-affiliation criteria, language, phenotypic expression or geographic location), the degree to which some of even the most frequently used surrogates actually correlate with genetic distributions is unknown, their definition is often variable, and their choice is highly subjective in the absence of prior testing and confirmation. Moreover, in instances of latent population structure, it may not be possible to objectively observe or access the internal organization of the mixed data. Often in such "cryptic" cases, proxies do not adequately represent the true populations or approximate meaningful groupings (Pritchard & Rosenberg, 1999; Pritchard, Stephens & Donnelly, 2000).

In the presence of these legitimate concerns over indiscriminate use of proxy information and the desire for an unbiased identification of underlying groups, it is wise to pursue other avenues for manipulating large scale genetic data sets, when individuals are taken to be both "unknown" and representative of a "mixture" of multiple "unobservable" groups. As a robust alternative, it may be prudent to adopt a cluster and classification approach, through which it is possible to (1) cluster individuals who are genetically similar in the absence of personal or group identifiers, (2) resolve the mixture of data into its constituent component[s] and thereby identify distinct *genetically inferred* clusters, and (3) when external, *a priori* information is available, evaluate if and how these biological groups correspond to personal demographic data. To achieve this cluster and classification goal, the choice lies between implementing one of two broadly-defined statistical methods for the detection of group structure and identification of individual memberships: (1) distance-based methods or (2) model-based methods. Distance-based methods are often appealing because they are easy to employ and can be visually

interpreted. While these have been historically dominant in population genetic literature and shown to be successful (c.f., Bowcock et al., 1994), these methods are being increasingly censured for their exploratory nature as they require the calculation of pair-wise distances, the graphical representation of the resulting matrices as trees or multidimensional scaling plots, and the identification of clusters by subjective "by eye" inspection (Risch et al., 2002). This interactive approach often leads to both method and user-induced subjectivity, resulting in low-confidence and allowing for the introduction of unacknowledged and, so, irreconcilable bias. In this analysis clusters are dependent upon the chosen distance measure and graphical display, cluster validity is difficult to assess, and the overall framework renders the incorporation of external information (e.g., sampling location) difficult. As Pritchard and colleagues (2001, p. 946) have argued, distance methods are, due to disadvantages such as these, poorly suited to "fine statistical inference." Finite mixture methods are, as discussed above, considerably more complex in their theoretical underpinnings and practical application. They have been frequently characterized as inaccessible and criticized for their computationally intensive procedures, as their full realization requires maximum-likelihood (EM) or Bayesian (MCMC) estimation. They are, however, especially advantageous because they are statistically rigorous. These mixture-methods work under the assumption that the observations from each cluster are drawn randomly from a mixture of underlying probability distributions and use information criteria, Bayesian methods and maximum-likelihood estimation to select the optimal model (parameterization and component-number solution), compute the cluster parameters, make individual cluster assignments and calculate their associated posterior probabilities of membership (McLachlan & Peel, 2000; McLachlan et al., 2004). As model-based clustering algorithms do not require sample data to be entered in predefined groups (so that the actual numbers of genetic clusters need not be specified before the procedures are executed), they also provide a robust identification option for groups with distinctive allele frequencies as they are

unaffected by arguments against the user and program induced biases introduced by *a priori* knowledge of, for example, population, racial, or ethnic affiliations, individual genetic origins, sampling locations, geographic distributions, phenotypic information, or surrogates for racial proxies (Rosenberg et al., 2002; Bamshad et al., 2003). As these methods are both more mathematically and conceptually satisfying than other heuristic clustering and distance-based techniques (on account of their parametric design, iterative implementation, and probabilistic, "model-based" approach to the clustering and classification of "unknown" data), they are swiftly emerging as the preferred methodological choice for genetic analyses (McLachlan, Do & Ambroise, 2004; Schlattmann, 2009).

With a good understanding of mixture modeling theory and a knowledge of the research goals and disciplinary expectations brought to the new genetic clustering and classification projects, fundamental similarities within all of the "*structure*-like" approaches can be identified. These common constraints, assumptions, design and architectural choices are tied either to the principles of population genetics (and its data) or to the statistical (mathematical, algorithmic, computational) framework that enables the basic execution of finite mixture analysis. In general, a guiding algorithm assembles clusters using a data set of multilocus genotypes by estimating for each individual specimen the proportion of its genome that belongs to each cluster. Population-informative clusters composed of like individuals are then identified, so that, through these patterned memberships, the underlying ("cryptic" but "true") population structure is exposed (Pritchard & Rosenberg, 1999). As the STRUCTURE program is considered to be the best known, easiest obtained (free and downloadable online[4]) and the most widely used of the software packages that make this approach accessible, it is the most relevant example to use

[4] The current version, STRUCTURE 2.3.3 (released 06-2010) is available for download at: http://pritch.bsd.uchicago.edu/structure.html

in a more detailed discussion of computer-assisted genetic clustering (see Pritchard, Wen & Falush [2009] for the most recent software documentation). The STRUCTURE software implements "a model-based clustering method for using multilocus genotype data to infer population structure and assign individuals to populations"(Pritchard, Stephens & Donnelly 2000, p. 945). Input data commonly consists of globally distributed polymorphisms that have been genotyped for a large number of individuals, for whom population identifiers have been removed or remain "unknown" (c.f., Rosenberg et al., 2002; Bamshad et al., 2003). Given the fluidity in the definitions of "populations" (in factors such as size or breadth of locales), STRUCTURE's analyses may also be carried out on intra-continental or even within-country data sets representing "major racial/ethnic groups" (c.f., Tang et al. [2005] for U.S. populations). STRUCTURE's procedures can be conveniently applied to various types of common markers (e.g., microsatellites, MSATs, restriction fragment- length polymorphisms, RFLPs, or single nucleotide polymorphisms, SNPs) and it has been recently extended to include more refined approaches, including the use of models with informative priors for improved fit (Pritchard, Stephens & Donnelly, 2000; Hubisz et al., 2009). In its general application, the STRUCTURE procedures are implemented to, first, identify individuals with similar allele frequencies and, second, use this information to sort these "unknown" individuals into *k* clusters (where *k* is varied across independent runs of the algorithm). The patterning observed in the final cluster solution may be ultimately informative of population substructuring within a diverse sample (i.e., composed of a mix of populations, ancestries, races, representing many sampling locations). The statistical framework defining the STRUCTURE software dictates that all of the genetic data that is sampled arises from one or more of *K* unobserved populations. Each of these populations is represented by a set of allele frequencies (*P*) at each locus, where *P* is a multidimensional vector with elements p_{jlk}, corresponding to the frequency of allele *j* at locus *l* in population *k*. It does not assume a particular mutation process, but it does require that loci are at Hardy-Weinberg and linkage equilibrium

within the [sub]populations (although version 2.0 and later can deal with weakly linked markers) and that a value for the number of possible clusters (*K*) is predefined (but which may be and should be varied over multiple executions using the same data set as its best value can be statistically determined). The program allows for the consideration of different model types including: (1) no-admixture model, where individuals hold membership in one population *or* in another, (2) admixture model, where individuals may hold multiple memberships, so that some ancestors in population *k* are believed to have contributed to individual *i* some portion of his or her genome, and (3) linkage model, where the admixture model is generalized to account for admixture linkage disequilibrium (Pritchard, Wen & Falush, 2009; for the special case of the LD model, see Falush, Stephens & Pritchard, 2003). As all of these models vary in their complexity, model choice effects changes to STRUCTURE's procedural conditions. For example, in the simplest "no-admixture" model, STRUCTURE follows a hard clustering solution, in that all of the genetic material from any one individual is presumed to have come from one population and, so, the prior probability for each population is 1/K. In the more realistic, flexible, but more complicated, model "with admixture," each individual is permitted to hold membership in more than one, and often in many, of the *k* populations and, so, is assigned a "mixed ancestry." When individuals are allowed to hold membership in multiple clusters, the degrees of association for each individual can be *a posteriori* assessed using robust probabilistic theory (assuming that membership coefficients sum to 1 across all clusters). In this case of clustering "with admixture," the patterns of assignment of the individual specimens to *k* population clusters can be used to determine the number of parental populations (K, the total optimal number of clusters) and to estimate individual mixing proportions. As the gene frequencies and the population of origin of each allele copy of each individual are typically (or are taken to be) unknown in advance, they must be probabilistically inferred from the data (here, using a computationally intensive Bayesian approach). STRUCTURE implements a Markov

chain Monte Carlo (MCMC) algorithm to integrate over the parameter space and produce the cluster assignments. It proceeds from an initial arbitrary configuration of parameter values, iteratively updates the subsets of the parameters to "new" values, which are conditional on the data and the other current parameter estimates, and continues over many runs until, in theory, the algorithm reaches convergence at the posterior distribution of all of the parameters given the data available (for issues of convergence and practice, see Pritchard, Stephens & Donnelly, 2000, or Falush, Stephens & Pritchard, 2003).

The resulting estimated proportions or posterior probabilities of membership may offer useful information on the ancestry or "racial" background of each individual as they represent the percent of admixture or contributions of ancestries within each individual as expressed by their genotype (but constrained by the context of the sample distribution). The driving force of the STRUCTURE procedure is the hypothetical number of ancestral populations or the optimal number of clusters (K), which must be specified in advance or "fixed by the user" (Hubisz et al., 2009, p. 1324). For this requirement, given its critical importance, the best approach to identifying preset cluster numbers remains under continuous debate (see Evanno, Regnaut & Goudet, 2005). While it is true that the value of K must be provided by the "user" to initiate STRUCTURE's clustering algorithm, heuristic methods for selecting the optimal number of components are regularly recommended. They are acceptable "as hoc guide[s] to which models are most consistent with the data" and which "seem to give sensible answers in practice" (Pritchard, Stephens & Donnelly, 2000, p. 949). When these model fit and K-choice "tests" are conducted, this optimal value is not actually user-defined but is statistically determined. For potential concerns over forced solutions, it is significant to note, as Bolnick (2008) has, that "the fact that structure identifies a particular number of clusters is insignificant: it does so simply because the user told it to do so. What is more important is that *structure* provides a way to determine the value of K that

is most appropriate for the data in question (i.e., the most likely number of clusters or populations represented" (p. 74-75). This optimal value of K is commonly defined as the component solution that maximizes the probability of observing a given data configuration. This best number may be derived internally using the model choice criterion, Ln P(D), implemented in STRUCTURE, where comparisons are drawn between penalized log likelihoods over independent runs with differing numbers of K clusters so that the optimal value of K is the number associated with the highest probability (see Pritchard, Stephens & Donnelly, 2000). The best component solution can be also identified using a model of linked loci when samples are large (Falush, Stephens & Pritchard, 2003) or independently validated by employing external software that may adopt slightly different approaches. For example, Evanno and coworkers (2005) offer a revised likelihood method, that can be easily executed with the StructureSum program (Ehrich 2006). In instances when this "best" number of components is determined to be quite large and the data set in question is, in logical turn, quite complex, STRUCTURE is known to produce substantially different solutions for estimated component membership coefficients across multiple clustering runs, even when the conditions of the analysis are kept constant, i.e., the same data set, the same choice of K, different probability values (Pritchard, Wen & Falush, 2009). This is in part because STRUCTURE [re]assigns all individuals to clusters in each execution of the mixture procedure. In the presence of a data set with a large number of individual observations and an overall complexity of internal structure (i.e., composed of many clusters), repeated clustering is computationally intensive and time-consuming. Variation in cluster assignment between each replicate may be attributed to STRUCTURE's inability to consistently make the best possible cluster allocations given insufficient burn-in length and run time for the execution of its iterative algorithm (i.e., lacking the time necessary to overcome the effect of the starting configuration and accurately estimate the parameters). Moreover, highly complex and nuanced data structure is often difficult to quantify, and, so, ambiguity in cluster

solutions may also arise from the possibility of several equally probable ways to partition the data into clusters. In consequence, different runs will yield different membership assignments for some individuals, even when K remains fixed. This ambiguity in multiple "ultimate" solutions can be easily remedied, however, by subjecting the results for a single data set obtained over many implementations of STRUCTURE's clustering procedures to the companion software, CLUMPP (Jakobsson & Rosenberg, 2007). This cluster-matching and permutation program works to resolve the often combined *problems of* "label switching" heterogeneity and the detection of "genuine multimodality" (i.e., truly different solutions) in STRUCTURE's results to enable their subsequent quantification, visualization and interpretation (Jakobsson & Rosenberg, 2007). CLUMPP can be used, therefore, to combine multiple STRUCTURE runs for a chosen (presumably optimal) value of K in order to yield average cluster membership values (c.f., Londin et al., 2010) and its output can be further manipulated in other partner packages, like the *DISTRUCT* program (Rosenberg, 2004) for cluster visualization (offering an accessible graphical display of the cryptic population structure uncovered within the original data set).

The "meaning" of the results obtained from STRUCTURE can be externally evaluated by principal component analysis (PCA), using, for example, the EIGENSTRAT method (Price et al., 2006) or finite mixture analysis (FMA) via MCLUST (Fraley and Raftery, 2002, 2007, 2010), to test if the "best" number of population clusters (K) obtained from the STRUCTURE program remains robust under additional scrutiny, when consideration is given to personal identifiers or to other common proxies, like geography. While a "model with prior population information" was introduced in the early versions of the software for instances of detecting strong population structure, isolating of hybrids, and making a direct correspondence between sample location and genetically inferred cluster (Pritchard, Stephens & Donnelly, 2000), the more recent versions of STRUCTURE allow for the incorporation of external "sample

group" information (i.e., Bayesian priors), but with special focus on "placing much more prior weight on clustering outcomes that are correlated with the sampling locations" and for particular use "in settings where the data are not highly informative"(Hubisz et al., 2009, p. 1323). By weighting the analysis with respect to some prior distribution (e.g., reference data on geography, the proportions of "races" or other discrete groups based upon phenotypic characteristics), STRUCTURE takes advantage of the valuable relationship between geno-type and group-type to offer more and better information on biogeographic ancestry and biosocial identity. These informative priors have also been shown to be exceptionally useful for improving STRUCTURE's final results, when the internal structuring of the data set is weak and, in turn, the program's cluster solutions and classifications are difficult to interpret. Because, as Hubisz et al. (2009) explain, "there is an enormous number of ways that N individuals can be partitioned into K populations" and "the basic STRUCTURE models assume that all partitions of the N individuals into K populations are equally likely, a priori, […] it takes a considerable amount of statistical evidence to provide strong support for any particular partition" (p. 1323). The inclusion of prior data helps STRUCTURE to solve this "rather difficult problem" and account for some of the noted incongruities in findings when different types of population genetic analyses have been compared to STRUCTURE (Hubisz et al., 2009, p. 1323). It explains, for instance, why significant among-sample F_{ST} values can be obtained for individuals assembled from different locations while the STRUCTURE program can fail to identify well-defined population structure using this same data. As this example shows and the emergence of many extension, companion, and duplicate packages suggest, the STRUCTURE software, with its model-based clustering approach, is growing increasingly capable in its ability to resolve earlier concerns over its method and output, to extend its scope to address models and data of varying complexity, and to respond with considerable flexibility to the changing needs and areas of interest in contemporary genetics. For evidence for the contribution of this mixture approach to

understanding the patterning of human genetic diversity in terms of biosocial categories and the estimation of biogeographic ancestry, it is possible can turn to the seminal studies by Rosenberg et al. (2002) and Bamshad et al. (2003) who used the STRUCTURE software to investigate world-wide population structure and continental group membership and to Tang et al. (2005), who have addressed finer levels of substructing by applying this same approach to populations within the United States.

In the interest of evaluating how well, or if at all, commonly referenced "populations" (whose definitions have been traditionally made on the basis of culture or geography) relate to real patterns of biological variation and speak to their utility in, for example, studies of evolutionary history or epidemiological contexts, Rosenberg and colleagues (2002) used the HGDP-CEPH Human Genome Diversity Cell Line Panel "to test the correspondence of predefined groups with those inferred from individual multilocus genotypes" (p. 2381). They subjected 377 autosomal microsatellite loci, sampled from 1056 individuals and representing 52 world-wide populations, to model-based cluster analysis, implemented with an early edition of the STRUCTURE software. In the absence of (or, in fact, withholding) prior knowledge of the individuals' genetic place of origin, sampling location, or phenotypic classification, Rosenberg et al. (2002) demonstrate how STRUCTURE's clustering algorithm is sufficiently sensitive and the internal data structure sufficiently strong to reveal distinct "genetic clusters." More remarkably, they report that, when their data set is analyzed under a fixed six cluster solution (the maximum value of K presented in their publication), these biologically-inferred groups have both geographic and social meaning. Of the six "main" genetic clusters identified with STRUCTURE, five are shown to correspond to the "major geographic regions" of Africa, Eurasia, East Asia, Oceania, and America (Rosenberg et al., 2002, p. 2382). These five large-scale divisions clearly reveal continental level structure: what Bolnick (2008) has recently equated with "races." The sixth component is composed largely of the Kalash of northwest

Pakistan and may instead represent a finer degree of [sub]structure; demonstrating, therefore, the ability of this genetic clustering approach to also capture more complex forms of population variance. As the Kalesh are a small-scale "isolated" people, characterized by their Indo-European language and limited geographic location, Rosenberg et al (2002) take their unique cluster allocation along with the fact that, unlike other Pakistani groups, they showed no membership with the East Asian component (at K=5), as an example of how this genetic clustering approach can detect signals of distinctive and divergent population history. Their findings, in effect, support previous external arguments that attribute European or Middle Eastern origins to the Kalesh. Rosenberg and coworkers also extend these smaller-scale discussions of origins and affinity to other predefined groups, as STRUCTURE also detected a variety of "subclusters," which were found to correspond "closely to either predefined regional or population groups or to collections of geographically and linguistically similar populations" (2002, p. 2384). The clusters with indications of mixed population histories were formed by STRUCTURE's "merging of subpopulations that share similar allele frequencies" and may be attributed to either recent admixture among founder populations or reflect shared ancestry before a divergence into different groups and without subsequent gene flow between them (Bamshad, 2003 p. 584; Li et al., 2008). It is assumed that, before settling upon a model and coming to these conclusions, Rosenberg et al. (2002) varied the possible "best" number of clusters (K) across program runs, in accordance with the STRUCTURE protocol (Pritchard, Stephens & Donnelly, 2000; Pritchard, Wen & Falush, 2009). It is unfortunate, however, that it is not possible to fully evaluate the weight of their findings as they do not (1) provide the probabilities of the observed data given each different value of K, (2) indicate which single value of K maximized the probability of the observed data, and (3) report any changes to the clustering patterns under repeated identical runs (i.e., the membership results for each of the replicates). They do, however, discuss briefly the broader meaning that may underlie the results

obtained for select cluster numbers (1<K ≤6). When K=2, the clusters were "anchored" by Africa and America, likely reflecting their well-known large genetic distance. Subsequent increases of K resulted in additional internal splits. At K=5, clusters were organized in agreement with major geographic regions and the Kalesh were identified as a unique group when K=6 (Rosenberg et al., 2002, p. 2382). While these explanations do provide reasonable support for their conclusion that "in the worldwide sample, individuals from the same predefined population nearly always shared similar membership coefficients in inferred clusters" (2002, p. 2384), it is only appropriate to acknowledge that their failure to identify which of these values of K represents the most likely (that is, statistically evidenced) cluster number creates difficulties in objectively assessing the reliability of the five or six cluster analysis (the solutions which they emphasize) and the strength of the presumed relationships between the reported cluster patterns and population identifiers. Rosenberg and colleagues (2005) do, however, report in their later reanalysis of this same data (but using an expanded panel of 933 markers) that more continuous sampling should not alter the "clusterdness" of the individuals and that the K=5 solution captures more of the genetic differences across geographic barriers. The documentation oversight in their original publication has, nevertheless, led to some strong methodological criticisms over which, if any, of their reported values of K represent the "best estimate" or demonstrate a measurably good model "fit" to their particular data set (see Bolnick, 2008). Even so, many subsequent projects have followed Rosenberg and colleagues' (2002) early study and these have grown in scope to incorporate more complex modeling (using STRUCTURE's recent extensions for, e.g., linked markers and priors) under the more difficult conditions of weak structure and limited data. Despite these added constraints and advanced computational requirements, these later projects have obtained equally good if not even improved evidence for the correspondence between genetic and population-informative clusters (c.f., Falush, Stephens & Pritchard, 2003; Hubisz et al., 2009). Moreover, other complementary approaches

to STRUCTURE's method have obtained similar results. Consider, for example, Li et al. (2008) who, using the same (HGDP-CEPH) panel as Rosenberg et al. (2002, 2005), found that "individual ancestry and population substructure were detectable with very high resolution" (p. 1100). They do report that some variation in cluster patterns did occur between their six and "new" seven component model and Rosenberg et al.'s (2002) previous findings: in particular, individuals from South/Central Asia and the Middle East were better distinguished. These differences in solutions should not come as a surprise, however, given that Rosenberg et al. (2002) already isolated these regions and peoples as having complicated demographic histories and admixture events. What is especially important to note is the support that this study gives to the consistency of mixture approaches (regardless of implementation, e.g., STRUCTURE versus the ML-based software, *frappe,* developed by Li and colleagues) for reliable population identification. As Li et al. (2008) conclude, "[m]any populations have been isolated from each other by geography or custom. The observation that they are genetically distinguishable suggests that self-reported ancestry is sufficiently accurate for assessing population stratification in disease studies, except for those involving recent admixture" (p. 1103-1104). For many, these more recent analyses offer, by example, *a posteriori* credence to the quality of Rosenberg et al's (2002) original findings and they reiterate the general utility of genetic clustering algorithms for detecting meaningful population structure with clarity of interpretation and statistical validity.

In a publication that followed shortly after Rosenberg et al. (2002), Bamshad and colleagues (2003) report on their similar use of the STRUCTURE program to aid in their detection of human population genetic structure, the identification of the number of genetic clusters present in their data set, and the evaluation of their clustering findings against the associated predefined populations. Their specific interest of inferring group membership is related to an acknowledged need for determining the utility

of common population proxies (e.g., race, ethnic affiliation, and continent of origin) for biomedical applications and epidemiologic studies and in clinical practice (e.g., disease prevalence, susceptibility and treatment protocols among and within certain groups). While their approach is exploratory, they frame their analysis around answering three interconnected questions: "how many data are required to identify population structure at a meaningful level of resolution, to what level can population structure be resolved, and do some proxies represent population structure accurately?" (Bamshad et al., 2003, p. 578). They assayed 100 *Alu* insertion polymorphisms in a diverse group of 565 individuals; 200 of whom were also typed for 60 microsatellite markers. In order to evaluate their ability to make unbiased and informative classifications, all individuals were stripped of any identifying information. Moreover, for their power analyses, their sample was subjected to repeated genetic clustering with the number of markers being varied across the runs of the STRUCTURE program. The results of their individual cluster allocations (i.e., inference of ancestry), and by extension, power analyses, demonstrate that significant genetic differentiation exists among the populations sampled. These test individuals were assigned to the correct continent of origin (Africa, Asia, or Europe) with a mean accuracy of at least 90% when 60 Alu or microsatellite markers were used. When the number of loci considered was increased to 100, correct classifications jumped to a remarkable 99%–100%. Even members of a "historically admixed" (Bamshad et al. 2003, p. 578) sample from southern India were assigned to the correct genetic cluster with a classification accuracy of 87%, suggesting, as Rosenberg et al.'s (2002) work did with the Kalash, that this mixture clustering approach is sufficiently powerful to allow for unique ancestries and particular patterns of gene flow to be detected and identified under the rubric of populational thinking. In corroboration of this case, previous studies have proposed that the genetic composition of the Indian population reflects contributions from European and Asian sources (Bamshad et al., 2001; Majumder, 1998, 2001).

When their STRUCTURE analysis is reviewed in detail for its detection of cryptic group structure, it is apparent that Bamshad et al. (2003) performed two different sets of analyses and constrained their clustering to evaluate the data under models using only values of K between 1 and 6. In their first test, they analyzed samples from Europe, sub-Saharan Africa, and Asia. In their second, they repeated the same analysis but excluded the Mbuti Pygmies from the African sample. For both cases, they provide the statistical evidence underlying their choice of clustering schema. When all samples/regions were included in the STRUCTURE analysis, the four component solution maximized the probability of the observed data; individuals were clustered into one of three geographic areas that corresponded to their continental origins (Europe, Africa, Asia), while the Mbuti, with three other Sub-Saharan Africans (one subject from each of the Nande, Hema and !Kung groups), formed their own distinctive group. In their second test, when the Mbuti were excluded from the analysis, STRUCTURE identified the best model fit as a three component solution. In this reduced configuration, groups were reliably formed in accordance with the three major continental regions represented (Europe, Africa, Asia), offering a *post hoc* validation the previous four cluster solution. As the second analysis confirms, the unique aspects of the Mbuti Pygmies were the driving force behind the increase in the cluster number, from K=3 to K=4. This fragmentation of the full Sub-Saharan African group into smaller subclusters is in reasonable agreement with external evidence of evolutionary significance that supports the existence of greater population structure and genetic diversity in Africa (Tishkoff & Williams, 2002; Tishkoff et al., 2009; Campbell & Tishkoff, 2010). It is interesting that the pattern observed here (main geographic clusters with a single small-population outgroup), closely reflects the structure detected by Rosenberg et al. (2002). This confirms, therefore, that STRUCTURE's mixture method has the ability to operate effectively at different levels of nested hierarchy while also suggesting that the original composition of the test sample may

affect the chosen number of clusters (K). It is probable that statistically derived estimates of K are in some degree susceptible to influential effects of sample size (N), population size (n), the number of loci used, and the degree of dissimilarity (e.g., as biological and geographic distance) between the populations (Bamshad et al., 2004). The effect of markers on clustering methods is suggested, for example, when Bamshad et al. (2003) claim for their clustering for classification that "[w]hen all 100 *Alu* markers were used, no individuals were misclassified" and that "the average probability of assignment of individual samples to the correct continent of origin was similar when the same number of *Alu* markers or microsatellites were used, although the performance of either set of markers varied among samples from different continents" (p. 584). Their findings are in keeping with the often conflicting levels of success reported by other contemporary projects. Some documented poorly defined distinct clusters but used only a modest number of markers and, in turn, likely suffered from low detection power (Romualdi et al., 2002). Other studies used a much larger complement of markers and showed stronger evidence of clustering (Stephens et al., 2001). Yet even the limited number of markers used by Bamshad et al (2003) still produced respectably high results: in the 70%-80% range using 20 markers (but with some questionably informative posterior probabilities). With 60 *Alu* markers or microsatellites the correct continent of origin was assigned for at least 90% of all samples.

The potentially high degree of *a priori* differentiation imposed upon the analysis by the use of samples from widely dispersed geographic locations has also emerged as a possible shortcoming to their study (see Bolnick, 2008), although Bamshad (2003) and coworkers had already identified this issue themselves, writing that:

> [o]ur analysis is based on samples from regions of Africa, Asia, and Europe that are widely separated from one another. Accordingly, these samples also maximize the

> degree of genetic variation among populations... Indeed, detection of population structure and assignment of samples to the correct genetically inferred cluster was less accurate for samples from geographically intermediate southern India. Importantly, the inclusion of such samples demonstrates geographic continuity in the distribution of genetic variation and thus undermines traditional concepts of race (p. 587).

To these points, it should be made clear that Bamshad's clustering of the Indian population nevertheless produced classification accuracies exceeding 80%, for both K=3 and K=4. These findings can be more profitably interpreted outside of the context of an outright validation or rejection of race, and, instead, first evaluated in terms of procedure, as Bamshad et al. (2003) explain in their closing discussion: "[t]his result suggests that information on the demographic and evolutionary history of a population is needed to determine whether a proxy can be used to more accurately infer population structure" (p. 587). Rosenberg et al. (2005) have also argued by example that when data reflecting more continuous geographic sampling is used (i.e., accounting for the overlapping patterns of trait distribution and clines), there is, in reality, little change to the cluster configuration. In light of these types of concerns, voiced over publication emphasis, procedural decisions and ambiguities in findings, Pritchard, Wen & Falush (2009) offer measured commentary on the demonstrated utility of these genetic clustering, and specifically STRUCTURE-based, approaches and the concomitant responsibility of the user and reader to bring to these, like for any other statistical analysis, an informed interpretation of their results: "[w]hile the computational approaches implemented here are fairly powerful, some care is needed in running the program in order to ensure sensible answers" (p. 3).

If this methodological discussion is extended to one that addresses the broader issues of effective procedure and proper interpretation in the context of the assessment of human genetic variation, it can facilitate a return to the problem of the traditional methods for assessing inter-racial diversity, using measure such as F_{ST}, and, now, offer further evidence for their poor relationship to the quantification of group differences by clustering and classification techniques. It is worth here reconsidering Rosenberg et al.'s (2002) study for its analysis of molecular variance. Using the same genetic data set representing 1056 individuals from 52 world-wide populations, they obtained estimates for the within-population component as 93%-95%, among populations within regions as 2% -5%, and among regions as 3%-5% of the total genetic variation, with the range depending upon the number of regions assumed (e.g., one, five or seven groups). They also determined that among the 4199 alleles observed, only 7.4% of these were found to be exclusive to one region. Moreover, these region-specific or private alleles were quite rare, with a median relative frequency of 1.0% in their region of identification. These values are in keeping with prior (if not even slightly more conservative) expectations for "low" values of "difference" and should be taken, in accordance with their traditional interpretation, as empirical evidence for the "non-existence" of human races. Nevertheless, Rosenberg et al. (2002) were also able to conduct, on the same data and present in the same study, a genetic cluster analysis which not only revealed latent, yet, well-defined group structure but also demonstrated a meaningful relationship between genetically inferred clusters and predetermined population parameters (labels of geography and identity). Rosenberg et al.'s (2002) analysis, among other recent studies of genetic variation at the individual level, convincingly suggests that the taxonomic implications of genetic information emerge most advantageously from correlations among many different loci in large samples sizes and indicate clearly how these correlations are what make classification possible. This new evidence for the biosocial significance of population-based classification, despite the estimation of small among-population

variance components (see also Shriver et al., 1997, 2003; Bamshad et al., 2003), has been used as support for mounting arguments (1) that the relative proportion of variation within-populations to among-populations serves no value as a criterion for judging the validity of races and (2) that bio-geographic and bio-cultural categories (like, races, ethnicities and ancestries) are not without some biological meaning (c.f., Long & Kittles, 2003; Bamshad et al., 2004; Jorde & Wooding, 2004).

The studies by Rosenberg (2002), Bamshad (2003) and their colleagues have effectively isolated population substructure at intra-continental levels. They have related genetically inferred clusters with success to their corresponding social (ethnic, population, race, ancestry) groups and have achieved reasonably high classification success when assigning individual memberships. These studies are, however, designed from a macro-population perspective under the assumption of global sampling, where emphasis was ultimately placed upon their ability to detect continental groups or differences between the three or four major geographic regions (e.g., Africa, Asia, Europe, and, only sometimes, the Americas). As it has been shown that sample composition can influence the outcome of the clustering procedures and rarely would practical uses of these results require a world-wide perspective but, in fact, would be better suited by studies that consider the particular demographic configuration of a given population-at-large, it is not only appropriate but necessary to also consider the application of these methods to more geographically limited data. The American population in particular makes for an ideal analytical choice as it offers a more direct correspondence in contemporary samples (drawing from the most populous U.S. groups or self-identified "races," e.g., African American, White, Hispanic) between genetic data used in these STRUCTURE or *structure*-like analyses and the phenotypic, specifically craniometric, data of interest to forensic anthropologists concerned with human identification in medico-legal contexts.

As obtaining a reliable estimate of the proportion of recent genetic ancestry that a given individual shares with members of one or more groups both enables more accurate predictions concerning disease and healthcare and increases the power of studies that focus on the environmental and genetic factors that are linked to the individual's expression of "health"-specific traits, the quantification of the relationship between self-identified race or ethnicity and the genetic structure that exists in the U.S. population is of growing concern. It is in light of the prevalence of case-control studies that are using racial data in the United States that Tang et al. (2005) have recently investigated this relationship using specifically American samples. Drawing upon samples collected for the Family Blood Pressure Program (FBPP) study on the genetics of hypertension, they obtained genetic data for 326 microsatellite markers that were genotyped uniformly in what they describe as "a large multiethnic population-based sample" (Tang et al., 2005, p.268). These subjects represent 15 geographic sampling locations, consisting of 11 recruitment sites in the continental United States, one in Hawaii, and three in Taiwan. All individuals in this study self-identified with one of four predefined racial or ethnic groups: White, African American, East Asian (Chinese and Japanese), or Hispanic (Mexican). Taking an approach similar to Rosenberg et al. (2002), but using it to investigate this primarily U.S.-based sample, they assessed individual genetic similarity at the level of these intra-continental populations with the STRUCTURE software. They limited their analysis to the "NOADMIX"option, thereby assuming that any individual's total genome was derived from a single homogeneous population (Pritchard, Stephens & Donnelly, 2000; Pritchard, Wen & Falush, 2009). They tested the clustering solution under sequential values for the hypothetical cluster number ($K_2...K_n$) but found that when K >4, STRUCTURE produced "unstable results" (e.g., empty clusters, inconsistencies in configurations). They therefore reported on their findings for 2, 3 and 4 clusters only. When K=2, the Asian sample (mixed Japanese and Chinese) emerged as a single, separate

cluster. When K=3, the African American group clustered independently from the White and Hispanic samples. When K=4, an almost exclusively Hispanic (99.8%) cluster emerged. All but two individuals had a posterior probability of membership approaching 1 for their true cluster of origin and only two subjects had a probability <0.95 (one self-reported as Hispanic and one African American, both of whom were misallocated to the White cluster). For all of the 3636 subjects, classifications made based on genetics and self-identified racial or ethnic group differed only for five individuals (0.14% of the total sample). From these configurations (with particular emphasis on the test where K=4), the authors conclude that their "cluster analysis results are completely consistent with previous theoretical predictions regarding the ease of separating these groups on the basis of the number of markers tested" (Tang et al., 2005, p. 272; c.f., Risch et al., 2002). They also posit that, because they chose not to evaluate their samples under the model with admixture (which would also necessitate their inclusion of subjects representing the parental populations, e.g. Africans, Native Americans), their results show "that the (admixed) groups included have approximated within-group random mating sufficiently long enough to give rise to distinct genetic clusters" (Tang et al., 2005, p. 272). In further explanation of the potential effect of differentiation time, they write that their "genetic cluster results indicate that older geographic ancestry—rather than recent geographic origin—is highly correlated with racial/ethnic categorizations and, thus, is the major determinant of genetic structure in the population" (p. 272). They do, however, suggest that additional studies attempting to detect within-group substructure are warranted for Hispanic subjects in particular given the known differences in source populations (with different geographic origins) for the many subsets that make-up the Hispanic population. The authors conclude that their study emphases the importance of personal population identifiers at all levels of population genetic analyses and in clinical contexts. They argue, specifically, that, while only "modest" geographic differentiation within each [sub]population was observed, when the "four major racial/ethnic groups"

within the United States are assessed, a clear and meaningful correspondence between genetically inferred clusters and the biomedically employed "SIRE" (self-identified race and ethnicity) categories was reliably identified using STRUCTURE's model-based clustering approach (Tang et al., 2005, p. 274).

While a considerable amount of time has been spent discussing the data, statistical procedures and the analytical premises upon or around which these three studies, and others like them, have been built, the larger current of thought on race and ancestry that is both running through these reports and influencing the ways in which their findings have been described, interpreted and assigned meaning by scholarly readers within the fields of population and anthropological genetics or by writers of popular science should not be ignored. Take, for example, Risch and coworkers (2002) who have argued simultaneously that (1) genetic differences among populations create clusters of discernable, major groups, which correspond to the classical definition of races based upon continental ancestry, (2) that "effectively, these population genetic studies have recapitulated the classical definition of races based upon continental ancestry..." and (3) that a case can be made for "great validity in racial/ethnic self-categorizations, from both the research and public policy points of view." (p. 1,3). It can be imagined how easily, with a poor understanding of their context, simple causal connections between genes and old views of racial biology can be drawn from statements such as these. The underlying complexity in the conclusions derived by these type of genetic clustering studies on the "meaning" of geographically structured, population-based clusters has been often underestimated. As a result, many of these studies have been misrepresented as providing biological cause for fixing racial boundaries along strict continental lines or defining artificially delimited populations. This thereby gives weight to interpretations that both trivialize the cluster findings, oversimplify what is known to be true about the inherent variability within the patterns of human diversity, and, as Bolnick (2008, p. 74) argues, validate

"traditional ideas about race" (c.f., Wade, 2002; Burchard et al., 2003; Mountain & Risch, 2004). A brief discussion of the state of scholarship on the meanings of ancestry and race and their use within the current trends in genetic literature is important to not only shed clearer light on the relationship between these concepts, but to also establish a frame of reference for their interpretations in genotypic discussions of population structure and classification that can be readily accessed when they are reconsidered in the context of forensic practices for the identification of human skeletal remains. To this end, it will be necessary to revisit Sauer's (1992) well-known question: "[i]f races don't exist, why are forensic anthropologists so good at identifying them?"

Teasing apart the concepts of ancestry and race in contemporary genomics is a difficult task: efforts in this direction are confounded by the language of "science" versus "society," the variability in the criteria used to define each term, and the problem of how the use of ancestry and race differs in theoretical discourse and actual practice. It is for these complexities that consistent, distinctive, and widely usable definitions for genetic ancestry and race have remained elusive in even the most rigorous of scholarly research and analysis (Koenig, Lee & Richardson, 2008; Whitmarsh & Jones, 2010). The political, philosophic and social issues surrounding the word "race" grew in their intensity during the 20th century as it became increasingly difficult to justify traditional concepts of bounded, isolated peoples in the presence of both the unprecedented degrees of global population exchange and the new approaches to biological science that were providing convincing arguments for the "idea of a 'single family of man'"(Fujimura et al., 2010, p. 169). There emerged, therefore, a strong reactionary movement within mainstream scholarship that advocated for the outright rejection of race. This wide-spread emphasis on its "non-existence" may have assuaged broader societal concerns, but it also introduced new problems of practical concern to academia. While race categories were losing their validity as meaningful

organizing principles in the biological sciences, an interest in and a need for a means to speak to the acknowledged variation in genotype and phenotype within the human "race" continued. There emerged, therefore, an obvious operational dilemma. Even outspoken proponents of the anti-race view, such as Lieberman and Kirk (2001), were sufficiently compelled by this quandary to ask, "now that you have taken race away from us, what will fill the void?" In its place, the concepts of ethnicity and ancestry were, and still are, promoted today, by, for example, institutions such as The Race, Ethnicity, and Genetics Working Group, as "alternative" and implicitly better ways to characterize, respectively, the [bio]sociocultural versus the [more purely] biological differences between human groups (Berg et al., 2005). Discourse on race has been, therefore, actively "sanitizing" and "depoliticizing" racial and related concepts by using "populationist terminology and frames"(Reardon, 2005; Fujimura et al., 2010, p. 169) in such a way that "in recent years, ancestry has been widely promoted as an objective, scientific alternative to race" in genetics and across the human [social] sciences (Bolnick, 2008, p. 80; Shields et al., 2005). Although ancestry is interpreted with consistency, when invoking its most fundamental meaning as some representation of the biological component of human variation, the full implications of its use and its relationship to race are often lost on researchers and readers alike. The causes and consequences of this are explained by Feldman and Lewontin (2008): "both the regional heterogeneity within major geographic regions and the widespread mixture of formerly isolated populations results in a confusion between race and ancestry that is critical and must be accounted for..." (p. 98). Jorde and Wooding (2004) have similarly argued that ancestry "is a more subtle and complex description of an individual's genetic makeup than is race"(p. S30) and, as such, it is both preferred for and preferenced in scientific discussions. The result of this confusion is that, in reality, this *meaningful term*, "ancestry," is used ubiquitously but without being clearly defined *for its meaning* (Bolnick 2008, p. 80). If a selection of sources that offer explanations of ancestry in genetics is compiled, however, it is possible to come to a

current consensus for its definition: ancestry refers to the geographic origins of an individual's ancestors (Collins, 2004; Jorde & Wooding, 2004; Bamshad et al., 2004; Berg et al., 2005; Bolnick, 2008; Fujimura et al., 2010; Fujimura & Rajagopalan, 2011). Because of this focus, which rests upon the legitimizing factor of a long-past time and place, ancestry is, therefore, seen as more objectively historical and scientific and more reliably estimable using the statistical genetic toolkit (Bamshad, 2005; Fujimura & Rajagopalan, 2011). The Race, Ethnicity, and Genetics Working Group, acting as a point of reference for genetic researchers, has, however, recently recognized alternative views on ancestry that encompass geographic (e.g., European, Asian, sub-Saharan African), geopolitical (e.g., Italian, Chinese, Zambian), and cultural (e.g., Brahmin, Kalesh, Apache) identities. These multiple lines of [bio]historical, [bio]social, and [bio]geographic emphasis are taken into account by more flexible definitions, which also bring to the fore a level of interchangeability in the terminology of race, ethnicity and ancestry and state more overtly that some genetic component exists within all three of these types of classifiers. Consider, for example, the definition given by Risch et al. (2002), where "ancestry refers to the race/ethnicity of an individual's ancestors, whatever the individual's current affiliation" and, so, "from the genetic perspective, the important concept is mating patterns, and the degree to which racially or ethnically defined groups remain endogamous" (p. 3). Here, the ways in which ancestry is not only distinguished from race and ethnicity by time, but also how it is, simultaneously and complexly, linked to race and ethnicity through its ideas of diachronic heritage and geographic origin are apparent. Moreover, this definition also presents ethnicity and race in their common use as freely exchangeable concepts (Oppenheimer, 2001), i.e., members of an ethnic group share in the belief of a common origin, which is, itself, a defining criterion for a race (Cornell & Hartmann, 1998). Even in its most strict geographic definition, ancestry remains problematized by the ambiguous parameters that inform its estimation and the limits that these impose on its ability provide reliable and useful classifications (Elliott & Brodwin,

2002). An explanation can be found in the American Society for Human Genetics (ASHG) outline of the factors that influence the accuracy of any inference of genetic ancestry: (1)" how underlying patterns of human genetic diversity are distributed among populations," (2) " how that diversity is surveyed (i.e., which genetic markers are used and how many), " (3) "which populations are used as references, " and (4) "the statistical methods used to interpret patterns of variation" (ASHG Ancestry Testing Statement, 2008, np). What is immediately apparent from this list is that there is considerable opportunity for interpretation, so that the factors that underlie these estimates of ancestry are, as Fujimura and Rajagopalan (2011) have experimentally observed, "*known only through the data and data producing technologies of the geneticists*" (p. 11). As ancestry is estimated by inference (e.g., by statistics, observation, self-identification), it is unavoidably weakened as a concept by the lack of procedural consensus. There is, for instance, little protocol for how geographic regions should be defined, which areas should be represented in a given study, and what samples are truly representative of the variation under study. There is equally little guidance on the depth of time that should be considered when invoking the ideas of geographic, cultural, linguistic "origins" (i.e., an evolutionary, historic, or a *T*-generational timeframe). These decisions are ones typically made with thought given to the questions of interest or the hypothesis being tested and they are mediated by the availability of resources. Few studies explicitly state how these choices impact their operationalization of the ancestry concept (Fujimura et al., 2010; Fujimura & Rajagopalan, 2011). Finally, it has been observed, both throughout the current body of genetic (and especially biomedical) literature and in the, often critical, appraisals of current anthropological and medical genetic practices, that, in its actual application, "ancestry is not that different from race" (Bolnick, 2008, p. 81) and, along with ethnicity, race and ancestry are "often used interchangeably" (Risch et al., 2002, p. 3). This fluidity is made possible because (1) most studies interested in population structure and in individual ancestry make use of continental groups and (2)

both anthropological/population genetic interpretation of ancestry and public perceptions of "race" are largely based upon these same types and levels of division (Bolnick, 2008). The one drop-rule remains perhaps the greatest impediment to this interchange between race and ancestry, because the former has historically dictated single group membership, while the latter allows for mixed origins. But, as the race "options" in the most recent U.S. censuses (2000, 2010) reveal, the conceptually similar idea of "mixed-race" or "admixture" is prevalent in both institutional and public understandings of individual and self-identity, so that it appears in the contemporary American "genetic melting pot," both conceptual and biological barriers between race and ancestry are easily transgressed (Feldman, *Lewontin* & King, 2003). As a result of this potential for exchange, professional and public terminology overlaps, concepts are shared, and meaning is blurred so that the lexical distinction between ancestry and race (and even ethnicity) does not necessarily exist in the label used but is made clear instead by the specification of the intention of its use, e.g., the estimation of ancestry, the identification of race.

While the focus here has been on studies of population structure and personal identification from the perspective of contemporary genetics, the importance of this material to all aspects of human biological variation should be imminently apparent. The promise shown by these analyses for the meaningful, testable, and statistically robust detection of population structure and individual biosocial or geographic classification using (1) genotypic traits and (2) the STRUCTURE software's finite mixture or model-based clustering methodology suggests the possibility of obtaining similar levels of resolution using phenotypic traits, like skeletomorphology and specifically craniometrics, which is already accompanied by an established statistical tradition of by-group discrimination and classification. The rapid and consistent emergence of studies like these and the considerable debate that they have inspired provide reasonable justification for pursuing a model-based cluster analysis of human craniometric variation at both the

global population (inter-group) and regional subpopulation (intra-group) levels of structure hierarchy and an assessment of the correspondence between skeletally-inferred population structure and the popular concepts of ancestry and race that are central to forensic casework.

Chapter 3: Finite Mixture Model-Based Clustering

Classification and Clustering

"'But, as you have already pointed out, we do not need any more disjointed clues,' said Bartholomew. 'That has been our problem all along: we have a mass of small facts and small scraps of information, but we are unable to make any sense out of them. The last thing we need is more.'"(Gregory, 2003, p. 1.1)

The current information society has led to an increasing availability of data for modeling and analysis: individuals, corporations and institutions alike can access with ease great quantities of such "small facts and small scraps of information"about themselves and their surroundings, and, in turn, compile large scale data stores. Although this data has the potential to offer insights into trends within the larger external environment and predict the developmental trajectory of variables of interest, the sheer quantity of scattered and seemingly unrelated information in databases, archives, and independent records has hampered efforts in data synthesis and frustrated investigators attempting thorough and meaningful analyses (Giudici, 2003; Larose, 2005). The issues of data organization, processing and interpretation are not however unique to the digital age. As the concern expressed by Bartholomew, a mediaeval physician cum forensic investigator, attests, the fundamental problem of *making any sense* of these numerous "facts", "scraps of information"and "clues"is one that transcends all temporal, spatial and technological boundaries, be they artificially defined or organically conceived. Despite ongoing arguments over the cognitive signatures of *Homo sapiens sapiens* and debates on the qualities that define being human, researchers in evolutionary psychology and the complementary fields of behavioral

ecology and genetics have recently amassed a growing body of literature that speaks to the role of pattern recognition in the early formulation of the human mind as well as to the evolutionary stability of information management as, what many consider, an innate aspect of the human behavioral condition.

The most essential, yet, most general of human motivations is the need to organize, exert influence upon and ultimately control the surrounding loci of interaction. In "bringing the environment into line with one's wishes,"human beings can manipulate their social, physical and biological environments to best support their own immediate survival, the reproductive success of their offspring, and the preservation of their genetic heritage by securing the welfare of their expanded kin groups (Heckhausen & Schulz, 1995, p. 285). This is so because reproduction and survival are inextricably tied to the acquisition of physical (e.g., shelter) and biological (e.g., food) and social (e.g., relationships) resources (Alexander, 1979; MacDonald, 1988). To this end, the human mind is designed with the key neural, perceptual and cognitive processes that permit the extraction of certain types of information from many, diverse environments. The manipulation of this information is essential to achieving the individual and species-wide goals that enhance survival and reproductive success. The adaptive challenges characteristic of the human evolutionary path can be understood as reflected in humans' species-specific data processing mechanisms, which serve as the interface between the evolutionary significant objectives of the individual and the environments in which these objectives are realized (Cosmides & Tooby, 1994; Tooby & Cosmides, 1990). Selection pressures should favor, therefore, the individuals who most effectively organize their universe in ways that best exploit their environments and sustain successful reproduction and survival strategies. Moreover, it should favor the development and maintenance of the cognitive and functional systems that support universal mechanisms for recognizing key information, for undertaking its organization, processing and interpretation, and ultimately

controlling the environments related to it (Geary, 1998). For example, ethnographic studies of naming conventions for folk biological taxa bear out these assumptions on the universality of organizational methodologies. As Berlin and colleagues (1973) note, "it has become apparent that, while individual societies may differ considerably in their conceptualization of plants and animals, there are a number of strikingly regular structural principles of folk biological classification which are quite general" (p. 214). Furthermore, there exists a similarity between the folk nomenclature of plants and animals among preliterate societies and the scientific taxonomy of the same organisms (Altran, 1994; Diamond, 1966). Everitt and coworkers (2001) have generalized this relationship by explaining the inseparable relationship between the act of naming and the act of classifying. Everitt (1993) even emphatically declares that "*Naming is classifying*" (p. 1). These observations are in keeping with the culture and language theory developed by Foucault (1966, 1969, 1971), where the basic human drive for organization is hierarchical in nature and always classificatory. If as Foucault posits, to speak is to classify and language is the action of thought, the importance of the universality of the human approach to the organization of data, to understanding new phenomena, and to learning new objects by focusing on the actions involved in classification cannot be overstated. A new identity for an object or a new understanding for a pre-acknowledged phenomenon is achieved by (1) identifying descriptive features, (2) comparing these features to those of known objects or phenomena, (3) isolating the areas of similarity or dissimilarity, (4) generalizing these differences (or distances) in terms of "proximity" to other "like" entities, and by (5) basing membership criteria on established measures, sets of rules or conditions. Drawing upon this universality, Xu and Wunsch (2009) consider classification as "one of the most primitive activities of human beings" (p. 1) and, so, they have assigned it "an important and indispensible role in the long history of human development." It is not surprising then that the classification of data into a set of categories is considered one of the most important approaches to data

analysis and that the issues of information organization, processing, and interpretation are resolved most naturally by capitalizing upon the innate human drive to apply some form, be it instinctive or statistically developed, of cluster analysis (Anderberg, 1973).

In its most general terms, "clustering" refers to "the process of assembling in various sets objects which behave similarly to each other" (Bock, 1994) and, so, it implies a concern for the identification of "homogeneous groups of objects" (Arabie et al., 1996). It follows, therefore, that a "cluster" can be broadly defined as "a set of entities that are alike," and, so, it also stands to reason that "entities from different clusters are not alike" (Everitt, 1980). The pervasive appearance of informal definitions such as these in otherwise rigorous statistical literature exposes what is often considered the fundamental weakness in cluster analysis: the unresolved problem of formulating a single formal, yet sufficiently inclusive, definition for the process of clustering, and, in turn, for its product, the cluster. Everitt (1993) even posits that, in fact, a "formal definition is not only difficult but may even be misplaced"(p. 6). What Das et al. (2009) have considered as an informality in technical language has also lead, for some, to an equal casualness or, even, an acceptable ambiguity in technique; explaining cluster analysis, as Kaufman and Rousseeuw (2005) have, as "the art of finding groups in data" (p. 1) or suggesting, like Bonner (1964) did, that the value judgment of the user is the ultimate criterion for evaluating the meaning of cluster and clustering. When scientific rigor is among the first concerns when deciding upon the most satisfying method, it is hard to disagree with Bock (1994), that "such a formulation is too vague" (p. 4). This project attempts to offer here a preliminary, widely applicable but still dynamic definition of both "clustering" and a "cluster," through a brief explanation of the procedure of cluster analysis: what clustering does, what it uses, and what it produces.

When considering a matrix of information or set of observations in need of organization, cluster analysis seeks to tease out hidden structure, i.e., patterns of association already present, but possibly latent, within the larger data mass. It does so by finding subsets of interest and establishing these as groups for further study. Particular clustering methods may also be employed to impose a data structure on a set of known homogenous observations. In this way, it is used to produce statistically derived partitions in order to split-up the given data into subsets defined, for example, by internal factors, like group size, external factors such as geography or income bracket, or, by some common feature, like a biological attribute. In clustering, the data set is partitioned into [sub]groups or "clusters,"so that each member of each subset shares some common characteristic, in the univariate case, or combination of traits, when conducting multivariate cluster analysis. These groupings are typically described by a multidimensional vector of *a priori*-selected continuous, discrete (including binary), or mixed variables representing any of the four levels of measurement (nominal, ordinal, interval, ratio). The boundaries of each cluster within the *d*-dimensional feature space are delineated by some subjectively chosen measure of similarity/proximity or dissimilarity/distance, test statistic, or information criterion. "Clusters" summarize the relationship(s) among the data and can be explained by their patterning, using properties such as "internal cohesion" and "external isolation" (Cormack, 1971; Gordon, 1980). Although clusters can be interpreted as continuous regions within the test space, their high density of points separate them from other regions with low levels of observation aggregation. Relationships among like-observations are, therefore, represented by groups within the full data set, which are smaller, relatively homogenous and well separated.

Even though clustering is acknowledged as a basic, instinctual, human activity that is necessary for information organization and retrieval and it is universally known as a key tool used in scientific inquiry,

Ereshefsky (2000) has pointed out that "cluster analyses vary: first, on the breath of similarities desired among members of a group, and second, on the relationship between similarity and theory" (p. 24-28). For this reason, Bock (1994) has argued that any attempt at engaging in formal classification activity, that is, "clustering" or conducting "cluster analyses," requires an investigation into the fundamentals of the procedure. It is worthwhile, therefore, to, first, pose the following questions:

1. What is a cluster? Is it a set of objects, a construct, or a concept?
2. How is a cluster or class characterized? Is it best defined by similarities among values across variables of interest, the presence of a special feature, the resemblance among group members, or by some likeness to a prototype?
3. What defines a classification system or the clustering structure? Is it a hierarchy, a network, or some other conceptual framework that is project-specific?
4. What structural relationships exist among the clusters? Is there overlap among clusters (fuzzy vs. hard partitions), a nested structure, or a set of semantic relations?
5. How is the quality of clustering defined? Are ideas of good vs. bad, weak vs. strong, meaningful vs. statistically significant cluster solutions useful to the practitioner and the audience?
6. What is the ultimate goal of the analysis? Is the clustering technique implemented for the purpose of information organization, retrieval or compression, classification, detection of latent structure, or for group discrimination?

While these questions may appear oversimplified, given the complexities associated with making the necessary procedural decisions (e.g., choosing the "best" algorithm, number of clusters, or model), they are still critical to producing a cluster solution and giving the appropriate interpretation to its results. It is very curious, therefore, that they continue to remain poorly resolved. A multiplicity of "correct"answers has been given for these questions and few of these may actually prove relevant to a

practitioner's interests or reasonable for the specific conditions of any given data set. This is because, in the absence of formal procedural rules, the protocol implemented in many cluster analyses is often heavily dependent upon the unique combination of disciplinary and philosophic positions of the theoreticians who explain the techniques and practitioners who make use of them. It should come as no surprise then that an equal, if not greater, number of ideas have been put forth that represent the singularly optimal approach to cluster-based analyses. "Best practice" is, as a result, highly subjective: there are several, frequently modified, personal or disciplinary defined "good" approaches for mining, organizing, and analyzing every type of data. Clustering has experienced, therefore, a lengthy and complex intellectual history. Understanding its developmental path, and being aware of the baggage that way collected along this journey, is key to appreciating the value in using the finite mixture modeling approach to clustering that has been adopted here for the analysis of craniometric data.

Theoretical discussions of clustering principles can be traced back to Classical philosophy, where classification is implicit in Platonian ideas of the "universal"and explicitly addressed in Aristotle's *Departibus Animalium* and *Historia Animalium* and among the works of his followers. Theophrastos, for example, produced the first comprehensive account of plant classification, the *Historia Plantarum*, for which he became known as the father of botany. More importantly, it was with this work that he laid the foundation for future studies that adopted an objective approach to determining the maximum number of groups of organisms, to defining group-specific physical, ecological, and behavioral characteristics, and, ultimately, to the ordering of species in terms of a super or sub-class hierarchy (Greene & Depew, 2004). It is no surprise then that clustering received extensive attention from the 18th century naturalists, including Buffon, Cuvier and Linnaeus, who used it in support of their interest in biological categorization and their pursuits in what has come to be known as "taxonomy." Its

importance is shown, for example, in the writings of Linnaeus on the practice of taxonomy in his work, *Genera Plantarum* (1737): "All the real knowledge that we possess, depends on methods by which we distinguish the similar from the dissimilar...[t]his is the great work, the important labour, for should the genera be confused, all would be confusion" (Everitt et al., 2011, p.1).

The utility of clustering was also recognized among practitioners in other areas of academic thought and in more recent history: its effect on science can be seen, for example, in Mendeleyev's structuring of elements in the periodic table and his interpretation of the atom in the 19th century, or in the early 20th century theories of stellar evolution given by Hertzsprung and Russell; while, in philosophy it has continued to play a key role among logicians in discussions of classification systems and concept theory (Sutcliffe, 1993; Hansen & Jaumard, 1997). Despite its ubiquitous use, it has experienced a long period of tacit acceptance. The formalized term, "cluster analysis,"appeared late in history, originating in the work of the American psychologist, R.C. Tryon, in 1939. Moreover, it is only in 1963, with the publication of "Principles of Numerical Taxonomy,"that Sneath and Sokal provide the first systematic treatment of the computationally intensive, algorithmic approach to cluster analysis. Their routines of self-organization and classification fully introduced the polythetic, or multi- character, system that has come to define the statistical tool that is now called "clustering." In doing so, they not only filled the methodological gap in allocating data objects into homogenous groups in the age of burgeoning computer technology, but they also encouraged the codification of the methodologies for identifying aggregations and, in turn, defining "clusters." Sneath and Sokal's mathematical programming, thereby, paved the way for the emergence of the new discipline of "numerical taxonomy," "automated classification" and "cluster analysis" proper, and, ultimately, for the establishment of the formal theories in computational statistics

that underpin the many iterations of clustering methods that are in popular use today, including the mixture likelihood approach presented here.

Mixture: Development

"I hope that you flourish in probabilities."(Edgeworth, 1893, correspondence with Pearson)

A lengthy developmental trajectory can be traced for the history of the mixture likelihood approach to clustering. The origins of finite mixture analysis, mixture modeling, or model-based clustering and the related latent class models can be located in the first half of the 19th century; they are evident in early developments in probability theory and distribution fitting, and saw then their first applications in the social sciences (Stigler 1986, 1999). While the work of Bernoulli and LaPlace foreshadowed the contribution of probability to the measurement of uncertainty in social data, it was, for example, Quetelet's concepts of the "l'homme moyen" (1835), "la courbe de possibilité" (1846) or "binomiale" (1870), and Poisson's (1837) reanalysis of a data set of French conviction rates from 1825-1830, using probability-based distribution models, that made for a directed effort. Through their studies, they set out to address the concerns expressed by Keverberg (1827) for demographic research as well as the need for widely applicable and sensitive methods for classification, comparison and the detection of [in]homogeneity in the emerging area of "social physics." Although ultimately inadequate and, in parts, flawed, the ideas put forth by Quetelet, Poisson, and others, like Compte, Poinsot and Bertillon, were nevertheless boldly provocative in their time and, so, very influential on the progression of statistical

thought. By expanding upon and refining these earlier attempts at population interpretation, organization and membership prediction, key figures such as Galton, Edgeworth, Newcomb and Pearson brought about a significant shift in the intellectual climate of the 1880s. Their movement towards the construction of an empirical, conceptual methodology offered a replacement for experimental control and, therefore, lifted the barrier that had been impeding progress in mathematical statistics and its application to questions of demographics, economics, and biometrics for at least a century. As Stigler (1986) has correctly argued, Pearson's research in particular "carried the seeds of a 20th century revolution" (p. 341). Pearson's legacy would come to fruition with mathematicians like Fisher in the 1920s and later with Fisher's own student, Rao, in the 1940s and 1950s.

In addition to his ground-breaking work on correlation, Pearson (1894, 1895) has been credited with laying the theoretical foundations for mixture modeling in his monographs on asymmetrical frequency or skew curves and, most importantly, with his development of the method of moments. Growing out of his interest in statistical graphing methods and relying heavily on the mathematical tools applied more typically to questions of mechanics, Pearson set out to develop a method for population-specific biometric data. His goal was to analyze a data set of forehead breadth to body length measurements taken on 1000 crabs, representing *Crangon vulgaris* and *Carcinus moenas*, sampled from the Bay of Naples and made available by Weldon (1892, 1893). These crown-rump measurements were of particular interest given that a "double-humped" frequency curve was previously observed and that Weldon (1893) has already given this statistical phenomenon an evolutionary interpretation: this skewness suggested intra-specific change and signaled a movement towards the emergence of a new crab subspecies. In the 1894 publication from his series, "*Contributions to the Mathematical Theory of Evolution*," Pearson successfully resolved this asymmetrical frequency curve into a mixture of two

normal heteroscedastic components using a moments-based fitting method. Although such curve dissection was implicit in Quetelet's work and explicitly discussed by Galton (1869) in "*Hereditary Genius*," Pearson was the first to extend a clear mathematical and, so, structured methodology to this approach. Pearson has explained in detail how he fitted to Weldon's data sate a mixture of two normal probability density functions with different means (μ_1, μ_2) and variances (σ^2_1, σ^2_2) and in two proportions (π_1 , π_2) in order to accommodate the apparent skewness of the data, which could not be resolved by the normal symmetric distribution. From these steps, he obtained his moments-based estimates for the five parameters of the mixture model as a solution of a ninth degree polynomial.

It is known that, for certain combinations of sampled data, the nonic equation may have multiple negative roots, and, as a result, multiple estimates may be produced (Cohen, 1967). Pearson's (1894, p. 89) analysis yielded two sets of results. It is possible to calculate from his figures, that 41.45% of the population was allocated to component one and 58.55% to component two, by his first solution, while 53.28% was allocated to component one and 46.72% to component two, using his second solution. Pearson has explained that, even though the compound curves and solutions are very close, the first solution is the "required one"as its estimates produced the closest agreement between the sixth central moment and the corresponding moment of the fitted compound curve, and it is in "good accordance"with Weldon's other measurements. Pearson's structured and mathematically explicit approach to addressing Weldon's evolutionary-developmental explanation for the asymmetry in the sample histogram places this paper among the earliest studies, if not, the "first ever," "to advocate statistical analysis as a primary method for studying biological problems."(McLachlan & Peel, 2000, p. 2). Despite the promise of Pearson's moments-based method, the complexity of the moment estimators and intensive computational demands underlying this initial foray into finite mixture analysis

discouraged its practical application. As Charlier (1906) remarked, "The solution of an equation to the ninth degree, where almost all powers, to the ninth, of the unknown quantity are existing, is, however, a very laborious task. Mr. Pearson has indeed possessed the energy to perform this heroic task in some instances in his first memoir on these topics from the year 1894. But I fear that he will have few successors, if the dissection of the frequency curve into two components is not very urgent"(cited in McLachlan & Peel, 2000, p.3). Charlier's predictions were in part correct – no notable studies were produced that adhered faithfully to Pearson's method of moments. Several attempts were made, however, in the early part of the 20th century to either simplify the calculus by eliminating, in particular, the need to find the root(s) of the nonic, or to apply his basic ideas towards the development of diagnostic plotting procedures for the graphical dissection of a mixture of distributions. Charlier, himself (1906), and, with Wicksell (1924), extended Pearson's method to the case of bivariate normal components; Doetsch (1928) considered its applicability to more than two univariate components; and Rao (1948) tested the use of Fisher's k-statistics. These studies all assumed normality for the true components and the special case of equal variances, so that, accordingly, only the first four moments or cumulants were necessary. It is Cohen (1967) who has been credited with being the first to translate Pearson's project into a usable method. His procedure circumvented the necessary estimation of a negative real root of a polynomial equation of the ninth degree by introducing an iterative process that, instead, involved solving a cubic equation for a unique negative root. Cohen was successful in simplifying the estimators, correcting for Pearson's problem of multiple solutions, and in offering a more practical extension of the method of moments beyond the special case of equal variances. In a similar but later attempt at simplification, Quandt and Ramsey (1978) extended Pearson's procedure for estimating the five unknown to moment generating functions. Despite their contributions, these techniques failed to achieve critical acceptance within the applied statistical community. Even a half century later, Charlier's

original concern still remained at the fore of criticism: "it seems to us that all these contributions suffer from over complexity and the necessity for lengthy calculations"(Preston, 1953, p. 460). The intractability of even the revised moments-based methods, especially in the absence of modern computer technology, refocused research attention away from fitting the model using formal estimation methods and, instead, onto the applicability of more informal graphical techniques (Fowlkes, 1979). For example, variations on the normal probability plot were used by Hald (1948, 1952), Harding (1948) and Cassie (1954) to obtain parameter estimates for a mixture of two normal distributions; Oka (1954) and Bhattacharya (1967) proposed the use of plots of the logarithms of the ratio of the observed frequencies in nearby classes; while Preston (1953) combined the method of moments with a skewness-kurtosis diagram. As all of these adaptations to Pearson's method are "rapid, and much less laborious"(Preston 1953, 460), a variety of moments-based and graphical techniques have continued to be developed for resolving mixtures of underlying probability distributions into their components: Tarter and Lock (1993) introduced the curve fitting, "λ", method; Roeder (1994) developed a graphical technique (the mixture detection plot, MDP) and accompanying "check"(the mixture detection test, MDT) designed to determine the number of components in a mixture using a kernel density estimator; and Dacunha-Castelle and Gassiat (1997) characterized the number of support points of the mixing distribution using the Hankel matrices of the first algebraic moments. Of particular interest to anthropological research is paper by Josephson and colleagues (1996), in which they argued that the method of moments is a better alternative to the coefficient of variation and mean methods when addressing questions of sexual dimorphism, when using anthropometric data. Nevertheless, moments-based and diagnostic plotting procedures have not undergone a concentrated, discipline-wide effort for development, improvement and refinement; consequently, they have not received their codification within the canon of standard statistical practice. Instead, it appears that, in very recent years, they have been largely displaced by a

growing trend to estimate the parameters of a mixture model by maximum likelihood (ML) methods (McLachlan & Peel, 2000). Its popular appeal reflects, in part, the ease of implementation of these methods given new developments in computational statistics, including high powered processors and efficient estimation via the EM algorithm, as well as the simple fact that ML approach yields better estimates for finite mixture models (Fryer & Robertson, 1972). The beginning of last quarter of the 20th century marked, therefore, the second major shift in the intellectual climate towards improved methodology. It affected both the theoretical approaches to the study of mixtures and the mechanics of the practical implementation of model-based clustering using finite mixture methods.

The increased interest in the identifiability of mixture models, the advent of high-speed computer technology, and the emergence of elegant iterative estimation procedures have dominated the intellectual climate of the last few decades. These new developments have also contributed the growing favor of using maximum likelihood estimation over other methods for the problem of decomposing mixtures. Nevertheless, the 19th century astronomer and statistician, Newcomb (1886), should be recognized for his contribution to the approaches popularized today. He was among the first to suggest an iterative reweighting scheme (i.e., the theory of weighted least squares) for computing the maximum likelihood estimate (MLE) of the common mean of a mixture of a finite number of univariate normal populations with known proportions and variances (Stigler, 1973; Butler, 1986). His work not only predated Pearson's early attempts at the method of moments but also foreshadowed the formalization of the EM algorithm that came almost a century later. Even so, it was not until 1932 that his approach was repeated by Jeffreys, who iteratively computed the mean estimates of two univariate normal populations with known variances and proportions. It was Rao (1948), however, who first fully exploited ML estimation in a mixture context, using Fisher's scoring method for a mixture of two

univariate distributions with equal variances. While Rao is most often credited, and highly praised, for the formal publication and dissemination of this approach, his MLE-mixture method was only revisited two decades later. Hasselbald (1966) returned to the problem of resolving a mixture of g univariate normal distributions with equal variances into its constituent components and later (1969) took up the issue of fitting mixtures of exponential distributions. Soon after, Behboodian (1970) reconsidered the univariate example; while, both Day (1969) and Wolfe (1970) worked on its multivariate extension. Day focused on the case of two normal homoscedastic populations fit using the same covariance matrix and Wolfe concentrated on both the solution for multiple populations with different covariance matrices as well as mixtures of multivariate Bernouilli distributions as they applied to Latent Class Analysis (LCA). Wolfe (1965, 1967) also offered innovative solutions for the implementation of the intensive computational requirements of ML analysis by exploiting emerging the computer technology for parameter estimation. Wolfe's free-access statistical software, NORMIX, compiled his algorithmic approaches, facilitated their application to multivariate normal mixture modeling, and encouraged their use within a larger research community. These authors, with a handful of their precursors and contemporaries (Hartley,1958; Rao, 1965; Beale & Little, 1975; Sundberg, 1974, 1976), produced solutions for the problem of mixture decomposition using iterative procedures: the missing information principle of Orchard and Woodbury (1972) deserves special note. Collectively, these solutions can be seen, in hindsight, as corresponding closely to the special applications of the now standard, or "generalized," version of the Expectation-Maximization (EM) algorithm of Dempster, Laird and Rubin (1997). The theory underlying of the monotone behavior of the likelihood and the convergence properties of the ML estimator was not fully appreciated until Dempster and coworkers (1977) formally defined the design of the EM algorithm and codified its use for computing maximum likelihood estimates in a variety of data conditions and on varying levels of generality. It was shown to be

especially useful for "incomplete data," and it was also applied with success to truncated or censored data, for variance component and hyperparameter estimation in Bayesian contexts, as an analogous technique to iteratively weighted least squares, and in factor analysis. Most importantly for this project, it has been essential to finite mixture modeling.

Over the course of many discussions on the development of the EM algorithm, it has been argued that much of the theory supporting Dempster, Laird and Rubin's (1977) version of EM algorithm had been established prior to their, now seminal, publication but that this information existed in relatively disparate parts which were "buried in the literature of special examples, notably in Baum et al. 1970." (Dempster, Laird & Rubin, 1977, p. 3). Dias (2004) is right, therefore, to point out that, had the publications on Hidden Markov Model (HHM) by Baum (Baum & Petrie, 1966; Baum et al., 1970) in the *Annals of Mathematical Statistics*, received due attention, or had Dempster and colleagues not collected and combined the preexisting EM-like ideas into a unified theory, this review would be tracing a very different developmental trajectory for the recent history of statistical. Even Baum remarked, in the discussion section of the paper by Dempster et al. (1977), that the incomplete data formulation of the EM algorithm is fundamentally the same as the HMM procedure or the Markov mixture model but without the constraints of the Markov-chain (c.f., Chib, 1996). Similar argument can be given for Bayesian statistics: its progress may have been accelerated and the approach to "estimation" transformed had the publication by Hastings (1970) been properly recognized. His Monte Carlo sampling methods using Markov chains have been called "the most important missed opportunity of the 1970s" (Dias, 2004, p. 1). As a result, foundational work on fitting mixtures within a Bayesian framework emerged late and only in earnest with Diebolt and Robert's publication on MCMC sampling in 1994.

It was, perhaps by chance then, that estimation research focused first on the general purpose maximum likelihood algorithm for missing-data problems, and Dempster and coworkers (1977) popularized this work with the EM algorithm. Regardless, their seminal publication set the stage for this simple, efficient and quite elegant iterative procedure to make its mark on the course of statistical thought, by forging the path for future trends in computational statistics. As Meng and van Dyk (1997,) have remarked "although we shall perhaps never be able to find out who really sang the first musical tone of the EM algorithm, we all agree that it was Dempster et al. (1977) who brought it into the all-time top 10 of statistics" (p. 512).The influence of the EM algorithm is obvious when its original focus is considered first. It clearly emerged out of research on missing data and as a result it had a great effect on the problem of by-case or by-variable deletion and on the choice of missing value replacement by estimation or imputation (e.g., Rubin, 1972; Little & Rubin, 1987; Schafer & Olsen, 1998). Its contribution to the development of model selection criteria by Akaike (1973, 1974, 1981) and Schwarz (1978) cannot go unstated either: these information criteria have emerged not only as essential instruments in the mixture modeling toolkit, but they are also among the most useful tools for evaluating statistical models in a variety of contexts. Improvements to the EM algorithm itself have also continued unabated and these too have led to further advances. By better understanding its convergence properties (Wu, 1983; Boyles, 1983), researchers were able to speed up rates of convergence and compute the standard errors of the parameters with greater ease by developing acceleration procedures, like, for example, Aitken's Acceleration (Louis, 1982). Additional simplifications to the EM algorithm have also been offered: e.g., the Expectation-Conditional Maximization algorithm (Meng & Rubin, 1993), Expectation-Conditional Maximization Either algorithm (Liu & Rubin, 1994), and the combined ECME and SAGE algorithm (Meng & van Dyk, 1997; after Fessler & Hero, 1994).

With further and, at times, peripheral developments in statistical theory towards the close of the 20th century, new avenues for improving and expanding upon mixture modeling methods were also opened up. The introduction of the bootstrap by Efron (1979) and its parametric application to the Likelihood Ratio Test (LRT) allowed for scholars, like Aitkin et al. (1981), McLachlan (1987), and later Feng and McCulloch (1996), to circumvent the problem of its unknown distribution and offer an alternative to the use of information criteria when assessing of the optimal number of components and for model selection. Advances in regression theory and the use of covariates, and most notably the introduction of the Generalized Linear Model (GLZ) (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989), encouraged the extension of finite mixture methods to a variety of mixture regression models, including Gaussian (Quandt, 1972; Quant & Ramsey, 1978; DeSarbo & Cron, 1988), Generalized Linear (Jansen, 1993; Wedel & DeSarbo, 1995), Choice (Kamakura & Russell, 1989), Count (Wedel et al., 1993) and Concomitant Latent Class models (Dayton & MacReady, 1988; Formann, 1992; Peng et al., 1996). The descriptive nature of the finite mixture model was, therefore, successfully transformed into an explanatory model by incorporating dependent and independent variables. Bartholomew (1980, 1984) and Lindsay (1983) also made significant advances in latent variable modeling and nonparametric maximum likelihood estimation, respectively. The 20th century closed with the publication of the definitive work on the EM algorithm, its properties and modifications, by McLachlan and Krishnan (1997), which has informed all subsequent refinements to finite mixture analysis and variations on the fundamental model-based clustering techniques. Even when considering the, often taxing but usually negotiable, problems associated with its asymptotic properties, finite mixture models with either a fixed number of components or a range are usually estimated with the expectation-maximization (EM) algorithm within a maximum likelihood framework (Dempster et al., 1977). The well-designed

estimation procedure of the EM algorithm and its extensions, therefore, explains a significant part of the success of the ML method and the radical increase in the use of finite mixture modeling in recent years.

As Efron (1979) has remarked, "Most of the statistical theory in common use was developed under the constraint of slow and expensive computation. Now computation is fast and cheap. It is not surprising that new theory is being developed, which takes the advantage of the high-speed computer" (p. 461). Much of the increase in popularity of the use of mixture models can be credited to advances in computer technology: e.g., multi-core processors, cheaper and faster solid state memory, amplified system speeds and faster clock cycles, and increased internet bandwidth for off-site cloud computing. The availability of easily accessible software applications (Haughton, 1997) have also played an important practical role: e.g., EMMIX versions 1.3 by Peel and McLachlan (1999) and 2.0 by McLachlan, Peel, Basford and Adams (nd), GLIMMIX by Wedel (2001), MULTIMIX by Hunt and Jorgensen (1999), SNOB by Wallace and Dowe (1994), or NORMIX by Wolfe (1970). The R environment for statistical computing (R Development Core Team, 2010) features several open source packages for mixture models, each of which have their own focus and unique strengths: fpc for mixtures of linear regression models (Hennig 2000), mmlcr for mixed-mode latent class regression (Buyske, 2003), and FlexMix for discrete mixtures of regression models (Leisch, 2004). The contributed R packages, MCLUST (Fraley and Raftery 2002, 2009, 2011) and EMMIX (McLachlan, 2011; McLachlan, Wang & Peel, 2011), are preferred here for fitting mixtures of multivariate Gaussian distributions and conducting supervised (full and partial) classifications. Finally there are several promising algorithms that have received less attention because they are not integrated within a larger software package or do not "run" as standalone programs. The approaches taken by COOLCAT (Barbará et al., 2002), for clustering categorical attributes

using entropy, or STUCCO (Bay & Pazzani, 1999), that mines data for a new concept called "contrast sets," are sufficiently innovative to warrant coding in an environment like R.

Robust finite mixture theory, coupled with fast and efficient algorithms for the implementation of its model-base clustering procedures, have now formed a powerful analytical tool whose utility transects all segments of research in the biological, physical and social sciences. Figure 1 represents the results from a query for the topics "finite mixture" or "mixture modeling"or "model based clustering"using the Web of Science citation database (Thomson Reuters, 2009). There is a significant linear relationship between year and publication type: adjusted R^2 = 0.9, p-value = 2.562e-16. This data not only shows a growing interest in this approach over the last thirty years but it also demonstrates an especially rapid rise in mixture-based publications in the last five years. This 5 year interval accounts for almost 60% of the total number of the topic-specific records returned (n=1413). McLachlan and Peel (2000) also observe a similar publication trend, which they believe demonstrates how "the importance of finite mixture models in the statistical analysis of data is underscored by the ever increasing rate at which articles on mixture applications appear in the statistical and general scientific literature" (p. xix). They attest to the "widespread use of mixture in recent times" by referring to their own book, *Finite Mixture Models*, in which close to 40% of its 800 citations were published after 1995.

Finite mixture analysis has received positive feedback for its contribution to research in areas as diverse as criminology (D'Unger et al., 1998; Brame et al., 2001), demography (Willekens, 1999; Gage, 2000, 2003), image segmentation (Wehrens et al., 2004; Tran et al., 2006) and marketing (Dillon & Kumar, 1994; Cadez et al., 2003). These examples are but a few of its many far reaching uses, which are both too numerous and discipline specific to present in detail; moreover, developments in mixture

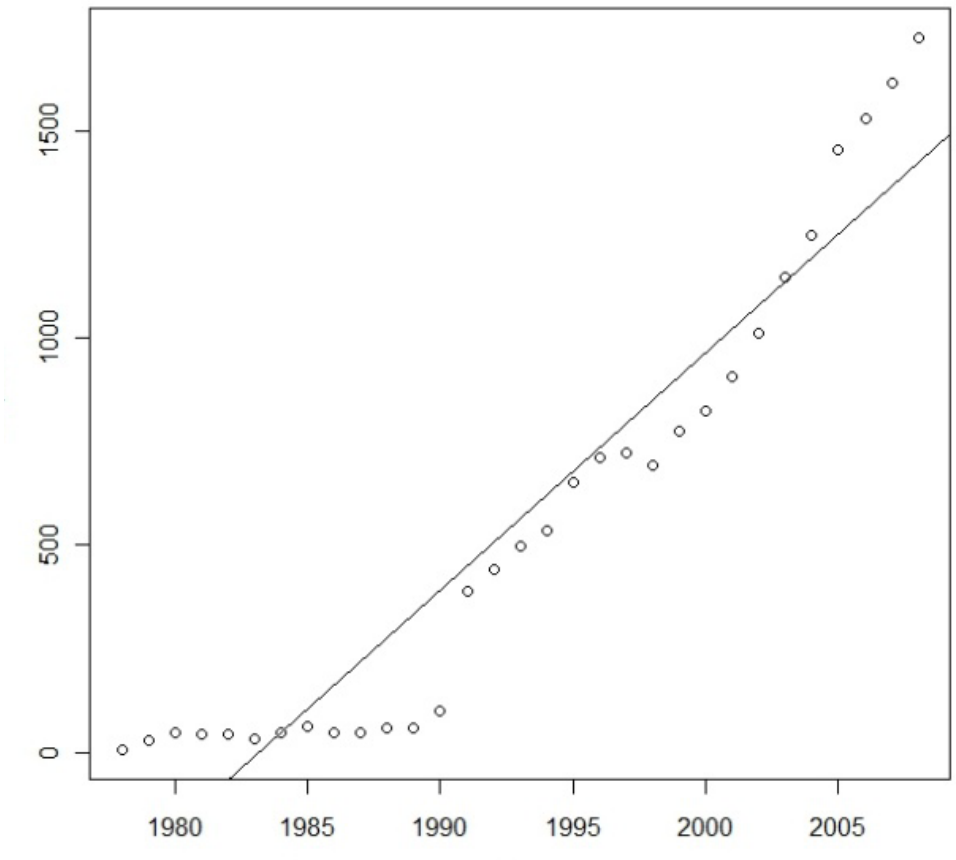


Figure 1: The annual percent distribution of publications referencing mixture-based methods (1980-2010).[5]

application and implementation continue to increase exponentially in the fields of machine learning and pattern recognition and in the biomedical sciences (c.f., Jajuga et al., 2002; Banks et al., 2004; Weihs & Gaul, 2004). Recent summary, yet lucid, discussions of mixture modeling within the general context of approaches to cluster analysis can be found in Basu et al. (2009), Xu and Wunsch (2009), Gan et al.

[55] The frequency is desiplayed relative to the total number of publications returned for a topic query of the Web of Science database. Th query parameters were: Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH; Topic="finite mixture" or "mixture modeling"or "model based clustering"; Time Span=1979-2009. Record count=1413.

(2007) and there are already several thorough treatments of the topic available that address the salient points of its history and offer detailed accounts of both its theory and extensions. Finite mixture methods in particular are well-treated in the comprehensive works by McLachlan and Peel (2000), McLachlan and Basford (1988), Titterington et al. (1985), and Everitt and Hand (1981). More focused discussions are also given by Lindsay (1995), on its geometric properties, by Schlattmann (2009), on its contributions to biostatistics, and Böhning (1998) on computer-assisted approaches to finite mixture techniques. Figure 2 presents the distribution of the total number of mixture-based publications (n=1413) by subject areas and clearly displays its breadth of use and pan-disciplinary impact. It is worth noting the relatively few contributions by "anthropology"; they make up less than 1% of the total records returned.

Although latent class (LC) models have experienced a long tradition of use in the social sciences (Lazasfeld & Henry, 1968; Goodman, 1974; Clogg, 1995; Vermunt, 2003) and mixture or model-based clustering approaches are common in anthropological, population and medical genetics (McLachlan et al., 2004; Pritchard et al., 2000), finite mixture analysis has received relatively little attention in anthropology. Only a decade ago, Kramer and Konigsberg (1999) remarked that "in our experience the method is almost completely unknown to most physical anthropologists" (p. 412). This low level of interest is further supported by a query of the Web of Science database for the years spanning 1979-2009 returns. It returned only 10 publications that apply mixture-based methods to anthropological questions. Of these, only four publications would be subsumed under the subdiscipline of physical anthropology; this is quite remarkable, given that 15,964 records were returned for the total number of articles on all topics published in the *American Journal of Physical Anthropology* alone, from 1979-2009.

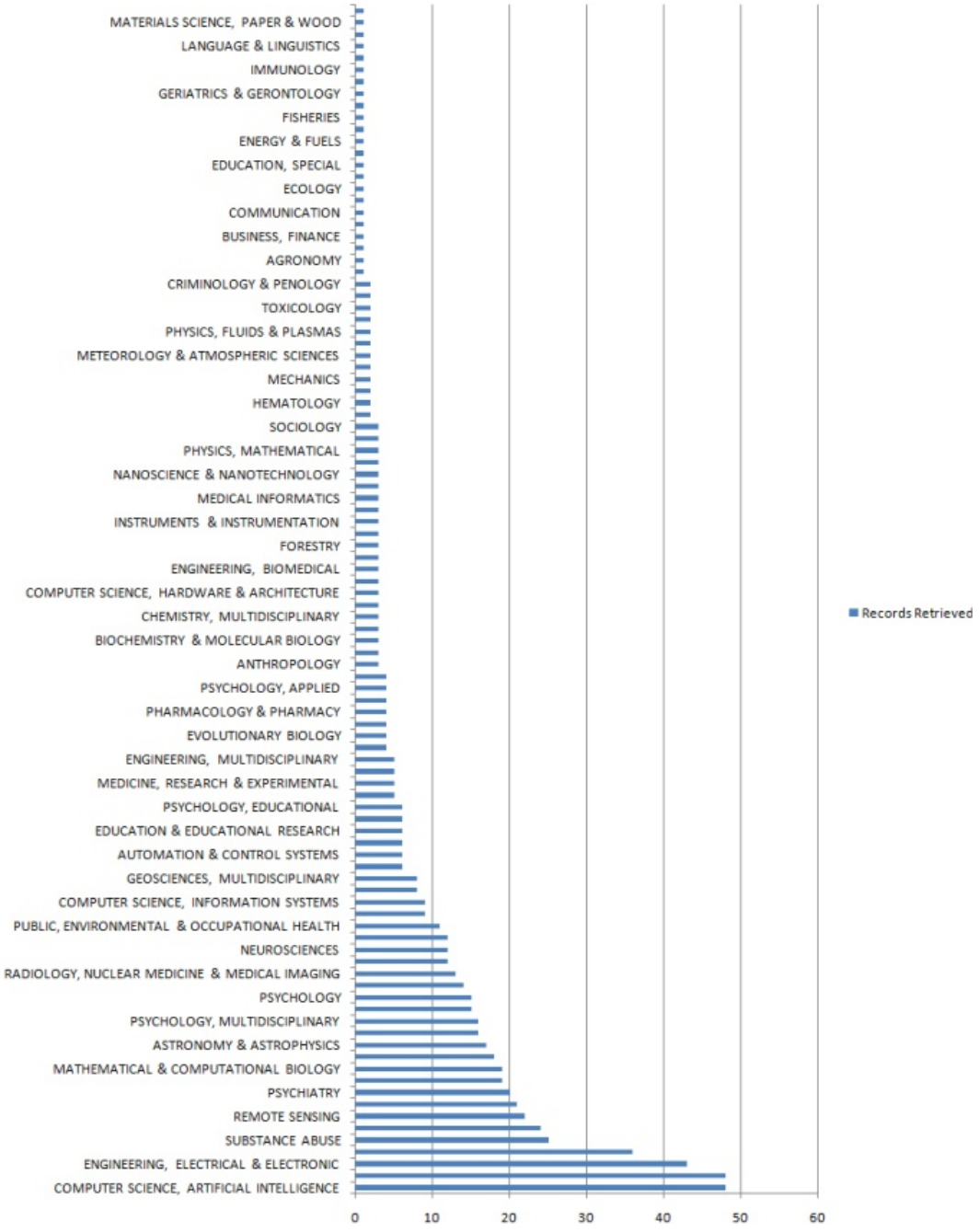


Figure 2: The by-subject distribution of publications referencing mixture-based methods (1980-2010)[6]

[6] **Query parameters: Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH; Topic="finite mixture" or "mixture modeling"or "model based clustering"; Time Span=1979-2009. Record count=1413.**

A secondary trend is also evident in a review of the literature dealing specifically with the application of mixture methods to the subject of skeletal biology in physical anthropology: the majority of published examples have been hitherto limited to topics of interest to paleoanthropologists, including issues of species identification and questions related modern human origins. More specifically, the majority of these studies have taken as their common focus the problems of population differentiation, with an interest in resolving questions of inter or intra-population differences and the number of groups represented in an unknown, often "fossil," sample.

Among the first forays into the application of mixture methods to anthropology is the relatively early paper by Josephson and colleagues (1996). In an attempt to provide a better approach, "an unbiased, reliable method" (1999, p. 191), to the quantitative evaluation of intraspecific morphological variation in extant and fossil species, they proposed a procedure for obtaining estimates of sexual dimorphism in samples of unknown sex using the Method-of-Moments (after Johnson and Kotz, 1970, p. 89). They assumed that, for any given metric trait, its distribution is Gaussian, bimodal, and sex-dependent. When taken in this way, their example data represents a mixture of two underlying normal populations, where one modal "hump" is composed of males, and the other of females. They used three moments around the mean of the combined male and female sex distribution to estimate both the means and common standards deviation for the two underlying distributions. Although they did not take up the iterative algorithmic approach that enables the regular application finite mixture modeling (i.e., for large high dimensional data sets), their findings demonstrate the merit of using even the most basic mixture procedure. These authors argue that their Method-of-Moments has several advantages over previous approaches to the estimation of dimorphism (e.g., Fleagle et al., 1980; Kay, 1982; Leutenegger & Shell

1987; Godfrey et al., 1993; Plavcan, 1994). It is easy to understand and simple to use. It does not require a prior knowledge of sex for the target specimens. It does not rely upon a reference sample or call for a living species analog. It defines meaningful, 95% confidence intervals, and it produces conservative estimates under many different conditions. While Josephson, Juell and Rogers (1996) concede that the Method-of-Moments performs best at decomposing the mixture when there is minimal overlap between the two (e.g., male and female) distributions and estimates are improved as sample sizes are increased, their simulation tests do attest to overall method accuracy and reliability. When these benefits are coupled with its simplicity, they support an argument for its use over other published approaches for estimating sexual dimorphism in paleoanthropological (and other) contexts.

Although Pearson and coworkers (Brant et al., 1990; Pearson et al., 1992) were among the first to propose the use of computationally intensive mixture models to "assign observations to latent categories representing unknown information such as...taxonomic group" in quantitative anthropological literature, it is Kramer and Konigsberg (1999) who undertook the first systematic evaluation of mixture or model-based clustering methods to questions of classification by skeletal measurements in physical anthropology. They tested not only the contribution of finite mixture analysis to questions of species recognition and the classification of fragmentary specimens with missing values, but they also assessed its utility as an alternative to discriminate function analysis, when *a priori* information of group membership is unavailable, unobserved, or, as in this "test," is purposefully omitted. Using a thoughtfully chosen subset of six craniometric variables and a collection of extant, large-bodied hominoids of known sex and species (*Pan troglodytes*, *Gorilla gorilla*, *Pongo pygmaeus*, and modern *Homo sapiens*), they constructed simulated, "unidentified," "fossil" samples of various species-groups and sample-size configurations to which they applied model-based clustering methods. They

tested the ability of the finite mixture procedure to assign these target "unknowns" to their correct taxonomic classification by estimating the probabilities of group membership directly from the internal characteristics of the measurement data itself; that is, when all indicators of the "known" or "true" group membership (e.g., sex and taxonomic designation) were excluded from the clustering and classification steps in their analysis. Further, in order to approximate the recovery condition of most fossil specimens, Kramer and Konigsberg took into consideration the confounding effects of taphonomic processes on specimen perseveration and, in turn, the limitations imposed upon landmark-based, metric data collection by accounting for the presence of and random variation in missing values. They stochastically removed both measurements and cases from their simulated "fossil" sample to approximate conditions with MCAR data. They then evaluated the utility of finite mixture methods to situations in which only "unknown" fragmentary specimens are available for study: when missing values are a problem, measurement collection is dictated by the condition of each case, and only incomplete data is available for cluster analysis. Kramer and Konigsberg obtained high percent accuracy results for all of their finite mixture tests. Most estimates were in the upper 90% range; they achieved 100% accuracy in the case of a mixture of 124 chimpanzees and humans. Even when accuracy dipped into the, still very acceptable, 80% range for the more difficult, small fragmentary samples, they were able to demonstrate how this model-based clustering method still consistently outperforms k-means clustering in making meaningful, correct classifications.

Despite the variation introduced by target samples composed of multiple species and mixed sexes, the fluctuations in sample sizes, and the added complication of clustering "fragmentary" specimens with values missing completely at random, Kramer and Konigsberg have shown how finite mixture analysis can, not only, produce very accurate assignments, even when it is required to accommodate for missing

data, but how it may also serve as a reliable alternative to discriminant function analysis and k-means clustering, which are well-known for their weaknesses due to algorithmic constraints and prior assumptions. Kramer and Konigsberg's findings attest to finite mixture's theoretical and practical strengths by demonstrating how this modeling method (1) can estimate missing data, (2) is probability based, and (3) is unbiased in its classification. This last point is of special interest: in finite mixture analysis, the posterior probabilities of group membership that inform the individual classifications are estimated without the use of *a priori* information on group membership. The assignments are unbiased because they are made in the absence of qualifiers, like sex or taxonomy, and the method does not require predefined "reference" populations of "known" cases from which classification functions are constructed, against which target specimens are compared, and into which the new members are allocated. Kramer and Konigsberg (1999) have offered very promising conclusions, stating positively that "[f]urther refinement of FMA models may lead to the development of a more rigorous procedure for identifying species in the fossil records" (p. 419).

Encouraged by Kramer and Konigsberg's promising results using mixture modeling and recognizing the limited application of cluster analysis in general to anthropological problems, Kramer and et al. (2009, 2010) have recently retested the utility of an unbiased clustering approach to group discrimination using skeletal data. In this study, they compared the quality of the classification solutions produced by both fuzzy clustering and finite mixture analyses using baboon craniometrics, specifically. The larger purpose of this study was to evaluate the applicability of these methods to their other, more-broadly based, questions of interest: (1) which of the clustering methods available are most useful to identifying a biological signature of hybridity, and (2) can the "best" clustering approach be applied profitably to the problem of resolving the presence or degree of admixture between Neanderthals and Anatomically

Modern Humans? Although this project is still in its preliminary stages, with a focus on the publication of the method tests, Kramer and colleagues have reconfirmed, using the baboon analogue as a proxy for paleo-populations, that finite mixture analysis is a powerful discriminatory tool that both yields statistically sound and interpretable classifications and has the sensitivity to detect complex internal data-structure that may be informative of between-group interaction, gene flow, and ultimately evolutionary history. Their ongoing work has produced results for each individual, by sex, in each cluster that support a reliable and logical classification scheme. In general, they have found that individuals were assigned to the group that corresponds most closely to their taxonomic classification based on high posterior probabilities of group membership and low uncertainty in the allocation of each individual specimen to the "correct" cluster. When misclassifications did occur, they observed a consistent pattern for both sex samples: the posterior probabilities were divided between the two groups into which the specimen showed the greatest morphological similarity. They also noted that these misallocations occurred almost exclusively between the *anubis* and *hamadryas* baboons, for both the male and female data. This is an important finding in terms of the "reality" of the cluster solutions, as these two groups are known to hybridize in the wild. These mixed assignments may in fact be telling of important biological information, under the assumption that, since hybrids share genetic material, they should equally share morphological characteristics. Kramer and coworkers suggest, therefore, that these mixture-based classification results should bring into question the biological meaning underlying the traditional multi-species taxonomic classifications for *Papio*: these findings may offer tentative support for the reorganization of baboon taxa to include select subspecific designations. While their analysis speaks to baboon craniometrics, their conclusions have broader methodological implications as they support the use of finite mixture analysis as a robust statistical tool for the decomposition of complex population mixtures and the separation of closely related specimens into statistically and

biologically meaningful groups. Perhaps, even more importantly, they demonstrate the potential of model-based clustering and its probability-based approach for detecting inter-group motility and for investigating questions of admixture or hybridity.

The work by McKeown (1996) on the use of finite mixture analysis for sex estimation and missing values replacement stands as the first thorough multivariate study of the application of mixture methods to human skeletal data in anthropological contexts. In order to test the utility of a FM approach to the determination of sex from craniometrics, McKeown selected the Hokkaido (Northern Japanese) and Zulu (South African) subsamples from the larger Howells data set (1996). These samples both meet the methodological requirements necessary for her analysis and offer specific advantages: they represent modern anatomical or dissection room specimens, individual sex is known, and the sex-based population proportions are close to equal. She also evaluated the effectiveness of the finite mixture method, as executed with EM algorithm, to compute expected values for unobserved data. After identifying a suitable MCAR (missing completely at random) pattern using metric variables observed for the Late Mississippian Averbuch skeletal sample (Berryman, 1984), she simulated the case of fragmentary (prehistoric) archaeological remains with an incomplete suite of craniometrics by imposing the pattern of missing measurements identified from the Averbuch crania (n=154) onto the two pre-selected Nothern Japanese and South African populations from the Howells data set. McKeown's results showed that under these controlled conditions, the finite mixture analysis provides levels of accuracy equal to discriminant function analysis (DFA) for the "relatively dimorphic" Hokkaido samples, even in the context of missing data (with an averaged correct classification of 85.7%). FMA was not as successful, however, when applied to the "less dimorphic" Zulu sample, neither when the full data set was used nor when variables were removed: the average FMA correct classification was 64.5%,

compared to 80.1% that was obtained using DFA. Interestingly, she returned good results (74% correct classification) when FMA was applied to a subset of the Averbuch sample directly (n=94); producing a sex ratio comparable to the true male and female proportion, despite the considerable number of missing measurements and the high prevalence (78%) of cranial deformation in her sample. McKeown concludes that finite mixture analysis worked best for samples that demonstrate high levels of sexual dimorphism, and it is especially responsive to differences in male and female cranial size. In this regard, FMA provided results as good as those of the DFA and had the added advantage of not requiring *a priori* data (e.g., reference samples, or true memberships) for making the classification. She proposes (1996), therefore, that "for samples where pelvic elements are unavailable for creating a training subset as necessitated by discriminant analysis, finite mixture may prove to be a more appropriate and equally accurate method..." (p. 50).

In a recent publication, Konigsberg, Algee-Hewitt and Steadman (2009) present an updated implementation of McKeown's (1996) finite mixture missing data approach as applied to questions of population affiliation and using the contributed R package, MCLUST. As this paper represents the third installment in a four-part series that addresses the statistics underling forensic practice, they consider mixture modeling as only a part of their larger commentary on the practice of estimating sex and "race" from craniometrics and the evidentiary significance of an assessment of these categorical biological variables in forensic anthropology. The focus of this project is not on testing the appropriateness of finite mixture analysis to forensic casework nor on evaluating its implementation via the MCLUST package in R. Moreover, it is not the intension of these authors to present a model-based clustering method for the estimation of sex or ancestry. The finite mixture approach is used only as a heuristic tool in support of the subsequent discriminant function analyses, as Konigsberg et al (2009) explain: "In the

analysis reported here, we do not consider more than two groups because it is not our intent to determine the number of groups, only to ascertain if there is structured heterogeneity within the data" (p. 79). Despite this controlled application of the finite mixture procedures to the question of sex, Konigsberg and colleagues clearly demonstrate the promise of finite mixture analysis for future applications to modern human craniometrics by using it to identify geographically patterned variation in the Howells data set and the Forensic Anthropology Data Bank. It is evident from their discussion that this clustering approach has further potential for answering questions in forensic anthropology and how, through extensions to their analyses, it could be put to use for research problems that seek, specifically, to tease out internal structure from biological data for the purpose of group discrimination, i.e., in the case of estimating ancestry or race. Pontikos (2004) has, in part, already addressed this potential with his application of finite mixture analysis to the question of world-wide craniometric variation: using the MCLUST procedure, he has demonstrated the ability of model-based clustering to identify multi-group structure in the Howells data set. When considering the 2,504 samples, representing Howells' 28 populations, and using all 57 variables, he obtained 14 groups which had, in his opinion, direct correspondences to what he calls "populations of 'known' racial affinity." Konigsberg et al (2009) have responded to his analysis by calling into question his procedural choices: "he used all 57 of Howells' original measurements which placed a strong penalty on models that allow for group differences in the variance-covariance matrices. It seems ironic then that his use of 57 measurements as a nonparsimonious descriptor for skulls led him to accept a simple variance-covariance model (equal for all groups) on the basis of parsimony (via the BIC)" (p. 83). Given these reservations regarding his sample and methodological rational, it is difficult to accept Pontikos' cluster analysis outright, as a meaningful statement on the utility of finite mixture methods, and the MCLUST package, for the decomposition of mixtures of biogeographic populations and the relationship of race to world-wide craniometric variation.

The recent work by Algee-Hewitt and colleagues on alcoholism-related changes to bone density represents a recent contribution to the literature on the application of mixture modeling in forensic anthropology and best-practices for casework of medico-legal significance (Algee-Hewitt & Wilson 2008; Moore et al., 2009). These authors argue that difficulties in explaining the effects of prolonged alcohol abuse on bone microstructure, the lack of consistency in the clinical diagnosis of observable bone changes and their symptomology among the living has encouraged *ad hoc* conclusions by the forensic community. This approach has hampered real efforts to make any reliable identifications for the dead. Algee-Hewitt et al. posit that, although it is common knowledge among some forensic anthropologists that chronic alcoholics frequently present boney characteristics very similar to those observed in individuals of advanced age or in "osteoporotic" skeletons, this information is often inherited from mentors or practically acquired while working with skeletal materials in "known" contexts for which medical history is reported. They note, as a result, that alcoholism has received little attention in discussions of forensic anthropological method and theory: suggestions of alcoholism are commonly relegated to side notes (e.g., "*possible alcoholic?*") when the standard age indicators do not support an age-related diagnosis for the observed skeletal demineralization or when other pathological explanations are not logical. Algee-Hewitt and coworkers have applied model-based clustering to post-cranial bone weights and DEXA scan data from a large sample of modern skeletons to (1) evaluate the assumption that middle-aged chronic alcoholics express skeletal degeneration in the form of bone weight loss (i.e., "secondary osteoporosis") that is significantly different from non-alcoholic members of their same birth cohort and not unlike older osteoporotic individuals for whom age is the primary risk factor, and to (2) test the forensic anthropologist's ability to differentiate between documented chronic

alcohol abusers and non-alcoholics of comparable ages using bone structure alone. Model-based clustering was chosen, first, for its lack of *a priori* assumptions and user or method induced biases: the bone data was stripped of all sub-population identifiers (such as alcoholic, non alcoholic, age), before it was analyzed. It was also believed to best approximate the laboratory conditions typical to a forensic identification case, whereby the practitioner is charged with the task of making meaningful statements regarding the identity and life history of the skeletal remains, in the presumed absence of prior personal information. In this test case, the mixture modeling procedure failed to identify significant internal structure within the bone data that either warranted multiple clusters, or generated patterned cluster memberships. The alcoholic specimens did not cluster uniquely or cohesively with members of the older age group; instead, they were distributed throughout the larger sample. As the "alcoholic" bones used did not possess measurable characteristics distinctive from the skeletal population-at-large, Algee-Hewitt, Wilson and Moore have concluded that a presumptive diagnosis of "alcoholism" provides little reliable information relevant to the construction of the biological profile for unknown skeletal remains. They have recommended, therefore, that forensic practitioners exert extreme caution when making statements regarding possible chronic alcohol abuse based upon bone weight or bone structure as the primary diagnostic criterion or the sole evidence for grossly-observed cases of secondary osteoporosis.

Mixture: Theory

"We scientists often miss things that are 'staring us in the face' because they do not enter into our conception of what might be true, or alternatively, because of a mistaken belief that they could not be true." (Medawar 1969, p. 52)

In the method-based literature on clustering, Jain and Dubes (1988) are often cited for their recommendations on how to manipulate, employ and extract reliable information using a well-equipped clustering toolkit. Their statement on best-practice for cluster analysis is frequently paraphrased: cluster analysis is useful tool for exploring data but it should be supplemented by techniques that aid in visualizing data. These recommendations are usually given in concert with references to Exploratory Data Analysis (EDA), which is well-known and whose best-practice discussions promote a similar approach. Although EDA has changed with recent trends in methodologies, it is still keeping true to the ideas of "attitude" and "flexibility" central to the pioneering work of Tukey (1977). A comprehensive, practical, and current definition of EDA is available from the National Institute of Standards and Technology's online Engineering Statistical Handbook (NIST/SEMATECH 2010), which is integrated with the free, downloadable DATAPLOT software for data visualization and preprocessing (Heckert, 2001): "EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret" (np). The use of clustering techniques in data mining contexts (when faced with large, often high dimensional data sets) and for [un]supervised or Machine Learning (ML) has very recently demonstrated just how important preparatory steps are to the ultimate success of any approach that is reliant upon the smooth operation and timely execution of an algorithm. As Kotsiantis et al. (2006) remind the practitioner, "[t]he representation and quality of the instance data is first and foremost" (p. 111). Data preparation through visualization and pre-processing techniques can include data cleaning, data abstraction, plotting or graphing, normalization, transformation, dimension reduction, feature extraction and selection, filtering

out any instances of irrelevant and redundant information, correcting for or adding in noise, and detecting and assessing the effect of outlying observations. Popular introductory texts on clustering procedures tend to agree with these preliminary, heuristic approaches, but they also encourage readers to adopt a standard research design; this may begin by the examination of simple graphical displays (searching for signs of clusters) and then by letting these findings influence or guide the direction of the subsequent analyses. Everitt (1993, p. 11) has even taken this one step further, stressing that "well chosen displays can enhance the understanding of the data and may also provide a partial anecdote to the dangerous habit of applying clustering techniques in a careless and uncritical way." Wand and Wang (1996) have argued that unreliable data represents some disconnect between the condition of the data set and the conditions of the real world; and Wang et al. (1995) have shown that the amount of erroneous data may account for 10% of any given database and, often, even more, depending upon the type of data collected, method of acquisition and its storage. While it is hard to disagree with the potential benefits that an initial careful, systematic, and "exploratory" review of data can bring to an analysis, the use of both simple statistics and graphics (e.g., means, standard, deviations histograms, scatterplots) and more complicated tests and coordinate projection techniques (e.g., outlier detection using Mahalanobis' Distance or Cook's D; principal component analysis) in this way introduce their own method-specific problems into the analysis, which may be misleading to the practitioner and detrimental to the course of analysis. These techniques are, therefore, not as curative as Everitt would suggest, for the either the univariate case or its multivariate extension: simple and quick statistical estimates, like the sample mean and standard deviation, can be significantly biased by individual outliers that are far away from the middle of the distribution; marginal distributions of multivariate data are, more often than not, poor reflections of their multivariate structure, so that the application of exploratory methods to each variable separately (i.e., examining univariate visualizations, such as the

stem-and-leaf plot, boxplot, or histogram) is rarely helpful in the context of interpreting the full data set; and, for multivariate measures of association, when using ordination methods, the choice of the viewing plain may mask the true variation in the data and cause patterns with clustering potential to be missed. Even a moderately experienced analyst could easily overlook hidden structure within a given data set by following the guidelines provided by Everitt (1993) for kicking the "habit" of poor applications by misjudging when to pursue clustering, what type of clustering method to apply, and how to critically evaluate any preliminary patterns observed. It is unwise, therefore, to make judgment calls based on preliminary, subjective, and possibly misleading assessments alone.

In the presence of unsupervised and machine learning theory, innovations in artificial intelligence, and the demands for making sense out of large, complex data sets in the biosciences, there is refocused attention in the field of cluster analysis (Mitchell, 1997; Pyle, 1999; McLachlan et al., 2004; Schlattman, 2009). This is likely because cluster analysis is now being used in serious health-related applications, including gene mapping, disease prevalence and identification, and even the practical administration of drugs or the design of clinical trials. Clustering is in the process of expanding its traditional definition as (1) a technique useful for finding groups in data, but requiring some user-defined parameters or the contribution of some a priori knowledge (i.e., knowing and inputting the predetermined number of *k* groups in order to partition an N-dimensional population into *k* sets, for *k-means* clustering) or (2) as method that is both unreliably exploratory and only confirmative of suspected or previously determined structure (see MacQueen, 1967; Anderberg, 1973; Hartigan & Wong, 1979). In contrast to Anderberg's (1973) suggestion, in his "principles" and "intelligent advice" for clustering applications, that "a set of clusters is not a finished result but only a possible outline" (p. 22), new opinions are being voiced that concentrate efforts on applications where the number of classes or clusters in a data set must be

discovered (when little or no prior information is available), when structure must be imposed upon a relatively homogenous data set (to establish "fair" divisions) and the partitioning of a data set is the end product (Kaufman & Rousseuw, 2005). These new opinions suggest that cluster analysis is in the midst of a methodological revolution that is encouraging the reevaluation of clustering's utility and the rehabilitation of its image despite the history of skepticism towards its procedures (as "data dredging"), its results (as artifacts of the algorithms), and the reality of any conclusions drawn (Fielding, 2007, p. 9). Clustering can be the method itself: alone, it can provide meaning for data with an otherwise inaccessible structure, without *a priori* requirements or potentially biased user-defined parameters, and it can yield results that are statistically strong and whose meaning can be easily accessed and assessed. It is, therefore, appropriate at this point to offer a revision to the common paraphrase given for Jain and Dubes' (1988) statement on clustering, so that it reflects what may be the true spirit of the statement, especially given their algorithm-intensive position: clustering is the analysis. The choice of how to apply it and which method to use can be determined statistically and not just manually, by a visual inspection using observer-defined criteria. This is not to imply that visualization techniques are not useful; when applied *post hoc*, for example, these graphical tools can be used to give external support to the clustering solution, just as p-values are produced to give credence to the results of frequentist hypothesis tests. A measured choice of a robust clustering method, like model-based clustering, and its thoughtful and constrained application can (1) tease out latent organization (clusters, components, groups, populations) from simple and small to high dimensional, nested, very large and/or complex collections of observations and (2) find heterogeneity within a data set that may otherwise appear as one homogenous point-cloud composed of many like data-points when the analysis relies upon preliminary exploratory tests or less clustering sensitive methods. In order to demonstrate these effects in practice, this project presents, in the discussion that follows, an example analysis that makes use of a

data set of well-documented modern human statures (Meadows Jantz, 2008). This test case serves as an introduction into the theory behind the probability model-based, distributional approach to the decomposition of mixtures, the statistical and practical advantages of adopting model-based clustering or finite mixture methods, and, ultimately, as a prelude to a project-specific explanation of the methodological choices made here.

Mixture: Stature Example

Often when clustering problems are reflected upon, some prior expectations are brought to the discussion. These expectations can be justified by past experiences with similar, "known" data, shaped by an accumulated knowledge of how certain types of data may act in particular contexts or react to particular conditions, or, simply, based upon common sense. As Bozdogan (1994) has remarked "it seems more relevant and logical" (p. 71) to consider or sample as a collection of observations arising from several different populations rather than from a single population; this is so when the individuals within a group or class differ from each other in some respect or quantity (Fleiss & Zubin, 1969). In this case, there is a concern for statistical distributions that can be expressed as, what Everitt and Hand (1981) have called, "superpositions" of component distributions (p. 1). Given their statistical properties, these superpositions can be equally called mixtures or compound distributions. To illustrate this concept more clearly, it is helpful to consider, as an example, the distribution of height in a population of children. This infinite mixture can be simply expressed as the following density function:

$$h(height) = \int g(height; age)\, f(age) d\, age$$

where $g(height; age)$ is the conditional distribution of height on age, and $f(age)$ is the age distribution of the children in the population. As the probability density function (*p.d.f.*) of height is expressed above as an infinite superposition of conditional height density functions, it can be called a mixture density. Using an example of adult stature for males and females, it is possible to also express a height density function in the form below, that has instead a finite number of components:

$$h(height) = h_1(height; male)\, p(male) + h_2(height; female)\, p(female)$$

where $p(male)$ and $p(female)$ are the probabilities that a member of the population is a male or female, and h_1 and h_2 are the height density functions for males and females respectively. Here, the density function has been expressed as a superposition of two conditional density functions or, put another way, as a mixture of a finite number of components. It is this latter, *finite mixture*, formulation that is of particular interest to this project and, so, the remainder of this discussion. It is of special importance because the type of skeletometric data of concern here has a limited number of defined, biological groups (i.e., sexes, ancestries). Although this finite form is elegant in its calculus and easily followed, the finite mixture is complicated by the absence of *a priori* sample information; that is, when the data are not available for each conditional distribution separately and, so, only the combined distribution can be immediately studied. In this case, it is assumed that the component label data is missing. As a result, it is not possible, for the example of adult height, to observe the underlying variable of sex that divides the total set of stature measurements into two, male and female, groups. Circumstances such as this call for the practitioner to focus her efforts on estimating both the mixing proportions, $p(male)$ and $p(female)$, and the parameters in the conditional distributions. Bozdogan (1994) provides a lucid explanation of the standard approach to mixture-model cluster analysis: "[g]iven a sample from the overall mixture-model, the problem from a clustering point of view is to determine and identify the number of mixture clusters, the parameters of the distribution characterizing each

mixture cluster, and to which cluster each individual belongs" (p. 71). This generalization of the mixture-clustering problem can be easily turned into one that explains finite mixtures by recalling that, for the finite case, there is some limit to the ultimate number of "natural" clusters that may be found or exist.

The univariate example of the finite mixture of adult stature by sex can be extended to respond to the earlier discussion of the issues of preprocessing and data exploration, considering especially the graphical step advocated by Everitt, and the benefit of using mixture-modeling as the primary (clustering) approach for this data's analysis, since the information on the true number of components making up the compound distribution of data and the group or cluster membership of each individual observation is (or is taken to be) unknown. The example of adult stature by sex is a particularly apt choice because (1) it offers the computational simplicity and didactic clarity of a univariate analysis, (2) it rests upon the well-known fact that there are significant differences in male and female height, (3) it uses reliable and recently collected data, (4) it provides a test comparable to this project's larger craniometric analysis as it is concerned with modeling continuous biological measurements, and (5) it adheres to the general approach taken for modeling heterogeneity using demographic variables, whereby continuous mixing distributions are generally assumed in population studies (Vaupel & Yashin, 1985; c.f., Dias & Willekens, 2005, for modelling sequential data in demography). Recorded statures for 523 modern American adult White males (n=378) and females (n=145) were selected at random from the William M. Bass Donated Skeletal Collection (Meadows Jantz, 2008), housed at the University of Tennessee, Knoxville (Forensic Anthropology Center, 2010). In order to the avoid the complication of correcting for the acknowledged ancestry specific variation in height and the problems of secular change in stature over recent time, only individuals of known identity (sex, ancestry, age) who died between 1983 and 2007 were included in this analysis (Meadows Jantz, 1970, 1995; Ross & Konigsberg, 2002;

Jantz et al., 2008; Wilson et al., 2010). Using all of the data combined, an exploratory histogram was generated first, in order to conduct an initial evaluation of its clustering potential (Figure 3).

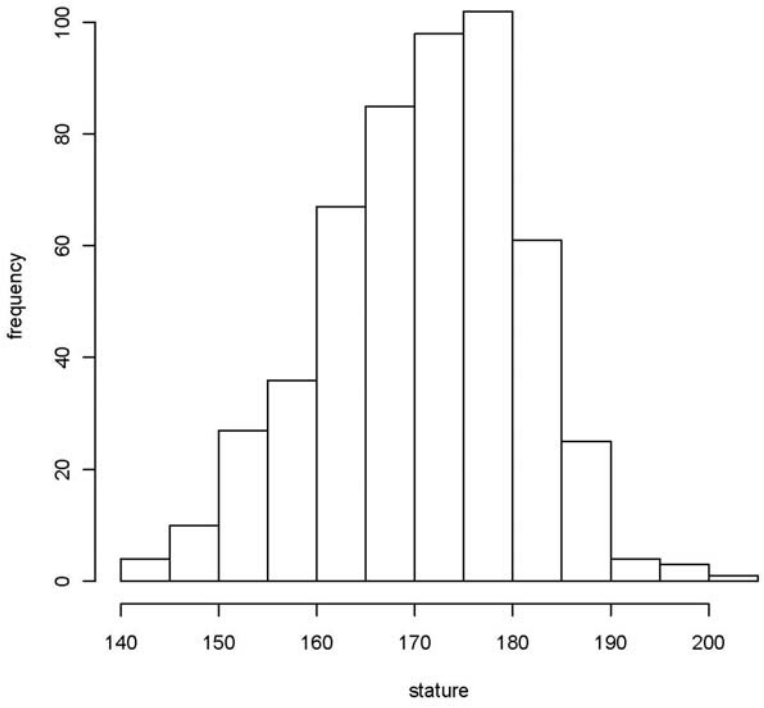


Figure 3: Histogram of the combined adult male and female statures from the William M. Bass Donated Skeletal Collection

A visual assessment of this histogram by a blind observer would suggest that the Bass Collection stature data represents a single homogenous group without internal patterning or structure; that is, a seemingly single unimodal distribution without mixture. After Everitt (1993), a situation such as this may "indicate that the data do not contain distinct groups, thus making the application of clustering less compelling" (p. 11). A hasty determination of homogeneity in this case would, however, be wrong. It is, in fact, by subjecting this stature data to model-based clustering that its true internal structure is revealed and can, in turn, be accessed. A univariate, normal finite mixture analysis of the full data set, including male and

female statures, but excluding the key group identifier (i.e., sex), was executed using the R package MCLUST version 3 (Fraley & Raftery, 2011). Following the recommendations of Fraley and Raftery (1998), the parameterization of the model and the optimal number of clusters were determined using the Bayesian Information Criterion (BIC), as integrated within the MCLUST software. The highest BIC value (-3885.523) supported an "V" (varying variance) model with a seven cluster partition, while the second highest value (-3899.462) corresponded to an E" (equal variance) model with a two component solution. As there is only a slight difference between the two best BIC values, the increase in parameters necessary for a considerably higher component mixture with greater complexity resulting from the change in the structure of the variance-covariance matrices, from equal to variable, cannot be justified: parsimony supports the two group, constrained covariance model. Additional tests have indicated that, while an increase in the number of components does offer improvement to the model fit, this better model fit (and, possible, "over-fit") is almost always achieved at the expense of the interpretability of the results. The decision to fit the two-group mixture with restricted normal components also agrees with statements made by McLachlan and Peel (2000, p. 81, 83) on the assumptions commonly made and accepted "in practice." In order to visualize the two cluster solution, a density plot was produced, which shows the single "unknown" mixture distribution, but identifies the location of the two component means with vertical lines (Figure 4).

For didactic purposes, this example data set was chosen because the univariate test can be simply explained and the true group (sex) membership for all of the cases is *a priori* known. This test allows for an ease of computation and interpretation, as well as the luxury of conducting *a posteriori* comparisons

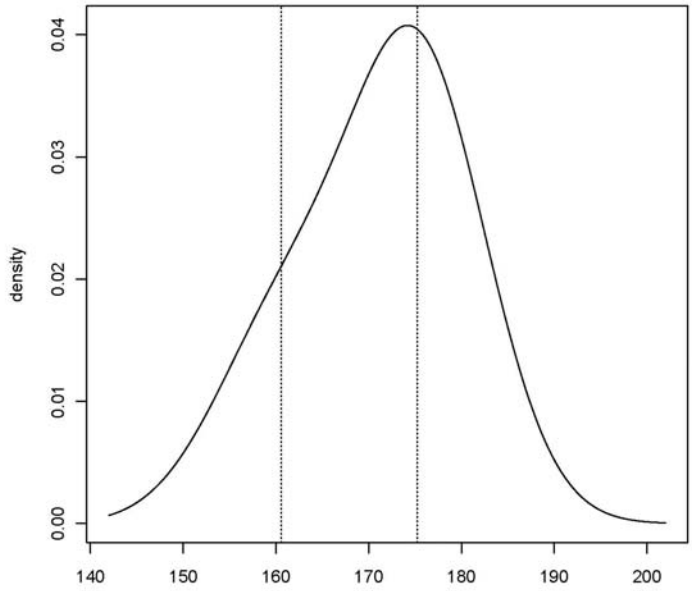


Figure 4: Two component mixture density plot for the combined adult male and female statures from the William M. Bass Donated Skeletal Collection.

and visualizations that evaluate the quality of the clustering classifications and offer validation for the clustering results. In most real-world applications the number of component clusters is not known and needs to be estimated (using some user-selected information criterion or likelihood test); similarly, the subpopulation composition is also unknown and, so, individual observations need to be allocated to the cluster to which they "belong" and these assignments must be supported by some measure of statistical certainty (such as a high posterior probability of membership). The two-group solution produced for the Bass stature data, when pooled by sex, identified 133 specimens as females (true n=145) and 390 specimens as males (true n=378). The true distributions for the male and female subsets are plotted separately using the *a priori* known group-membership information and the calculated individual group means: average stature for males is 174.81cm, and for female is 160.94cm (Figure 5 and Figure 6).

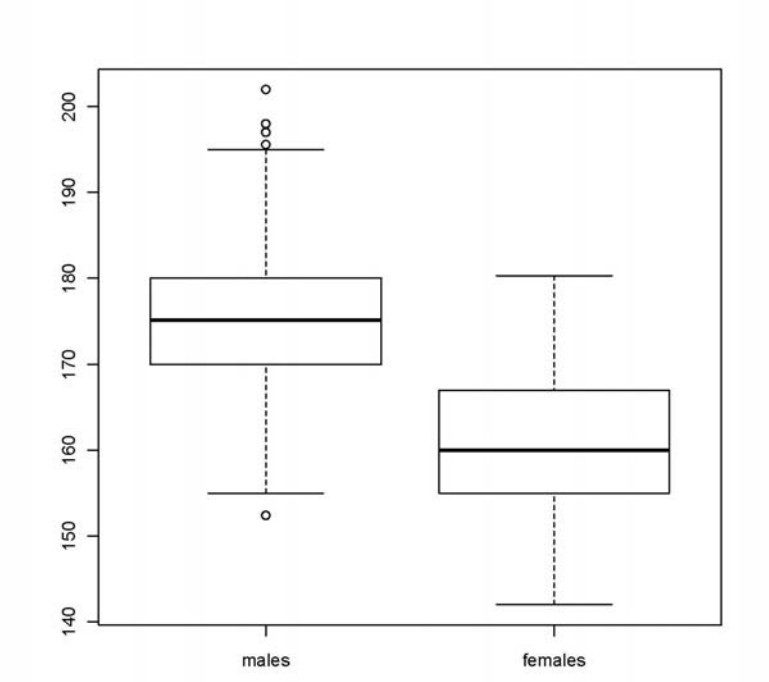


Figure 5: Boxplot of the individual adult male and female statures from the William M. Bass Donated Skeletal Collection, when sex is "known."

It is clear from these results that the male and female distributions overlap considerably. If these findings are considered in light of the earlier work on the identification of sexual dimorphism, notably by Josephson and colleagues' (1996) use the Method of Moments, and by Godfrey et al.'s (1993) application of univariate finite mixture analysis, it is obvious why these authors cautioned their readers against the uncritical application of mixture method to (combined sex) data, that have similar group-values. Although these previous mixture-based studies reported results that were better than those generated by the other more traditional methods tested, they still argued that when true levels of dimorphism are low (i.e., groups are very similar in their measured characteristics), weak estimates are more likely to be produced and only poor support for the latent, true bimodal structure of the mixed population can be given. It is remarkable, therefore, that, when the *a priori*, "known," group densities

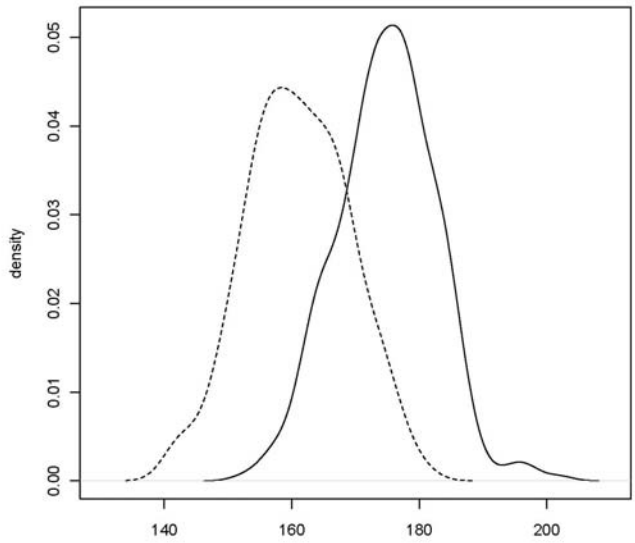


Figure 6: Density plot of the individual adult male and female statures from the William M. Bass Donated Skeletal Collection, when sex is "known."

and their true sample means (Figure 5, Figure 6) are compared against the component density plot produced from the finite mixture analysis (Figure 4), it can be seen that the robust model-based clustering procedure was, nevertheless, able to successfully decompose the compound "mixed" distribution into the two sex-based clusters. These model-based clustering results offer valuable information on the internal structure of the stature data set that would have been otherwise inaccessible if only the raw height measurements were available and observed using traditional heuristic or graphic methods (Figure 3). This easily interpretable example of the well-known relationship between sexual dimorphism and stature (1) attests to the utility of finite mixture analysis as a powerful clustering tool in univariate contexts with hidden structure, (2) supports testing of its multivariate extension to *G* groups with *J* attributes, and (3) lends credence to its application to problems that consider, specifically, modern human skeletometrics.

Defining the Finite Mixture Model

"The mixture model has long been a challenge to the statistician, whether beginner, practitioner or theoretician." (Lindsay, 1995, p. 1)

Mixture models are powerful tools for data analysis as they provide a principled probabilistic approach to unsupervised clustering, while their complex distributional functions have also recommended them for density estimation problems. Lindsay (1995) has correctly argued that the "simplest and most natural" (p. 2) interpretation of the finite mixture model is to believe that all of the sampled observations are collected from, or belong to, some number of unobserved or unknown groups, populations, subpopulations, or latent classes. Attempts may be also made to better define their inter-group relationships by assuming that these (two or more) groups are either mutually exclusive (with sample heterogeneity coming from a biological variable, like sex) or may exhibit varying degrees of overlap (being, not mutually exclusive, as one might expect for a qualitative variable, such as race). As the method's name, *finite mixture*, implies, these observations are believed to take the form of a mixture, regardless of group-type. It follows, then, that the intention when modeling (i.e., clustering) a given data set is to "unmix" a presumed compound sample and, in doing so, discover the underlying structure or latent group present within it. For the mixture likelihood approach to clustering, the hope is to obtain meaningful data-partitions or meaningfully populated clusters. Using robust probability-based statistics and a model-based, or parametric approach, to cluster analysis, the assumption that the sample population is composed of a single group (which is often the "visually" observed pattern, e.g., from simple plots or summary statistics) is weighed against the presumed existence of two or more hidden groups. Therefore, for a g-component mixture, the FM model is employed to (1) reveal within

the given data the underlying component structure and (2) isolate these components individually, as unique clusters of [sub]populations within the mixture of data points (g externally existing groups, $G_1, \dots, G_g$, in some proportions, $\pi_1, \dots, \pi_g$) that make up the full sample data set (superpopulation G, consisting of a finite number of g groups). A formal definition of the standard finite mixture (FM) model is given here; a more detailed statistical analysis of the specialized or modified FM models, and their extensions, is included in the comprehensive accounts of the theory and application of modeling via mixture distributions provided by McLachlan and Peel (2000), Lindsay (1995), McLachlan and Basford (1988), Titterington et al. (1985), and Everitt and Hand (1981). Gan et al. (2007) and Böhning (1999) offer useful discussions of the FM algorithms, while Fraley and Raftery (2011) and McLachlan et al. (2011) provide information on the implementation of model-based clustering using the contributed R packages, MCLUST and EMMIX, respectively.

When defining the finite mixture model, it is assumed that a sample of n observations or individuals, each of which is denoted by $i(i = 1, \dots n)$, is characterized by J attributes or variables. The jth attribute of the individual i is represented by Y_{ij} and the sample value by y_{ij}. The vector $\mathbf{Y}_i$ consists of the elements Y_{ij}; $\mathbf{y}_i$ can be similarly defined so that $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ represents the sample data. It is assumed here that $\mathbf{y}_1, \dots, \mathbf{y}_n$ are observed values, drawn from a random sample, n, and, so, they are the realized values of n that are independently and identically distributed ($i.i.d.$). The finite mixture model, having G components (hereafter, called groups, [sub]populations, clusters, or latent classes) for $\mathbf{y}_i$, is defined by the composite density (i.e., the probability density function, $p.d.f.$)

$$f(\mathbf{y}_i; \varphi) = \sum_{g=1}^{G} \pi_g f_g\,(\mathbf{y}_i\,; \theta_g)\,.$$

The mixture proportion, π_g, represents the prior or posterior probability that the data observed for a particular individual comes from the component or group, g, and can be interpreted as the relative size of the g components. These mixing proportions or weights, $\pi_1, \ldots, \pi_G$, must respect the conditions that $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$. It is further expected that, within each component (i.e., conditional on belonging to component g), observation $\mathbf{y}_i$ is characterized by the density $f_g(\mathbf{y}_i; \theta_g)$ and φ stands for all of the parameters in the FM model. As this project's interest lies primarily in the FM model as it applies to clustering, the components or groups ($G_g, G = 1, \ldots, G$) are taken to define a partition of the data, and $i \in G_g$ indicates that observation i belongs to cluster g. In some formulations of the mixture model, the number of components, G, is taken as not only finite in number but also fixed, so that the test population is a mixture of g pre-defined groups that are, therefore, *a priori* known to exist. In, for example, the now very common use of FM in clinical studies, the problem may be to estimate disease prevalence (the mixing proportion, π_1) in a population, G, consisting of two pre-selected groups (patients with the disease, G_1, and those without, G_2), on the basis of some measured feature vector (McLachlan et al., 2004). This project applies FM models to a different goal. The intention here is to determine, first, if there is any structured heterogeneity in the samples of craniometric data. Is it possible to isolate clusters? If so, then, how is this variation distributed? Are there meaningful cluster memberships? Does this patterning have any relationship to pre-existing population definitions? Specifically, do the clusters produced correspond to known biogeographic ancestry or biosocial racial groups? To this end, this project assumes that the optimal number of clusters (i.e., the value of G) is finite but unknown for all of the test samples and it accepts that this component number must be determined along with the estimates for the mixing proportions, π, and parameters, φ. This flexibility (selection with constraint) in the number of clusters for finite mixture should not be mistaken for

component "adaptivity," as in a Gaussian Mixture Sieve model, where the estimate of the number of components can increase with the sample size n (Geman & Huang, 1982; Priebe 1994).

Unlike most other clustering methods, which are heuristic and exploratory, finite mixture models are themselves parametric density estimation models that can be productively fit using likelihood methods. Although the density for the continuous univariate case can be easily approximated using a nonparametric approach (e.g., by plotting a simple sample histogram), when the data is symmetrical and unimodal, it may be better to assume a normal distribution and adopt a parametric approach (Silverman, 1986). The distribution is, henceforth, taken to be "known" but the form of the mixture parameters must still be specified: for the normal distribution, it is necessary estimate the mean, μ, and the variance, σ^2. A similar parametric estimation approach can be adopted for multi-modal continuous (multivariate, high dimensional) data by approximating the density using a finite mixture of normal distributions; so, the unknown density can be seen as a weighted sum of parametric mixture components. The number of these components governs the negotiation between the efficiency of the parametric unmixed density (where, $G = 1$) and the flexibility of nonparametric methods (by allowing the component, G, to grow). This FM method can be said, therefore, to occupy a convenient space between parametric models and kernel density estimators (Jordan & Xu, 1995). It is because of this parametric approach (that is, because it is necessary to specify the form of the density of an observation in each of the underlying groups) that finite mixture analysis is alternatively called *model-based* clustering or mixture *modeling* (Banfield & Raftery, 1993; Fraley & Raftery, 2002). Finite mixture's elegant implementation via the EM algorithm and its use of probabilities for robust clustering give both support for its general utility and an explanation for its recent popularity. It is, however, the "model" in finite mixture *modeling* that has given it an invaluable practical advantage when data sets are complex

and highly structured, as well as considerable philosophic or theoretical appeal within certain scholarly circles (Basford & McLachlan, 1985). Hawkins and colleagues (1982) have commented critically on a trend in cluster analysis literature, where authors "lay more stress on algorithms and criteria in the belief that intuitively reasonable criteria should produce good results over a wide range of possible (and generally unstated) models" (p.353). They discuss, for example, the trace W criterion, its common misapplications and the willful violations of its key assumptions of normality and spherical clusters. It is not surprising that strong opinions have emerged that ardently encourage a model-based approach to clustering. Aitkin and coworkers (1981) have asserted that "it seems to us that, when clustering samples from a population, no cluster method is *a priori* believable without a statistical model" (p. 460). For finite mixture methods, they (1981) have written very positively that "[c]lustering methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory" (p. 421).

The mixture of multivariate normal (MVN) components or Gaussian distributions is the usual choice out of the available distributional options for most discussions of finite mixture analysis in a parametric clustering context; this is because, as McLachlan and Peel (2000) have already noted, "a common assumption in practice is to take the component densities to be (multivariate) normal" and "it is typical to proceed on the basis that any nonnormal features in the data are due to some underlying group structure" (p. 81). It is also important to realize that this choice is made given the fundamental and, arguably, irresolvable problem of testing for normality when using "unknown" samples. Hawkins (1981) has designed a reasonably good, combined test of the assumption of normality and homoscedasticity; it is an exact test for complete data when n_i are small and it has power equivalent to using the likelihood ratio test statistic in combination with the multivariate coefficients of skewness and kurtosis (after

Mardia, 1974). In its direct application to a mixed data set, it is only able to assess model fit for the specialized case of mixture with "known" or partially classified data, as in the framework of discriminant analysis, when assumptions are tested for each individual group. Muller and coworkers (1982) have proposed a modified application of Hawkins' (1981) original test. In this case, the data is clustered following the regular mixture likelihood approach, classifications for the individual observations are obtained as usual, and then Hawkins' (1981) test is applied to the clusters resulting from the FM analysis. Such *post hoc* testing is, however, only appropriate under the assumption that the mixture-based classification results are "correct" or if some prior information is available, which may aid in validating these cluster memberships. This revised procedure, and the methodological circularity that it seems to demand, beg the question: if reference data is required to verify the clustering solution, why would one not simply choose to apply, in the first place, some supervised classification procedure, like discriminant analysis, which is well-known for its ability to differentiate groups with measurable accuracy? McLachlan and Basford (1988, p. 60) have offered apt commentary on this issue of normality, which has been recently reiterated by McLachlan and Peel (2000, p. 85). In their opinion, "it is self-serving to cluster the data under the assumption of normality when one is subsequently to test for it, but it is brought about by the limitations on such matters as hypothesis testing in the absence of classified data" (McLachlan & Peel, 2000, p. 85). Given the present lack of adequate solutions for testing distributional assumptions for the situation in which the sample is "unknown" or completely unclassified, this project has chosen to follow this "practical" protocol and test for both world-wide and subpopulation group structure using craniometric data under the assumption that each component, cluster or group has its own multivariate normal distribution characterized by a vector of means and an unconstrained variance-covariance matrix.

For the multivariate normal (MVN) mixture, the conditional density for component g is taken to be the J dimensional normal distribution

$$f_g(y_i;\ \theta_g) = \varphi_g(y_i;\ \theta_g) = (2\pi)^{-\frac{J}{2}}|\Sigma_g|^{-1/2}\exp\left[-\frac{1}{2}(y_i - \mu_g)^T\Sigma_g^{-1}(y_i\ -\ \mu_g)\right]$$

where $\theta_g = (\mu_g, \boldsymbol{\Sigma}_g)$ with the mean vector, μ_g, and the covariance matrix, $\boldsymbol{\Sigma}_g$. The domain is $\mathcal{R}^J$ and the multivariate normal probability density functions form an exponential family. The covariance matrix for univariate data reduces to a scalar variance. Assuming a Gaussian mixture model for data consisting of observations n with G multivariate components, the corresponding likelihood is

$$\prod_{i=1}^{n}\sum_{g=1}^{G}\theta_g\varphi(y_i;\ \mu_g, \Sigma_g).$$

Data produced by mixtures of multivariate normal densities can be defined graphically: (1) clusters are centered at the means, μ_g, (2) there is an increased density in points as they approach the mean, and (3) the surfaces of constant density are ellipsoidal in distribution. The geometric attributes of MVN clusters are governed by the covariances, $\boldsymbol{\Sigma}_g$, which may be parameterized to impose restrictions across the components, as each component is modeled by the normal distribution (Dias, 2004; Celeux & Govaert, 1995; Bozdogan, 1992; Hathaway, 1985). For this multivariate Gaussian case, commonly used models with different parameterizations of $\boldsymbol{\Sigma}_g$ include:

1. $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_G = \sigma^2 I$, with σ^2 unknown
2. $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_G = \text{Diag}(\sigma_1^2, \ldots, \sigma_J^2)$ where $\text{Diag}(\sigma_1^2, \ldots, \sigma_J^2)$ is a diagonal matrix whose diagonal elements, $(\sigma_1^2, \ldots, \sigma_J^2)$, are unknown
3. $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_G = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is an unknown positive definite symmetric matrix
4. unrestricted covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G$

When $|\mathbf{\Sigma}_g| = 0$ for at least one component, the FM model is not defined; constrained parameterizations typically increase estimation stability. As the g-component normal mixture model with unrestricted covariance matrices is highly parameterized with $\frac{1}{2}n(n+1)$ for each component-covariance matrix $\mathbf{\Sigma}_g(g = 1, \ldots, g)$, choosing the formulation of the covariance structure is an important step towards improved estimation. This specification, i.e., the "constraint" or "full restriction," aids in model simplification. It helps to avoid over parameterization, especially when sample sizes are small, since there are $G(J+1)(J+2)/2$ parameters to estimate for the multivariate normal model. Moreover, it reduces the risk of model and irregularities by eliminating, for example, the problem of singularity. A general parameterization of the covariance matrix for multivariate normal mixtures using eigenvalue decomposition (that is, based on a variant of the standard spectral decomposition of Σ_g) has been proposed by Banfield and Raftery (1993). It takes the following unconstrained form, after Scott and Symons (1971):

$$\Sigma_g = \Lambda_g D_g A_g D_g^{\mathrm{T}}$$

where $\Lambda_g$represents a constant of proportionality, $D_g$is the orthogonal matrix of the eigenvectors, and $\mathrm{A}_g$is a diagonal matrix with elements proportional to the eigenvalues. As these factors in Banfield and Raftery's decomposition are considered to be independent sets of parameters, they can be treated individually, thus, allowing for variation among components, or they can be constrained uniformly across all components. Through parameterization, clusters can be defined with different or special-case features and, when parameters are fixed, shared-geometric properties can be assigned to all of the

components of the mixture: Λ_g determines cluster size (i.e., the volume occupied by the cluster in p-space, not its membership size), $D_g$controls the orientation of the gth component of the mixture, and $\mathrm{A}_g$defines cluster shape. In this way, the form, $\Sigma_g = \Lambda\, DAD$, denotes, for example, a model with the same constrained characteristics for all components (Friedman & Rubin, 1967). This parameterization can be alternatively expressed in clustering language (c.f., the MCLUST package; Fraley & Raftery, 2011) by using a model "identifier." This three letter notation encodes the model's geometric properties as "E" (equal), "V" (varying), or "I" (the identity). The model identifier, "EEE," indicates that the volumes, orientations, and shapes of all clusters are equal (E) and, therefore, corresponds to the mathematic expression, $\Sigma_g = \Lambda\, DAD$, given above. The options currently available in MCLUST (version 3) for the parameterization of the covariance matrix, Σ_g, and the relationship between the model "identifier" and the model form are summarized in Table 1 (after Fraley & Raftery, 2011).

Table 1: Available model parameterizations (reproduced from Fraley & Raftery, 2011, p.6)

identifier	Model	HC	EM	Distribution	Volume	Shape	Orientation
E		•	•	(univariate)	equal		
V		•	•	(univariate)	variable		
EII	λI	•	•	Spherical	equal	equal	NA
VII	$\lambda_k I$	•	•	Spherical	variable	equal	NA
EEI	λA		•	Diagonal	equal	equal	coordinate axes
VEI	$\lambda_k A$		•	Diagonal	variable	equal	coordinate axes
EVI	λA_k		•	Diagonal	equal	variable	coordinate axes
VVI	$\lambda_k A_k$		•	Diagonal	variable	variable	coordinate axes
EEE	λDAD^T	•	•	Ellipsoidal	equal	equal	equal
EEV	$\lambda D_k A D_k^T$		•	Ellipsoidal	equal	equal	variable
VEV	$\lambda_k D_k A D_k^T$		•	Ellipsoidal	variable	equal	variable
VVV	$\lambda_k D_k A_k D_k^T$	•	•	Ellipsoidal	variable	variable	variable

This variety of possible parameterizations, although highly advantageous, nevertheless, raises three important questions. Which of these available options represents the best choice of model? How should the covariance matrices be defined or restricted? And, when implementing an analysis in MCLUST,

which combination of geometric properties (E, equal, V, varying, or I, the identity) is "optimal" and how can this be determined? Although McLachlan and Peel (2000) have said that a normal mixture model with unconstrained component covariance matrices is often too general for practical use, they are hesitant to propose a blanket restriction of equal component covariance matrices for finite mixtures of "unknowns" (p. 97). Unlike the normal homoscedastic solution-based classifier (c.f., Fisher's Linear Discriminant Function) which is robust against slight departures from equal variances, they have found that similar mild departures may adversely affect the estimation of the component means and mixing proportions in the context of data without *a priori* identifiers or known patterns of the "true" memberships (c.f., Basford & McLachlan [1985] or McLachlan & Basford [1988] for their two component bivariate clustering of hemophilia data). Several solutions have been proposed, including the application of Hawkins' (1981) combined test for normality and homoscedasticity, Banfield and Raftery's (1993) parameterization method, or the use of factor-analyzers for high dimensional data with a relatively small sample size (McLachlan & Baek, 2010; Baek et al., 2010). As the factor-analyzer approach is concerned with issues of dimension reduction and overfitting (issues that are addressed in Chapters 2 and 4), only the first two possibilities are considered now.

For the question of [pre]testing for homoscedasticity, the same arguments made against the use of Hawkins' (1981) test and its revised procedure (Muller et al., 1982) to assess unclassified samples for normality can be reiterated here, since the conditions of the data and the purpose for which the FM model is being fitted have not changed: the goal remains to discover hidden structure, without using prior knowledge of sample composition. Reusing the words of McLachlan and Peel (2000), this approach would be equally as "self-serving" in this context. To this concern is the added problem of increasing model complexity. As mixture model with "unrestricted" component-covariance matrices in its normal

component distributions is highly parameterized, with $\frac{1}{2}p(p+1)$ parameters for each component covariance matrix $\sum_i$ $(i = 1, \ldots, g)$, this freedom comes at the loss of computational tractability, at the cost of parsimony in the analysis, and at the risk of failure in the Expectation-Maximization algorithm's computations, i.e., resulting from matrix singularity or shrinking components (McLachlan & Peel, 2000; McLachlan & Baek, 2010). Although assuming equal covariances may be an oversimplification of the true data structure, when a sample is large, when it is characterized with many variables and when it is likely composed of several underlying groups, this generalization may be the better and certainly more practical approach when computational requirements are considered; also, it may well be more satisfying in terms of the interpretability of the model's solution.

Fraley and Raftery (1998, 2007) have nevertheless shown how their mixture clustering approach can offer the advantage of flexibility in model fitting, so that it is possible consider finite mixture models with separate component covariance matrices. Banfield and Raftery (1993) have proposed a convenient method for model selection (including the choice of parameterization of the covariance matrices) that occupies the theoretical space between the most general or flexible model and the restriction of a common covariance structure. In this framework, practitioners are not required to make assumptions on the structure of the covariance matrices based upon arguments of parsimony and practicality or from the results of pre-processing steps (e.g., following Muller et al's routine of clustering, testing, accepting homo or heteroscedasticity, and then re-clustering), but instead enables model selection by a statistical measure of best-fit, using an "information criterion." Banfield and Raftery (1993) recommend the Bayesian Information Criterion (BIC), which has been shown to provide consistent and conservative estimates for a variety of theoretical conditions and practical examples (see Leroux, 1992; Roeder & Wasserman, 1997; Campbell et al., 1977; Dasgupta & Raftery, 1998) and which Fraley and Raftery (2011)

have conveniently integrated into the MCLUST package. In the MCLUST formulation, both parameterization and component-number selection (equally critical steps in model-based clustering) are combined in a single estimation procedure: the "optimal" solution is identified by first fitting models with different parameterizations and numbers of clusters to the "unknown" and "raw" (still mixed) data by maximum likelihood, and then by applying a BIC selection method (after Schwarz, 1978). Fraley and Raftery's formal selection procedure dictates that the highest BIC value generated is used to identify the best-fitting model, although in practical application, it may be reasonable to opt to choose among the top competing models if their BIC values are very close and the parameterization or component number associated with one model may provide a more realistic solution over another model with a closely comparable BIC estimate. Given these (ongoing) advances in the ease of selection using statistical procedures and increased computational power, McLachlan and Baek (2010) have recently advocated for the use of the BIC method and Peel and McLachlan (1999) have recommended the use unconstrained variances when the true relationships among the sampled observations are (or are taken to be) unknown.

Despite the differences between model selection by logic or prior expectations and the information criterion approach, in both cases, there remains flexibility in how the model is designed or determined. If these methodological arguments are extended to represent a philosophic point-of-view, it can be argued, as Konishi and Kitagawa (2008) have, that

> a statistical model is not something that exists in the objective world; rather, it is something that is constructed based on the prior knowledge and expectations of the analyst concerning the modeling objective, e.g., his knowledge based on past experience and data and based on the purpose of the analysis, such as the specific type of

information to be extracted from the data and what is to be accomplished by the analysis (p. 3).

There is, therefore, no singular recommendation for best-practice in model selection. Since this project adopts, in general, the theoretical position that preferences statistical testing over user decision-making, it is desirable to ground the restrictions on the covariance matrices (or the geometric properties of the components) in statistical arguments using a computationally realistic approach and an easily executed procedure. For these reasons, the BIC-based procedure, as it is available in MCLUST (Fraley & Raftery, 2011) is applied here in all "baseline" analyses. It is important, however, to take note of the words of caution offered by Konigsberg and colleagues (2009) on making indiscriminant use of this approach: "we cannot recommend turning mclust loose [...] adoption of the BIC in any analysis is largely a philosophical issue depending on one's thoughts about the desirability of model parsimony"(p. 83). It is felt here that this concern is most relevantly directed towards the issue of model selection in the service of identifying the optimal number of components for the clustering solution instead of parameterization.

While the choice among the competing models for the parameterization of the covariance structure contributes in varying degrees to the quality of model-fit, deciding among mixture models that span a range of possible numbers of components (i.e., testing for the number of g clusters or groups present in the mixture and used in its final solution) is, as McLachlan and Peel (2000) have aptly described, "an important but very difficult problem which has not been completely resolved" (p. 175). Its importance lies in the fact that, when applying mixture models to real data in practical situations, the actual number of groups is unknown (hence, structure is "hidden" or "cryptic") and must be inferred from the data itself. The number of components chosen dictates how this unobserved structure, that lies latent within in the mixed data set, is resolved into a statically determined pattern of clusters and, accordingly, how

the assignments of cluster membership are made. When the finite mixture model is used, as it is here, for the purpose of model-based clustering (that is, when the ultimate clustering goal is to obtain for each observation its probabilistic allocation to one of g groups), how the user comes to settle upon a "correct" model and the assumptions are made about the optimal number of components, G, in a g-cluster mixture have profound consequences for the quality of the clustering, the "reality" of the results in terms of their representation of a data set's true heterogeneous structure (regarding either group number or group composition), and the utility of these conclusions in any subsequent discussions of classification using the preselected traits or measured attributes. The specification of cluster number is the force that informs and, so, drives the implementation of the "mixture clustering" algorithm; the approach to its selection is therefore of paramount concern. It is not surprising then, that a survey of mixture literature yields an array of published methods and opinions on how best to estimate G; each with varying utility in practical applications. McLachlan and Peel (2000), Burnham and Anderson (2002), and Konishi and Kitagawa (2008) provide helpful overviews of the methodological options. *Fraley* and Raftery (1998), present, in particular, the theory that underlies the BIC approach in MCLUST, while McLachlan (1987) gives justification for his resampling method, which can be swiftky implemented in EMMIX and made to work using MCLUST. In general, all of the component selection methods can be ordered into four types, according to the theoretical framework and philosophical standpoint out of which they have emerged:

(1) Hypothesis Tests

- In this case, the smallest number of components for a "good" model is identified by successively testing of the null hypothesis of G components against the alternative hypothesis of $G + 1$ components, using the likelihood ratio test statistic, LRTS (e.g., bootstrapping the LRTS after

McLachlan [1987] and Feng & McCulloch [1996]; c.f., the Modified LRTS after Lo et al. [2001] and Lo, [2005]).

(2) Information Criteria

- Penalized likelihood criteria can be used to weigh the benefit of obtaining a "better" fit with an increase in the number of components against the expense of having to estimate many more parameters for the models with more clusters (e.g., using Kullback-Leibler estimators, like AIC, after Akaike [1973]; Bayesian criteria, like BIC, after Schwartz [1978]; or consistent criteria, like ICOMP, after Bozdogan [1994]).

(3) Classification Criteria

- These methods use an entropy or classification likelihood statistic to evaluate the degree of separation among the estimated posterior probabilities (e.g., Es after DeSarbo et al. [1992]; C and CLC after Biernacki & Govaert [1997]; c.f., also probabilistic and fuzzy indices, like NFI, after Roubens [1978].)

(4) Minimum Information Ratio and its related criteria

- In this less common case, information criteria are used but they are specific to the mixture modeling method, they are based upon properties of the information ratio matrix, and they are calculated as a function of the convergence rate of the EM algorithm (e.g., MIR after Windham and Cutler [2002]).

If the reach of this discussion is further extended, into the more specialized areas of mixture analysis (such as discriminant function and exploratory data analysis), it can be seen that there are other applicable procedures, which do not fit neatly within these larger categories. These options include, for example, non parametric graphical diagnosis tools (e.g., Roeder, 1994) and cross-validation techniques

(e.g., Smyth, 2000). This project adopts only the first two methods: although very different in philosophy, both of these approaches estimate the order of the mixture model (i.e., its number of clusters) by evaluating the likelihood. Hypothesis testing is chosen because of a general familiarity with the larger concept of "significance," its ease of interpretation (e.g., using test statistics, confidence limits, and p-values), its long history and scholarly currency in statistical literature, and its popular appeal in the social sciences (Ziliak & McCloskey, 2008). The information criterion approach is equally supported, because it provides a cohesive and objective methodology for a broad class of empirical problems, it allows for an extension of likelihood theory, and, in actual practice, it is quite straightforward in its computational demands and is easily applied, while still being steeped in robust statistical theory (Burnham & Anderson, 2002). Here, emphasis is placed upon on the bootstrap of the likelihood ratio test statistic (LRTS) and the Bayesian Information Criterion (BIC) specifically because they both have received considerable review: their strengths and weaknesses are well-known for a variety of data contexts and under many different statistical conditions. Furthermore, they have been incorporated, respectively, in the model-based clustering software, EMMIX and MCLUST, and, so, they can be implemented simply and effectively using mostly predefined functions in R. It is important to acknowledge, however, that despite this accessibility, the ease of application of these methods in actual practice will, nevertheless, vary with the individual practitioner's statistical and programming background.

Given the general popularity of significance testing and the common stress placed upon the importance of p-values in traditional statistical discourse, a natural approach to determining the optimal number of components in MVN mixture model is to answer the question of *how many components should be used* or predict *how many clusters exist* using hypothesis testing. In designing a test for this optimal number,

it would be desirable, for practical applications and with a concern for tractability, to address this problem in terms of identifying the smallest number of components in the mixture that is compatible with the data and to make use of the well-known and simply applied likelihood ratio test statistic (LRTS), defined as $-2log\lambda$. This approach should allow the practitioner to select a parsimonious but robust model and, ultimately, yield reliable clustering results. To this end, it would be logical to carry out successive hypothesis tests of the null hypothesis,

$$H_0: g = g_0$$

against

$$H_1: g = g_1$$

for some $g_1 > g_0$, assuming that $g_1 = g_0 + 1$. In this way, additional components can be introduced incrementally, up to the point at which g reaches some threshold and the likelihood begins to fall away. In simpler terms, it can be said that: the homogenous (null) 1 component model is tested against the heterogeneous (mixed) model with 2 components, then the 2 component model is tested against a model with 3 components. This method is continued accordingly until a predefined stopping point is reached, at which time the number of components takes value the of g_0 in H_0. The use of the LRT, with an asymptotic chi-squared (X_k^2) null reference distribution (with k degrees of freedom equal to the difference between the number of parameters under the null and alternative hypotheses) is, however, problematic in this case as the regularity conditions of Cramér on the asymptotic properties of the maximum likelihood estimator do not hold for the finite mixture model: under the null hypothesis the general mixing proportions are on the boundary of the parameter space and the parameters are not identifiable under the null model (Wolfe, 1971; Hartigan, 1985; Ghosh & Sen, 1985). A wide variety of experimental, simulation studies have actively engaged this problem, since its recognition in the recent history of mixture theory, and have, in turn, inspired the development of corrective solutions for the use

of the LR test in the context of mixture models. Kiefer and Wolfowitz (1956) concluded that the global MLE does not exist for a normal mixture whose component distributions have unequal variances; while Day (1969) attributed spurious local maximizers to the case when one component distribution has a small variance relative to the others in the mixture, and when its data points are few in number and not close together. Hartigan (1985) demonstrated that, when testing a single component MVN mixture against its two component solution with a shift in mean, the LRTS tends towards infinity in probability at a very low rate when the parameter space is unrestrained. Aitkin and Rubin (1985) proposed a solution which imposes a prior distribution on the vector of mixing proportions so that any null hypothesis for the number of components is specified in the interior of the parameter space. Quinn et al. (1987) have subsequently demonstrated, however, that the standard asymptotic results still do not hold for the Aitkin-Rubin approach. Goffinet et al. (1992) considered the example of a mixture distribution for which the mixing proportions were known *a priori* and they obtained results suggesting that the asymptotic distribution of the likelihood ratio test statistic is dependent upon the mixing proportion, the dimensionality of the problem, and if the within-component variance is known. Feng and McCulloch (1994) tested the claim that the simulated null distribution of the LRTS, when comparing a single normal against two normal distributions, is tied to the restrictions placed upon the component variances. McLachlan and Peel (2000) have recently reproduced this study and have obtained different results. They found, instead, that the simulated null distribution of the LRTS is close to X^2_{16} for $\acute{c} = 10^{10}$. Feng and McCulloch (1996) later extended the work of Redner (1981) and Self and Liang (1987) to establish that the maximum likelihood estimator is consistent when the true parameter lies on the boundary of the parameter space and in a nonidentifiable subset. Lo et al. (2001), after Vuong (1989), demonstrated that the LRTS is distributed asymptotically as a weighted sum of independent X^2 random variables with one degree of freedom and they obtained results from simulation studies indicating that the LR test can

be applied to a homoscedastic normal mixture with an adjusted LRTS. Lo and coworkers (2005) have also, more recently, proposed a LRTS procedure that tests for the number of components in a MVN mixture when variances are unequal. Aitkin et al. (1981) were first to present a simulation method to test two mixture models against one another, although it allowed for the realization of this likelihood approach in a limited practical application (only for the case of 1 versus 2 clusters). It was McLachlan (1987), however, who suggested to use likelihood ratio test statistic, $-2log\lambda$, with a parametric bootstrap procedure to estimate the corresponding p-value. His approach has been proven to be both feasible and effective when employed in real-world mixture contexts and it has continued to gain methodological weight from subsequent validation tests. For example, Soromenho (1994) recently compared the performance of five competing methods for determining cluster-number (e.g., Wolfe, 1970; Orlov as described in Celeux, 1987; Aitkin & Rubin, 1985; Celeux & Diebolt, 1985; McLachlan, 1987) and found that McLachlan's bootstrap approach consistently produced better results, both in terms of correctly identifying the true model and yielding high empirical power. Although effective, McLachlan's method has seen but limited use in applications outside of statistical or scientific literature; this is probably due to its multi-step, computationally demanding requirements. The procedure defining McLachlan's parametric bootstrap method involve the following steps: (1) proceeding from the beginning, under the null hypothesis, H_0, the value of the LRTS ($-2log\lambda$) is calculated and the parameters for both the one and two component FM models, where $g = g_0$ and $g_1 = g_0 + 1$, are estimated, (2) simulated data sets are generated from the parameters obtained for the one component model, (3) the one and two component models are fit to the simulated data and the LRTS is computed for these simulated data sets individually, (4) the observed LRTS is evaluated against the distribution of the LRTS calculated for the simulated data sets, and (5), finally, an approximated "p-value" can be produced, by taking the proportion of the simulated LRTS greater than or equal to the observed as its

estimate under the null hypothesis. McLachlan and Peel (1997) have admitted themselves that this procedure is intensive and time consuming, stating that "Even with a limited replication number, K, the amount of computation is still considerable" (p.261). McLachlan et al. (1995) have developed an algorithm for implementing the resampling approach in the case of normal mixture models. For the simplification of its use, Peel and McLachlan (1999) have included within the stand-alone version of the mixture software, EMMIX, the option to implement a bootstrap of the likelihood ratio test statistic at each step (i.e., for every test of $g + 1$, when fitting a range of numbers for g) and, ultimately obtain the desired estimate of the p-value. These same calculations can also be made independently, in, for example, R, by using the BOOT package (Canty & Ripley, 2009), which generates R bootstrap replicates of a statistic applied to any given data and allows for the use of either parametric or nonparametric resampling methods; R also permits the user to code a problem-specific function, which, in this case, would implement the desired parametric bootstrap according to the theory provided by, for example, Efron and Tibshirani (1993) or Davison and Hinkley (1997).

Although it has been shown (1) how resampling-based methods can offer the results (e.g., p-values) that are familiar to the general academic community and are often commonly expected as a sign of "good" scientific practice, and (2) that there are available software options for executing these procedures, the computational demands of carrying out a hypothesis test, based on successive runs of the clustering algorithm, which also involves bootstrapping, are considerable and frequently serve as an impediment to the use of this approach. Information Criteria (IC) offer another option for assessing model fit: using IC the practitioner can simply and quickly determine the optimal number of clusters as well as the model's parameterization. This information statistic approach evaluates the improvement in fit obtained from an increase in component complexity against the burden of estimating the larger number of

parameters for a model with (many) more clusters. As with all approaches based upon likelihood inference, it is possible to obtain a quantitative evaluation of the strength of the evidence supporting the estimates of the parameters in a given model (Royall, 1997). Although information criteria exist outside of the framework of hypothesis testing and, so, do not provide the customary measures of significance, they have the advantage of being computed and interpreted without relying upon standards based in subjective judgment or requiring external or additional user-defined information, e.g., α-levels or Bayesian priors. Information Criteria are also commonly favored because they are simple to apply. They require only the maximum likelihood attainable within a given model, rather than the likelihood throughout the parameter space. The BIC approach, for example, is frequently preferred for these practical benefits.

It is possible to assess for any mixture the strength of the evidence supporting the identification of the potentially "best" model from a collection of well-defined *a priori* candidate models, which differ in both their parameterization and number of components. In short, the BIC method, implemented via MCLUST, is easy to use and interpret, given that it produces tangible values; moreover, these values give the user the advantage of working with several ranked models that can be objectively identified as "good." In general, however, this ease of use does come at the cost of accepting a variety of assumptions, particularly normality or near-normality of the posterior distribution, which may be, unavoidably, violated in real-world contexts. Burnham and Anderson (2002) nevertheless highlight the obvious advantages, remarking that (1) most software routinely produce the maximum value of the log-likelihood and, so, the chosen information criterion can be easily calculated by hand, (2) the MLE of σ^2 can be calculated from the "standard output" provided by LS programs and the information criterion can be calculated from this estimate, and (3) many statistical packages generate values for select criteria

automatically without user-coding or calling to a function or procedure (p. 100). An estimate of the AIC is especially common among the values output by many popular programs, the BIC method for model selection is built into MCLUST as a batch-process, and EMMIX will report values for AIC, BIC and AWE.

The Bayesian Information Criterion (BIC) or Schwarz's Information Criterion (SIC), as proposed by Schwarz (1978) and included in MCLUST's mixture-clustering procedures (Fraley & Raftery, 2011), allows for model selection (i.e., for parameterization and cluster-number estimation) within a framework of Bayesian inference (Jeffreys, 1961). Despite its typical classification among Information Criteria or Statistics, the BIC does not in fact have a fully information-theoretic justification: true information criteria, like AIC (Akaike, 1974), are estimates of the (relative) Kullback-Leibler distance (Kullback and Leibler 1951) and, so, rely upon information entropy as a measure of information lost when a particular model is used in place of the (unknown) true model. As the BIC takes Bayesian evidence as its focus (to which it is in fact an approximation), candidate models are evaluated in terms of their posterior probability. According to Bayes' theorem, this posterior probability represents the probability of the data being generated from the ith model when the data x_n is observed. A logical extension for model selection is to assume that if one "best fit" model is to be selected from all candidate r models (those with varying combinations of cluster number and geometric properties), it should be the model with the largest posterior probability. If it is also accepted that prior probabilities across all model are equal, then, it is reasonable to select the model that maximizes the marginal likelihood. This marginal likelihood or marginal distribution of the data x_n, associated with the posterior probability of the model, can be approximated by using the Laplace method for integrals (Davies, 2002). This approach and, specifically, the utility of Laplace's method as a tool for approximating Bayesian predictive distributions, Bayes factors and Bayesian model selection criteria has been well-demonstrated (Davison 1986; Kass &

Waserman, 1995; Kass & Raftery, 1995; O'Hagan, 1995; Konishi & Kitagawa, 1996). The BIC is then formulated from the resulting expression as an approximation to twice the log Bayes factor and the value of the maximized log-likelihood, with a penalty being imposed for the number of parameters in the model (Kass & Raftery, 1995, Fraley & Raftery, 1998). The BIC is given, therefore, by

$$-2 \log f\left(x_n \middle| \hat{\theta}\right) + \ p \log n$$

where $f\left(x_n \middle| \hat{\theta}\right)$ denotes a model estimated by the maximum likelihood method (and, so, $\log f\left(x_n \middle| \hat{\theta}\right)$ is the maximum log-likelihood value), p represents the number of independent parameters of the model and $\log n$ is the penalty term. This approach can be generalized for questions of model fit using an information statistic approach by selecting the model with the number of components that minimizes the criterion, after supplying the corresponding penalty term, α, as it is commonly notated in the global equation: for example, α = 2 for AIC (Akaike, 1974), $\alpha = \log n$ +1 for CAIC (Bozdogan 1987), and α = 3 for AIC3 (Bozdogan, 1993). From these terms, it can be seen that the BIC and the CAIC have the added benefit of being "invariant" to the sample-dimension; that is, they identify the correct model with probability one as the sample size grows (Bozdogan, 1987). Such penalized criteria also offer other advantages that recommend them for practical use: (1) they permit comparisons of both nested and non-nested models (the BIC, in particular, has been shown to work well for logits, covariance structure and linear regression models; see Raftery [1995]), (2) they tend to favor the most parsimonious model (unless the gain in fit from a more complex model is very large), and (3), despite their simple definition and application, they are not *ad hoc* statistics but are rooted in established theory, so that, as Kuha (2004) has argued, "a selection based on such a criterion can thus be regarded, at least approximately, as a choice of the best model according to some underlying definition of optimality (p.190)."

For model-based clustering, the BIC facilitates a quick and easy comparison of models that differ in both geometric properties and number of components (and so does not require the component-wide restriction of equal variances) and has been conveniently integrated within the MCLUST software. For this reason, this project has chosen to follow the recommendations of Fraley and Raftery (1998) and make use of MCLUST's integrated BIC procedure as the primary method for choosing among the candidate mixture models. In this practical application, the highest BIC value is taken to identify the "best" parameterization and component solution, but additional "good" competing models (that favor parsimony) are also considered. This approach is adopted with the knowledge that actual biological phenomena are complex, correlated, and in many cases subject to the influence of other (potentially unknown or untested) variables. It is believed, therefore, that it is important to assume a philosophic position for model selection that best negotiates the balance between the concern for whether a given statistical model accurately represents the true structure of the data set and whether it is effective as a tool for teasing out useful information on the hidden heterogeneity within the mixture of observations. For this reason, it is wise to consider both statistical significance and biological meaning when employing this clustering approach and evaluating its results. It is important, therefore, to take into account arguments for parsimony, as a relatively simple model may prove to be more suitable both computationally and biologically, and to consider relaxing selection criteria when logic or experience advises a different, but still statistically "good," model choice. Graphical techniques are also recognized as helpful devices for visualizing and, then, evaluating the behavior of this selection criteria (e.g., the calculated BIC values) over different parameterizations and cluster numbers (c.f., Xu & Wunsch, 2009).

The problem of obtaining values for the model parameters (e.g., the mixing proportions, component means, and variances for a set of independent observations) that define a mixture's populations and their associated densities has been the focus of a large and varied body of literature in the last century; it is only in the last three decades, that the method of maximum likelihood (ML) has received its due attention and that rigorous ML work has been conducted in earnest. The 1970s marked a turning point in the future of statistics towards robust estimation methods: before then, maximum likelihood estimates "were often the subject of wishful thinking" but, since then, the maximum likelihood method has emerged as "the most widely followed approach to this problem, thanks primarily to the advent of high speed electronic computers" (Redner & Walker, 1984, p. 195). Moreover, in the frequentist framework, ML methods have been shown consistently to give better estimates for FM models (Fryer & Robertson, 1972) than those obtained using the often clumsy and exhaustive procedures like moment generating functions (Quandt & Ramsey, 1978). This is especially true when the EM algorithm is employed in maximum likelihood estimation: its effective use in the presence of incomplete data (i.e, for the E-step) and for the purpose of clustering (with iterations of the E and M-steps) has been clearly demonstrated. Moreover, it offers the advantages of responding well to acceleration procedures that speed-up convergence and of being reliably and easily implemented using a variety of software packages and programming environments (McLachlan & Krishnan, 1997; MacLachlan & Peel, 2000).

In the context of normal mixtures, the estimators or estimates (MLEs) of the model or mixture parameters ($\varphi = \pi_1, \dots, \pi_G, \mu_1, \dots, \mu_G, \Sigma_1, \dots, \Sigma_G$) are derived by maximizing the log-likelihood. The likelihood and log-likelihood functions are the basis for obtaining these estimators for the model parameters, given the observed data. While these two functions differ in their shapes, they have their maximum point at the same value, which is defined as the Maximum Likelihood Estimate (MLE). There

may or may not be, however, a unique value of φ that maximizes the log-likelihood function, $\ell(\varphi; y)$, and, in a non-identifiable model, there may be more than one (even ∞) "best" solutions. The ML estimate of φ can be obtained using some iterative procedure, which may include for FM models of unknown classes and latent class (LC) analysis: e.g., the EM algorithm (Dempster et al., 1977; McLachlan & Krishnan, 1997) and its extensions (c.f., Celeux & Govaert [1992] for the Classification EM "CEM" algorithm or Celeux & Diebolt [1985] and Wei & Tanner [1990] for the Stochastic EM "SEM" algorithm, which is itself the same as the Monte Carlo EM algorithm with a single replication), the Newton-Raphson algorithm (Everitt, 1987), the Wolfe-Day iterative method (Wolfe, 1970; Goodman, 1974) or alternative methods like global optimization using Simulated Annealing "SA" techniques (Brooks & Morgan, 1995).

Most "applied, computational statisticians" consider the EM algorithm to be, as Biernacki and coworkers (2003) have stated, "the standard tool for finding maximum likelihood solution" (p. 564). Meng and van Dyk (1997) calculated that, in the short 20 years following its introduction, over 400 publications have taken the development and improvement of the EM algorithm as their subject. The EM algorithm is undoubtedly very popular given its appealing properties: it is simply interpreted (allowing for its straightforward coding), it is quickly implemented in a wide-variety of software packages (with differing levels of user control), it is easily generalized to apply to many different contexts and can be specialized to aid in particular types of problems (like FM models for classification), it boasts numerical stability, and it can provide fitted values for missing data (in the E-step) without requiring further computation or the application of additional imputation methods. Its major limitation has been the problem of slow convergence (which is related to the amount of "missing" data to be estimated), but, in recent years, improved technology and extensions to the base algorithm have offered solutions that are both corrective and reasonably employed.

Although the underlying theory of the expectation-maximization procedure emerged out of the missing information principle proposed by Orchard and Woodbury (1972), and later advanced by Beale and Little (1975), the original formulation the EM algorithm proper was presented by Dempster et al. (1977). Subsequent discussions of its convergence properties followed by scholars such as Boyles (1983) and Wu (1983), and attempts have been made to increase its speed or rate of convergence and correct for computational "pathologies" (using techniques like conjugate gradient acceleration or quasi-Newton methods after Jamshidian & Jennrich, [1993, 1997]; or by recasting the EM into a proximal point framework, after Chrétien & Hero [1998]). McLachlan and Peel (2000) have addressed thoroughly the application of the EM algorithm to both FM analysis and the specific problem of model-based clustering when prior data on cluster number or composition is unavailable and when a priori identifiers of the sample observations have been withheld or are not known. As McLachlan and Krishnan (1997) have described the EM algorithm, its extensions, and related procedures in detail, only a summary of the EM algorithm is presented here in general contexts. McKeown (1996) has successfully demonstrated the utility of the EM algorithm in the replacement of MCAR craniometric data, showing its direct applicability to skeletal data sets in physical anthropology.

The EM algorithm is frequently associated with missing data methods because of its basic framework: a hidden space (y, z) is constructed, so that the MLE of the parameters, φ, is more easily computed in this space; it is assumed that there is some complete set of observed data (y, z) with the density $p(y, z|\varphi)$, where φ is a vector of parameters; φ, is estimated, when only y is observed, by maximizing $p(y|\varphi) \equiv L(y|\varphi)$. The EM algorithm can be characterized as, first, a multi-purpose method for optimizing the likelihood functions, which is useful to situations in which data is missing or when other optimization

methods have proven unsuccessful. In this general implementation of the EM algorithm, φ^0 is taken to represent an initial estimate of the parameters, φ; the algorithm then cycles between the expectation ("E"-step) and maximization ("M"-step), so that, for the E-Step, it calculates the posterior probability that the ith observation belongs to the gth component, given the current values of the parameters; it then computes the "missing values" (z_i) from the data with the current parameter estimates; in the M-Step, the algorithm then updates the parameter estimates using the estimated posterior probability and likelihood equations (c.f., Everitt & Hand, 1981); in this way, the complete-data log-likelihood is maximized, with each z_{ij} being replaced by its current conditional expectation, with respect to the parameters. It is because the posterior probabilities are unknown that an iterative method must be used to solve these equations. The two-step, E and M, process must continue, therefore, until the estimated values converge and a (MLE) solution is achieved. In most applied contexts, convergence is redefined in terms of reaching a specified tolerance level. Due consideration must be given to stopping criteria in order to realize the best possible execution of the estimation procedure. A common choice for the EM stopping rule is to consider either the magnitude of the relative change in the parameter estimates or the log likelihood. While this approach has been shown to be practically useful, it does come at the cost of refocusing the concern away from the question of the algorithm's actual convergence and onto, instead, a measure of its lack of progress; its practical advantage may, therefore, come at too great a philosophic price for some users (Lindstrom & Bates, 1988). Böhning et al. (1994) have proposed a stopping point for the EM algorithm that exploits Aitken's acceleration procedure; they suggest that iteration is terminated if

$$\left| l_A^{(g+1)} - l_A^{(g)} \right| < \text{tol}$$

where *tol* specifies a limiting value or tolerance. This stopping criterion is especially applicable to approaches that require resampling, including, for example, McLachlan's (1987) method of assessing the number of components in the mixture model. EMMIX offers the choice of using Aitken's acceleration, while MCLUST allows the user to define a tolerance level, accept the program's default (1.e-5, 1.e-5) or to impose a limitation on the number of steps or EM iterations. In general, imposing a restriction on the number of iterations is not recommended.

Validating the Mixture Solution

Marriott (1974) has remarked that the mixture likelihood approach to cluster analysis "is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model, investigates it by well-established statistical techniques, and provides a test of significance for the results" (p. 70). As his statement implies, one of the important steps after fitting a good mixture model to a given sample is to discern if this chosen model has validity. The practitioner must ask if the statistically inferred clusters are artifacts of some random fluctuations in the data set or if they represent a "true" organization of the observations into meaningful groups (McLachlan & Peel, 2000). It is necessary, therefore, to, not only, evaluate the statistical significance of the clustering choices (e.g., BIC or other IC values, p-values for the bootstrapped LRTS) but, to, also, inspect the clusters themselves for meaningful patterns of memberships. The way in which the model-based cluster solutions are validated may differ by project or practitioner as the choice of techniques is closely tied to the intention of the FM analysis and the goal of the user. McLachlan and Basford (1988),

for example, have noted that, in many instances, the data under FM analysis does not represent a true mixture of populations; fitting a model with g-components can be a statistically valid way to organize this homogenous data into g-groupings. In this case, there is no external reality against which the clustering results can be compared and, in turn, validated. In other applications, when the data is considered "unknown," the researcher may opt to evaluate the clustering patterns *post hoc* ; that is, only after the "best-choice" model has been fitted and a solution has be obtained. In this case, the practitioner interested in seeing if or how the organization of the data into g-components can be related to externally existing groups, e.g., "real" [sub]populations. By extension, it may be reasonable to investigate whether a given cluster solution reveals previously unidentified groups, suggests overlap or admixture among groups, or supports a reconsideration of how some preexisting populations are currently defined. When membership information and group identifiers (e.g., sex, population, age, subset A, sample 1, etc.) are available, it may be useful to, first, assume that the sample is "unknown" for clustering purposes and, then, later, conduct *a posteriori* tests that link the true group information to the "unbiased" clustering results and evaluate how well they correspond. These two sources of information on "membership identity," when combined, can be used to determine whether real heterogeneity exists within or among certain groups and if any of the observed (cluster or "true") membership patterns have useful meaning. For example, it is possible to: (1) evaluate FM method utility given particular constraints imposed upon the model or condition of the data, (2) validate or call into question previously established ideas on the true structure of the sample as represented by preselected variables, and (3) investigate the classificatory power of these variables (e.g., the characteristics, traits, measures, and metrics) commonly used in *a priori* analyses to assess the identifications made using supervised classification approaches. This procedure is applied to the current project in such a way that: (1) the mixture of craniometric data is clustered without prior knowledge of population membership, (2)

unbiased classifications are produced, and (3) the observed cluster patterns are compared against the known or "true" population affiliations. The results produced by the finite mixture analysis can be used in turn to determine (1) if statistically valid biological differences among groups can be detected by applying model-based clustering methods to standard craniometric data representing globally distributed samples, (2) if nested structure can be identified using local, regional, and continental groups, and (3) if predefined biogeographic and biosocial categories are both meaningful and realistic population descriptors when groups are formed based on craniometric variables alone, i.e., without the use of *a priori* identifiers or best-fit reference populations.

The strength of the clustering generated by the mixture modeling procedure can be assessed by evaluating the size of the fitted posterior probabilities of individual membership (McLachlan & Basford, 1988). It is assumed that the observed data $(y_1, \dots, y_n)$, although "unknown" in its classification or strength of group memberships, originates from a mixture (G) of a specified number of groups or populations $(G_1, \dots, G_g)$, each of which have their own MVN distribution and variance. As the FM model is probability based, it is further assumed that these observations can be effectively combined into clusters using their estimated posterior probabilities, $\hat{\tau}_{i1}, \dots, \hat{\tau}_{in}$ with respect to G_i $(i = 1, \dots, g)$. To this end, each individual observation within the full mixed data set is allocated to a particular cluster if the posterior probability of membership for that cluster is higher than that same value estimated for all other possible clusters. Therefore, the observation y_j is allocated to G_t if $\hat{\tau}_{tj} > \hat{\tau}_{ij}$ $(i = 1, \dots, g; i \neq t)$. From this allocation rule, it can be seen that posterior probabilities close to 1 imply well separated clusters and high classification success: in other words, the mixture assayed can detect sufficient internal structure within the data to cluster the available observations into g-distinct groups with a high degree of certainty. Low posterior values indicate, in turn, that the components are too closely related

to obtain reliable memberships. Although Ganesalingam and McLachlan (1980) and Basford and McLachlan (1985) have offered thoughtful discussions on possible estimators of the allocation rates for cluster membership, the suggested measures are known to overestimate classification strength, giving, as McLachlan & Basford (1988) have noted, "an optimistic assessment" (p. 127) even when a multi-step bias correction is introduced. It is not surprising then that approaches yielding more conservative estimates are typically presented in the relevant validation literature, including measures of uncertainty and cross-validation techniques.

From a review of the matrix of posterior probabilities associated for a given sample, individual memberships can be evaluated for any observation across all possible groups: it is reasonable to look for patterns in these probability values that may be considered "signatures" of complex biological processes and indicative of unique events (e.g., admixture). It is important to clarify that low posterior probability values do not necessarily offer evidence against heterogeneity; they are but indications of the mixture's ability to resolve the data at hand into well separated clusters using the information input. Individuals may indeed form cohesive clusters but these clusters may be overlapping. The practitioner's choice of variables, for example, may be capturing qualities shared among members of other clusters instead of emphasizing the features that differentiate the groups. A great advantage of probabilistic clustering is that it allows the user to obtain a statistically meaningful measure of what has been called the "uncertainty" in assignment (Fraley & Raftery, 2011). It is possible to evaluate the level of uncertainty in the classification of each individual from a sample, with respect to the assigned cluster membership, for any given model using the calculated posterior probabilities. It is reasonable to also generate the quantiles of uncertainty for the entire sample to assess of the overall quality of the classification, but, recalling that, when groups intersect, some uncertain classifications should be expected in the

overlapping regions, even when the overall clustering is considered very successful. In the context of a true "unknown" sample, these uncertainty values can serve as the measure of the clustering results. MCLUST offers a convenient graphical function that allows for an easy visual assessment of the relative uncertainty of any misclassified observations (Fraley & Raftery, 2011).

When true observation-origin is known for either a subset or all of the sample data, it is desirable to assess if, in fact, the configurations of the clusters produced by the chosen mixture model have biological meaning. In order to make this assessment, the posterior probabilities for group membership obtained for the "unknown" sample is compared against the true "known" affiliation documented for each individual (Bouguessa et al., 2006; Granzow et al., 2001; Handl et al., 2005). With known samples, it is possible to further assess the cluster results using supervised classification (where the number of classes, G, is taken as known) by calculating percent error rates; for example, clustering solution can be verified using mixture-based discriminate function analysis (Ripley, 1996; McLachlan & Peel, 2000; McLachlan, 2002; Fraley & Raftery, 2002). As posterior probabilities and error rates provide information on the quality of the classifications produced by the algorithm employed, they may speak to the utility of model-based clustering given the general parameters of the data set (e.g., continuous, large, multivariate, high dimensional), the overall type of data (e.g., human biological, skeletal, metric), and the specific properties of the individual data set(s) under analysis (e.g., world-wide distributed, within-population subclasses, selective sampling by age, cases of archaeological, modern or forensic significance).

Implementation: Mixture Software

In a time when the pencil, paper, and once requisite calculator, of mathematical calculus has been effectively replaced by the fast and efficient processors of statistical computing, any discussion of finite mixture models and the algorithms that support them requires a review of the software available for their implementation. Although cluster analysis has been commonly considered a lesser field within the discipline of statistics, being looked upon with skepticism for its presumed subjectivity, the growing popularity of the robust probabilistic and parametric mixture modeling approach to cluster analysis has resulted in a rapid development of model-based clustering tools. Built-in software options and stand-alone programs are becoming increasing available for the implementation of model-based clustering using finite mixture methods. This mixture-based approach to clustering is particularly suited for such an expansion as it can be flexibly defined: methodological choices can be made alongside programming decisions to develop an approach that meets the need of the users and supports the ideological standpoint of the program engineers. There are, therefore, just as many executable options as there are "mixture" programs. A variety of mixture-based models can be fit simply using proprietary statistical programs by calling, for example, to the pre-set procedures in the SAS/STAT software, like Proc FASTCLUST, Proc MIXED and NLMIXED, and by implementing third-party macros (SAS 9.2, 2010). Users wanting to make limited modifications to preexisting functions, while still working within the comfort of a well-supported software package, may be better suited by an interactive proprietary program like MATLAB (V.7.10; The MathWorks, 2010). Martinez and Martinez (2005) provide the simple commands necessary to employ the predefined mixture functions (see mbcfinmix, agmclust, gemix) using MATLAB's 2002 Toolbox and GUI. Several stand alone packages for finite-mixture analysis are available with differing levels of complexity: popular examples of these include EMMIX by Peel and McLachlan

(1999), GLIMMIX by Wedel (2001), MULTIMIX by Hunt and Jorgensen (1999), SNOB by Wallace and Dowe (1994),and NORMIX by Wolfe (1970). As the majority of these applications have been written to accompany cluster analysis textbooks or in support of publications on new or modified methods and algorithms, they are limited in their utility outside of their focused design and often there is a lack of flexibility in the implementation (i.e., no accessible code). Unlike the large propriety software programs which have been programmed for mass-market consumption, documentation for these types of applications is frequently sparse, updates are rare and bugs are common, and there are often problems of compatibility with different operating systems and platforms. At the same time, however, they offer quick and easy solutions to specific problems and are frequently posted online as free-access downloads or shared among colleagues. Model-based clustering may be most profitably implemented using the R environment for statistical computing (R Development Core Team, 2010). R features several open-source packages that have been specifically designed for mixture modeling, including: fpc (Hennig, 2000), mmlcr (Buyske, 2003), FlexMix (Leisch, 2004), EMMIX (McLachlan, 2011; McLachlan, Wang and Peel 2011), and MCLUST (Fraley & Raftery, 2011). For the purposes of this project, R and its extension packages, including MCLUST and EMMIX, are used to execute all mixture analyses and their graphical visualizations. It is argued, here, that R offers considerable practical and philosophical advantages for users who prefer greater control over how their analyses are being conducted, who advocate for transparency in methodology, and who support shared resources. As R is defined as "Free Software," it compiles and runs on a number of different operating systems, such as Unix/Linux, Windows and MacOS, and it undergoes regular updates in response to changes in technology. R is a GNU project ("GNU S") that is designed around the computer language, S. This open-source, editable structure allows users to augment program functionality by having free-access to the underlying source code, which can be modified or expanded upon using R's native language, C, C++ or Fortran. The programming savvy user

can, therefore, tailor any analysis to reflect personal interests and goals, write extension packages, and define new functions. As the reference system is dynamic, user-friendly documentation is continually updated in accordance with changes made to the base packages or functions. MCLUST and EMMIX were chosen for this project over the other available packages because of their stability and functionality: they can handle large data sets, they allow for batch-processing or partitioning of the procedures, and they have a variety of built-in options which can be replaced at any time with user-defined procedures or other R functions. Moreover, MCLUST and EMMIX rely upon well-documented methods and are supported by extensive statistical and programming literature: e.g., initial parameter values can be determined by model-based agglomerative or k-means clustering, models are fit by Maximum Likelihood Estimation using the EM algorithm (for which tolerance levels or stopping criteria can be easily specified), a variety of model parameterizations are available to choose from, the likelihood based methods are used to determine the optimal number of clusters and best model, posterior probabilities, cluster membership and uncertainty values are easily obtained, and a variety of graphical options are available as part of the core packages. MCLUST and EMMIX also have the added ability to facilitate related methods, such as discriminant function analysis and density estimation, or specialized procedures, like model regularization with a prior or the introduction of noise. Despite these obvious computational advantages, R, and, by extension, the MCLUST and EMMIX packages have been commonly criticized over their inaccessibility: that is, in order to execute these procedures properly and with any required manipulation, the user is expected to be knowledgeable in the specifics of R coding and general programming. For this reason, this project also tested the stand-alone, EMMIX, software (Peel & McLachlan, 1999) for normal mixture modeling. Although its options are pre-defined, EMMIX does cover all of the procedures relevant to FM methods and, through a simple user interface, facilitates the individual selection of procedures, thus, allowing for both easy execution and thorough analyses.

The user can, for example, adjust the upper limit for all arrays to accommodate large data sets, simulate a sample from a normal mixture model, assess standard errors and the number of components by bootstrap methods or information criteria (BIC, AIC, AWE), use Aitken's acceleration when bootstrapping, fit a normal or *t*-component mixture model for a specified number of clusters or a range of numbers, fit models by Maximum Likelihood via the EM or Stochastic EM algorithm, define the covariance structure, obtain starting values based on a variety of initial clusterings (using hierarchical and k-means methods and random partitions of the data), generate random seeds, define stopping rules or thresholds, specify partial classification, perform discriminant analysis and obtain predictions for target samples. In addition to offering a comprehensive list of standard options, the EMMIX software returns easily interpretable results and allows for convenient exportation of data to external programs for graphing and further analysis. This EMMIX software is also advantageous as it corresponds to what are considered the most important recent publications on finite mixture models (McLachlan & Bassford, 1988; McLachlan & Peel, 2000), the EM algorithm (McLachlan & Krishnan, 1997), and discriminate function analysis (McLachlan, 1992).

Advantages of Finite Mixture Methods

It is hoped that, over the course of the historical, theoretical and methodological discussions, it has been made clear that the statistical framework supporting a model-based clustering methodology offers several advantages over the more traditional clustering techniques:

1. FM methods provide a (semi) parametric approach to clustering that is constructed upon sound probability theory. Model-based clustering, therefore, supports the identification of sample heterogeneity and the estimation of cluster memberships with some measureable degree of statistical certainty.
2. *A priori* group identifiers are not required in order to generate probabilistic cluster memberships, evaluate individual allocations, and validate results. FMA provides a statistically rigorous, unbiased, alternative to supervised learning methods that require the use of known groupings of some measures (i.e., a "reference" sample) to categorize others groups or individual observations (the "target" sample) and infer the structure of the data as a whole.
3. Knowledge of the "true" number of groups hidden within the mixture of sample data and, in turn, the number of clusters to be used in the analysis is not required prior to running the FM analysis. The optimal number of components can be determined instead *in tandem* with the full analysis (that is, during the execution of the batch-clustering procedure) using either rigorous tests (e.g., a parametric bootstrap of the LRTS) or well-known information criteria (e.g., BIC, AIC, AWE).
4. Missing values are not only accommodated by finite mixture methods but can even be estimated (assuming that they are MCAR, "missing completely at random," after Rubin [1976] and Little [1988]) within the clustering procedure itself (at the E-step of the EM algorithm). As the EM algorithm was initially developed under the rubric of missing data estimation methods and is frequently employed for imputation purposes alone, if offers robust estimates grounded in well-documented statistical theory.
5. Variables with different measurement scales scan be easily accommodated, predictive variables can be included as covariates, and profiling variables can be incorporated into any model.

6. Multilevel structured data (e.g., in the form of three-way arrays), complex sampling surveys, and repeated observations or measurements can be handled without preprocessing before clustering (e.g., without their reduction to a one-mode or two-mode two-way array and without repetition).

The statistical and computational properties of FM mixture models offer a robust alternative to supervised classification methods. Mixture models display sufficient sensitivity (1) to detect heterogeneity in even large and complexly structured data, (2) to differentiate, when warranted, among relatively similar groups and identify subclasses within these groups, and (3) to classify unknown observations based upon degrees of similarity or population membership. All of these procedures are conducted in the absence of prior identifiers or "true" information on individual associations and in the presence of very mathematically satisfying, probabilistic, model-bound, and efficient algorithmic approaches.

Chapter 4: Materials, Methods and Analysis

Project Rationale

"The present work is [...] a search for specific distinctions between the populations of different major regions. This is of course analogous to the idea of 'racial' differences of past anthropology, but on an objective and systematic basis, not on typology. The hope is to arrive at valid comparative descriptions of the several populations, such as might be useful in problems of the origins of recent humanity in general."(Howells, 1989, np).

As this opening statement to Howell's (1989) investigation into the relationship between "skull shapes and the map" reiterates, coming to terms with the patterns of human variability has been a central focus of biological anthropology throughout its disciplinary history. The collective interest in the phenotypic differences long observed in both past and present populations was first driven by a desire to characterize *the* "human races" and to identify their origins (Coon, 1962, 1965; Garn, 1965). In contemporary anthropology, this interest has been repositioned to place emphasis on determining the broadly-based evolutionary, demographic, and environmental forces that have actively shaped the distribution of this variation (Relethford & Harpending, 1994; Relethford, 1994, 1997, 2002, 2004; Harvati & Weaver, 2006; Martínez-Abadías, Esparza & Sjøvold, 2009; Nakashima et al., 2010). Ongoing developments in population theory also continue to support smaller-scale analyses. Current researchers are taking up carefully focused questions of biological distance, history, and ancestry for target groups or geographic areas (Buikstra, Frankenberg & Konigsberg, 1990; Shriver et al., 1997; Jantz & Owsley, 2001; Shriver & Kittles 2004; Westcott & Jantz, 2005; Hanihara, Yoshida & Ishida, 2008; Hanihara &

Ishida, 2009; Kawakubo et al., 2009) and, at the same time, population-specific knowledge is being used to satisfy the practical demands of the medico-legal community for biologically-based methods of human forensic identification (Sauer, 1992; Ross, 2004; Ousley & Jantz, 2005; Spradley et al., 2008; Konigsberg, Algee-Hewitt and Steadman, 2009; Ousley, Jantz and Freid, 2009; Kahn, 2010; Sankar, 2010). While there are few if, arguably, any quantitative traits that display the "extreme global differentiation and environmental correlations" (Relethford, 2009, p. 17) necessary for deriving simple explanations of human adaptation and interaction, mean craniofacial measurements are known to differ in size and shape between [bio]geographically and [bio]socially defined groups (Howells, 1973, 1989, 1995; Sjøvold, 1984; Hanihara, 1993, 1996, 1997, 2000). When assessed together, these craniometric variables give quantitative signals of ancestry that are frequently well retained, even despite the immediate forces of extrinsic factors, like climate or changed environments; these signals, in turn, offer evidence supporting a significant genetic component to cranial morphology (Devor, 1987; Gill, 1998; Vidarsdóttir, O'Higgins & Stringer, 2002; Sparks & Jantz, 2002; Roseman, 2004; Roseman & Weaver, 2004; Betti et al., 2009; Logan & Jantz, 2010). It is for this general neutrality that craniometrics have been used with measurable success to examine morphogenetic differences, among and within human groups, and to address questions of evolutionary, paleodemographic, and modern forensic concern. To this point, Steele and Powell (1999) have acknowledged that there exists a general acceptance "that multivariate analyses of the cranium and the dentition can be as accurate in reflecting genomic relationships as the analysis of blood and serum protein" (p.118), while Relethford (2009) has remarked upon the value of this data in a recent discussion of race, stating that "the average patterns of global craniometric variation often provide a reasonable reconstruction of population history" (p. 17).

Either from the global perspective of nested hierarchical population substructuring or, more specifically, in the applied context of group discrimination and ancestry estimation for individual identification, studies addressing human cranial variation are extensive and continue to be widely discussed (see Edgar & Hunley, 2009; Konigsberg, Algee-Hewitt & Steadman, 2009; Ousley, Jantz & Freid, 2009). Despite this ubiquity, a full appreciation of the contributions to the field made by these studies has been frequently overshadowed by both philosophic and methodological concerns that have been tied to research design. Many studies have been, often unavoidably, constrained in their scope of analysis by the limited availability of representative samples and have been readily criticized for their use of statistical techniques that rely upon predefined reference populations, involve external supervision, and require the input of *a priori* information or user-assumptions (e.g., Brown & Armelagos, 2001; Armelagos & Van Gerven, 2003; Williams, Belcher & Armelagos, 2005; Albanese & Saunders, 2006). In response to these methodological issues, which are tangible and testable, a multi-part analysis is presented here. This project makes use of three well-documented sources of craniometric information: the Howells (Howells, 1996) data set, the Hanihara (Hanihara, 2008) data set, and the Forensic Anthropology Data Bank (Jantz, 2008). These samples collectively permit large-scale analyses and tests of population structure both at a variety of partitions (i.e., at continental, regional, local levels) and for differing investigative goals. This project implements the model-based methods of finite mixture analysis as alternatives to the supervised clustering or discriminant function approaches that have been traditionally employed in population studies using craniometric traits (McLachlan & Peel, 2000; Ousley & Jantz, 2005).

The multi-sample model-based clustering approach adopted here has clear advantages for the recognition of latent, population-specific patterns using craniometric data and for their interpretation in terms of geography, ancestry, and race. First, this statistical method offers a means to reveal cryptic

population structure within a given mixture of "unknown" individuals, for whom all identifiers of true group membership or proxies for genetic relationships (e.g., sex or population, race, ancestry, ethnicity) are not available *a priori* or have been intentionally removed for the purpose of conducting uninformed or "unbiased" analyses. As finite mixture modeling algorithms are designed to operate without the inclusion of external information on group affiliation for "test" data, and do not necessitate the incorporation of comparative reference or "training" samples, the practitioner can effectively circumvent the arguments of user-induced bias, forced allocation, and the use of "*a priori* racial criteria" that are frequently levied against supervised classification techniques (Armelagos & Van Gerven, 2003, p. 60-61). Second, model-based clustering allows for a close comparison between the population-cluster results obtained from genetic analyses using popular mixture software, like STRUCTURE, and the findings from phenotypic tests using craniometric data. This symmetry in methodology is of interest as past studies have been argued to "show that global patterns of craniometric variation reflect largely underlying patterns of genetic relationship, which in turn reflect geographic structure" (Relethford, 2009, p. 19). A comparative analysis using mixture methods has, therefore, the potential to provide the skeletal biologist with preliminary information on the relationship between the biogeographic patterns of genotypic and phenotypic data in a way that differs from earlier tests using F_{ST} (c.f., Relethford, 1994, 2004; Roseman, 2004; Roseman & Weaver, 2004). Third, this mixture method is grounded in robust probability theory and supports convenient and mathematically satisfying tests of best-fit for the model parameterization and the selection of the optimal number of components in the mixture. Moreover, it allows for the empirical evaluation of the individual classifications produced accordingly from the chosen cluster solution. As Marriott (1974) has noted, this mixture likelihood-based approach "is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model, investigates it by well-established statistical techniques,

and provides a test of significance for the results"(p. 70). As model-based clustering permits estimation and hypothesis testing using widely accepted and easily accessible procedures of "standard" statistical theory, it is possible to also avoid the well-known weaknesses in studies that attempt to validate cluster results derived from the classic exploratory data (EDA) methods. Familiar statistics can be simply extracted from mixture analyses and, often, regular procedure-output automatically includes p-values for tests of significance for an increase in the component number, log-likelihoods for any given solution, information criteria values for model optimization, posterior probabilities of membership for each individual, component means, and matrices of the mixing and variance parameters for a chosen model. The value of "testability" and finite mixture's *model-based* approach from an evidentiary standpoint is obvious when attention is paid to Aitkin, Anderson, and Hinde's (1981) cautionary statement: "when clustering samples from a population, no cluster method is, a priori believable without a statistical model"(p. 460). Finally, the use of craniofacial measurements drawn from the large, world-wide data sets collected by Howells (1973, 1989, 1995, 1996) and Hanihara (1993, 1996, 1997, 2008) ensures robust statistical tests of population differences, both in terms of sample size and breadth of geographic representation. The use of individuals obtained from the Forensic Anthropology Data Bank (Ousley & Jantz, 1997; FDB, 2005; Jantz, 2008) allows for investigations into the importance of biosocial race and biogeographic ancestry in forensic anthropology using context-specific data, which most accurately reflects the craniometric diversity present in modern American casework. The combination of these craniometric data sets with demographic information obtained from the United States Census Bureau (2010, 2011) serves as a means to contextualize these samples within the range of the racial diversity represented in the American population-at-large. A further account of the samples used as well as the methodological considerations and procedures relevant to this project's analyses are provided in detail below. As it is argued here that data access and methodological transparency are key factors in good

scientific practice, this project makes use of samples and software that are freely available online or by request.

Global Patterns of Craniometric Variation: The Howells and Hanihara Samples

The intention of this project is to assess the usefulness of finite mixture methods to answer the broadly-based questions of the accuracy, reliability, and biological validity of metric group discrimination using craniofacial traits. The application of model-based clustering to a sample of "mixed" craniometric data is taken as the first step in the investigation of population-specific patterns in cranial variation. The goal, here, is to reveal cryptic population structure within this data and ultimately resolve the underlying population distributions into statistically and anthropologically meaningful clusters, in which all members share more common biogeographic ancestry or biosocial affinity than individuals assigned to other clusters. To do so effectively, a large multivariate data set is required. It must be composed of several well-documented cranial series that have been thoughtfully chosen for their diversity, geographic coverage and degree of completeness. It is expected that "best-choice" data adheres to the following criteria: (1) true group affiliations are known *a priori* for testing and validation (2) estimates of sex are reasonably certain, (3) the range of morphology in modern human crania is represented, and (4) world-wide geographic variation is adequately (if not equally) sampled. This project's primary global sample is drawn from the WW Howells Craniometric Data (Howells, 1996). This full final data set[7] was compiled from 1965 to 1980 and, as used here, incorporates both his preliminary measurements on 17

[7] The cranial series presented in Howells' (1989) Appendix A is referred to here as his "final" data set (2524 individuals and 28 populations); an additional 524 specimens, including potentially problematic cases as well as paleoanthropological specimens and casts collected during the same data collection trips, are distinguished from his original data set hereafter as his "TEST" data set.

"core" "regional" populations, as already reported on in his 1973 study, *Cranial Variation in Man,* as well as his later extensions, which incorporated additional "local" groups drawn especially from the Far East and Pacific for the purpose of targeting otherwise underrepresented areas. This "Howells" sample includes a total of 2504 crania obtained from the full series (n = 2524, 57 variables) of the 28 male and 26 female populations reported in Appendix A of Howells' 1989 monograph, *Skull Shapes and the Map* (Table 2). It is believed that this data set is especially well-suited for tests of population structure, both for the quality and system of data collection, as well as for the history of use of Howells' measurements. Howells' data set has been considered in several foundational papers on the global distribution of modern human craniometric variation (Howells, 1973, 1989, 1995, Relethford, 1994, 2001, 2002, 2004) and makes up a part of the reference skeletometric database included within the forensic identification software, FORDISC (Ousley & Jantz, 2005).

Howells collected data on complete specimens for three local populations within each of six predefined geographic areas: Europe, sub-Saharan Africa, Australia and Melanesia, Polynesia, the Americas, an the Far East (see Howells, 1989, Appendix A). For Howells (1989), these 18 regionally defined populations "are meant to serve as a sort of core of modern humanity, distributed 3 to a region, to give some kind of central estimate of means and interpopulation variation" (p. 2). While he felt that these samples represented "the range of modern human variation" (1989, introduction), he also recognized the value of covering the "pattern of human variation reached by local differentiation" (1989, p. 2) and, so, included within his final data set 10 additional "local" series. These local populations present cohesive, comparatively unique, and, often, unusual, morphologies (e.g., extremes in size and shape) and often do

Table 2: Sample composition of Howells' craniometric data set

Population Designations (Howells 1989, 1996)		*Data Set with 57 variables*			*Data Set with 12 variables*		
Name	Number	Females	Males	Total	Females	Males	Total
AINU	25	38	48	86	38	48	86
ANDAMAN	24	35	35	70	35	35	70
ANYANG	28	0	42	42	0	42	42
ARIKARA	13	26	42	68	27	42	69
ATAYAL	19	18	29	47	18	29	47
AUSTRALIA	7	49	52	101	49	52	101
BERG	3	53	56	109	53	56	109
BURIAT	26	53	55	108	54	55	109
BUSHMAN	23	48	41	89	49	41	90
DOGON	5	47	46	93	52	47	99
EASTER ISLAND	11	37	49	86	37	49	86
EGYPT	22	47	58	105	53	58	111
ESKIMO	27	54	53	107	55	53	108
GUAM	21	27	30	57	27	30	57
HAINAN	18	38	45	83	38	45	83
MOKAPU	10	49	51	100	49	51	100
MORIORI	12	51	57	108	51	57	108
N JAPAN	16	32	55	87	32	55	87
N MAORI	30	0	10	10	0	0	0
NORSE	1	55	55	110	55	55	110
PERU	15	54	55	109	55	55	110
PHILIPPINES	20	0	50	50	0	50	50
S JAPAN	17	41	50	91	41	50	91
N MAORI	29	0	10	10	0	0	0
SANTA CRUZ	14	51	51	102	51	51	102
TASMANIA	8	42	44	86	42	45	87
TEITA	4	35	31	66	50	33	83
TOLAI	9	54	56	110	54	56	110
ZALAVAR	2	45	53	98	45	53	98
ZULU	6	43	55	98	46	55	101
Grand Total		*1122*	*1364*	*2486*	*1156*	*1348*	*2504*

not represent a real population unit and time span. Consider, for example, Howells' Egyptians, who, although "African," share few features with the sub-Saharan groups and instead best resemble Europeans, with retracted faces, prominent nasal projections, and limited prognathism; the Inugsuk Eskimo of Greenland, while often considered "North American" by geographic affiliation and "Native American" or "American Indian" by origins, are both unique and extreme in their "Super-Mongoloid" morphology; or the San Bushman, who collectively share a pattern of general African morphology (with low faces, broad nasal apertures and large interorbital distances), but, as a series, were spread-out over a wide geographic area (Howells, 1989, p. 13-15). It is for these reasons of inconsistency that certain groups, even when well geographically defined or having good morphological concordance, were excluded by Howells as typical representatives of a "core"area. Howells' method of exclusion results in the reduction of overall variation when consideration is given to only his six core populations. This approach offers the opportunity to conduct a conservative test of craniometric differences and provides, therefore, as Relethford (2001) has noted "an ideal sample with which to contrast variation within and among regional samples" (p. 630). It is believed, however, that it is important to address the maximum range of variation that is testable using all of Howells' series, even at the risk of complicating the results by the inclusion of these local and atypical groups that may be seen as morphological, spatial or temporal outliers and may contribute noise to the data. Rarely in practical contexts of group discrimination do anthropological practitioners have the luxury of artificially restricting a sample to cases with "normal" morphologies. In reality, it is often necessary to account for low typicalities among individuals inhabiting the extreme tails of population distributions or areas of distributional overlap (attributable to, e.g., pathologies, complex histories, admixture) when making personal identifications and membership allocations. Moreover, as this project adopts finite mixture methods for the express purpose of conducting a blind analysis of population substructuring, by stripping Howells' data of its

identifiers, it would be not only self-serving but counterintuitive to cull the available groups based on perceived expectations of "normal" or "acceptable" levels of variation. To a similar point, although sex is known with certainty for only the Zulu and North and South Japanese samples, both males and females for all populations are included in this analysis. While knowing sex with some measurable and high probability is an important factor when building a case-specific biological profile in forensic anthropology, Howells' sex designations by personal inspection are not problematic for tests of world-wide variation: the separate male and female group samples are sufficiently large ($\approx n \geq 50$) and the reliability of Howells' assignments can be taken with a reasonable degree of confidence to assume that the introduction of a few misidentifications would not adversely influence the quality of the clustering results. The decision was made, therefore, to pool the sexes and correct for differences in size by scaling Howells' data using the geometric mean, following the method proposed by Darroch and Mosimann (1985). In keeping with Howells' (1989, p. 3, 9-11) recommendations, this mixed-sex approach better approximates the true range of human variation and avoids issues of sex-based bias by accounting for sexual dimorphism. Howells' individual sex estimates were also considered in sex-specific, scaled and unscaled, analyses in order to detect general trends in allocation, which may suggest that the cluster solutions are being driven by uniquely male or female shapes. Further consideration is given to this issue of "known" sex only when samples are chosen from the Forensic Anthropology Data Bank (Jantz, 2008) for evaluating the importance of race in making positive identifications.

Howells' data set contains a maximum of 82 craniometric measurements; 57 of these are available for 2486 individuals, when all 28 males and 26 female populations are used (Table 3). For the majority of this project's analyses, a subset of 12 standard craniometrics variables (Table 3), as defined by Howells (1973, Appendix B), are used. These "base" inter-landmark distance (ILD) measurements were chosen

first for the fact that they are well-published, traditionally collected, and easily and reliably taken. They also have the advantages of being common among the statistically significant predictors selected by algorithmic methods, corresponding well with the architectural features and dimensions of the human craniofacial complex, and accounting for the morphological characteristics that are considered key to group separation and the estimation of population affinity. In order to respond to common concerns over the most informative variable subset and, in particular, recent studies that have used or advocate for the use of very many or all available craniometric variables (e.g., Relethford,1994, 2001, 2009, Pontikos, 2004; Hubbe & Neves, 2007; Elliott, 2008; Strauss & Hubbe, 2010), Howells' data (1989) was also evaluated for evidence of [sub]structuring using model-based clustering with all 28 populations, pooled sexes and the full complement of 57 size-corrected metric variables. As this secondary analysis

Table 3: Craniometric variables selected (after Howells 1973)

57 Craniometric Variables				*14 Variables*	*12 Variables*
GOL	NLH	EKB	PAF	BBH	BNL
NOL	JUB	DKS	OCC	BNL	BBH
BNL	NLB	IML	OCS	BPL	BPL
BBH	MAB	XML	OCF	DKB	NPH
XCB	MDH	MLS	FOL	FRC	NLH
XFB	MDB	WMH	NAR	GOL	OBH
ZYB	OBH	SOS	SSR	NLB	OBB
AUB	OBB	GLS	PRR	NLH	NLB
WCB	DKB	STB	DKR	NPH	DKB
ASB	NDS	FRC	ZOR	OBH	FRC
BPL	WNB	FRS	FMR	OCC	PAC
NPH	SIS	FRF	EKR	PAC	OCC
ZMB	FMB	PAC	ZMR	XCB	
SSS	NAS	PAS	AVR	XFB	
			VRR		

enables a direct comparison between the clustering results produced under the single changed condition of many versus few predictor variables, it is possible to assess any corresponding changes in

algorithmic stability and trends in allocation accuracy and to offer informed recommendations for future applications of finite mixture in craniometric studies.Howells' craniometric data has received considerable use since its final version was first fully described and documented in his 1989 report. Interest in it has continued to grow over the last 15 years, after it was made publically available online and the Department of Anthropology at the University of Tennessee became its webhost (Howells 1996).[8] The final data set has undergone extensive pre-processing, screening and validation, as Howells has explained: "Possible errors in recording had been looked for by checking all measurements against predictions from multiple regression equations, and questioned figures were tested by querying kind friends at the institutions housing collections or by re-inspecting specimens myself"(p. 441). Missing data and measurement accuracy are, therefore, not of concern for Howells' (1989) final full data set. The analysis of this data set is also unaffected by issues of sample proportion (i.e., balanced sex ratios and equal representations of each group) as Howells took great care to ensure consistent numbers among and within each series, selecting when possible 50-55 specimens per sex, per population (Howells, 1989, p. 1-3). This, in fact, remains the case for all of this project's data, as finite mixture analysis has been shown to operate reasonably well regardless of sample proportions, a factor which is otherwise known to adversely affect the quality of analysis of many other, more popular classification techniques (Holden & Kelley, 2010).

The relatively recent but active history of use of Howells' craniometrics for questions related to the world-wide distribution of craniofacial variation, biological affinity and population-specific morphologies

[8] The WW Howells Craniometric Data are presently maintained online at http://konig.la.utk.edu/howells.htm. The cranial series presented in Howells' (1989) Appendix A is available for download under the file name HOWELLS (2524 individuals and 28 populations); the additional 524 specimens making up the "TEST" data for Howells' 1995 monograph are also available for download at the same URL, under the filename HOWTEST.

(e.g., Relethford, 1994; Roseman, 2004; Roseman & Weaver, 2004; Ousley & Jantz, 2005; Strauss & Hubbe, 2010) recommends this data set as a reliable and applicable resource for the current project's test of global population [sub]structuring. Nevertheless, studies using Howells' data in population discrimination, ancestry estimation and individual identification contexts have received notable criticism. For example, Leathers and colleagues (2002) have argued that "it is clear that Howells was intent on providing known groups that would be distributed among the continental 'racial' groups" (p.100) and that "classifying populations, whether by geography or by 'race', is not morphologically or biologically accurate because of the wide variation even in homogeneous populations. Howell's samples lack the distinct morphology necessary to make Fordisc 2.0 a useful tool for classifying unknown crania." More recently, Relethford (2009) has acknowledged a potential weakness in his own use of Howells' core data; specifically addressing the results of his UPGMA (Unweighted Pair Group Method with Arithmetic Mean) cluster analysis using a phenotypic distance matrix for 18 of Howells' populations. While his analysis did produce clusters with "clear separation into the six geographic regions sampled by Howells," he qualifies his findings with the caveat that "these clusters reflect the fact that craniometric variation (like genetic marker and DNA marker variation) is geographically structured and that Howells sampled regions at far distances from each other rather than a random geographic sampling" (p.18). Similar concerns for uneven, insufficient or biased sampling across geographic regions have also be voiced in the genetics literature by critics who question the validity of the "racial" or continental clusters identified in recent tests of population [sub]structuring (see Bolnick, 2008).[9] This ascertainment issue is addressed here, as it relates to modern craniofacial diversity, by combining Howells' full 1989 data set

[9]In an expanded test, however, Rosenberg and colleagues (2005) demonstrated little change in cluster resolution with increased and more widespread sampling strategies; this suggests that, for molecular data at least, population clusters are readily detectable despite a presumed strong clinal patterning to the distribution of global genetic variation.

with the extensive sample of population-specific craniometrics collected by Hanihara, which have served as the data for much of his research on the world-wide distribution of modern human cranial variability (e.g., Hanihara, 1993, 1996, 1997, 2006, 2009, 2010). Hanihara's data was chosen for its meticulous documentation, large sample and variable sizes, and most importantly for its significantly greater geographic distribution, particularly in areas where Howells' data may be seen as deficient, i.e., western and southeast Asia and the Pacific. By incorporating Hanihara's data and weighting the final test sample by the areas of greatest population density, the range of possible variability is expanded, so that it more closely approximates the conditions of the global population-at-large. With this increase in sampling density, the potentially confounding factor of clinal variation is introduced; in doing so, this project can effectively respond to the concerns over artificial cluster patterns resulting from limited geographic representation and the selective consideration of spatially (and, in turn, more morphologically) distant groups.

Cranial metrics on a total of 11757 (3225 females, 8532 males) representing 19 broad geographic groups were selected from the full data set obtained personally from Dr. Tsunehiko Hanihara of the Kitasato University School of Medicine in 2008 (Table 4). While this data set has been described as "the largest and most diverse cranial data set available as yet" (Hubbe, Hanihara & Harvati, 2009, p.1720), its application to questions of craniometric diversity has been largely constrained to publications by Hanihara (1996, 2006, 2009, 2010), which have placed special emphasis on understanding the morphologies and population histories of Asian peoples (Hanihara, 1993, 1997, 2006; Hanihara & Ishida, 2009; Ishida et al., 2009; Kaburagi et al., 2010). While current trends in new publications are suggesting an increasing interest in its contribution to broadly-based topics of evolution and environmental effects, which benefit from large and diverse samples (Hubbe, Hanihra & Harvati, 2009; Betti et al., 2009, 2010),

it has rarely been combined with other "modern" data sets (c.f., Carson, 2006) and continues to be disseminated in its raw data form by personal request only (Hanihara, 2008). As a result of its rather low public exposure, there is little precedent for its manipulation (e.g., the extraction of population and sex-based subsets, or variable selection) and it has yet to receive consistent external data validation in terms of pre-processing (e.g. standardization, missing value deletion or replacement) and data screening (e.g., detecting recording errors, duplicates, and outliers). For the purpose of this combined analysis, uniformity in data collection practices, correspondence among alternatively defined variables, instances of overlap in sampling locations and missing values are of primary concern. As the decision to combine Hanihara's data with Howells' is driven predominantly by a desire to improve the breadth of geographic representation, to increase the density of the groups sampled and to capture a potentially greater range of craniometric variation, these populations are not differentiated by sex or subjected for formal tests of male versus female clustering patterns. It is believe that the sex-specific analyses of the Howells data set should be sufficient to reveal any potentially confounding trends and to signal any need to further address shape variation, in terms of sex-specific morphologies, among or within the other test populations. While the isolation of sex in largely "unknown" data sets (i.e., males and females are identified by estimation or with limited documentation) is problematic for conducting formal tests under the (likely violated) assumption of error-free designations, concern over sex in this particular case is wholly unnecessary. It can be argued, here, again, that when the research focus is on large-scale and, so, generalized, population-based patterns of human variation and not on the accuracy of classification at the level of a target individual, a small percentage of misidentified sexes have little effect on the cluster findings overall. Sex-based size differences in the combined Howells-Hanihara data set are, however, corrected for by standardization using the geometric mean; scaled and unscaled mixture analyses were carried out in order to justify this decision. In order to further retain maximum variation, possible

outliers were not excluded; only those cases arising from obvious measurement error were flagged and monitored. In removing the extreme tails of a particular population's or full data set's distribution, artificial constraints are unnecessarily placed upon the "true" (observable, available) range of morphologies. Moreover, density-based outlier detection has been extensively studied in the context of data mining and a variety of finite mixture methods have been proposed for use as purely detection techniques in practical data screening applications and for combining in a single mixture-based procedure outlier detection with clustering (Yamanishi et al., 2004; Tao & Pi, 2009). Outliers can be easily managed, for example, with the addition one or more components, which represent a different distribution for the outlying data (Fraley & Raftery, 2002). A mixture clustering approach is, therefore, not only well-suited to handling instances of extreme or atypical individuals but may also offer valuable information on understanding the place of outliers within their population of origin based upon their assigned cluster.

When viewed as two independent samples simply combined, a blended Howells-Hanihara data set would contain data on well over 15,000 measured crania. Some loss of data is, however, unavoidable in the process of integrating the two collections into a single workable data set. The Howells and Hanihara data sets are combined here in a way that is careful to retain the maximum useable number of individuals; nevertheless, the total sample size is subject to reduction because of duplicate observations, variable incompatibility and, for most analyses, missing values. The data merging procedures used here are outlined below. The steps taken for sample amalgamation are presented first and the decisions made regarding variable selection and missing data imputation are given second. This process began with the identification of duplicate observations. These repeated cases are of concern as the intention is to produce a final combined data set that maximizes the available samples and, so, includes skeletal

series that were measured by both Hanihara and Howells. Known overlap between these observers includes, for example, specimens from the geographic regions of New Britain Island, the Philippines, Tasmania, Japan, China, and Kenya, as well as individuals representing specific tribal or ethnic groups, such as the Zulu, Buriat, and Bushmen. The simplest approach to isolating possible duplications may be to first compare the identification criteria reported by Howells (1989) and Hanihara (2008) for each population in question (e.g., specimen identification number, museum or collection source). Then, when it is reasonably certain that the same sample had been measured by both recorders, one of these samples can be excluded from the combined data set. While logical, this approach is problematic for instances when a sample's or a specimen's provenience is not reported with sufficient detail to enable a clear distinction. This is, for example, the case for Hanihara's Tasmanians, whose source information is not documented in the data sheets made available by Hanihara (2008). A computational method for identifying sample overlap was developed in response to this specific problem. A routine was written that parses the data matrix in order to seek out individuals with similar row values and to flag these observations for inspection. These flagged results are then manually filtered for true duplicates by comparing the "matched" individuals against their individual identifiers (sex, population, origin, source, sampling location, identification number) and recorder information. Finally, the decision to retain or delete the flagged (and now inspected) observations(s) is based upon the following pre-established criteria: (1) if duplicate or multiple records are returned for the recorder and all cell values for the measurements and specimen metadata are determined to be identical, all but one of the flagged records are deleted, (2) if multiple records are similar in all values except for sex or population, both records are deleted, (3) if duplication is reasonably certain, but the records were produced by two different recorders, Howells' data is retained, or (4) if the records were similar between the two recorder's data, but an exact match cannot be made, both records are deleted. Using this method, 76

records were marked for deletion, representing only 0.4634% of the combined data set. There were no observations removed from Howells' contribution to this combined data set. While there is some precedent for population-wise deletion when overlap in data collection is suspected, this project's identification approach, although conservative, appears to require fewer exclusions. In justification of these findings, the observations deleted using this computation approach were compared against results obtained by Carson (2006), who combined a smaller subset of Howells' and Hanihara's samples (5632 individuals in total) using manual comparison and series-wide deletions. While she found (2006) that "in many cases... Howells and Hanihara had measured different skeletal samples from a particular population or geographic locale, and both could be included in the analyses" (p. 104), Carson still made considerable deletions, excluding, for example, all of Hanihara's Ainu data, and, she was ultimately unable to determine for certain if an additional 20 observations were indeed duplicates.

Howells, in his 1996 description for the online version of his final data set, offers important commentary on his now well-known and very popular data collection procedures:

> The measurements are 57 in number for all cases (though most are provided with a slightly larger number, not used in my reports). These mostly coincide with standard measurements as defined by earlier workers, but various of them were modified and redefined for reasons of my own...Not originally anticipating wide use by others, I never intended to establish a new standard set of measures, whatever may actually have happened. My choice of measures was partly limited by the need to confine instruments to a set which would be highly portable and durable, without regard to any suite of measurements that might be made working entirely in one's home laboratory (1996, p. 441-442).

It is clear from this description and from personal practice (i.e., using other measurement standards), that Howells (1973) largely defined the system that he used for collecting his cranial measurements. Hanihara appears, however, to have more consistently adhered to Martin's (1928) traditional protocol: over half of the 45 measurements collected by Hanihara (2008) correspond to Martin's numeric designations. While there is substantial synchrony between the two standards, there are points at which Howells and Hanihara differ in landmark definitions and measurements. If unaccounted for at the time of variable selection, this mismatch would certainly obscure true cluster patterns and compromise the quality of the mixture results. In order to maintain consistency among the world-wide craniometric tests, the intention is to repeat the base set of 12 craniometric variables previously selected for use with the Howells data alone (Table 3). All 12 of these variables are present, in name at least, among those that appear in both Howells' and Hanihara's data sets. When these two (recorder-specific) sets were verified against the definitions and protocols that guided their collection, only one of these inter-landmark distances, OBB (orbital breadth), was identified as problematic. Hanihara, following Martin (1928), defines a different location than Howells (1973) for the landmark, ectoconchion. As a result, the combined data set was modified to exclude OBB but retain all of the remaining (now 11) base measurements and to incorporate three measures of length and breadth, GOL (glabello-occipital length), XCB (maximum cranial breadth), and XFB (maximum frontal breadth). Although these subsequent additions are not included in the list of ILDs given by Konigsberg, Algee-Hewitt and Steadman (2009), they have been traditionally considered important for craniofacial discrimination (Table 3). In order to retain this full suite of 14 preselected variables and maximize the usable sample size, missing values were estimated by multiple imputation using mixture modeling (Schafer, 1997; Steele, Wang & Raftery, 2010). In a second analysis, the problem of incomplete data was addressed by eliminating all missing values using case-wise deletion (Little & Rubin, 2002) in order to conduct cluster

analyses with and without replacement using the mixed Howells and Hanihara data set. When the combined sample was treated in this way, a reduced subset of 12,948 individuals was produced. In all tests of the combined data set (full and reduced), sexes were pooled and size adjustments were made using the geometric mean, following the method proposed by Darroch and Mosimann (1985).

From the perspective of population studies, merging Howells' and Hanihara's samples is profitable for its increased representation, variation, and size. For the purpose of statistical analysis, it is advantageous for the recovery of underlying structure, as data sets composed of very many true groups have been shown to often yield the best cluster solutions and classifications (Kuiper & Fisher, 1975; Milligan, Soon & Sokol, 1983). Given the known instances of data collection overlap and similarities observed in naming conventions across Howells' and Hanihara's cranial series, a blended Howells-Hanihara data set should be reasonably straightforward to produce. It is logical to assume, in theory, that correct matches can be made between the population designations given by Howells (1989) and Hanihara (2008), since, for all cases, there is some amount of documentation available for the demographics, geographic origins, and collection locations of their respective skeletal groups. There are nevertheless problems of uncertainty when endeavoring to combine Howells' 54 with Hanihara's 478 sex-specific groups and, as a result, several key questions have emerged. When is it appropriate to lump very local populations and sexes together? When should the prior integrity of a "unique" series be maintained by retaining its separate designation? Upon what criteria should these decisions be based, and how will different configurations of these "new" groupings affect the identification and explanation of population structure, if at all? A review of Howells' and Hanihara's Tasmanian data, for example, may help to shed further light on these concerns. This population is of special interest because it has been already recognized as a problematic "regional" series, with questionable "known" identifications for many of the available crania. Howells' (1989, p. 96-97) has provided substantial reference information on this series, as well as specific data for

the individual specimens that he recorded: 45 males, 42 females, ca. 19th century, 14 source institutions, 31 males and 27 females with problematic identifications. There is, however, no comparable metadata in Hanihara's data set (e.g., origins, date, additional tribal or ethnic affiliations, source or sampling location). Only weak identifiers, including some specimen numbers, but no museum or collection tags, are given for the few (17 male and 7 female) crania making up Hanihara's (2008) sample of "Tasmania Aborigines / Tasmanians." There is both too little specimen information and insufficient observations upon which to base a correspondence between these two sets of crania or from which distributional similarity can be reliably assessed. While arguments can be made for supposing that associations based on population name are sufficiently reliable, this approach is troublesome for several reasons, all of which can be extended beyond this example to other cases. Howells (1989) has noted, in a discussion of the selection process for his Tasmanian series, that "the problem of assembling a Tasmanian series is acute" as there is the "danger that skulls of other ethnic varieties may be labeled Tasmanian because of the rarity and Romantic interest...European skulls can be readily detected, even though labeled Tasmanian. The real hazard lies in the fact that Melanesians, Maoris, and above all, Australian aboriginals, did come to this island in the 19th century" (p. 96). In this case, it is important to recognize that there may be considerable variation among individuals identified as Tasmanian by either recorder. In a broader context, it is also important to consider the potential for inherent error when a questionably documented regional series is taken as a cohesive, testable group (like "Tasmanian"), under the assumption that these craniometrics accurately reflect (as best as any skeletal population can) the craniofacial characteristics of the once living population. In the absence of thorough explanations, the exact meaning of any designations given to a specimen or series can be called into question. At first pass through both data sets, it would be easy to mistakenly combine, for example, Howells' and Hanihara's "Anyang" into a single series without appreciating their significant

geographic and historical differences. Howells' (1989) crania are said to "derive from the sacrificial burial pits in and round the imperial tombs of the Shang Dynasty at Anyang, Honan Province" (p. 119). Hanihara's specimens represent, instead, members of the African Anyang, an ethnic group from Cameroon. It is also noted in his metadata that they share affinity with the Edia, Ibea, Mandingo, Duala, Boki, Kumabembe, Mabea, Bawa, and Yaunde peoples of that same region; they are, therefore, certainly unrelated to Howells' group. The risk of even greater confusion becomes apparent when a closer look is given to the naming conventions used within Hanihara's own data set. Hanihara (2008) has, for example, two sets of individuals identified with the designation, "Santa Cruz Island": one group seems to correspond with Howells' (1989, p. 105) American Indian series from California, while the other appears to represents Melanesian peoples from the Solomon Islands. The reality of merging Howells' and Hanihara's data set is obviously complex and problematic. Since this project's finite mixture approach does not dictate the prior designation of group membership, model-based clustering analysis can be run using all of this combined data, without needing to consider, either, Howells' or Hanihara's individual identifiers or arrive at some blending of the two. As the cluster solutions produced by mixture analysis are unaffected by how the test data is partitioned into sexes and/or populations (if it is at all, e.g., in the case of unidentified remains), the "true" group designations, or, in fact, any number of artificial grouping schemes, can be used to evaluate *a posteriori* the relationship between the "statistically inferred" clusters and the "known" population memberships. As the combined Howells and Hanihara data set is of interest here for revealing global trends when using sufficient sampling to address clinal variation, retaining the original naming system for the individual populations is of less importance. The problem of possible confused identities can be, therefore, easily mitigated by simplifying the class terminology and pooling the sexes. The regional designations defined by Hanihara (2008), which appear in his data set as prefixes to the more specific group labels, are used to collapse

Howell's and Hanihara's local, tribal and ethnic population names into broad geographic categories. This classification system, containing 19 regional divisions (Table 4), will serve as the basis for the interpretation of the results obtained from the analyses of the combined Howells and Hanihara data set.

Table 4: Combined Howells and Hanihara data set, sample composition by sex and region

Population by Regional Designations (after Hanihara 2008)	*Recorder/Source*			*Recorder/Source*			*Sample*
	Hanihara (2008)		Total	Howells (1989)		Total	Grand Total
Sex	F	M	F/M	F	M	F/M	F/M
AL: Aleut	126	171	297	0	0	0	297
AT: Australia and Tasmania	57	243	300	91	97	188	488
CA: Canada	29	81	110	0	0	0	110
CSA: Central and South America	230	413	643	55	55	110	753
EA: East Asia	220	613	833	167	269	436	1269
ES: Eskimo	385	469	854	55	53	108	962
EU: Europe	207	952	1159	153	164	317	1476
IS: Indian Subcontinent	60	418	478	0	0	0	478
MEL: Melanesia	140	381	521	0	0	0	521
MIC: Micronesia	12	42	54	27	30	57	111
NA: North Africa	105	563	668	53	58	111	779
NEA: Northeast Asia	127	276	403	54	55	109	512
PNG: Papua New Guinea	125	182	307	54	56	110	417
POL: Polynesia	79	459	538	137	177	314	852
SEA: Southeast Asia	131	711	842	35	85	120	962
SSA: sub-Saharan Africa	175	760	935	197	176	373	1308
SWA: South and West Asia	47	292	339	0	0	0	339
UK: United Kingdom	102	339	441	0	0	0	441
U.S.: United States	868	1167	2035	78	93	171	2206
Grand Total	*3225*	*8532*	*11757*	*1156*	*1368*	*2524*	*14281*

While the model-based cluster analysis of Howells' craniometric data set serves as the primary test of population structure in this project, it has been already suggested how, in some ways, the composition of Howells' (1989) final full data set may be problematic for this analysis. Potential weaknesses in

Howells' sampling strategy have been already noted by others (c.f. Leathers et al., 2002; Relethford 2009) and issues related to specimen availability were reported by Howells himself. Reflecting upon his 1989 study, Howells (1995) wrote the following:

> The total corpus of this carefully selected material includes 1,348 male crania and 1,156 female crania. Necessarily, this universe is not ideal and species-wide but only a partial one, limited by the availability of collected materials; it has spacious interstices for which no information is included, notably, all of southern Asia. This becomes especially evident when trying to place skulls whose natural place would be in such lacuna. There is no cure for this (p. 3).

The pairing of Howells' 28 different cranial series with Hanihara's much larger and more diverse collection offers, a reasonable solution to these problems, by giving breadth to the analyses of world-wide population structure. In combining these data sets, a minimum of 10,424 more cases (almost five times Howells' maximum sample size) were made available for analysis and many "new" populations were introduced. This new, combined data set now covers entire geographic regions which were not well, if at all, represented by Howells' populations (e.g., especially the Indian subcontinent and Southeast Asia), includes cases of both sexes for most areas (e.g., adding in females for the Philippines), and samples areas that were previously characterized by only limited examples (i.e., expanding Howells' European data into the Mediterranean, with the addition of Greece, Italy, Portugal and Spain).

This project's analysis can also be further expanded to speak to a second "sample" issue, related to the data collection strategy adopted by Howells when compiling his 1989 data set. It begs the question: do the cluster solutions found using Howells' (1989) data set reflect Howells' sampling strategy more than the true structure of the global craniometric data? As with any case of non-random or selective

sampling, whether intentional or predicated upon access to collections or specimens of potential interest, the trait distributions available for analysis may fail to reflect the true patterns of variation that exist within a given skeletal population. In the context of world-wide human craniometric data, analyses may reveal artificially inflated group differences when samples are chosen for their geographic distance, are drawn from isolated areas, or represent biohistorically unique populations. Unnatural structure may also be imposed upon a developing data set by several key choices made by the data collector: (1) selective culling based upon measurement protocol to meet requirements for complete specimens or specific elements, (2) non-random selection in order to satisfy sampling criteria, such as balanced sex ratios or the representation of pre-determined cohorts, and (3) the elimination of atypical or divergent morphologies, which are believed to lie outside of the range of "normal" variation for a particular group or geographic region. These factors may suppress the range of variation within a composite sample, which can, in turn, influence test results and/or their interpretation: either by decreasing the overlap between similar populations (and inflating discrimination) or by excluding valuable traits based on low, observed, frequencies in limited samples (thus, weakening discriminatory power). These concerns can be addressed, as they apply to Howells' (1989) data set, by introducing an additional 226 cases drawn from the over 500 "target skulls" or "object specimens" that make up the "TEST" data referenced by Howells in his 1995 monograph, *Who's Who In Skulls. Ethnic Identifications of Crania from Measurements*.

The individuals included in Howells TEST data (n = 524) are of particular interest as they represent, in Howells' (1996) words, "extras or rejects from the above [28 populations] samples but largely crania from a wide variety of sources, including prehistoric specimens or casts thereof" (p. 441). By incorporating these irregular "test" specimens into the analysis of Howells craniometric data, sample

variability is enhanced. Modeling complexity is also increased by adding in the problem of clustering highly atypical and problematic morphologies. As these test specimens are "preponderantly male" (Howells, 1995, p. 5) and correspond to only some of Howells' 28 "base populations," their inclusion also works to destabilize the balanced sex ratios and the consistent sample sizes. By considering these problematic cases as well as an out-group consisting of "prehistoric skulls" (Table 5), it is possible (1) to evaluate the strength of the data structure by testing the ability of the mixture clustering methods to make correct allocations for the atypical specimens, (2) to compare Howells' visual assessments against the statistical allocations in order to confirm or reveal new information on "affinity" for the cases where *a priori* identifications are uncertain, and (3) to use the paleoanthropological out-group to evaluate the sensitivity of the mixture method to these outliers and to characterized its behavior in the presence of morphologically extreme groups. In all instances, interest lies in observing whether, if, with the inclusion of the TEST data set's "rejects," the cluster patterns formed are consistent with the results obtained using only Howells' (1989) "base" data set, or, if some or all of these TEST specimens are identified as "different" and allocated into their own unique cluster(s). This TEST data also serves as an ideal target sample for use with Howells' (1989) standard data set in an investigation of population structure using mixture density-based discriminant function analysis. Together, these test and target data subsets enable *post hoc* assessments of the utility of this method and its comparability with this project's core approach of using fully-unsupervised model-based clustering (c.f., McLachlan, 1992; McLachlan & Peel, 2000; Fraley & Raftery, 2002, 2006).

Table 5: Combined sample composition of Howells' base (1989) and "TEST" (1995) data set

Population Designations (Howells 1989, 1995)		*Data Set with 57 variables*			*Data Set with 12 variables*		
Name	Number	Females	Males	Total	Females	Males	Total
AINU	25	39	49	88	39	49	88
ANDAMAN	24	37	36	73	37	36	73
ANYANG	28	1	49	50	1	49	50
ARIKARA	13	26	43	69	27	43	70
ATAYAL	19	21	43	64	21	43	64
AUSTRALIA	7	52	56	108	52	56	108
BERG	3	53	57	110	53	57	110
BURIAT	26	56	55	111	57	55	112
BUSHMAN	23	55	48	103	56	48	104
DOGON	5	47	47	94	52	49	101
EASTER ISLAND	11	38	50	88	38	50	88
EGYPT	22	50	64	114	56	64	120
ESKIMO	27	58	58	116	59	58	117
GUAM	21	27	32	59	27	32	59
HAINAN	18	38	52	90	38	52	90
MOKAPU	10	53	54	107	53	54	107
MORIORI	12	53	68	121	53	68	121
N JAPAN	16	32	56	88	32	56	88
N MAORI	30	0	11	11	0	1	1
NORSE	1	60	60	120	60	60	120
PALEO	31	4	20	24	4	20	24
PERU	15	74	75	149	75	75	150
PHILIPPINES	20	0	50	50	0	50	50
S JAPAN	17	41	52	93	41	52	93
S MAORI	29	0	10	10	0	0	0
SANTA CRUZ	14	56	57	113	56	57	113
TASMANIA	8	45	50	95	45	51	96
TEITA	4	36	31	67	51	33	84
TOLAI	9	55	62	117	55	62	117
ZALAVAR	2	48	62	110	48	62	110
ZULU	6	44	55	99	47	55	102
Grand Totals		*1199*	*1512*	*2711*	*1233*	*1497*	*2730*

As these additional TEST specimens were collected by Howells' using his standard measurement protocol (often at the same time as the samples reported in his 1989 publication) and are mostly complete cases, the combined "full" 28 population (1989) and "TEST" (1995) data set can be evaluated using both this project's 12 base variable subset (n = 2730) as well as an enlarged suite of 57 measurements reported by Howell's (n = 2711) (Table 3). In accordance with the procedures already discussed for the combined Howells (1989) and Hanihara (2008) data set, missing data is addressed by case-wise deletion, sexes are pooled, and Howells' base (1989) and TEST (1995) measurements are corrected for variation in size using the geometric mean, following the method proposed by Darroch and Mosimann (1985).

The Forensic Significance of Race: The Forensic Anthropology Data Bank

The Howells and Hanihara craniometric data sets are invaluable resources for investigating the question of global population substructuring. They provide the necessary sample breadth and sample size to ask if, in the absence of any a priori group identifiers, this project's model-based clustering method can organize "mixed" skeletal samples into clusters that are both statistically significant and anthropologically meaningful. These data sets, therefore, offer the opportunity to speak to the problem of determining the nature of the distribution of craniometric diversity: it is possible to investigate if the observed variation is clinal or populational, or, weakly or strongly structured. Answers to these questions may contribute to a very basic, but critical, understanding of the theory of human morphological variation and aid in the interpretation of biogeographic ancestry. Since these data sets

contain samples from many geographic regions and at different time depths and include only limited information on the individual specimens (e.g., providing sex, but lacking age, birth year, birthplace, pedigree), they are less useful for addressing research interests that are context-specific, which often require finer levels of group identification, greater availability and certainty in demographic identifiers, and restricted time periods. As this project also sets out to address the problem of race, as it is either biosocially or biogeographically defined, within the American population and, specifically, as it applies to forensic anthropological identifications in medico-legal investigations, it requires a testable data set that is appropriate for these unique research goals. In this case, an ideal craniometric data set should (1) represent the diversity in self-identification and the variations in morphology expected for the United States and (2) closely approximate the population-at-large from which forensic casework is expected to be drawn. For these two important reasons, this discussion of "race" in forensic anthropology is framed around the model-based clustering analysis of a large craniometric sample of modern Americans drawn from the Forensic Anthropology Data Bank (FDB). A subset of the FDB data available was generously provided by Dr. Richard L. Jantz (2008) and the Forensic Anthropology Center (FAC) at the University of Tennessee. This "full" data set comprises a maximum of 65 linear distance measurements for a total of 1145 positively identified individuals (392 females, 752 males), representing five biosocial races and six documented cases of admixture. The unique qualities of the FDB make its craniometric and demographic information especially useful for discussions of the biosocial construction of "identity" and its relationship to the scientific quantification of craniofacial morphology.

Forensic anthropology is an applied, case driven, goal oriented subfield within skeletal biology (Ubelaker, 2008). As practitioners, forensic anthropologists seek to aid the medico-legal community in making identifications for unknown, mostly skeletal and often fragmentary, human remains by

developing biological profiles (ABFA, 2010). Using a combination of grossly observable and measurable morphological features on the available skeletal elements, forensic anthropologists rely upon technical expertise to both estimate demographic parameters, as they relate to the population-at-large, and isolate unique traits that characterize an individual as *the* individual (such as sex, race or ancestry, age-at-death, stature, and pathology). They also look to identify those taphonomic variables that help to explain the circumstances surrounding the unknown individual's death, deposition, and decomposition (including markers of peri- and post-mortem trauma and approximations of time-since-death). Forensic anthropologists estimate these features with some degree of statistical certainty based upon the application of standard criteria, case-work experience, and advanced techniques supported by inter-disciplinary collaboration. The development, testing, and application of forensic criteria for the estimation of the biological parameters that are key to personal identification necessitate well-documented collections of comparative skeletons or accurate and extensive reference databases. Good estimation requires the practitioner to assume, correctly, that the "target sample" (i.e., the unknown sample or case, comprised of the individual[s] for whom estimate[s] of biological parameters are desired) comes from the "same" (or at least similar or comparable) population as the "reference sample" or "prior." For explicitly Bayesian methods, these estimates for the unknown individual or sample are framed in terms of probabilities that are informed by what is known about the distribution of interest (e.g., of trait, sex, race, age, etc.) in a documented reference collection or data set; for this reason, forensic analyses are often said to be conditional upon prior information (c.f., Konigsberg, Algee-Hewitt & Steadman, 2009). These expectations are frequently problematic because large scale collections of modern documented skeletons are few in number and it cannot be assumed that the identification criteria obtained from past populations are directly, if at all, applicable to modern Americans (c.f., Wilson, Herrmann & Jantz, 2010). Making what may seem to be logical assumptions on

sample similarity, in the absence of appropriate testing, and choosing "best-fit" reference collections can have singularly compromising effects on not only the estimates produced but, ultimately, on the chance of success in arriving at a correct personal identification. Usher (2002) is certainly correct: marrying populations with proper reference collections is, "one of the most overlooked, but basic, sources of error"in skeletal estimation (p. 29). For reasons such as these, as Jantz (2001) has explained, "the original purpose of the forensic anthropology data bank was to provide up-to-date identification criteria for estimation of ancestry, sex, age, and stature, rather than having to rely on 19th century anatomical collections" (p. 784).

The Forensic Anthropology Data Bank is quite remarkable for both the variety of demographic information available for many of the cases on file (e.g., race, sex, age, birth year, place of birth, medical history, occupation, living stature, and weight), and the quantity of the corresponding morphological data that is routinely collected. Skeletal information stored in the FDB typically includes cranial and postcranial metrics, three-dimensional coordinate data, non-metric cranial trait presence/absence, phase-aging and categorical indicator scores, evidence of trauma, congenital abnormalities and pathologies, and dental observations (FDB, 2005). Of particular interest is the fact that this data bank contains a mixture of individuals from actual forensic casework and modern skeletal collections (FDB, 2005). The submission of entries by expert forensic practitioners working in all areas of the United States is extremely important as it provides the geographic breadth, racial diversity, and recovery contexts that is critical to studies concerned with the distributional realities of contemporary casework; while the information obtained from skeletal collections contributes to a more complete characterization of the composition, trends and secular changes within the American population-at-large (c.f., Wilson, Algee-Hewitt & Jantz, 2007; Algee-Hewitt, Wilson & Jantz, 2007; Meadows Jantz &

Jantz, 2008). The FDB is seeing ever increasing use as an appropriate prior for statistical analysis using modern American samples and as a source for representative and reliable data for studies concerned with new method development and validation tests relating to all aspect of skeletal analysis (e.g., Konigsberg, Algee-Hewitt & Steadman, 2009; Langley-Shirley & Jantz, 2010; Wilson, Herrmann & Jantz 2010).

The Forensic Anthropology Data Bank is especially valuable to this dissertation because of the way it documents information on individual identity: the FDB uses the biosocial language of race, rather than the biogeographic terminology of ancestry estimation. With this information, it is possible to avoid the problems of making arbitrary "lumping" or "splitting" decisions (i.e., dividing or amalgamating groups based on continental, regional or local geography) or choosing among predefined population names, which are often highly variable in their meaning and usage. Relethford (2009) has already voiced this concern for Howells' core data set, writing that "There is no justification based on classification accuracy for preferring five, six, seven, or eight groups. From the perspective of racial classification, if we take these different groups as equivalent to races, then we see that the number of races, and the geographic cutoffs used to define them, are subjective decisions" (p. 20). The FDB's documentation gives the researcher the advantage, therefore, of (1) making unambiguous (i.e., not practitioner or recorder-defined) associations between the available craniometric data, the model-based cluster assignments or "classifications," and reported race, and (2) using personal identifiers that are relevant to the American environment (e.g., Hispanic, as a uniquely North American label), from which the majority of U.S. forensic cases are drawn.

This project's analysis of the individuals sampled from the Forensic Anthropology Data Bank (2008) is directed towards the ongoing controversy over the appropriate use of "race" in forensic anthropology. It takes special interest in addressing classification and estimation in response to the FORDISC "debate"that continues to dominate current literature on personal identification, and in light of the question first posed by Sauer (1992), while still in the early stages of forensic method development: ″if races don't exist, why are forensic anthropologists so good at identifying them?" Because the issue of population structure is being addressed from the perspective of the forensic estimation of ancestry and the translation of these estimates into the language of race for use in medico-legal contexts, it is important to keep the necessary concern for "significance" and "evidence" clearly In mind when making putative identifications. This is especially true when consideration is given to the many new (and changing) requirements for expert testimony and the rules for the admissibility of evidence in many U.S. courts that have followed from the landmark case, Daubert v. Merrell Dow Pharmaceuticals, Inc., in 1993 (c.f., Sapir, 2002; Christensen, 2004; Wiersema et al., 2009; Kimmerle et al., 2008; Grivas & Komar, 2008). For these reasons, the FDB data considered here represents a sample culled from the full data bank. It includes only positively identified complete-data cases, with known sex, race, and age or birth/death year. Missing data was handled exclusively by case deletion because it was felt that the "estimation"and "imputation" procedures that enable data replacement introduce an additional level of uncertainty that cannot be easily quantified and which may influence the outcome of this project's analyses in unpredictable ways. In order to limit the number of case-deletions and to maintain consistency among the data sets under analysis, a standard suite of 12 variables are considered here: these correspond to the variables used in the tests of Howells' (1989) cranial series and to the measurements already employed with success by Konigsberg, Algee-Hewitt and Steadman (2009) in their test of a mixed sample which included FDB data (Table 3). The final FDB data set defined here

consist of 925 individuals, representing both sexes, five races (White, Black, Hispanic, East Asian, and Native American) and six individuals identified as "admixed" (White/Hispanic, White/Black). It is believed that this sample configuration approximates the population diversity of the United States as it includes the five most common racial groups, in quantities that reasonably correspond in their proportions to the statistics on self-reported race from the U.S. Census Bureau (2011) (see Table 47). It is also assumed, with confidence, that this data set is sufficiently large in size to support its partitioning into subsets for analyses that are race specific, e.g. American Blacks, Whites, and Hispanics (n = 848), or American Blacks and Whites (n = 828). Since the key demographic variables of birth year and sex are available for all of the cases making up the FDB sample, it is possible to also evaluate the effect of secular change and the presence of sex-specific variation on the cluster solutions returned by finite mixture analysis. This analysis of secular change is controlled by pooling the sexes, but retaining their sex-specific identifiers for later separation, and dividing the data into Black (n = 210) and White (n = 618) subsets. This discussion is limited to these two modern American groups because secular change has not yet been well-documented in the other, Hispanic, East Asian and Native American, groups that make up the remainder of the FDB data set. Sex-specific trends in the Black (84 females, 126 males) and White (222 females, 396 males) are sought out by comparing known sex against the cluster number assigned to each individual by finite mixture analysis. This project does not test for sex effects in the remaining three groups independently because of the limited, unbalanced samples sizes available for the Hispanics (3 females, 17 males), East Asians (1 female, 53 males), and Native Americans (6 females, 10 males). Chapter 2 of this dissertation paid special attention to the controversy over the validity of craniometric classification in the context of the FORDISC debate. From this discussion, two important points should be recalled: (1) that considerable emphasis has been placed upon the presumption of "forced allocation" when discriminant function analysis is used as a tool for investigating population affinities or

aiding in human identification for medico-legal casework and (2) that the Howells data set has served as a convenient, common source of cranial metrics for testing FORDISC and its statistical approach (e.g., Brown & Armelagos, 2001; Armelagos & Van Gerven, 2003; Williams, Belcher & Armelagos, 2005; Albanese & Saunders, 2006; Hubbe & Neves, 2007; Elliott & Collard, 2009). In response to this recent history of concern, the modern "race-based" sample from the Forensic Anthropology Data Bank is used here, both alone and in combination with Howells' (1989) craniometric data set, to make comparisons among the findings obtained from tests of unsupervised model-based clustering, model-based clustering using partial classification, discriminant function analysis via model-based clustering, and traditional linear discriminant analysis (LDA). In the case of the combined FDB and Howells data set (n = 3426), this analysis is structured in two ways. The cluster solutions and classification success are evaluated, first, by keeping both the population names and races associated with the respective data sets, and, second, by relabeling the groups according to the "race" and "ethnicity" designations used in the reporting procedures for U.S. Census 2010 (see Table 58). The races defined by the Federal Office of Management and Budget (OMB) and the U.S. Census Bureau (2010) used here include: White, Black or African American, American Indian and Alaska Native, Asian, and Native Hawaiian and Other Pacific Islander. Only these five "single" races are consider in these analyses for reasons of clarity (there are in fact 57 combinations using the available race categories) and because they correspond to the summary tables included within the 2010 Census Briefs (c.f., U.S. Census Bureau 2010). Moreover, recent published statistics indicate that "in the 2010 Census, 97 percent of all respondents (299.7 million) reported only one race" (Humes, Jones & Ramirez, 2011, p.4). Hispanic individuals from the FDB data set are considered to be "White" when using the U.S. Census categories, since (1) the U.S. Census Bureau (2011) defines "Hispanic" as an "ethnicity" and not a race and (2) "the 2010 Census racial distributions of the

Hispanic population and of the non-Hispanic population differ" but "over half of the Hispanic population identified as White and no other race" (Humes, Jones & Ramirez, 2011, p. 6)

Sample and Model Assumptions

Most common clustering techniques are based upon heuristic procedures, and, as such, are typically included within the family of exploratory data analysis (EDA) methods. These approaches are frequently favored, in practice, for their ease of access and implementation: the majority of clustering methods in commercial statistical software are of this type and the procedures that they employ are reasonably intuitive (i.e., highly interactive, user-guided, and visual) and, so, simply interpreted. New areas of research using large, high dimensional data sets, for which the true identities and relationships of the cases are frequently unknown (e.g., data mining of web and social networks, analysis of gene expression data from microarrays and image segmentation for biomedical applications) have brought about a shift in the foci of interest for cluster analysis. Recent reviews of developments within this area have noted especially a push for increased "systematic guidance" for "solving the basic practical questions that arise in cluster analysis" and an emerging dissatisfaction with the fact that "the statistical properties of these methods are generally unknown, precluding the possibility of formal inference" (Fraley & Raftery, 2000, p. 1).

Finite mixture methods have gained "new" interest because their well-established theory and probabilistic, model-based approach provide a sound mathematical framework for the "statistical

modeling of a wide variety of random phenomena"(McLachlan, Ng & Peel, 2003, p. 141). Finite mixture's model-based procedures (1) offer the advantage of a probabilistic clustering, which permits overlapping clusters that correspond to the components of the mixture model, (2) provide uncertainty measures for the cluster allocations in terms of the fitted values for the posterior probabilities of component memberships, (3) allow for the critical question of the "true" number of hidden classes to be addressed by evaluating the number of components required for the best model fit, and (4) permit the question of model choice (i.e. parameterization and component solution) to be resolved using the likelihood function. Much of the theory that informs these model-based clustering procedures assumes that the full data set of interest is composed of a mixture of underlying probability distributions, with each component in the model corresponding to a different data cluster or [sub]population. As it is assumed that these [sub]groupings are distributed according to a particular parametric form, it is necessary to specify the component densities (e.g. normal or *t*-family) of the model before implementing *model-based* clustering (McLachlan & Peel, 2000; Peel & McLachlan, 2000). Since Scott and Symons (1971) first popularized solving mixture problems by assuming normal components for continuous multivariate data, considerable attention has been paid to clustering criteria using this theoretical approach (c.f., Symons, 1981; McLachlan & Peel, 2000; Fraley & Raftery, 2006). The multivariate normal (MVN) mixture approach is now clearly formalized, with well-known, flexible, and highly useful features: data generated by a mixtureof normal distributions are characterized by clusters centered at the means μ_k (with greater density near the mean), the corresponding surfaces of constant density are ellipsoidal, and the geometric features (shape, volume, orientation) of the clusters are determined by the covariances Σ_k, which may also be parameterized to impose different cross-cluster constraints (Murtagh & Raftery, 1984; Banfield & Raftery, 1993; Fraley & Raftery, 1998).

It is not surprising that McLachlan and Peel (2000, p. 81) have observed what current trends in publications and software also support, (e.g., Leish, 2004; Fraley & Raftery, 2006; Benaglia et al., 2009) that "a common assumption in practice is to take the component densities to be (multivariate) normal." Gaussian mixtures are frequently preferred for their computational tractability and broad applicability. Multivariate normal densities are, as McLachlan and Baek (2010) have remarked, "a convenient choice" (p.2) because they can be iteratively fit by maximum likelihood (ML) using the expectation-maximization (EM) algorithm (Dempster, Laid & Rubin, 1997; McLachlan & Krishnan, 1997). Normal mixtures are also forgiving of their violation and the hypothetical cost of relaxing modeling criteria is most often outweighed by the computational advantages gained; unless the components are very heavily tailed, maximum likelihood estimation assuming normal components has been shown to outperform other approaches in practical contexts (Hunter, Wang & Hettmansperger, 2007). Assuming normality is an equally theoretically satisfying choice for most mixture-clustering contexts. When clustering large data sets, the use of Gaussian mixture methods is easily justified by invoking the central limit theorem (CLT): it is known that (1) when large independent samples are drawn randomly from a population, the means of those samples will be normally distributed and (2) even if the distribution of the within-population data is not normal, the distribution of population means will approach normality as sample sizes increase (c.f., Strait, Moniz & Strait, 1996; Kallenberg, 1997). Furthermore, it is reasonable to presume from the outset that any non-normal features within the data are related to the underlying group structure that the mixture method is being used to detect. For these reasons, it is appropriate to adopt normal mixture procedures as good standard practice in the absence of formalized tests of normality. The problem of logic in testing for normality under the restrictions of clustering "unknown" data has been already addressed by McLachlan and Basford (1988) and most recently by McLachlan and Peel (2000, p 85), who wrote that "it is self-serving to cluster the data under the assumption of normality

when one is subsequently to test for it, but it is brought about by the limitations on such matters as hypothesis testing in the absence of classified data." It is assumed here, therefore, that each component in the mixture-model has its own multivariate normal distribution, which is characterized by a corresponding vector of means and an unrestricted variance-covariance matrix. Methodological precedents, practical advantages and logical arguments adequately support the choice to fit multivariate normal mixtures exclusively to all of the craniometric data sets.

These arguments can be extended to support, specifically, the choice not to conduct other formalized, pre-processing, tests. It stand to reason that, in the context of clustering of "unknown" data sets, sample composition, distribution and individual specimen identifications cannot be used as *a priori* guides for structuring analysis when the intent of that analysis is to approximate the difficult conditions of detecting cryptic structure within a mixed population represented by complex biological information. It is assumed, therfore, that the data collected by Howells, and Hanihara and by contributors to the Forensic Anthropology Data Bank (c.f., FDB, 2005) exists without measurement or recording error. Formal detection tests are not conducted here because it has been convincingly shown elsewhere that finite mixture analysis is sufficiently sensitive to isolate extreme observations during the clustering procedure itself. These unique observations can be identified easily by a *post hoc* inspection of the mixture patterns for clusters with single or few allocations. If they are relegated to their own cluster(s), they are effectively being removed from the clustering of the other data and, so, they should neither influence the ultimate cluster solution for the "normal" or "typical" specimens nor alter the sample-wide test of structure. If they are not isolated but clustered within members of their true population by the mixture procedure, then they are sufficiently similar to the other specimens to warrant equal consideration in the analysis. It is equally assumed that sex and population assignments were made without documentation or estimation error. Although it is well-known that self-reported information is

especially problematic for its variability in the consistent use of biosocial group identifiers, this information cannot be reasonably verified. For assignments obtained by estimation, it is believed that any classification errors made by an expert practitioner or experienced data recorder in the identification of sex or population (race, ancestry, ethnicity, tribe) would likely be repeated by other observers. The inclusion of this uncertainty into this project's model-building process only serves to enhance the reality of its analysis. Special consideration is given to this issue of "known" sex or ancestry only when samples are chosen from the Forensic Anthropology Data Bank (Jantz, 2008) for evaluating the importance of race in making individual positive identifications. Finally, in accordance with the principles that govern the analysis of incomplete multivariate data and the treatment of missing values using multiple imputation (MI) finite mixture (FM) methods, the final set of imputations produced are taken to represent true craniometric values. All subsequent analyses, using data obtained by this missing value replacement, are conducted here with this understanding in mind.

Size Correction

Regardless if the intention of a given project is to investigate the evolution or the current distribution of morphological variation, to focus on the level of variation above, between or within species, or to address either trait heritability or environmental plasticity, it must be able to reliably estimate the magnitude of phenotypic differences among the populations of interest (McCoy et al., 2006). Whether conducting studies in comparative biology or biological anthropology, most quantitative analyses of morphological differences begin by making adjustments for variation in body size. The importance of

"factoring out" or "controlling for" size in morphometrics is well known: Penrose is often credited for being among, if not, the first who, in 1954, "attempted to separate size and shape computationally" (Howells 1989, p. 9). The problem is conceptually simple, as Jungers, Falsetti and Wall (1995) have quite aptly demonstrated: "if we were to take a suite of the same measurements on a mouse lemur and on a gorilla and did nothing to adjust for gross differences in scale between the two, we would probably discover little beyond the obvious fact that Gorilla is larger than Microcebus" (p. 138). The difficulty in size adjustment for continuous data has historically rested in the choice of how to most meaningfully define relative size and shape. Mosimann and Malley already made this point clear in 1979. In one of the now foundational reviews of size and shape variables, they remarked that "there are a number of precise, interesting and quite different possible meanings and these need to be distinguished" (Mosimann & Malley ,1979, p. 175). Even in the early years of what has become known as the "revolution in morphometrics" (Rohlf & Marcus, 1993), interest in size and shape was high. It was of considerable concern among statisticians interested in the theoretical problems of size versus shape analysis, and to researchers in the biological sciences engaged in the practical problems of methodology and who were taking on topics like allometry or the partitioning of biological distance into size and shape components (c.f., Blackith & Reyment, 1971; Mosimann & James, 1979; Relethford, 1984; Darroch & Mosimann,1985;Bookstein, 1986; Corruccini, 1987; Howells, 1989). While in more recent years, focus has shifted towards making advancements in geometric morphometric approaches (e.g., Bookstein, 1997; Dryden & Mardia, 1998, Adams, Rohlf & Slice, 2004; Slice, 2005), discussions have continued in physical anthropology on best-practices for the decomposition of "form" into its constituent parts when using traditional landmark data. Residual, ratio, c-scores and principal component based approaches have all received attention. It was, however, in a comparative analysis conducted by Jungers, Falsetti and Wall (1995), that the ratio (Mosimann shape variable) approach was

singled out for yielding the best results. In this discussion, they (1995) made the important points that (1) "only variables in the Mosimann family of shape ratios allowed us to identify different sized individuals of the same shape" and that (2) "ratio shape differences may be largely independent of size in narrowly intraspecific or intrasexual data sets" (p. 137-138). There is now a growing precedent for the use of Mosimann shape variables following the "geometric mean" method suggested by Darroch and Mosimann (1985) in craniometric studies. Roseman and Weaver (2004) and, most recently, Konigsberg, Algee-Hewitt and Steadman (2009) made use of this approach for their size-scaled analysis of Howells' (1989) craniometric data specifically. The utility of this method to projects concerned with craniometric discrimination in explicitly forensic identification contexts has been shown by Ross (2004), for example, who investigated regional variability in craniofacial shape using modern Balkan samples.

In order to account for sexual dimorphism and standardize the linear distance measurements for differences in size, this project has adopted this robust "scaling" technique. The method of Darroch and Mosimann (1985) is applied to the craniometric samples compiled from Howells' (1989, 1995, 1996) and Hanihara's (2008) data sets and the Forensic Anthropology Data Bank (Jantz 2008), following the established procedure. First, a new size variable is produced by calculating the geometric mean for all variables, p, given that:

$$size = (\prod_{i=1} X_i)^{1/p}$$

The adjustment for size is then made by converting the raw craniometric measurements recorded for each individual within the data set(s) into Mosimann shape variables by dividing each original variable by the geometric mean of all measurements for that individual. As Mosimann shape variables are ratios, they provide a measure of the size of a particular region relative to the overall size of the cranium and are informative of the "geometric similarity"among populations (Jungers, Falsetti & Wall, 1995). The

geometric mean of each specimen also has analytical value; it has been retained as a measure of overall size, so that variation in size within and among the sampled populations can be reintroduced for "unscaled" analyses.

Missing Data Treatment

One of the many appealing feature of Howells' (1989) craniometric data is its large sample size with complete cases. While Hanihara's data set offers the advantage of evaluating many more populations with greater geographic coverage, it suffers from a plethora of incomplete data: 3,376 individuals are missing values for 1 to 14 variables. This incomplete data represents approximately 20% of the measurements in the combined data matrix, but is exclusive to Hanihara's contribution to this blended data set (Table 6). So far, this project has justified its methodological choice to combine Howells' and Hanihara's data sets by arguing that the inclusion of Hanihara's skeletal series augments the world-wide sample size, provides better coverage to areas that were otherwise sparsely represented and more accurately reflects the full range of modern craniometric variation. It would be counter-productive, therefore, to disregard, from the outset, any cases or variables that contain missing data and accept an immediate loss of either many observations and even full groups, or predictors of interest. From a statistical perspective, a significant reduction of sample size may also introduce an unacceptable loss of power as well as complications in covariance matrix estimation (Shafer, 1999; Little & Rubin, 2002). Unfortunately, there has been but limited discussion of how best to handle these missing values given the data conditions unique to cranial morphometrics and the particulars of multi-population, mixed-sex data sets (but see McKeown, 1996; Carson, 2006). For this reason, this project established its own

missing data protocol and restricted replacement procedures to the combined Howells and Hanihara data set.

Table 6: Missing data patterns

Missing Variable	*Number of Cases Missing Craniometric Variable*	*Number of Missing Values*	*Number of Cases with Missing Values*
BBH	1758	14	843
BNL	1737	13	50
BPL	2386	12	58
DKB	1619	11	48
FRC	1207	10	70
GOL	1316	9	101
NLB	1564	8	120
NLH	1565	7	124
NPH	1925	6	165
OBH	1378	5	151
OCC	1888	4	313
PAC	1304	3	286
XCB	1390	2	411
XFB	1433	1	636
		0	12948

The estimation of missing craniometric values has been a historically problematic endeavor for reasons related to the nature of skeletal data and to the choice of an optimal replacement method. Skeletal biologists interested in quantitative measures of human craniofacial variation are faced with the challenge of dealing with a difficult type of data. Craniometrics represent a human biological form, with some degree of heritability and plasticity (c.f., Cheverud, 1982; Sjøvold, 1984; Devor, 1987; González-José et al., 2004; Sparks & Jantz, 2002, 2003; Gravlee, Bernard & Leonard, 2003; Carson, 2006; Martínez-Abadías et al., 2009; Logan & Jantz, 2010). It is shaped by the intrinsic forces of genetics, e.g., ancestry

and sex, and can respond to the extrinsic pressures of selection, like climate and environmental stress. Given the unique individual interaction of these factors, the measurements that are taken to represent cranial form are inherently variable among and within groups and between even closely related individuals. The cranium is a composite integrated complex, and, so, its metric values are correlated and often collinear. Moreover, these linear distances fail to capture complete information on the cranium's volumetric shape and within this "lost" data there may be information that is critical to correct missing data estimation. Prediction is further complicated by the very theoretical premise that underlies the present project, drives its methodological decisions, and supports its clustering goals. The goal of this analysis is to test for the presence of latent population structure using model-based clustering. To do so, this project must assume that the craniometrics data set under analysis contains a mixture of populations, each of which has their own distribution of measurements. Given this "knowledge" of identifiable groupings, it would make the most sense to estimate missing data using a population-specific approach: as each group is presumed to be different in the configuration of their measurements from all other groups, imputation should be based upon the data available for the group into which the individual with missing data shares the greatest membership. The fallacy in the logic of this approach should be readily apparent: it is not reasonable to predict missing data using the same multi-population model that the project seeks to subsequently test. This approach is self-serving in the case of *a priori* "known" data and outright impossible for mixtures of individuals for whom their true population membership is "unknown." The researcher must be resolved, therefore, to estimate missing values assuming no hidden data structure and, in doing so, accept the risk of reducing the true among group variance and homogenizing the data to some "untestable" degree. A reasonable argument can, nevertheless, be made for partitioning the data set into males and females in order to conduct sex-specific missing data imputation. There should be no disagreement over making the assumption that the

global population-at-large contains both male and female sexes in equal proportion: this project is neither "testing" for sex-based structure nor "contesting" the existence of "sexes." The individual designations of sex made by outside, experienced observers should also be taken as largely accurate, as "male" or "female"identification can be made using credible external information. i.e., data other than the craniometric values under analysis, like pelvic morphology, post-cranial metrics, DNA, or personal documentation. Given these constraints and arguments over acceptable assumptions, this project has chosen to divide the combined Howells and Hanihara data set into males and females according to the *a priori* identifiers provided by each recorder and to use this sex-specific pooled-population data as the information upon which the predictions of individual missing values are based.

With the growth of digital imaging technologies for data acquisition and advances in data storage procedures over the last 35 years, researchers in the biological and social sciences have been able, with increasing ease, to obtain unprecedented amounts of information on study groups or collections containing many subjects. Large-scale data sharing and public-access to these databases has become a topic of interest among statisticians, with issues of data preprocessing for easy analysis appearing at the fore of their concerns (Rubin, 1978, 1987; Madow, Nisselson and Olkin, 1983). As any preliminary preparation of these types of data sets inevitably involves the consideration of missing values, new approaches to inference and their comparison with the otherwise standard solutions for incomplete data analysis have emerged alongside developments in database creation, management and mining (Rubin, 1996; Schafer, 2000; Little, and Rubin, 2002). Three solutions are commonly proposed "to handle the missingness in the data": (1) observations with missing data are omitted by row deletion resulting in a "complete case analysis," (2) variables are removed for which there are any number of missing values (i.e., column elimination), or (3) a computational technique is implemented, which

substitutes replacement values based upon some predetermined criteria (e.g., using column means or modes), or infers values based upon an underlying statistical paradigm, like regression, maximum likelihood estimation, or simulation, and using tools like the EM algorithm and Markov Chain Monte Carlo methods (Steele, Wang & Raftery, 2010, p. 352; Schafer, 2002). Of these, the first two may be the quickest and simplest of options but are the least satisfactory in approach: they are inefficient for large data sets, bias-prone, and wasteful. Deletion methods are acceptable choices only under the restricted conditions of having proportionally few missing values and randomly distributed case deletions; that is, when the user is assured that the missing data mechanism is MCAR (missing completely at random), that the problem observations do not differ systematically from the rest of the sample but make up a "random subsample" of the data set at large (Schafer, 1999, p.7), and that the case or variable eliminations do not remove "a significant fraction of the data collected" from the analysis (Steele, Wang & Raftery, 2010, p. 352). As pioneers in the area of incomplete data analysis, like Rubin (1976) and Little (1988), have long argued that the imputation of missing values will yield better results than any deletion method, the third of these options is the most satisfying from a statistical point-of-view. Estimation methods have, however, historically operated at high level of computational intensity and required a large degree of technical knowledge that has discouraged their practical application by many non-specialist users. It is perhaps due to this problem with broad accessibility that missing data estimation methods in anthropology, and specifically for craniometric data, while inspiring interest, have failed to gain the popularity needed to generate guidelines for best-practice or to establish criteria for method selection (see McKeown [1996] for a finite mixture approach). In the past decade, the multiple imputation method (MI), first introduced by Rubin (1978, 1987) for use with large-scale survey data sets and later generalized for wider application by Schafer (2000), has become increasingly accessible and is now commonly used across many fields. Multiple Imputation is not only "intuitively appealing and

relatively straightforward" (Steele, Wang & Raftery, 2010, p. 352) but also has the practical advantage of relatively easy implementation, either by using the accompanying standalone and free access software, NORM 2.02 (Schafer, 1999), or by calling to functions built into popular statistical environments, like the MI procedure in SAS 9.2 (SAS/STAT(R), 2009), or as part of the mix package, version 1.0-8 (Schafer, 2010) in R (R Development Core Team, 2010). The standard, Bayesian, approach to multiple imputation has been Rubin's (1987) Monte Carlo technique: missing observations are replaced by $m > 1$ simulated copies (with an effective number of imputations being reasonably small, e.g., usually between 3 and 5 runs) and the results of the analyses of each of the simulated complete data sets are then combined to produce individual parameters estimates with confidence intervals that speak to the question of uncertainty. While Rubin's method is both principled, not overly demanding in computational requirements and has been shown to produce useful results, it is not the only option for obtaining good missing data estimates. More efficient and problem-specific approaches can be developed that do not rely upon approximation (by simulation); such is the case for fully parametric models, where maximum-likelihood estimates can be computed directly from the incomplete data via the EM algorithm (Dempster, Laird & Rubin, 1977). While statistically optimal, these methods are typically difficult in their practical implementation (i.e., there are no published programs) and efficiency is gained at the loss of simplicity and generality (Schafer, 1999). Recently, Steele, Wang and Raftery (2010) have proposed a new approach to multiple imputation that approximates the marginal posterior distribution of the parameters or values of interest by a mixture of normals. This method is both conceptually familiar for its use of the well-known theory of density estimation (c.f., Ferguson, 1983; West,1993; Escobar & West, 1995; Fraley & Raftery, 2002) and practically advantageous for the fact that it has been shown to give narrower confidence intervals and to perform better than the traditional *t*-based solution (Steele, Wang & Raftery, 2010). Moreover, this mixture method is readily accessible as it can be simply implemented

using the contributed R package, MImix 1.0 (Steele, Wang & Raftery, 2010). Missing values in the combined Howells and Hanihara data set are treated within the framework of standard multiple imputation, but using the mixture method discussed here.

For the combined Howells and Hanihara data set, 3376 individuals are identified as missing 1 to 14 variables, of these cases nearly half (1333 individuals) are missing only 1 to 3 variables. This analysis takes this midpoint as its limit for imputation and, so, any cases with more than 3 missing values are eliminated. A total number of only 2119 individuals were marked for deletion when all of the cases meeting the missing-data or duplicated-data criteria were combined. This threshold for imputation/deletion chosen here is argued to effectively negotiate a balance between the problem of increasing uncertainty in estimation with additions to the number of missing values and the desire to retain a reasonably large, if not the maximum possible, sample size. When the data set is treated in this way, the ratio of missing data to complete cases falls close to Schafer's (1999, p. 7) definition of a "small" rate of missing information (5%) and for which a single imputation "may be fairly accurate." Since it is known that single imputation may overestimate precision by disregarding the between-imputation component of variability (Shafer, 1999), this analysis considers three imputed copies. This choice of runs also agrees with the recommended target range of three to five imputations by Schafer 1999. The edited Howells-Hanihara data set was subjected to multiple imputations using the imputeData procedure, implemented with the contributed R package, mix (Schafer, 2010). To display the location of the missing and nonmissing values for select imputations, pairs plots (by variables) were produced using the function, imputePairs, available through the R package, MCLUST (Fraley & Raftery, 2006, 2011). The final complete case data set was obtained by following the method described by Steele, Wang & Raftery (2010). The corresponding R package, MImix 1.0 (Steele, Wang & Raftery, 2010)

was used to summarize the results of the analyses of the three imputed data sets using a mixture of normal distributions.

This missing data replacement procedure is limited here to the combined Howells and Hanihara data set. Preference is given this imputed data set only if it is determined that this enlarged sample provides a more robust data set for stable estimation and classification. Otherwise, missing values are addressed by case-wise deletion. This data reduction approach is used (1) in a second cluster analysis of the combined Howells and Hanihara data set for purpose of comparison, (2) when a data subset has very few missing values and only a minimal number of cases would be lost, (3) for any test of structure that is concerned with cluster composition by population or classification accuracy at the level of the individual, and (4) in any other situation when the poor estimation of single values may produce erroneous results and misguide specific interpretations (c.f., Little & Rubin, 2002). The data sets representing all or some Howells's global cranial series (1989) and his "test" specimens (1995) as well as different subsets of the Forensic Anthropology Data Bank (Jantz, 2008) are treated in this way as they meet the fundamental criteria for reasonably adopting case-wise deletion. As summarized by Schafer (1999), it is an acceptable approach when "the discarded cases form a representative and relatively small portion of the entire data set" (p. 7).

Variable Selection

While there are ongoing discussions over the question of variable selection, both in terms of establishing standards of data collection and variable-use procedures for human craniometric analyses as well as the choice of the optimal variable selection method among many competing techniques, there has yet to be a consensus on best-practice. This disagreement largely spawns from a single fundamental question: is it most appropriate to maximize the use of the information available by exploiting all variables in a given data set or to make use of some computational procedure that subjects the available data to a selection algorithm, which statistically identifies a subset of the total variables that has the highest discriminatory power? From philosophic as well as practical perspectives, both options have their faults. The use of many variables has been repeatedly demonstrated to greatly improve prediction and rates of allocation accuracy across many data types (Breckenridge, 2000; Hubbe & Neves, 2007; Lubke & Muthen, 2007). Pontikos (2004) has even shown for Howells' data set that when all 57 variables are used, meaningful clusters that reflecting classic anthropological groupings of cranial morphology and prior knowledge of population histories are reliably obtained. These findings should not be surprising given that this many-variable approach allows for a large degree of correlation, which, as Gower (1972) has argued, is often the source for the information that enables classification. Any increase in classification success gained by the use of many variables must come, however, at the cost of parsimony in the analysis, at the risk of matrix singularity, and at the loss of samples. These three points are of particular concern given the mechanics of the finite mixture method advocated here. As this mixture-clustering approach offers the advantage of flexibility in model fitting, this project considers finite mixture models with separate component covariance matrices. With this freedom comes, however, an increase in computational complexity and intensity: a mixture model with "unrestricted" component-covariance matrices in its

normal component distributions is highly parameterized, with $\frac{1}{2}p(p+1)$ parameters for each component covariance matrix $\sum_i \ (i = 1, \dots, g)$ (McLachlan & Baek, 2010). Since the number of parameters increases with the addition of more and more traits, an already complex analysis now using many craniometric variables (especially in the numbers suggested by Hubbe and Neves [2007] and [2010], or used by Pontikos[2004]) would be very susceptible to failure in the Expectation-Maximization algorithm's computations, i.e., resulting from singularity or shrinking components. It is possible to specify a conjugate (Bayesian) prior to regularize data fit for normal mixture models and remedy many of the problems of degeneracy in estimation without adversely affecting the clustering procedure (c.f., Fraley and Raftery [2007] for Bayesian regularization of BIC-based mixtures). There is, however, no statistical quick-fix for the disconnect that arises between the methodological assumptions that underlie both of these positions.

Whether the final component solution is selected using information statistics, like the Bayesian Information Criterion, or by a test for significant increases in the log-likelihoods between progressively more complex cluster models (Fraley & Raftery, 1998; McLachlan & Peel, 2000, section 6.1 and 6.9), the practitioner is engaged in negotiations between the best model fit and the lowest level of model complexity. This point is made explicit in the advice given by McLachlan and Khan (2004) for model specification: "it is sensible in practice to approach the latter question of the number of components g in a mixture model in terms of an assessment of the smallest number of components compatible with the data" (p.95). As the purpose of model selection is, typically, to identify the most parsimonious solution that adequately represents the internal structure of a given data set, it is hard to justify the indiscriminant use of many highly correlated and often collinear variables. Moreover, any approach that dictates the use of a large number of variables is simultaneously self-limiting: it reduces the options for

acceptable samples as, in accordance with well-known rules for case to predictor ratios (Huberty, 1994; Fielding, 2007), very large and complete data sets would be required. For craniometric studies specifically, Jantz and Ousley (2005) have also adopted these general statistical guidelines, suggesting that a "reasonable recommended maximum number of variables seems to be the minimum sample size among all groups divided by three" (p. 49). These expectations are frequently unrealistic for a field whose data is typically highly fragmentary and limited in its availability: paleoanthropological specimens are few, bioarchaeological populations are usually poorly preserved and small in scale, and forensic casework is frequently damaged. These methods are also conceptually problematic by encouraging extreme sample reduction (assuming that observations with missing data are dealt with by-case deletion in order to preserve all variables) and, in turn, tacitly condoning the exclusion of potentially valuable variation and wasteful data practices. Methods that use selection algorithms to reduce predictor dimensionality, like the step-wise selection procedure advocated by Ousley and Jantz (2005) and commonly employed in the context of forensic identification, are not without their own problems. While it is true that they offer an empirical means to identify noisy variables that may offer little improvement to the discrimination or classification model, selection techniques on some level restrict the range of a given sample's variation. When they are considered for use with model-based clustering, they typically impose artificial constraints on all possible parameterizations, potentially reducing the clustering method's effectiveness and calling into question the logic for its use.

Recently, a mixture-specific method for variable selection was proposed by Raftery and Dean (2006) which can be used to identify the locally optimal subset of variables with group/cluster information available. This technique evaluates the evidence for each variable's contribution to the model by testing the difference in the value of the Bayesian Information Criterion (BIC) between (A) the model using the

currently selected clustering variables (including the variable under investigation) and (B) the sum of BICs for the model with clustering but with the variable in question having been removed. Although this technique is conceptually simple, easily implemented, highly effective and, most importantly, tailored specifically for its combination with model-based clustering (c.f., the contributed R package *clustvarsel*, [Dean, 2009]), it is potentially problematic for the fact that it is designed around the exclusive use of the Bayesian Information Criterion. As a result, it dictates the adoption of this information statistic approach for all stages of the analysis, including model selection and clustering, even though it may be desirable to make use of an alternative technique, like McLachlan's (1987) resampling method, for determining the optimal cluster number (c.f., Scott & Symons, 1971). Because this project (1) has chosen to also employ this latter technique (i.e., bootstrapping the likelihood ratio test statistic, LRTS) as an alternate approach to model decision-making, (2) endeavors to maintain consistency in the suite of variables considered across comparable analyses, and (3) prefers not to precondition the data set used in the resampling (LRTS) analyses on a technique whose variable selection criteria is fundamentally different in its philosophy, this BIC-based variable selection algorithm is inappropriate for use here.

Despite the statistical underpinnings of any methods chosen for variable selection and the arguments for or against their appropriate application, it is wise to recall that all such techniques are restricted to choosing among the input variables that define a particular data set. They are always already constrained in their effectiveness; therefore, their value can be only ever as great as the information returned from the most discriminating variable. As these selection methods identify, among the available options, the predictors that meet some predefined statistical criteria, in many cases, the suite of variables found to be significant for a given test data set may be problematic, especially in the context of comparative analyses or repeatability tests. That is, they may have some, little or no external

applicability, when different values are used and different conditions are imposed. There is also no guarantee that the variables selected are the most familiar and commonly collected, the most biologically meaningful or relevant to population-based questions, or the most frequently preserved and easily transferred across analyses and between data sets. Given these concerns and the detrimental effects that common selection algorithms have on finite mixture methods, it is advocated here, and in general, for the adoption of user-selected variables (at least for preliminary tests). A knowledgeable practitioner may make equally good if not theoretically better variable choices than most *ad hoc* techniques. A well-selected craniometric data set will give consideration to the nature of the complex under investigation (for the full cranium, choosing variables that represent its key architectural features), the current practices for data collection and publication (which variables are most commonly tested and readily available), and to the ultimate goal of the analysis (including variables, for example, that are known to correspond to differences among groups, within groups or between the sexes, or are most susceptible to environmental effects). In bringing forward prior knowledge of the context and conditions of analysis, it is possible to effectively circumvent most of the acknowledged statistical and practical problems related to automated variable selection while also yielding results that are immediately meaningful as well as useful to future work.

In this way, a base suite of 12 craniometric variables have been pre-identified for use in this project (Table 3). These interlandmark distances will serve as the foundation for this project's core cluster and classification analyses, with only slight modification in accordance with changes to the composition of the world-wide and modern forensic samples. This base subset is reasonable in number given the individual sample sizes available for the Howells (n = 2504), Hanihara (n = 11757) and FDB (n = 925) craniometric data. It is also sufficiently small in number to accommodate the addition of variables key to

specific tests or enhanced analyses and to permit its use with subsamples derived by partitioning any of the three test data sets based on population, sex, geography, or sampling strategy. These 12 variables also have the advantage of reasonable external validity. They are drawn from the standard list of cranial measurements defined, collected and extensively used for population-tests by Howells (1937, 1973, 1989, 1995), they have become a key part of current data collection guides (Moore-Jansen, Ousley & Jantz, 1994) and forensic reporting procedures (c.f., FDB data forms, 2011), they are among the predictors used for the determination of ancestry in FORDSIC 3.0 (Ousley & Jantz, 2005), and they are in keeping with the number of variables (≥ 10) recommended by Ousley and Jantz (2005, p. 49) for making "reliable comparisons" among modern human crania. The central role that these craniometrics variables play in the context of actual medico-legal casework has been already shown by Steadman and Konigsberg (2003), who addressed sex and race for personal identification using posterior probabilities derived from FORDISC 2.0 using, as their real-time example, a single (closed) forensic case. The utility of this particular metric subset has been recently affirmed in a follow-up publication by Konigsberg, Algee-Hewitt and Steadman (2009, p. 83). They show how, when 11 of these 12 base measurements[10] are used in a discriminant function analysis with informative priors, "Mr. Johnson" is correctly classified for sex and race: he is identified as male with a posterior probability of 0.999 and as "American White" with a posterior probability of 0.6978 and a typicality of 0.6329. This 12 measurement subset gains even further external support when these user-defined variables are compared to those selected by automatic statistical procedures, and specifically in craniometric studies. Ousley, Jantz and Freid (2009) have recently reported their classification results for a sample of American Blacks and Whites using

[10] Konigsberg, Algee-Hewitt and Steadman (2009) considered 11 cranial measurements previously collected for the forensic identification of "Mr. Johnson": BNL, BBH, BPL, NPH, NLH, OBH, OBB, NLB, FRC, PAC, OCC. In the current project, DKB (interorbital breadth), as defined by Howells (1973, p. 178), is included as an additional, twelfth, measure.

standard craniometric data from the FDB. They found that when stepwise selection is used, only seven variables are needed to discriminate between these two groups, with a classification accuracy of 95%. Of these seven variables identified[11], all but three (AUB, biauricular breadth, BPL, basion-prosthion length, and MAB, external alveolar breadth) are common to the base set of 12 variables chosen for this dissertation's standard mixture analyses. The combined Howells and Hanihara data set[12] also includes BPL as well as two measures of cranial breadth, XCB and XFB, which are similar in their information to AUB. What is especially important to note is, as these authors do, that all of these variables "are ones that can be visually appreciated by forensic anthropologists" and that "nearly all of these measurements represent morphological configurations mentioned in forensic anthropology texts as valuable in determining race visually: prognathism, cranial index, nasal breadth, and orbit shape." This good correspondence in "significance," between the best computationally identified variables (determined by unbiased statistics) and those that are thoughtfully selected by the researcher, is argued here to justify giving weight to context, practice and utility, and experience when compiling or culling craniometric data sets and designing reasonable subsets of predictors for use in multivariate analyses.

Diachronic Trends and Secular Change

A brief accounting of the publications on skeletal differences returns, as one would expect, a large collection of work concerned with change in body form (size and shape) over time (e.g., Aiello, 1992;

[11] Oulsey, Jantz and Freid's (2009, p. 71) stepwise selection algorithm identified BNL, BPL, AUB, NLB, MAB, OBH, BBH; the first 6 are in order of selection, the position of the seventh variable (BBH) was not provided by the authors.
[12] This project tests the combined Howells and Hanihara data set using 11 of 12 base variables (excluding OBB) and also considers GOL, XCB and XFB, as defined by Howells (1973, Appendix B, p. 178).

McHenry, 1992, 1998; Lahr & Wright, 1996; Ruff, 2002; Bruner, Manzi & Arsuaga, 2003; Schaefer, 2004; Mitteroecker et al., 2004). Considerable attention has been paid to quantifying the differences in sexual dimorphism and defining the morphogenetic boundaries between and within species from the perspective of human origins (c.f., Lieberman, Pilbeam and Wood, 1988; Relethford, 1994; Karmer et al., 1995; Lahr, 1994; Lockwood et al., 1996; Bookstein et al., 1999, 2003; Miller, 2000; O'Higgings, 2000; Lieberman, McBratney and Krovitzm, 2002; Bruner et al., 2004; Hlusko, 2004; Wolpoff & Lee, 2006; Weaver, Roseman & Stringer, 2007). Built into these analyses is a sizeable time depth, which is clearly important to the type of broad, often inter-specific, questions being asked about the trajectory of, say, primate evolutionary history. Studies of diachronic developments in skeletal morphology in the more limited context of "our species" alone have also shown the importance of the relationship between body form and time. Many of these have documented "an apparently global trend for directional reduction in body mass, skeletal size and robusticity, and tooth size in the post-Pleistocene period" (Brown and Maeda, 2004, p. 29). While this research has focused largely on trends observed in the masticatory system and their association to key cultural developments (e.g., relaxed selection for tooth size relative to the increased consumption of cooked food), or as a secondary effect of reduction in body size (i.e., allometric scaling) in post-Pleistocene populations (c.f., Sofar, Bailit & MacLean, 1971; Sofar, 1973; Le Blanc & Black, 1974; Frayer, 1977; Carlson & Van Gerven, 1977; Kennedy, 1984; Formicola, 1987; Brace, Rosenberg & Hunt, 1987; Brown, 1992; Jacobs, 1994; Christensen, 1998), changes in the craniofacial skeleton have equally been detected across many geographic regions. For example, a reduction in facial height and prognathism is paired with an increase in orbit height and a decrease in orbit breadth in skeletal populations from Australia, North Africa and East Asia (Calcagno, 1986; Calcagno & Gibson, 1988; Brown, 1992). Demonstrating a similar link between quantitative differences in the craniofacial complex and the passage of time has, however, proved to be difficult in studies of

"modern"populations. For this reason, it is important that this project states its position on the question of diachronic change and acknowledges the effect it may or may not have on the mixture analyses of, first, the Howells (1989) and Hanihara (2008) data sets and, second, the contemporary American samples drawn from the Forensic Anthropology Data Bank (Jantz, 2008).

In addition to providing a great breadth of geographic range, the cranial series sampled by both Howells and Hanihara vary greatly in the time-periods covered. Howells (1989) has included, for example, Bronze Age burials of the Shang Dynasty Chinese at Anyang, Egyptian series "E" crania from the 26th-30th Dynasties near Gizeh, Hungarian "Zalavar" cemeteries dating to the 9th-11th century CE, and Norse specimens from Mediaeval Oslo. Of these, the "archaeological"Zalavar and Norse series make up two of the three populations that are used by Howells to characterize "European" morphology. Hanihara (2008) subsequently measured crania from many of these same populations as well as other samples of substantial age, e.g., the Neolithic Jomon of Japan, the Pre-Contact Mokapu (ca. 1385-1778CE) of Hawaii, Early Metal Age Thai dating to 3000-6000 years BP, and the Medieval burials from the site of the Priory of St. Mary Spital, its cemetery and charnel house (the "Spitalfields," East London) in England. Hanihara's extensive data set also includes many additional "recent" series, which either extend culturally from or overlap spatially with his "ancient" populations. This breadth of sampling provides "modern" coverage for all of the major geographic regions and gives balance to this data set's temporal representation. The sheer diversity and the size of Hanihara's data set, either when taken alone or in combination with Howells' series, equip its users with sufficient cases to address any potentially confounding effects of time by choosing to omit "problem" [sub]samples; that is, simply excluding skeletal populations coming from a particularly "early" cultural period or dating prior to a predefined "cut-off." There is already a precedent in recent literature for controlling temporal effects in this way.

When testing relationships among climate, geography and phenotypic diversity using the Howells and Hanihara data sets, Carson (2006) removed all "samples from populations or geographic locales that could be definitively dated to earlier than 1700 AD" (p. 104). Betti et al., (2009) considered "only skulls younger than 2000 years" (p. 809). Manica et al., (2007) also excluded "skulls older than 2000 years" in order "to avoid any bias in the quality of the material" (p. 346). This concern for "bias" and "quality" in these studies may have less to do with the problem of diachronic change in morphology and may be better related to their interest in testing environmental variables. Carson (2006), for example, has explained the difficulties of pursuing climate-oriented studies in terms of time: "In selecting environmental data to match the cranial samples, I was confronted with the problem that more contemporaneous historical measures of climate and atmosphere were not available for the worldwide distribution of crania. This was compounded by the fact that the dates for the skeletal samples are not exact, and in many cases can only be estimated within a 200 or 300 year period. Thus, selecting the temporally corresponding value for a particular environmental measure was difficult"(p. 108)..."More serious is the assumption that environmental data from the mid- to late-1900s can be correlated with crania that are somewhat older in age" (p. 104).

If Howells data is considered alone and earlier discussions on the large body of work that has made use of the Howells data set in tests of global population differences are recalled, it is easy to see how an opposite precedent has been established for dealing with Howells' "past" populations. Instead of eliminating members of his core groups, most users of Howells' s data set have carried out extensive and varied analyses without making "corrections" to his mix of recent, ancient, and even extinct cranial series (e.g., the Tasmanians). It is also worth reiterating that Howells' groups have been used extensively in arguments over the forensic estimation of ancestry using craniometrics; there is, therefore, an

expectation that any findings from Howells' data can be extended toward questions that affect current casework of medico-legal significance. This is certainly true for the test of FORDISC by Williams, Belcher and Armelagos (2005), which is among the most popular and thoroughly discussed of these "forensic" [re]analyses and opinion pieces. Their conscious choice to evaluate the utility of the FORDISC software using a sample of Meroitic Period Lower Nubian crania clearly implies a general acceptance of an easy correspondence between contemporary skeletal material and older specimens. More to this point is the fact that FORDISC itself has several archaeological (e.g., Howells') populations incorporated within its reference database and it, nevertheless, produces good results for modern forensic cases. Evidence external to the FORDISC debate also offers support for assuming temporal stability in morphology. For example, Rösing's (1992) study of diachronic trends in Egyptian craniometrics provides very valuable conclusions:

> The core result of this calculation experiment is a largely unchanged cranial morphology through the whole era of the ca. 3500 years of ancient Egyptian history, starting with the early dynasties and ending in the late period, the last days of independence and of continuous own culture [...and....]important [historical] events had little influence on population structure as we can see morphologically (p. 194).

More important even than this clear evidence of long-term stasis is the fact that these findings of morphological continuity are not limited to Egypt, Antiquity, or the Old World. Analyses of the archaeological site of Indian Knoll, Kentucky, suggest a similar consistency in cranial form through time in the Americas (c.f., Hermann, 2002). Although this Archaic burial site was in use from approximately 4500 to 3500 BP, Jantz (2004) has demonstrated how the morphological variation observed over this 1000 year period of population history does not exceed the amount estimated for modern Native

Americans tribal groups, e.g., the Arikara, Blackfoot, Peru, Santa Cruz and Zuni. While Ruff (1980, p. 101) did detect significant "craniofacial expansion with aging" among "younger" and "older" cohorts, he ultimately concluded that the "trends observed at Indian Knoll are most likely to reflect true craniofacial growth in size among the adult male inhabitants of the site, rather than secular trends." McKeown (2000) has also reported results from her analysis of a series of Arikara crania from the Great Plains that suggests a similar long-term stability within prehistoric American groups. She interpreted instances of variation in morphology observed among several archaeological populations as "small scale localized change" resulting from admixture or gene flow between Native American sites. She found in general that "there are no statistically significant temporal trends" (p. 162-5). Furthermore, current studies on punctuated change in craniofacial shape have demonstrated convincingly that episodes of morphological variability within an otherwise stable trajectory are mainly environmental, highly context specific, and involving only one or two generations (Smith et al., 1986; Cameron et al., 1990). For these reasons, short-term changes are likely very limited in the influence that they exert over studies concerned with broad population patterns, and especially those using many overlapping cranial series. Moreover, Sparks and Jantz (2002) have shown for Boas' (1910, 1912, 1928, 1940) classic immigrant data that, in fact, the effect of the changed "new" American environment is small when compared to the degree of cranial variation that was retained among the ethnic groups surveyed. It is known that both genetic and environmental components can exert control over the magnitude and direction of cranial change through time; for example, trends in brachycephalization have been observed from Europe to Japan (Suzuki, 1969; Nakahashi, 1987; Schwidetsky & Rösing, 1990). Explicit tests of selection versus population structure as explanations for the diversity observed in human cranial morphology have shown that population processes not "cranial plasticity" (c.f., Larsen, 1997) best account for the long-term patterns of variation that characterize most populations (e.g., Kobyliansky & Livshits, 1985;

Relethford, 1994; Roseman, 2004; Betti et al., 2009). To this point, Jantz (2004) has demonstrated in the context of past population studies and modern forensic identification methods that "selection or developmental plasticity do not erase or even obscure underlying population structure and history" (p. 9). Given the consistency in the evidence suggesting morphological stability through time, regardless of temporal period or geographic local, time depth should not be a confounding factor in global analyses of among group craniofacial variation: the "early" morphological signatures of ancestry in "past" populations can be taken to be, both, reasonably representative of "later" skeletal trends and applicable to issues of "current" interest. It is neither theoretically necessary, nor methodologically wise, to reduce the global samples considered here based solely upon the age of the cranial series in either Howells's (1989) or Hanihara's (2008) data sets. An important exception is made, however, when this project takes up the question of differences in craniofacial shape among the contemporary American groups sampled from the Forensic Anthropology Data Bank (Jantz, 2008). It is argued here that the emerging phenomenon of "secular change" warrants special consideration by any study that seeks to thoughtfully interpret differences in "current" craniofacial forms and to identify [sub]population signatures at the level of group and for the individual in the interest of medico-legal investigations.

Since Boas (1910) first documented changes in bodily form among "American" descendants of European immigrants at the beginning of the 20th century, environmental plasticity has been a topic of interest in physical anthropology. It has remained an important point of debate in discussions of the anthropometric differences detected among very recent and otherwise cohesive groups. As it has been shown that growth and body size respond both positively and negatively to extrinsic factors (Komlos, 1994; Larsen 1997), it should not be surprising that the novel environmental conditions developing over the last 200 years have been associated with morphological changes observed in all parts of the skeleton

and among biogeographically distant populations. Demographic transformations (Boldsen, 2000), like increased urbanization (Vogele, 1999), decreased infant and childhood mortality and disease prevalence (Mercer, 1990; Wilmoth, 2000), decreased activity with increased nutrition (Flegal et al., 1998, 2002; Armstrong, Conn & Pinner, 1999; Ogden et al., 2006), and increased adult longevity (Riley, 2001, 2005), have all been cited as the key "environmental cues" precipitating the recent morphometric changes detected among populations from "developed countries" (Weisensee & Jantz, 2011). While cranial plasticity "is usually considered less labile than height or other body dimensions" (Jantz & Meadows Jantz, 2000, p. 327), there is a growing body of literature identifying unprecedented modification to the craniofacial complex in widespread geographic areas, including Africa and its Diaspora (Cameron et al., 1990; Spradley, 2006), Europe (Buretić-Tomljanović et al., 2004; Buretić-Tomljanović, Ostojić & Kapović, 2006; Weisensee & Jantz, 2011), Japan (Kouchi & Koizumi, 1992; Kouchi, 2000), Mexico (Little et al., 2006), and the United States (Angel, 1976, 1982; Jantz & Meadows Jantz, 2000; Jantz 2001, Wescott & Jantz, 2005).

Of particular interest here is the evidence for craniofacial secular change in the United States, as these trends may have significant implications for the ways in which the cluster results of the FDB data can be most meaningfully interpreted. Jantz and colleagues have consistently demonstrated "that certain cranial variables exhibit striking change over the past 1.5 centuries in American blacks and whites" (Jantz, 2001, p. 784). These secular trends include modifications to cranial vault height, base length, total length and both facial and vault breadth (Jantz & Meadows Jantz, 2000; Wescott & Jantz, 2005). Of these key dimensions, Jantz (2001, p. 784) has noted in particular that the distance from basion to bregma is the single measure "showing the most remarkable change" over time. Using geometric morphometric techniques, Wescott and Jantz (2005, p. 234) have further refined the current

understanding of these craniofacial developments, by demonstrating that "secular change in American crania is concentrated on the base and posterior aspect of the skull." These shifts in cranial architecture are said "to reflect a major restructuring of the cranial vault, yielding a high, narrower vault in modern people" (Jantz, 2004, p. 11). In addition to these general patterns of what can be called collectively as modern American secular change, Jantz (2001) has identified clear subpopulation (sex and race-based) distinctions. When considering, for example, vault height and birth year, he (2001) explains that: "the strength of association is higher in whites than in blacks and highest in white males. These correlations are astonishingly high; they imply that between 30 and 50% of the variation of the cranial canonical score can be explained by year of birth" (p. 784).

This evidence of quantifiable changes in the morphology of the cranial skeleton among recent Americans has significant implications for this project's application of mixture cluster analysis to the craniometric data obtained from the Forensic Anthropology Data Bank (Jantz, 2008). There is potential for these clustering procedures to capture two levels of complexity in the FDB data structure. The patterns produced by the mixture analysis may not just reflect among groups differences but they may also represent latent within-group partitioning resulting from secular change. Ousley, Jantz, and Freid (2009, p. 74) have, for example, noted the effect of these temporal differences in the results of a discriminant function analysis using modern data: they report that "white males born between 1840 and 1890 can be separated from white males born 1930 to 1980 very well, and they are distinguished by time." As key demographic information is available for the individuals sampled from the Forensic Anthropology Data Bank, the consequences of this short-term change over time on the resolution of the finite mixture analysis can be reliably explored. As it is not the intention of this project to restate or retest previous findings of secular change in modern American data, the effect of time is evaluated here

only in terms of the results of the mixture analysis. The relationship between birth year and cluster membership is visualized graphically for Black and White subsets of the FDB data. These exploratory methods are adopted over formal tests of the correspondence among race, sex, birth year and cluster assignment, as the mixture components are statistically inferred and, so, lack the external reality necessary to conduct other standard analyses.

Even though there is some correspondence between the populations sampled by Howells and Hanihara and the documented evidence of secular change in specific regions or among particular biogeographic or cultural groups, secular change over "recent" time is not investigated for these data sets. This decision is justified for the following reasons: (1) many of these cranial series are of considerable age and even those that are described as "modern" are rarely sufficiently "current," (2) individual birth years or age cohorts are not available, (3) the anatomical transformations that have been taken to represent secular change in particular groups (e.g., secular change in the craniofacial dimension in children in rural Oaxaca, from 1968-2000) have yet to be adequately investigated for the applicability of their findings to similar but different samples and broader population contexts (e.g., related urban centers, Southern Mexico, all of Mexico, or Latin America), (4) the changes highlighted in studies of, specifically, recent American Blacks and Whites are most likely indicative of a unique U.S. phenomenon, driven by the unparalleled environmental conditions of 20th century America, and (5) this project's analyses of Howells' and Hanihara's world-wide data sets are focused on detecting the presence of population structure and, in turn, teasing out broadly-based patterns and global trends.

Tests of Structured Variation: Howells' World-Wide Data Set

"In the later generations multivariate statistics have provided a much better sort of model for variation within and between populations, in the kinds of overlapping groupings which can be detected in a multivariate space. These are satisfying, but we must always inquire how true such fits are. We wish art to imitate nature, but it is up to the anthropologists in particular to see that nature does not imitate art." (Howells 1984, p. 4)

Finite mixtures analysis is being recognized more and more as a robust approach to modeling the distributions of a wide variety of random phenomena. With the ongoing advances in statistical computing, the algorithmically intensive methods of model-based or mixture clustering are growing in their ease of accessibility, making finite mixture analysis a more viable option for practical applications in recent years. These mixture-likelihood clustering methods are becoming increasingly available as part of the free and open-source software environment, R (R Development Core Team, 2010). This platform, by its design, affords the user familiar with the statistical language, S, considerable freedom in the implementation of finite mixture methods and the opportunity to maximize their utility for particular data types, unique conditions and project-specific goals. In the interest of method transparency and resource accessibility, this project has restricted its resources to publically available data sets and open-source software; the core statistical analyses were performed in R, version 2.13.0, using its well-supported "R packages." For reasons related to large sample sizes and complex data structures, and with a concern for maximum efficiency in the execution of the finite mixture procedures, the contributed R package, MCLUST version 3 (Fraley & Raftery, 2011), was used to implement all of this project's model-based clustering analyses. This R extension provides a suite of functions that permit the user to combine

the EM algorithm with the Bayesian Information Criterion (BIC) and produce comprehensive strategies for clustering, density estimation and discriminant analysis via Gaussian mixtures, as well as graphical visualizations of model options, clustering patterns, and classification results (Fraley & Raftery, 1998, 2002). The needs of this project's data and the nature of its test questions dictated when the analyses conducted using MCLUST (Fraley & Raftery, 2011) were supplemented with comparable tests using the alternative R mixture package, EMMIX (McLachlan, 2011; McLachlan, Wang & Peel, 2011). The EMMIX package includes of all of the same procedures that are still freely available in the standalone software, EMMIX versions 1.3 (Peel & McLachlan, 1999) and 2.0 (McLachlan, Peel, Basford & Adams, n.d.), but offers the added functionality of working within the R environment. The EMMIX package allows the user to iteratively fit multivariate mixture models by maximum likelihood (ML) using the EM algorithm for normal distributions or the ECM algorithm for skew t-distributions (as described in McLachlan and Peel [1998]) and conduct density estimation, partial classification and fully supervised discriminant function analysis. Perhaps its greatest advantage is that it also provides a convenient means for determining automatically the best mixture model with the smallest number of components using the likelihood-based test and resampling approach of McLachlan (1987). The pre-release version (1.0.18) of EMMIX discussed here was kindly made available by Dr. Geoffrey McLachlan (2011) for use in this project.

This project's search for cryptic population structure began at the level of global craniofacial variation by applying finite mixture methods to the full skeletal series represented in the WW Howells Craniometric Data Set (Howells, 1996). All 28 populations (26 for females) and 2504 individuals were considered. The mixture analysis was, however, limited to a preselected subset of 12 standard craniometric measurements (Table 3), which had been previously scaled to adjust for differences in size (c.f., Darroch & Mosimann, 1985; Jungers , Falsetti & Wall, 1995). Model-based clustering was implemented in R, and the MCLUST package. It has been shown in many varied applications that, when using a finite mixture

model-based approach, the clusters produced are very frequently elliptical in shape, even for complex high-dimensional data. It is common practice, therefore, to consider fitting mixtures of elliptically symmetric component densities, and, to choose specifically within this class, the multivariate normal density (McLachlan & Baek, 2010). These normal mixtures offer considerable computational convenience, as they can be easily handled in MCLUST: for any given number of multivariate normal components, the parameters are estimated by maximum likelihood (ML) via the EM algorithm. The algorithm itself is straightforward in its programming and can be simply partitioned into separate functions as it proceeds iteratively in two steps, E, for expectation, and, M, for maximization, until the likelihood (or the parameter estimates) meets a predefined threshold or changes by a sufficiently small amount to achieve convergence (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 1997). Because, with mixture models, the likelihood is known to have multiple maxima (and so, the likelihood equation will have multiple roots), "starting" values have to be provided for the vector of unknown parameters to initiate the expectation or E-step in the first iteration of algorithm (McLachlan & Peel, 2000). To fit a mixture model in this way, nonrandom partitions of the data can be produced in advance using some alternative clustering procedure (for k-means, c.f., Konigsberg, Algee-Hewitt and Steadman [2009]; for non-random procedures in general, see Colman et al. [1999]). In all cases of clustering using MCLUST, the package's default method was adopted: the EM algorithm is initialized by the classification information obtained from a prior hierarchical clustering of the input (craniometric) data using an unconstrained covariance model (Dasgupta & Raftery, 1998; Fraley & Raftery, 2011). While MCLUST includes a provision for considering only a subset of the data in the initial hierarchical clustering phase (before applying the EM algorithm to the full data set) when the number of observation is large, it was possible to consider the full data set (n= 2504) without loss of functionality (Fraley & Raftery, 2011).

In order to decrease the dimensionality of a data set, it is common procedure to perform a principal component analysis (PCA), although it is well-known that projections of the feature data on the first few principal axes are not always helpful in revealing the true structure of the data or the relationships of interest between groups (see McLachlan & Peel, 2000, section 8.2). Similarly, in order to reduce the number of unknown parameters requiring estimation in a finite mixture model, a typical approach in many clustering programs has been to limit the component-covariance matrices by forcing them to be equal or diagonal (e.g., AutoClass by Cheeseman and Stutz [1996]). Banfield and Raftery (1993) have proposed instead a less restrictive framework for the geometric constraints in multivariate normal mixtures which involves the reparameterization of the component-covariance matrices (Σ_i) in terms of their eigenvalue decompositions, i.e., based on a variant on the standard spectral decomposition of $\Sigma_i (i = 1, \dots, g)$. In this case, the decomposition factors are taken as independent sets of parameters, which are either constrained to be the same for each component or are given the freedom to vary among components; when the parameters are fixed, the components share certain geometric features (e.g., shape, volume and orientation). Some of the more common covariance structures include, for example: $\Sigma k = \lambda I$, where all components are spherical and the same size; $\Sigma k = \Sigma$, where all components share the same geometry but are not necessarily spherical; or the unrestricted Σk, where each component may have a different geometry. Although this flexible approach is subject to the potential problem of singularity under certain data conditions (typically, if p is large relative to the sample size n), it offers the important advantage of fitting this project's finite mixture models under the condition of maximum freedom, where it is possible to determine the "best" unbiased choice for the given data from a variety of covariance structures. As Peel and McLachlan (1999) have advised, "unless the user has some prior knowledge of the covariance structure arbitrary covariances should be used" (p. 8). It also allows for comparisons to be made between the fit obtained from the mixture models produced under

the more flexible parameterizations and the simple "equal convariance" models that are characteristic of the more restricted or "supervised" clustering and classification methods. In these constrained contexts, the user predefines the form of the parameterization or the assumptions of the method limit the user to a particular configuration, as is the case for, e.g., linear discriminant function analysis. Ten covariance structures are currently available for implementation in the MCLUST software and are given the following identifiers: EII, VII, EEI, VEI, EVI, VVI, EEE, EEV, VEV, VVV. Further explanation of their individual model forms, distributions, and geometric properties are provided in Table 7 (Fraley & Raftery, 2011, p.6).

Table 7: Available model parameterizations (Fraley & Raftery, 2011, p.6)

identifier	Model	HC	EM	Distribution	Volume	Shape	Orientation
E		•	•	(univariate)	equal		
V		•	•	(univariate)	variable		
EII	λI	•	•	Spherical	equal	equal	NA
VII	$\lambda_k I$	•	•	Spherical	variable	equal	NA
EEI	λA		•	Diagonal	equal	equal	coordinate axes
VEI	$\lambda_k A$		•	Diagonal	variable	equal	coordinate axes
EVI	λA_k		•	Diagonal	equal	variable	coordinate axes
VVI	$\lambda_k A_k$		•	Diagonal	variable	variable	coordinate axes
EEE	λDAD^T	•	•	Ellipsoidal	equal	equal	equal
EEV	$\lambda D_k A D_k^T$		•	Ellipsoidal	equal	equal	variable
VEV	$\lambda_k D_k A D_k^T$		•	Ellipsoidal	variable	equal	variable
VVV	$\lambda_k D_k A_k D_k^T$	•	•	Ellipsoidal	variable	variable	variable

Given the availability of the option to test for differences in covariance structure in MCLUST, prior constraints were not imposed on the covariance matrices but, instead, maximum flexibility in parameterization was allowed. The best models were determined by "user-free" criteria, based on information theory. The Bayesian Information Criterion (BIC) was used exclusively for model selection in MCLUST. It is an approximation to the Bayes factor (Schwarz 1978), which adds a penalty to the

loglikelihood based on model complexity, i.e., on the number of model parameters (Banfield & Raftery, 1993; Kass & Waserman, 1995; Kass & Raftery, 1995; Raftery, 1995). The calculated BIC values associated with the mixture-models tested serve collectively as the fundamental decision method for MCLUST's automated, and often batch, processes (Fraley & Raftery, 1998; Fraley & Raftery, 2011). In order to determine the optimal parameterization and component solution for the data, MCLUST was used to execute, in a single step, a comprehensive model selection and data clustering procedure. The "Mclust" function (1) fitted several models with different configurations (varying the relationships between the covariance matrices as well as the number of possible clusters) by maximum likelihood, (2) computed the Bayesian Information Criterion (BIC) given the maximized loglikelihood for the models, the data dimensions, and the number of components, (3) identified among the models tested, the parameterization and numbers of cluster that produced the largest values for this criterion, and (4) returned the "optimal" clustering solution for the particular mixture analysis (Fraley & Raftery, 2011). The approach operates under the assumption that the greater the BIC, the stronger the evidence in support of a good model fit, and the "better" the clustering results. As the maximum BIC is taken over all of the models and the numbers of components considered, the cluster solution identified by MCLUST is contingent upon the predefined value of G or the range of g-components, where G is taken to correspond equally to the maximum number of g possible "groups." The limitations placed on the number of the components tested may have a significant and potentially detrimental effect on the final cluster pattern, when the true number of groups is neither known nor intentionally specified by the user. While MCLUST compares, by default, BIC values for parameters optimized for up to nine components and the ten available parameterizations, the range for the number of the components in the mixture model to be fitted and "tested" (using the BIC) was intentionally broadened. For all clusterings, whether executed in MCLUST or EMMIX, a minimum of 1 to a maximum of 30 components

were considered. This range is arguably large by most clustering standards and especially for biological data: recall that Rosenberg et al., (2002) only reported STRUCTURE's results when K<7 for their cluster analysis of worldwide molecular data (c.f., Bolnick, 2008). The maximum value of 30 components was chosen because it exceeds the number of true populations in Howells' (1989) data set, i.e., 28 groups when sexes are pooled. It should, therefore, account for any mixture solutions that may produce, in the presence of extremely strong data structuring, a direct or close correspondence between the optimal number of clusters identified and the maximum number of true populations.

For this test of Howells' data set, MCLUST successfully fit all of the necessary parameter values using the default initialization. Despite the large sample size (2504 individuals) and the large maximum value of G (30 groups) considered in the analysis, reliable BIC values were obtained for the full range (1:30) of the g-components tested. Figure 7 displays the set of models (numbers of clusters and parameterizations) and their corresponding BIC values fit for this specified range. The model associated with the highest BIC value was chosen as the optimal clustering solution: "EEE" model with 16 components and a BIC value of 110765.65. Out of the 30 models and 10 parameterizations evaluated, the top 10 models were isolated and investigated these in greater detail for their cluster patterns and uncertainties in allocations; the number of components, parameterization, and BIC values for each of these models are provided in Table 8.

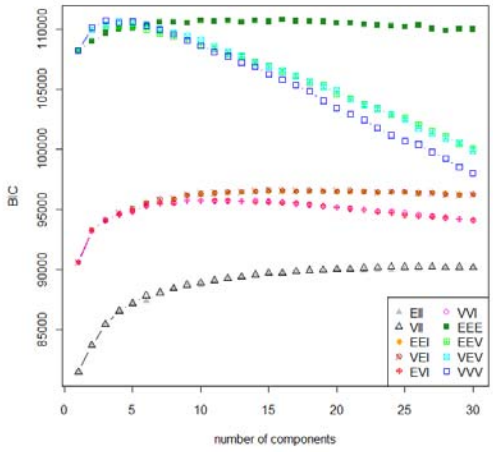


Figure 7: Model selection by BIC

Table 8: Top 10 models in order

Component Number	*Model Form*	*BIC*
16	EEE	110765.65
14	EEE	110725.492
12	EEE	110705.013
10	EEE	110701.919
3	VVV	110685.856
17	EEE	110677.28
18	EEE	110637.659
4	VEV	110626.893
11	EEE	110623.736
5	VVV	110610.704

If the BIC graph is reconsidered in light of this subset of top models, it is not only helpful in providing a visual representation of the behaviors of the "best" model over the range of components but it is also useful for identifying general modeling trends. In this case, it clearly shows a concentration of six "peaks" (points with high BIC values) at cluster numbers near the chosen 16-component solution (e.g., at 10, 11, 12, 14, 17, 18 clusters). Moreover, it highlights the fact that all of these seven models share

the same simple elliposidal, equal-covariance model parameterizations. This observed pattern offers external support for the strength of the relatively large, and, in turn, more complex component solution. It can also be taken as good evidence for accepting the highly constrained "equal" covariance structure over the more parsimonious alternatives (3, 4, and 5 clusters), for which the distributional characteristics (orientation, volume and shape) were allowed to vary freely between clusters (e.g., "VVV" or the slightly more restricted "VEV" models). These findings are further bolstered by the trends observed in the BIC values when the full matrix of all components and all models are considered. The 10 least supported or "worst" models overall all include component numbers between 1 and 5, and are split between the "EII" (spherical, equal volume, no shape) and the "VII" (spherical, unequal volume, no shape) models. The BIC values associated with these weak models range between 81504.95 and 87203.11, indicating a distance in BIC values between the best and worst models of over 29206.70 points. A mere 40.158 points separate the top two models and less than 80 points divide the best complex model from the best parsimonious. If further investigations are restricted to the BIC values given for the class of "EEE"models, under which the optimal 16 cluster solution falls, it is found that the models with the lowest BIC values correspond with not only these same small cluster numbers (1-5 components) but also with the largest numbers of components (25, 27-30 clusters), with only slightly improved BIC values. On average, these EEE models are still separated from the 16 cluster solution by a remarkable 1000 points. The significance of the high and low BIC values can be better appreciated when these differences are contextualized in terms of what Raftery (1995) has called the "grades of evidence" (p. 139) corresponding to between-model differences in BIC. While Raftery proposes for "BIC Difference" a more conservative version of Jefferys' (1961) rules of thumb for interpreting Bayes factors, the outcome of its use when adapted to the present analysis is still quite striking: when comparing values produced for two models, M1 and M2, Raftery (1995, Table 6) identifies a difference of >10 in BIC

values or a Bayes factor of > 150 (corresponding approximately to a BIC difference of 2 log 100 or the posterior odds of about 150:1) as "very strong" evidence in support of M2 and against M1. Following these now common guides (c.f., Kass & Raftery, 1995; Wasserman, 2000), there is little uncertainty left in the form and cluster number of the optimal model produced in MCLUST; even the improvement in fit between the BIC value estimated for this top model (110765.65) and the second best (110725.492) is unambiguous (with a calculated BIC difference of 40.158), following the standard "conservative values" given by Raftery (1995, p. 139). In addition to the validation that they offer for the results of this cluster analysis, these findings are of special interest for their broader interpretive value. It is useful to consider the constrained parameterizations reported by MCLUST: of the 35 BIC values that are >110000 (i.e., in the same order of magnitude as the best model), 23 of these are in the class of "EEE" models, with equal covariance matrices (they also include seven of the top ten models). This consistency in the algorithm's choice of the EEE model-form shows that, even when maximum flexibility is permitted in model parameterization, Howells' craniofacial data is still best fit by this restricted "equal" covariance structure. These findings tentatively suggest that it is reasonable to impose, in the absence of prior testing, equal covariance matrices across all groups, when investigating metric variation among modern human cranial series using model-based clustering. They also offer preliminary support for the common assumption that homoscedastic conditions are met, or can be violated without observable consequence, in other methodological contexts of interest to biological anthropologists, like the standard practice of conducting ordinary linear discriminant function analyses (LDA) using traditional craniometric variables (McLachlan, 1992; Huberty, 1994; Ousley & Jantz, 2005). These model selection results also have notable implications for the interpretation of the strength of the "best" (G = 16) mixture-clustering solution, and, in turn, the strength of the data structure that underlies Howells' (1989) multi-group craniometric data set: 28 populations, representing 6 distinct geographic regions and seven local series.

Encouraging findings for the utility of the BIC in model selection were reported early on in the development of methodological protocols for model-based clustering and density estimation (c.f., Campbell et al., 1997; Dasgupta & Raftery, 1998; Roeder & Wasserman, 1997; Fraley & Raftery, 1998) and the production of BIC values as standard output (alongside, e.g., the loglikelihood, or the AIC, ICL and AWE criterions) by many recent finite mixture software packages attests to the consistency of its current use and its accepted reliability (Fraley & Raftery, 2011; McLachlan, 2011). BIC's utility has been especially noted in tests of model selection criteria when taking an information theoretic approach: for example, it has been shown that "the penalty term of BIC penalizes complex models more heavily than AIC" and "it reduces the tendency of the AIC criterion to fit too many components" (McLachlan & Peel, 2000, p. 209). With the recent movement in statistical research towards improving the application of finite mixture methods for clustering large, complex and high dimensional data sets (e.g., using mixtures of factor and t-analyzers), BIC has also been singled-out as the preferred choice for testing the number of components (or factors) in the mixture (as in McLachlan and Baek [2010]). Nevertheless, concerns have been voiced over the tendency for BIC-based methods to fit too few components (c.f. Celeux & Soromenho, 1996). Raftery (1995) has explained this concern in terms of model-building strategies for research using human data:

> One apparent difficulty with the approach outlined here is that when a parsimonious but ill-fitting mode, M_1 is compared with a highly over-parameterized model M2, BIC often prefers the more parsimonious model, even though it may be clearly sociologically unacceptable. When forced to choose between two unsatisfactory models, BIC tends to choose the one with fewer parameters. This had led some researchers to worry that BIC is biased in favor of parsimony over fit (p. 152-3).

Given this known tendency, it is important that, in the case of model selection for Howells' data set, BIC favored relatively large cluster numbers for seven out of the ten top models and returned exceptionally low values for the most parsimonious models with very few components. These findings would suggest that this data's structure is sufficiently strong to override BIC's predisposition for solutions with low-complexity. In turn, they provide statistically robust quantitative evidence for the presence of craniometric [sub]structuring that cannot be related to sexual dimorphism (as clusters exist in numbers >2) and cannot be reasonably explained away by arguments that invoke user-induced biases or methodological problems involving forced allocations. It is also worth noting that the BIC values do not support a partitioning of the Howells data set into clusters that correspond directly in number to the known groups, given the cranial series available and the limited suite of metric variables considered here. This disagreement suggests, not unexpectedly, that the highly variable terminology used by Howells and in population-based self-identifications (which have been historically based upon on the problematic concepts of races, ethnicities, tribes, and linguistic, religious or cultural groups) do not accurately reflect the patterns of variation extractable from the craniometric data alone. The fact that this mixture approach has detected high-level structure, which can be resolved into organized clusters, nevertheless, demonstrates the promise of using model-based clustering results (i.e., the patterns of intra-component memberships and inter-component relationships) to tease out latent information on affinity. The patterns of group relatedness produced by the cluster analysis may better reflect the true structure of human variation than predefined population naming conventions. The statistically derived associations may also reveal meaningful links among the clusters, true associations and common biogeographic or biosocial identifiers. In order to address these possibilities, it is necessary to evaluate the results of this cluster analysis of the full Howells (1989) data set in greater detail. This discussion is focused on the mixture results from the optimal solution identified by MCLUST using the Bayesian

Information Criterion: 16 components with an equal covariance structure assumed across all groups.[13] This model is also compared against the 3 component VVV solution, the 5th best choice as identified by its associated BIC value. Although this model is complex parameterization,[14] as it allows for separate variance-covariance matrices for each group, it represents the most parsimonious of the "good" models produced (i.e., having the largest BIC and the fewest clusters) and for this reason merits further investigation.

It is useful to begin this discussion of the 16 component solution by first reviewing this model's clustering patterns and statistical properties. The number of individuals assigned to each of the clusters is presented in Table 9. While there are some fluctuations in the distribution of the number of cases across these 16 clusters, each cluster is well populated: even the smallest group (cluster 7) contains 12 individuals. The absence of clusters containing one to only a few data points implies that this data set is free of outlying observations, e.g., cases representing very abnormal morphologies and errors in measurement or transcription. They also suggest that this analysis is free of "spurious clusters," which represent statistical artifacts and serve as signals of poor estimation (McLachlan & Peel, 2000). This conclusion is further supported by the fact that, while cluster 7 contains less than 0.5% of the total sample, its pattern of population membership is consistent with other larger clusters. Cluster 7 contains 6% of the Bushmen, 5% of the Teita, and 1% of the Dogon. Whenever these three populations appear in meaningful numbers, they are grouped together, e.g., cluster 5 (14% of the Bushmen, 16% of the Teita, 23% of the Dogon), cluster 6 (29% of the Bushmen, 32% of the Teita, 40% of the Dogon), cluster 10 (13% of the Bushmen, 13% of the Teita, 9% of the Dogon), and cluster 14 (24% of the Bushmen, 24% of the

[13] "EEE" model: ellipsoidal, equal volume, shape, and orientation
[14] "VVV" model: ellipsoidal, varying volume, shape, and orientation

Teita, 17% of the Dogon). As Bushmen skulls are noted by Howells (1989, p.13) as being "very small" and sharing "a general African pattern" with the Teita and Dogon, it is likely that the individuals allocated to cluster 7 are not outlying observations but represent morphologies that both lie at the tail of their own distributions and occupy positions at an extreme end of the full data set. It can be assumed, with confidence, therefore, that the cluster patterns resulting from the full distribution of the 2504 individuals by cluster number is an accurate (statistical) reflection of the latent structure present within Howells' (1989) data set.

Table 9: Classification table

Cluster Number	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Number Of Allocations	30	551	375	24	573	383	12	21
Percent Representation	1.198083	22.00479	14.97604	0.958466	22.88339	15.29553	0.479233	0.838658
Cluster Number	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Number Of Allocations	54	58	42	110	63	149	33	26
Percent Representation	2.15655	2.316294	1.677316	4.392971	2.515974	5.950479	1.317891	1.038339

The next logical step in a review of these clustering patterns is to consider the strength of the clustering implied by the estimated parameters. Standard tests of significance for classification results (like chi-square statistics) are not applicable to most finite mixture analyses as, even when true memberships are known *a priori*, the number of the components of the mixture rarely correspond to the number of actual groups. It must be assumed, therefore, that the clusters produced relate to *g* external groups. The existence of these groups "in reality" must also be taken as "true" for the purpose of evaluating the overall clustering success of the chosen model and in order to determine the mixture-classification "errors" (McLachlan & Peel, 2000).

Both Ganesalingam and McLachlan (1980) and Basford and McLachlan (1985) have proposed a strategy for determining error rates for allocations made by mixture clustering that is based upon the fitted posterior probabilities of component membership. They assume that, if these posterior probabilities are near to one for the majority of the observations, the clusters are "well separated." McLachlan and Peel (2000) have explained how, from this assumption, "error rates of the outright clustering" can be obtained "by assigning the observation data point y_i to the component to which it has the greatest (estimated) posterior probability of belonging" (p.51). This method has subsequently become the basis for assessing the implied error rates in finite mixture analysis (c.f., McLachlan 1986; Basford & McLachlan, 1988; chapter 5; McLachlan & Peel, 2000, chapter 2). For an analysis using MCLUST, Fraley and Raftery (2011) have proposed that a good indication of "the uncertainty in the classification associated with conditional probabilities [...] can be obtained by subtracting the probability of the most likely group for each observation from 1" (p. 20). Ousley and Jantz (2005) have also discussed assessment criteria for evaluating the strength of craniometric-based classifications produced by linear discriminant function analysis. They (2005) have provided the users of FORDISC with a simple rule that defines meaningful posterior probability thresholds: "A high posterior probability (>0.9) indicates the individual is more similar to that group than to any other and hence is more likely to be correct than a low posterior probability (<0.7)" (np). Given the compatibility of these approaches, it is reasonable to assess the mixture solution using a variety of measures.

The clustering produced under the 16 component EEE model can be evaluated by calculating the per-case uncertainty values and summarizing these findings using the "quantiles of uncertainty" (Table 10) and the corresponding "uncertainty plots" (Figure 8 and Figure 9), after Fraley and Raftery (2011). The coordinate projection (Figure 8) shows the relative uncertainty of classification for the Howells' data set

using this 16 cluster EEE model-based classification. The size and color of the circles visually represent the relative degree of uncertainty in classification: the large filled symbols represent the 95% quantile of uncertainty and the smaller open symbols, the 75%-95% quantile. The ellipses are the multivariate analogs of the standard deviation for the mixture components. These sources of information on "uncertainty" together served as a measure of the overall quality of the MCLUST classification. The uncertainty profile plot displays the test observations in order of increasing uncertainty (Figure 9): while it appears that uncertainty is high, this is not surprising given what has been written about clinal variation in human biological data, in general, and what is known specifically for cranial metrics (Relethford, 1994; Hanihara, 1996). Overlap in craniofacial morphology should be expected when analyzing globally distributed populations and, as Fraley and Raftery (2011) have explained, "when groups intersect, uncertain classifications would be expected in the overlapping regions."

Table 10: Quantiles of uncertainty

Quantile	*0%*	*25%*	*50%*	*75%*	*100%*
Estimate	0	0.062712	0.147786	0.29319	0.651102

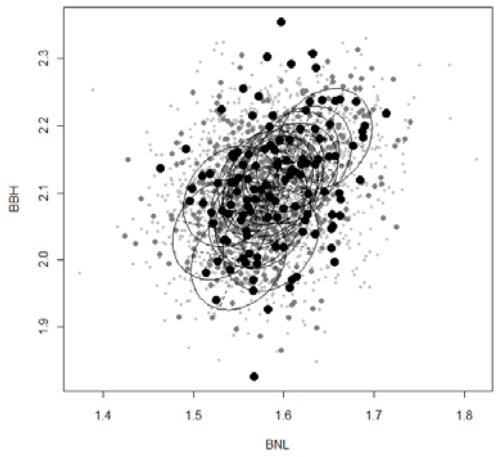


Figure 8: Coordinate projection showing uncertainty

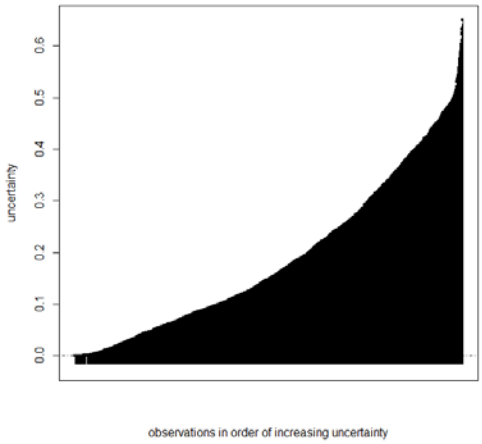


Figure 9: Uncertainty profile

In order to better contextualize this "uncertainty," it is useful to also evaluate these classifications at the level of the population. Ousley and Jantz's (2005) thresholds were integrated with the practices previously described by McLachlan and colleagues: it is only logical to estimate, either the correct classification or error rates for the mixture-based allocations, using the standards of strength or measures of classification accuracy that have been already established for craniometrics ($0.7 < \text{posterior probability} \geq 0.9$). The threshold is set, first, at the highest possible value to consider outright assignments: 38% of the total cases had posterior probabilities equal to or very nearly approaching 1. When the threshold was decreased to follow the guidelines proposed by Jantz and Ousley (2005) and posterior probabilities greater than 0.7 are taken as "correct" classifications, 76% of the total data was correctly allocated. Table11 displays the number of individuals per population with "correct" allocations using both thresholds.

It is interesting to note that the Egyptian group is among the populations with lowest outright allocation rates (29%), especially, since this obviously problematic group has seen common use in recent

discussions of the utility of FORDSIC. A considerable improvement was observed in the percent of the Egyptian crania that were correctly classified when the assessment criteria was relaxed: using the threshold of > 0.7, classification success for this series jumped to 76% - now matching the average (percent-correct) estimate obtained for the full sample. These findings nevertheless call into question the choice of this or comparable Egyptian series as appropriate and informative materials for method-tests concerned with, for example, the validity of discriminant function analysis in studies of human craniometric variation and forensic identification. The weak outright allocation rate of 29% may reflect Egypt's complex population history (c.f., Keita, 2007) and the difficulty in reconciling the relationship that this series has to its geographic origins in North Africa and its documented morphological affinity with cranial series from Europe. As Howells (1989) has written, "Egyptians are generally like Europeans" and "there seems to be no likeness with Sub-Saharans except a very low bizygomatic breadth and somewhat narrow base" (p. 13). It is worth noting that, even under the highly restricted criteria for outright allocations and despite the lowered rate of "success," there are detectable patterns in the among-population and population-to-cluster relationships that speak to this affinity. For Howells' Egyptian series, 39% of the total number of individuals were allocated to cluster 2, 38% to cluster 3, 18% to cluster 5, and 10% to cluster 12. Members of the three series making up Howells' "European" region were also found in those same clusters in reasonably large proportions: the Berg were distributed across clusters 2, 3, 5, 6 and 15, the Norse across clusters 2, 3, 5, 6 and 12, and the Zalavar across cluster numbers 2, 3, 5, 6 and 12. Moreover, Howells' sub-Saharan populations (the Dogon, Teita and Zulu) were poorly represented in clusters with high Egyptian and European contributions: cluster 3 contained less that 1% of each of the three African series, cluster 2 contained less than 10%, and cluster 5 contained less than 25%. None of the Dogon, Teita or Zulu were allocated to cluster 12. The higher value associated with the "Africans" in cluster 5 is perhaps best explained by the cluster's general

composition. As the largest of all of the 16 clusters, it includes 578 individuals and accounts alone for 23% of the full data set. Despite the fact that cluster 5 is so heavily populated, the Egyptian contribution is relatively small (only 18 individuals; 11 females and 7 males) and closely corresponds with the number of Berg (21), Norse (24), and Zalavar (28) crania that were also assigned to this group. These findings correspond well with the Mahalanobis D distances (based on discriminant scores) reported already by Howells (1989, p. 188-189): Egypt is most closely related to the Norse (3.4) and Zalavar (3.7) and only moderately related to the Teita (5.3), Zulu (5.3) and Dogon (5.8). The cluster results presented here suggest, however, a greater relationship between the Egyptian and Berg series than Howell's calculated distance of 5.4 implies.

When the total allocation rates are examined in order to identify the "best" results, Howells' Eskimo series clearly stands out as the population that has been allocated by the mixture clustering algorithm with the greatest statistical certainty: 88% of these cases were clustered "correctly"when assuming a threshold of > 0.7 and outright classifications were obtained with "success" for 64% of the 108 crania. The mixture analysis distributed the Eskimo sample over six clusters in relatively similar percentages: 10% in cluster 2, 19% in cluster 3, 16% in cluster 8, 27% in cluster 12, 26% in cluster 13, and 2% in cluster 6. This pattern varies from the trends observed for many of the other populations. There is a general tendency for (1) reasonably large subsets of the total number of individuals within a population to be assigned to a few different clusters and for (2) small scale contributions of only a few individuals to be made by each population across most, if not all, of the 16 clusters (see Table 12). What is especially remarkable about the behavior of the Eskimo series is how this group's morphology has influenced the 16 component solution identified using MCLUST. A review of the percent composition of the clusters themselves shows how the Eskimo appear as the dominant population for many of the 16 clusters: they

make up 80% of cluster 8, 44% of cluster 13, 26% of cluster 12, 8% of cluster 16, and 6% of cluster 3. These results are even more interesting when consideration is given to which of Howells' other populations are partnered with the Eskimo in forming these clusters (see Table 12). The Eskimo are consistently grouped with populations that would be considered Asian, either in terms of modern biogeographic designations or when the history of population interactions and evolutionary origins are considered (c.f., Bodmer & Cavalli-Sforza, 1976, pp.559-574; Howells, 1989). Take, first, for example, the unique make up of cluster 8. It is largely composed of Eskimo and it has the fewest overall contributors, i.e., 80% Eskimo, 10% Buriat, and 10% North Japanese. This same population triad is also repeated in clusters 12 and 13, albeit in changed percentages; these fluctuations in percent composition result from the incorporation of other related populations. Although the Eskimo still make up the largest part of cluster 12 (at 26%), the within-cluster pattern is much more varied, with an increase in the prevalence of Asian, Native American and cold adapted populations: e.g., Moriori at 15%, Arikara at 6%, Norse at 6%, Peru at 6%, North Japan at 5%, Buriat at 4% and Guam at 4%. In cluster 13, there is a shift in composition, resulting in an increased presence of Eskimo (44%), but many of the same groups are repeated, including the Buriat (16%), Moriori (11%), Norse (5%), and the North Japanese (3%).

These results for the Eskimo should not be especially surprising given what is known from Howells' (1989) analysis of this series. Although the Inugsuk Eskimo are considered North Americans by geography, Howells (1989) writes that "they are emphatic in their particularity, which is doubtless not unrepresentative for Eskimos generally" "(p. 14). It is the Eskimo's extreme morphology that increased their likelihood of correct classification: they are sufficiently different to be easily "well-separated" from the other populations using finite mixture methods (McLachlan & Peel, 2000, p. 51). This good separation was manifested in the clustering solution by the rather unique Eskimo "pattern" of isolation:

this series was distributed in large proportions across very few clusters. Even when combined with other populations, the Eskimo clearly dominated the make up of "their" clusters. The pairing of the Eskimo with other series, which were found to be either Asian in continental, historical or evolutionary origins or northern in geography, may be explained by (1) Howells' (1989, p. 14) characterization of the Eskimo as "Super-Mongoloid" in overall morphology and (2) the presence of features consistent with cold adaptation. These results offer, therefore, evidence in support of the population relationships that Howells' may have believed to be true but was unable to fully substantiate in his 1989 monograph: the classic Mahalanobis D distances separating the Eskimo from the other 27 groups were large and limited in both their range and variability (Howells, 1989, p. 188-189) and the cluster diagram based on these distances produced, in Howells' words, an "odd result." He writes that "except for American Indians, Buriats, and Eskimos, all 'Mongoloids' group nicely [...] American Indians, plus Buriats and Eskimos, join with Europeans" (Howells, 1989, p. 65).

These findings for the Egyptian and Eskimo series collectively suggest that, despite the complex population history of the Egyptians and the "emphatic particularity" of the Eskimo, the patterns in morphology and the trends in affinity identified by Howells (1989) in the summary of his "analysis by comparison of C-scores" still hold true (p. 13-16). What have changed, however, are the patterns of relationship obtained when using the different, model-based clustering, approach. Howells' cluster analysis, based on principal component analyses of the raw scores and Mahalanobis D distances, produced dendrograms that supported the "coherence"of his core regional groupings. He found for "non-core groups which lack companion populations," like the Eskimo, that they generated "varied and uninformative positions" (1989, p. 70). These finding led Howells (1989) to ultimately conclude that there are no "persuasive constant branchings among the main areas, other than shifting connections

seen in the various dendrograms" and to hesitantly state that "only in the case of the 'Mongoloid' areas (i.e., including Polynesians and Americans) may their be some support for overall connections, though if the populations were all to be blind-labeled (called Populations A,B,C, etc.) such connections might be difficult to see" (p. 83). The discussion of these two series given here serves as an example of how this project's model-based clustering method was able to tease out and exploit the craniometric signatures important to each population and obtain reasonable levels of success in classification (by mixture criteria, but using standards recommended for craniofacial studies) and identify patterns of relatedness in the absence of any knowledge of their true population designation or expected patterns of association.

In the discussion above, it has been established that the optimal 16 component model has both internal statistical validity and an acceptable level, if not an arguably high degree, of classification accuracy, when the difficulties of working with complex biological data are taken into consideration. The results inspire confidence that any relationships observed among the 16 clusters and any patterns detected in the individual component memberships reflect legitimate trends in Howells' (1989) data. They are neither artifacts of the chosen statistical procedures nor are they predicated on some set of assumptions about the true distribution of human variation. The methodological success and interpretive promise of this mixture model-based clustering cannot be uncritically dismissed as the manufactured results of user-directed, biased analyses. This project is justified, therefore, in extending its discussion of this 16 component solution to include a detailed review of its cluster patterns. The goal is to determine if these real "statistical" trends have equally real biological significance, by exploring how the 16 statistical clusters correspond to the populations represented in Howells' data set. From an anthropological perspective, this is the most critical stage in the analysis of the mixture solution.

Through these comparisons, it may be possible to uncover patterns that speak meaningfully to the questioned relationship(s) between morphology, geography and biosocial identity.

Table11: Allocation rates

Population (Howells 1989)	*posterior probability > 0.7*	*percent correctly allocated*	*posterior probability → 1*	*percent correctly allocated*	*total number of individuals*
AINU	60	70	19	22	86
ANDAMAN	48	69	21	30	70
ANYANG	27	64	14	33	42
ARIKARA	50	72	19	28	69
ATAYAL	31	66	8	17	47
AUSTRALIA	75	74	37	37	101
BERG	86	79	43	39	109
BURIAT	82	75	45	41	109
BUSHMAN	78	87	45	50	90
DOGON	80	81	45	45	99
EASTER ISLAND	65	76	33	38	86
EGYPT	84	76	32	29	111
ESKIMO	95	88	69	64	108
GUAM	45	79	20	35	57
HAINAN	66	80	26	31	83
MOKAPU	70	70	28	28	100
MORIORI	82	76	45	42	108
N JAPAN	68	78	27	31	87
NORSE	87	79	47	43	110
PERU	76	69	38	35	110
PHILIPPINES	37	74	18	36	50
S JAPAN	67	74	25	27	91
SANTA CRUZ	75	74	27	26	102
TASMANIA	66	76	43	49	87
TEITA	68	82	44	53	83
TOLAI	81	74	37	34	110
ZALAVAR	73	74	31	32	98
ZULU	79	78	53	52	101

As an initial step in evaluating the relative positions of the 16 populated clusters, the parameters values estimated for the mixture were used to generate a matrix of pair-wise scatter plots (
Figure 10) to visualize the allocation of the 2504 individuals over the 16 clusters, using the first five of the 12 base craniometric variables. A coordinate projection using the first two variables (Figure 11) and a set of random projections based on three different seeds were also produced for ease of inspection: each cluster is represented by a colored point cloud, the corresponding component means and standard deviations are indicated, and the ellipse is defined by the covariance matrices of the model.

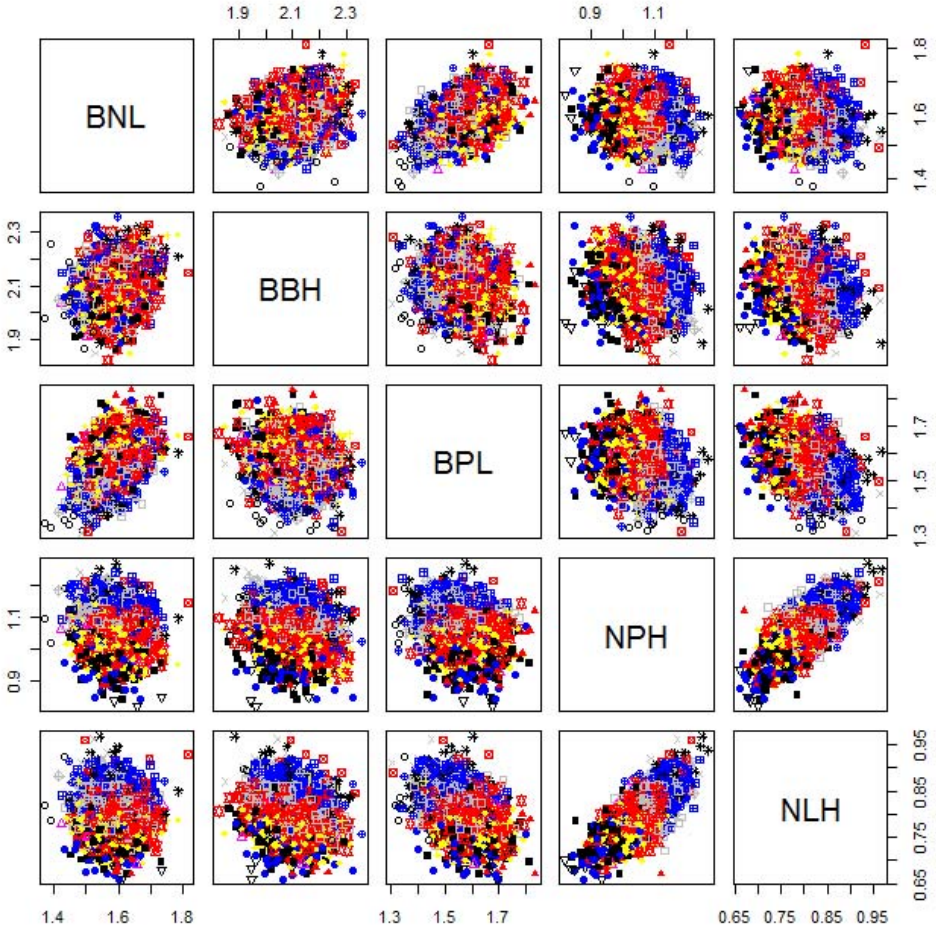


Figure 10: Pair-wise scatterplot matrix showing clusters

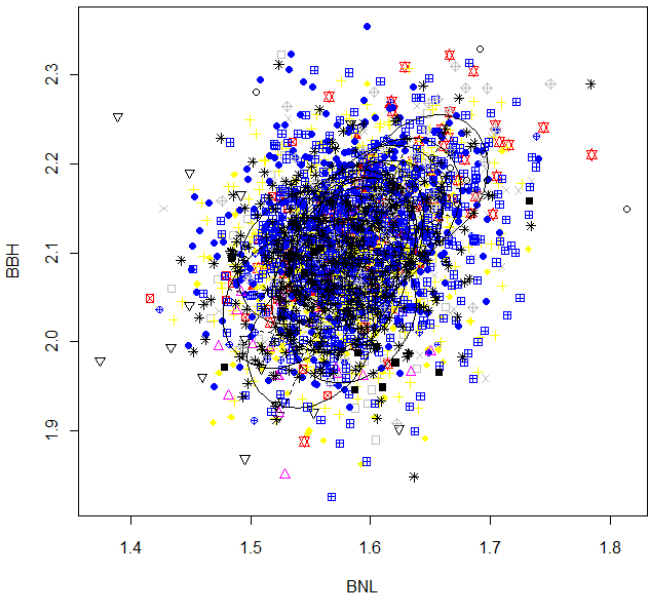


Figure 11: Coordinate projection showing clusters

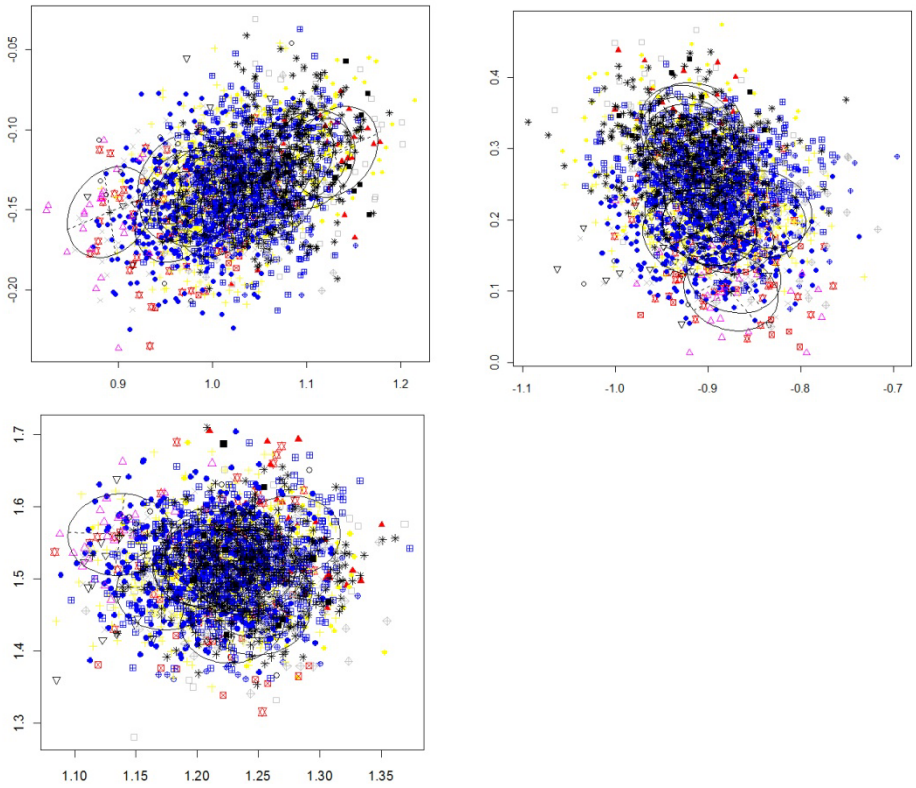


Figure 12: Random seed projections showing clusters

As can be seen from the relative positions of the 16 clusters, there is considerable overlap among the groups, which likely reflects the well-known gradations in craniofacial morphologies when global patterns of trait distributions are considered. On first inspection, these observations may appear to contradict the evidence previously obtained from the conditional (posterior) probabilities, which, suggested that these 16 clusters were reasonably well separated and the individual memberships were strongly supported (especially when assuming a threshold of > 0.7). These graphs provide, however, only overviews of the complex cluster relationships and, for this reason, should not be used in isolation to evaluate the strength of the data's internal structure or to identify meaningful separations. In order to assess the specific relationships existing within and between the clusters, it is necessary to investigate the cluster compositions in greater detail and using the language of Howells' populations. To this end, a numeric breakdown of MCLUST's 16 component solution was produced: Table 12 and Table 13 report the distributions of the cluster memberships for each of Howells' 26 female and 28 male populations respectively. These raw numbers are both interesting and obviously important for quantifying the results of the best MCLUST analysis (e.g., providing the [total] numbers of allocations by population, sex, or cluster). They are, however, less helpful when the goal is easy pattern recognition. This project chose, therefore, to further optimize the interpretive value of this numerical data by producing simple graphical representations using relative frequency distributions of the populations and cluster numbers. Two sets of pie-charts were created: one series displays the within-cluster partitions (i.e., giving the percent division of each cluster into its constituent populations) and the other represents the within-population contribution (e.g., showing what percent of each population is assigned to any of the 1-16 clusters). In the discussion that follows, a subset of these pie-charts are presented in detail.

Table 12: Cluster memberships by population (females)

Population	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*Total*
Females	*10*	*239*	*173*	*8*	*251*	*190*	*7*	*13*	*31*	*24*	*21*	*47*	*31*	*78*	*13*	*20*	*1156*
AINU	0	9	4	0	15	6	0	0	1	1	0	0	2	0	0	0	38
ANDAMAN	0	2	0	0	17	10	0	0	2	0	0	0	0	3	1	0	35
ARIKARA	0	9	8	0	3	1	0	0	0	0	2	2	0	0	1	1	27
ATAYAL	0	5	1	0	7	1	0	0	1	0	0	0	0	3	0	0	18
AUSTRALIA	1	7	2	1	21	12	0	0	0	1	0	0	0	4	0	0	49
BERG	0	11	12	0	10	10	0	0	1	3	1	0	0	1	4	0	53
BURIAT	0	14	10	5	4	0	0	2	7	0	2	1	6	1	0	2	54
BUSHMAN	0	3	1	0	7	14	2	0	0	5	0	0	0	16	1	0	49
DOGON	0	4	0	0	12	20	1	0	1	4	1	0	0	9	0	0	52
EASTER ISLAND	0	3	1	0	15	10	0	0	2	1	3	0	0	2	0	0	37
EGYPT	0	21	12	0	11	1	0	0	0	0	0	6	1	1	0	0	53
ESKIMO	0	6	12	0	0	0	0	10	1	0	0	10	14	0	0	2	55
GUAM	0	13	2	0	6	2	0	0	1	0	0	3	0	0	0	0	27
HAINAN	0	7	10	0	8	7	0	0	0	0	2	2	0	1	1	0	38
MOKAPU	0	4	12	0	15	2	0	0	9	0	2	2	2	1	0	0	49
MORIORI	0	19	17	0	1	2	0	0	2	0	0	7	3	0	0	0	51
N JAPAN	0	10	10	1	4	3	0	1	0	0	1	2	0	0	0	0	32
NORSE	0	12	16	0	13	2	0	0	0	0	2	4	2	1	2	1	55
PERU	0	19	15	1	3	2	0	0	0	0	1	3	1	0	1	9	55
S JAPAN	0	17	6	0	15	1	0	0	0	0	0	0	0	0	0	2	41
SANTA CRUZ	0	18	15	0	8	3	0	0	1	0	2	0	0	0	1	3	51
TASMANIA	8	1	0	0	6	20	0	0	1	1	0	0	0	5	0	0	42
TEITA	0	3	0	0	4	18	4	0	0	6	0	0	0	15	0	0	50
TOLAI	0	9	0	0	19	18	0	0	0	0	0	0	0	8	0	0	54
ZALAVAR	1	10	6	0	17	3	0	0	1	0	0	5	0	1	1	0	45
ZULU	0	3	1	0	10	22	0	0	0	2	2	0	0	6	0	0	46

Table 13: Cluster memberships by population (males)

Population	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*Total*
Males	*20*	*312*	*202*	*16*	*322*	*193*	*5*	*8*	*23*	*34*	*21*	*63*	*32*	*71*	*20*	*6*	*1348*
AINU	0	9	3	0	21	10	0	0	2	0	0	2	0	1	0	0	48
ANDAMAN	0	8	3	0	13	4	0	0	0	1	1	1	0	4	0	0	35
ANYANG	0	6	7	0	14	10	0	0	2	0	2	0	0	1	0	0	42
ARIKARA	0	15	13	1	4	2	0	0	0	0	2	5	0	0	0	0	42
ATAYAL	0	6	3	0	9	8	0	0	0	0	0	0	0	2	1	0	29
AUSTRALIA	2	6	1	0	23	13	0	0	0	1	0	1	0	5	0	0	52
BERG	0	11	16	0	11	8	1	0	0	0	1	1	1	1	5	0	56
BURIAT	0	13	11	13	2	3	0	0	2	0	1	3	4	1	2	0	55
BUSHMAN	5	2	0	0	6	12	3	0	0	7	0	0	0	6	0	0	41
DOGON	1	2	0	0	11	19	0	0	0	5	1	0	0	8	0	0	47
EASTER ISLAND	0	7	1	0	28	7	0	0	0	0	2	0	0	3	1	0	49
EGYPT	0	18	26	0	7	2	0	0	1	0	0	3	0	0	1	0	58
ESKIMO	0	4	9	0	0	0	0	7	0	0	0	19	14	0	0	0	53
GUAM	0	10	6	0	8	3	0	0	0	0	1	1	1	0	0	0	30
HAINAN	0	14	7	0	14	6	0	0	1	0	0	1	0	2	0	0	45
MOKAPU	0	21	9	0	14	2	0	0	2	0	2	1	0	0	0	0	51
MORIORI	0	19	17	2	3	0	0	0	1	0	0	10	4	0	0	1	57
N JAPAN	0	23	10	0	9	2	0	1	1	1	0	4	2	0	1	1	55
NORSE	0	17	13	0	11	5	0	0	0	1	0	2	1	2	3	0	55
PERU	0	20	12	0	7	3	0	0	5	0	1	4	1	2	0	0	55
PHILIPPINES	0	9	0	0	22	11	0	0	1	1	1	0	0	2	2	1	50
S JAPAN	0	19	11	0	12	4	0	0	3	0	0	1	0	0	0	0	50
SANTA CRUZ	0	12	13	0	13	0	0	0	1	0	5	2	1	1	0	3	51
TASMANIA	5	3	1	0	9	9	1	0	0	5	0	0	0	12	0	0	45
TEITA	0	3	0	0	9	9	0	0	0	5	1	0	0	5	1	0	33
TOLAI	7	8	1	0	16	18	0	0	0	3	0	0	0	3	0	0	56
ZALAVAR	0	20	9	0	11	4	0	0	0	0	0	2	3	2	2	0	53
ZULU	0	7	0	0	15	19	0	0	1	4	0	0	0	8	1	0	55

The first of these two series allows for easy review of the percent composition of each of the 16 clusters by population. Cluster 1 (Figure 13) is largely composed of Tasmanians (43%), Tolai (23%), Bushmen (17%) and Australians (10%). The obvious geographic similarities between the Australo-Melanesian groups suggest strong consistency in the character of this cluster. While there is a large percentage of Bushmen, who are South African in origin, the inclusion of these individuals does not detract from the cohesiveness of this cluster. Howells (1982, p. 14), for example, has already identified morphological similarities between the Australoid and African series (e.g., broad upper face). The fact, however, that other African populations are conspicuously absent, with the exception of a single Dogon representing the more distant region of West Africa, suggests that the pattern observed in cluster 1 may speak more to the nature of the Bushmen sample than to a strong African trend. The Bushmen represent a "local series" which was assembled by Howells (1989, p. 113-114) over a large geographic range. For this reason, this series may contain greater morphological variability and, in turn, greater apparent affinity with other, similar or proximate, but not necessarily African, groups.

Cluster 2 (Figure 14) is the second most populated of the 16 clusters: it contains 551 individuals, accounts for the allocation of 22% of the total data set, and is the most diverse in its memberships. As a result, many populations are represented in cluster 2, but all in equally small percentages. While this variable pattern makes it difficult to parse this component in a meaningful way, cluster 2 is not without some information value. This cluster will prove useful in later discussions of the second series of pie charts, from which relationships are drawn using the patterns of shared component memberships. It will be important to note which individuals appear in cluster 2 and to what extent, as the inclusion or exclusion of a particular population with respect to given cluster will alter its observed (possibly unique) membership signature and, in turn, influence its apparent similarity to other groups.

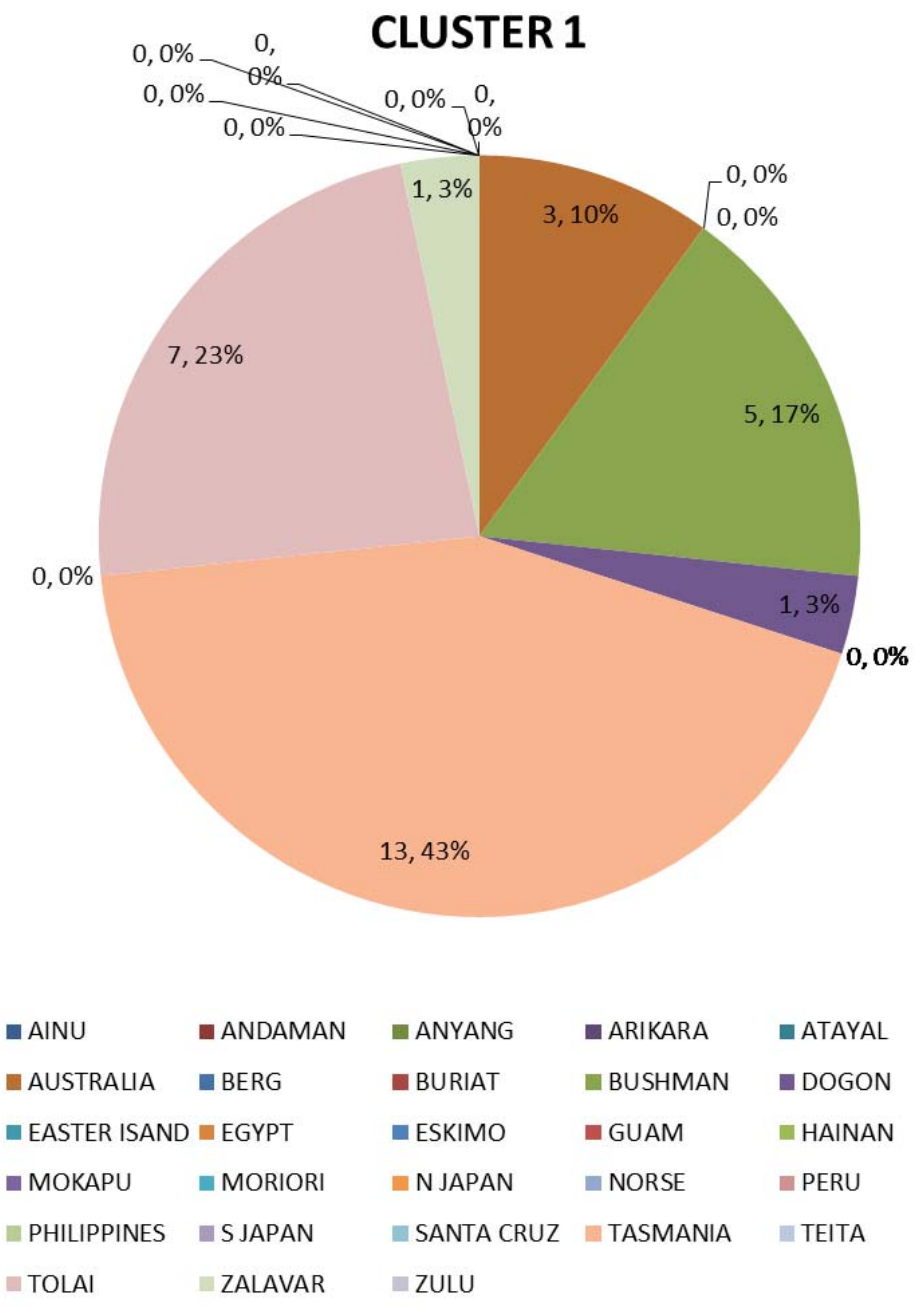


Figure 13: Cluster 1 composition

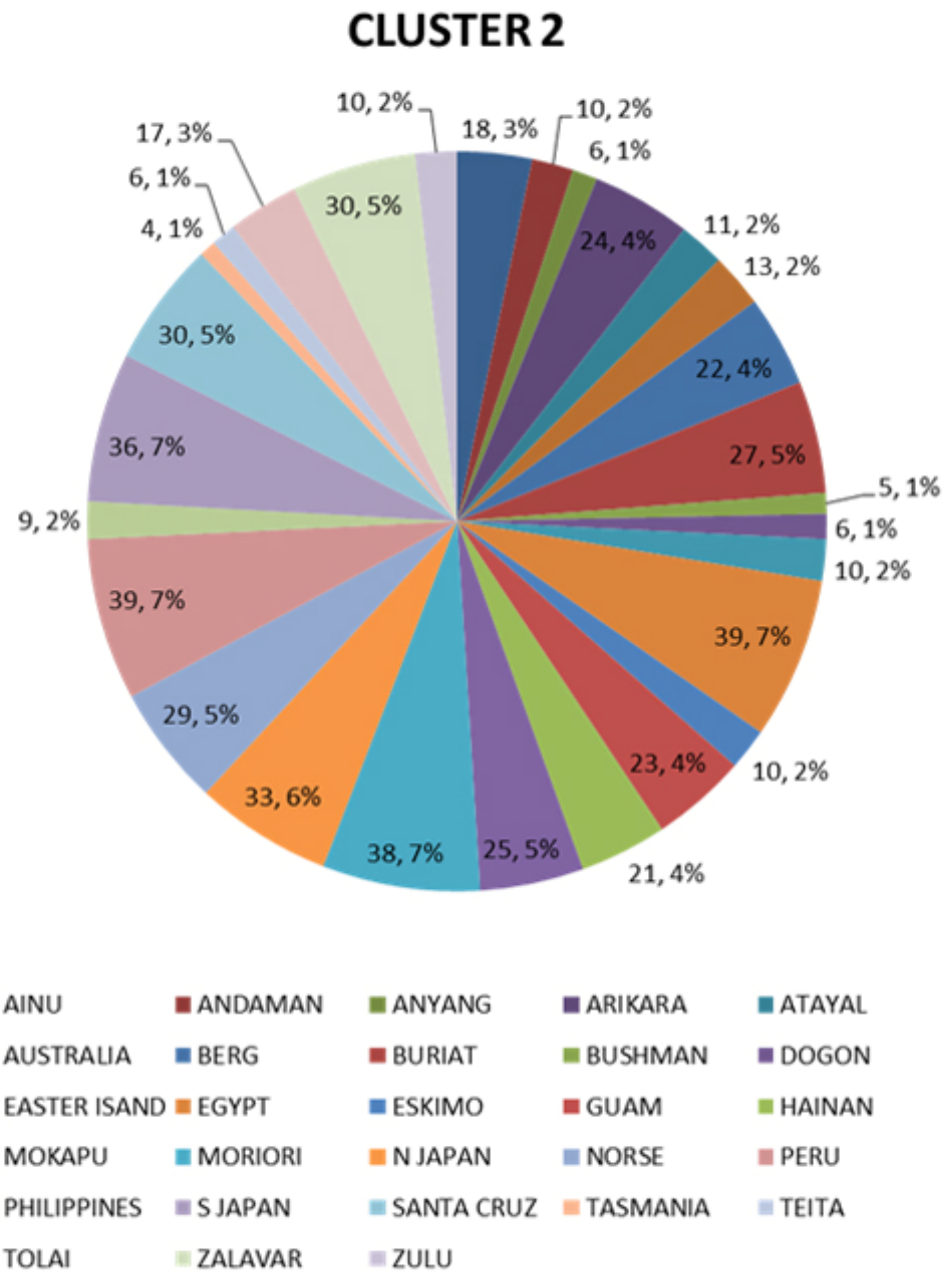


Figure 14: Cluster 2 composition

Cluster 3 (Figure 15), although smaller than cluster 2, is also considerably mixed. What is interesting here is that, upon closer scrutiny, a patterned divide is detectable. While the European (including Egypt), Native American and East Asian populations are all well represented, there is an obvious absence of any African, Australo-Melanesian, and Pacific Island groups. Cluster 3 may be interpreted, therefore, as capturing one part of a global north-south separation: this may correspond to the more traditional expectations for morphological clines and the difference in data structure when considering groups outside of Africa. Cluster 4 (Figure 16) is almost exclusively populated by the Buriat (75%), with only limited (singular) contributions made by the Native American and Asian series. The dominance of the Buriat group within what may be considered its "own" cluster attests to this Siberian population's "extreme and isolated" morphology; these statistical findings corroborate Howells' (1989, p.15, 118) claim that the Buriat represents an outlying "local" series. Cluster 5 (Figure 17), not unlike cluster 2, is large (including 23% of all cases) and contains, in general, a mixture of most of the populations. It is distinguishable from cluster 2, however, in that the Australians and Easter Islanders stand out as prominent groups within this overall diversity (each contributing 8% of the total cluster size) and the Eskimo are completely absent. Cluster 6 (Figure 18) shares many populations with cluster 1 (Figure 13), although it better captures the relationship between the African and Australo-Melanesian groups, which when combined make up 50% of this cluster. The key individual contributions to cluster 6 include the Zulu (11%), Dogon (10%), Tolai (9%), and Tasmanians (8%), as well as the Australians, Bushmen, and Teita in equal proportions (7% each). Cluster 7 (Figure 19) accounts for a particular range of African diversity: it is formed, almost exclusively, by a mix of the San Bushmen (42%) and the East African Teita (33%).

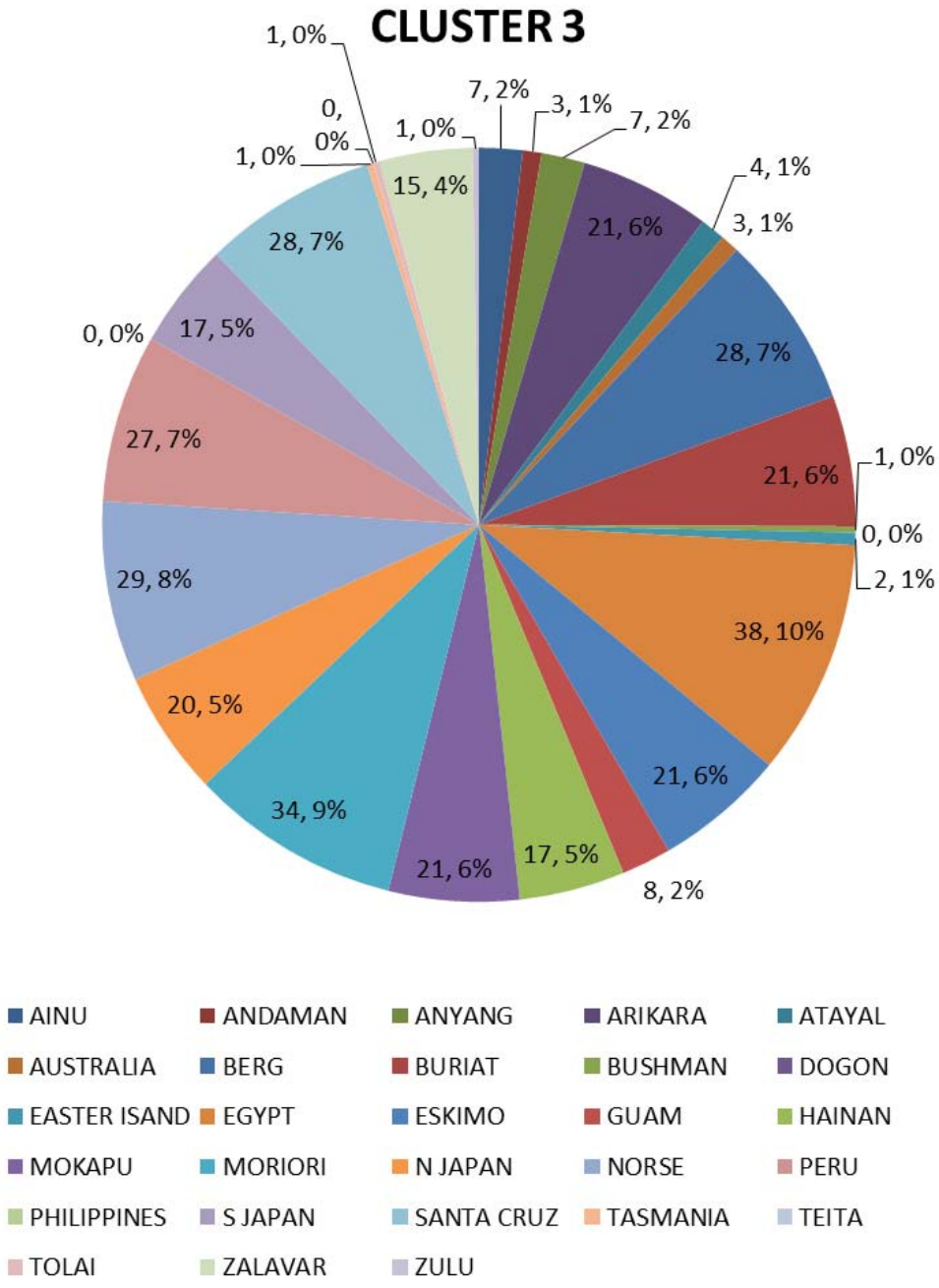


Figure 15: Cluster 3 composition

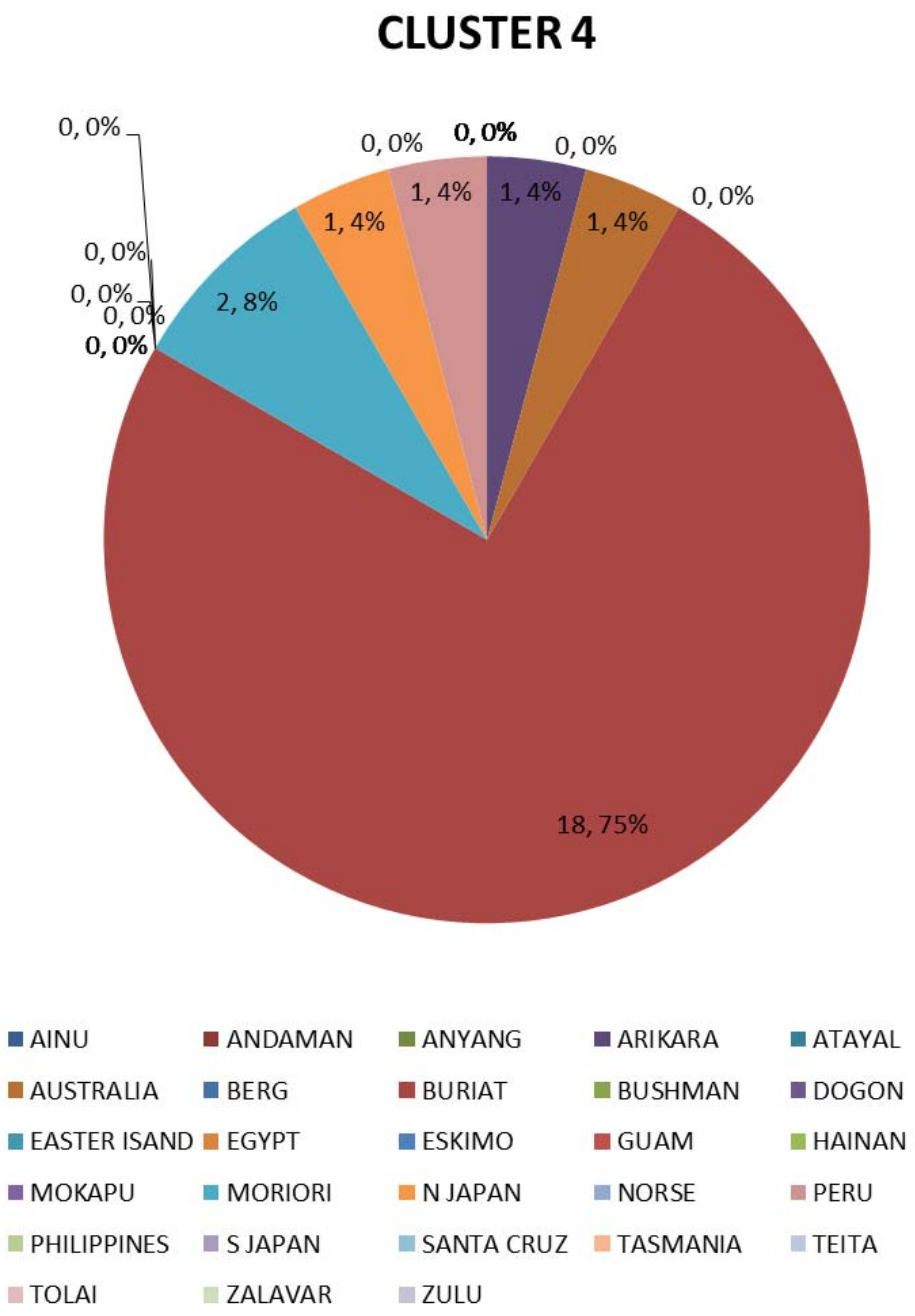


Figure 16: Cluster 4 compostion

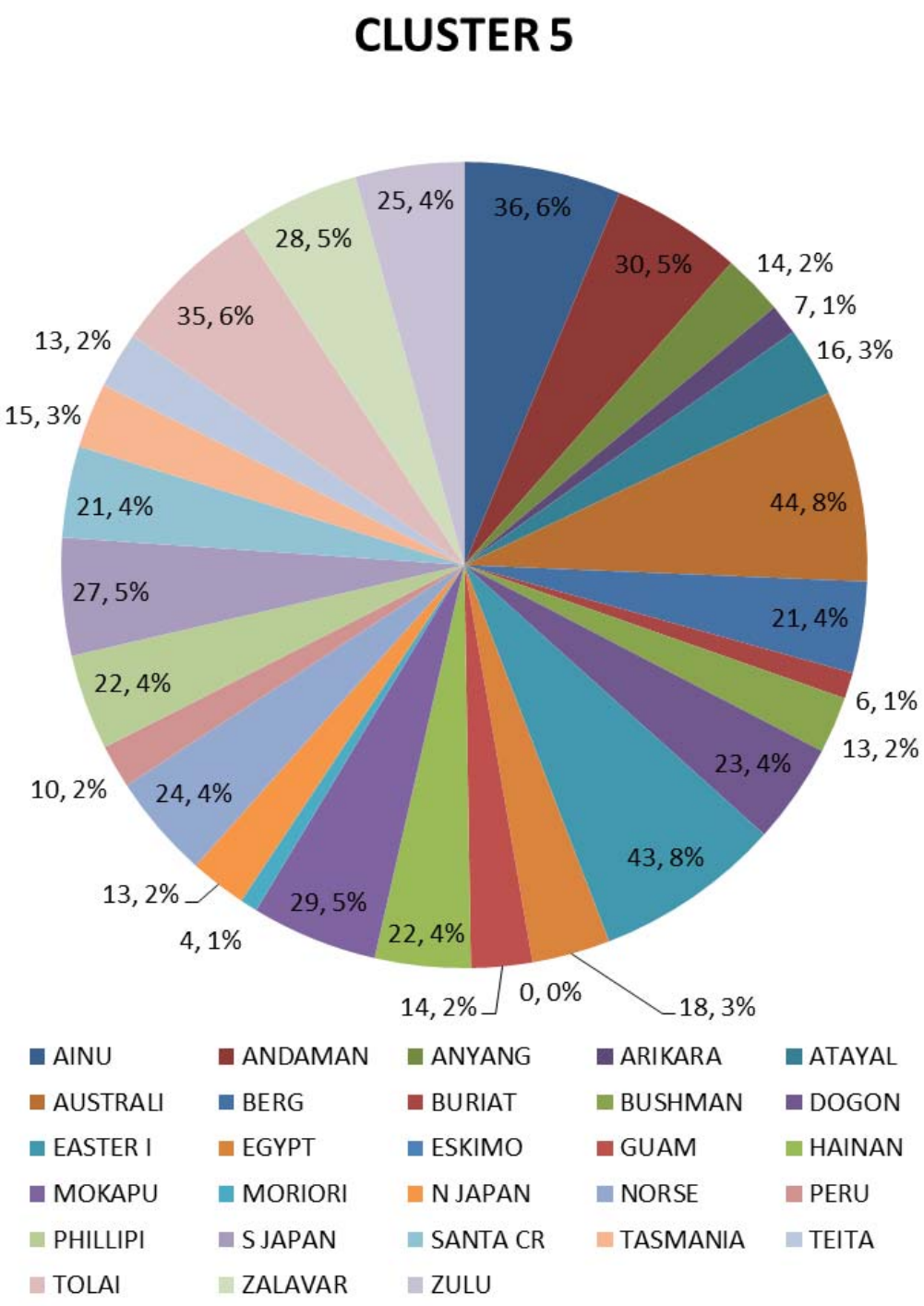


Figure 17: Cluster 5 composition

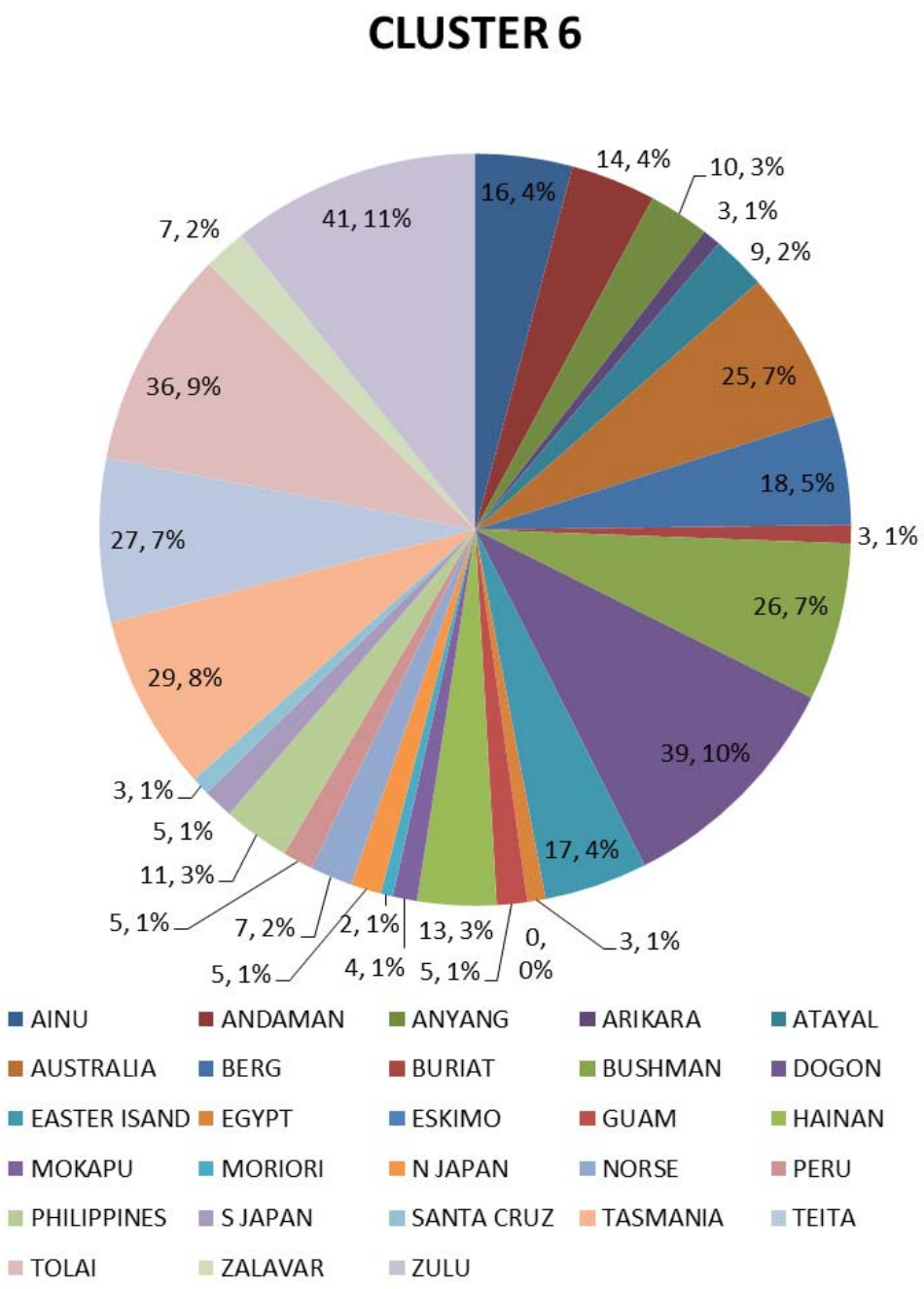


Figure 18: Cluster 6 composition

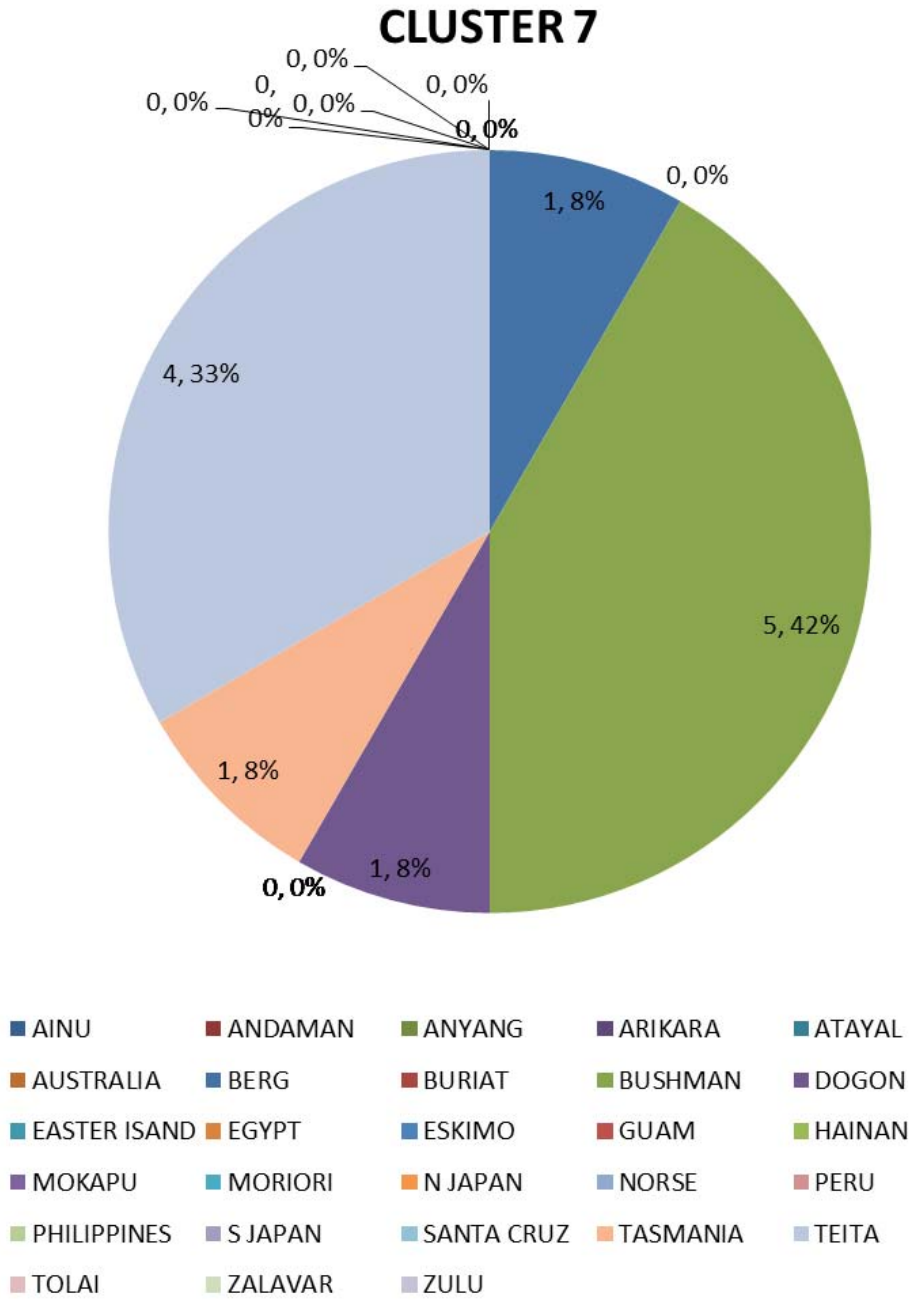


Figure 19: Cluster 7 composition

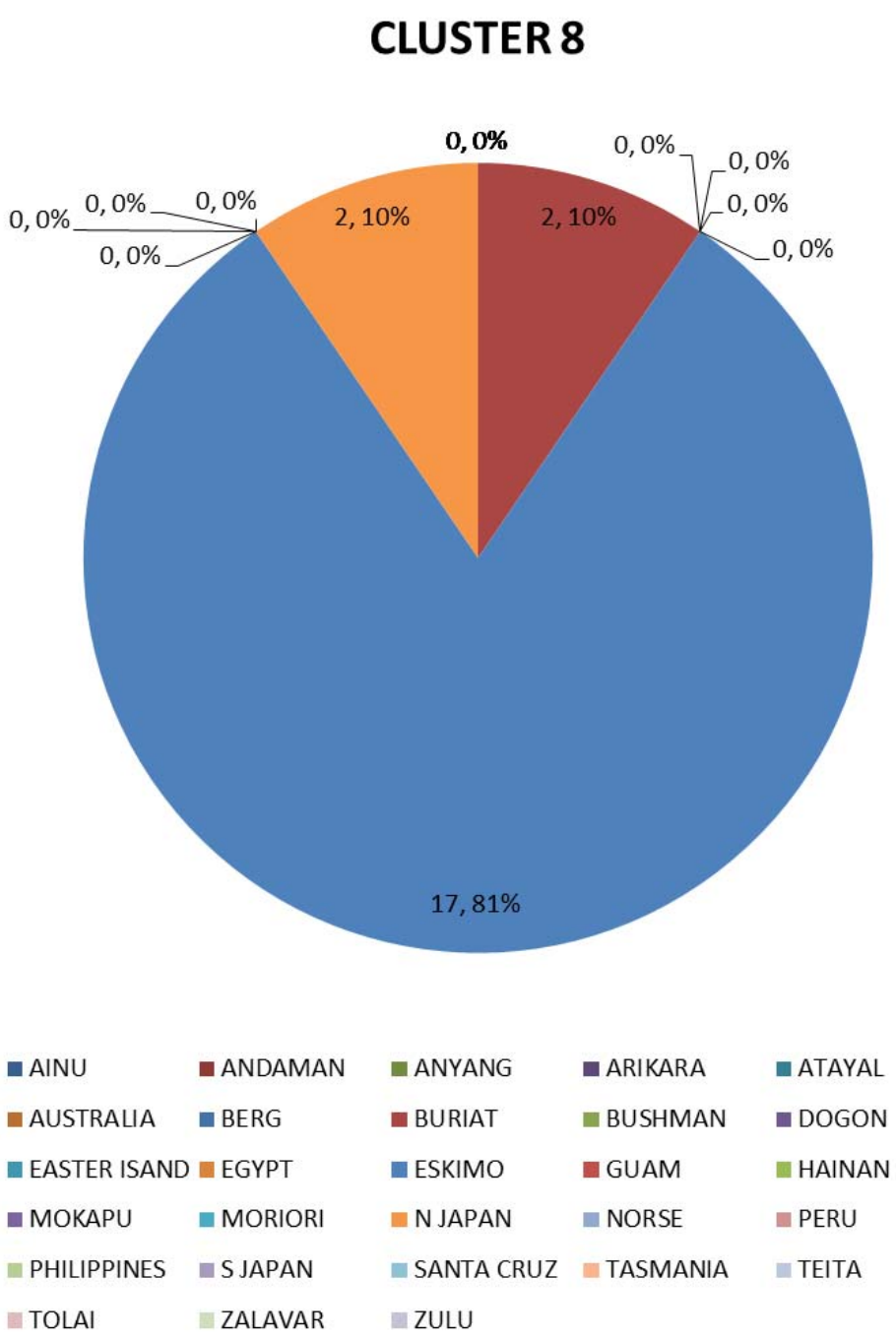


Figure 20: Cluster 8 composition

Cluster 8 (Figure 20) is clearly dominated by the highly atypical Eskimo group, which represents alone a remarkable 80% of the cluster. The remaining 20% is distributed between the North Japanese and Buriat series. This pattern is not entirely unexpected given prior knowledge of their evolutionary relationships and shared (cold) climate adaptations (Howells, 1989, p. 14-15). Cluster 9 (Figure 21) is composed of predominantly Asians or populations of Asian descent, including Howells' Mokapu (20%), Peru (9%), Buriat (7%), South Japanese (6%), Moriori (6%), and Ainu (6%). Cluster 11 (Figure 23) is dominated by both Native American and Pacific Island populations: Santa Cruz (17%), Easter Island (12%), Arikara (10%), and Mokapu (10%). As these groups share common Asian origins, cluster 11 can be taken as a focused subset of the variation that was seen already in cluster 3 (Figure 15) and cluster 9 (Figure 21). The partitioning of these series over several clusters corresponds well with Howells' (1989, p. 14) characterization of the discernable but "not strikingly uniform" craniofacial morphologies of the American Indians. It also highlights the concomitant difficulty in characterizing the interrelationships observed between a composite, geographically diverse and morphologically variable super-group (here, the "Native American") and interpreting the ways in which its sub-groups relate individually to other populations or regions (like, Europe and the Far East). If cluster 7 (Figure 19) represents a specific range of African groups, then cluster 10 (Figure 22) clearly captures the maximum sub-Saharan diversity. The Bushmen (21%), Teita (195), Dogon (16%), and Zulu (10%) make up this "African" cluster in reasonably large and comparable percents. The inclusion of the Tasmanians (10%) in equal numbers again reiterates the morphological affinity between Oceanic aboriginal groups and southern African tribes that was already recognized by Howells (1989, p. 13-14). Cluster 12 (Figure 24) is a mixture of several groups and is close to the average cluster size (110 individuals) for the 16 component solution. Despite its overall diversity, both the Moriori and Eskimo are present in notable proportions.

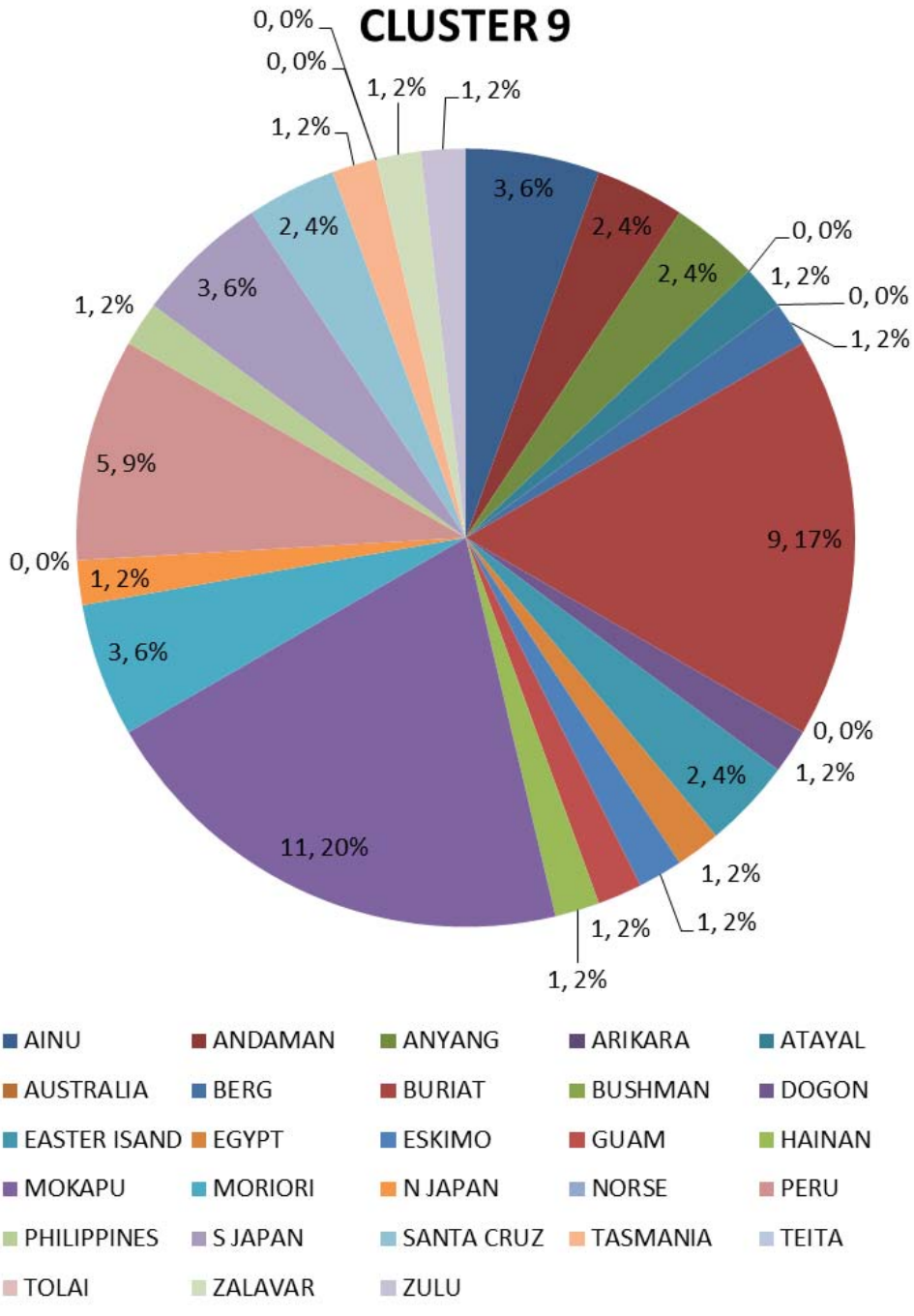


Figure 21: Cluster 9 compostion

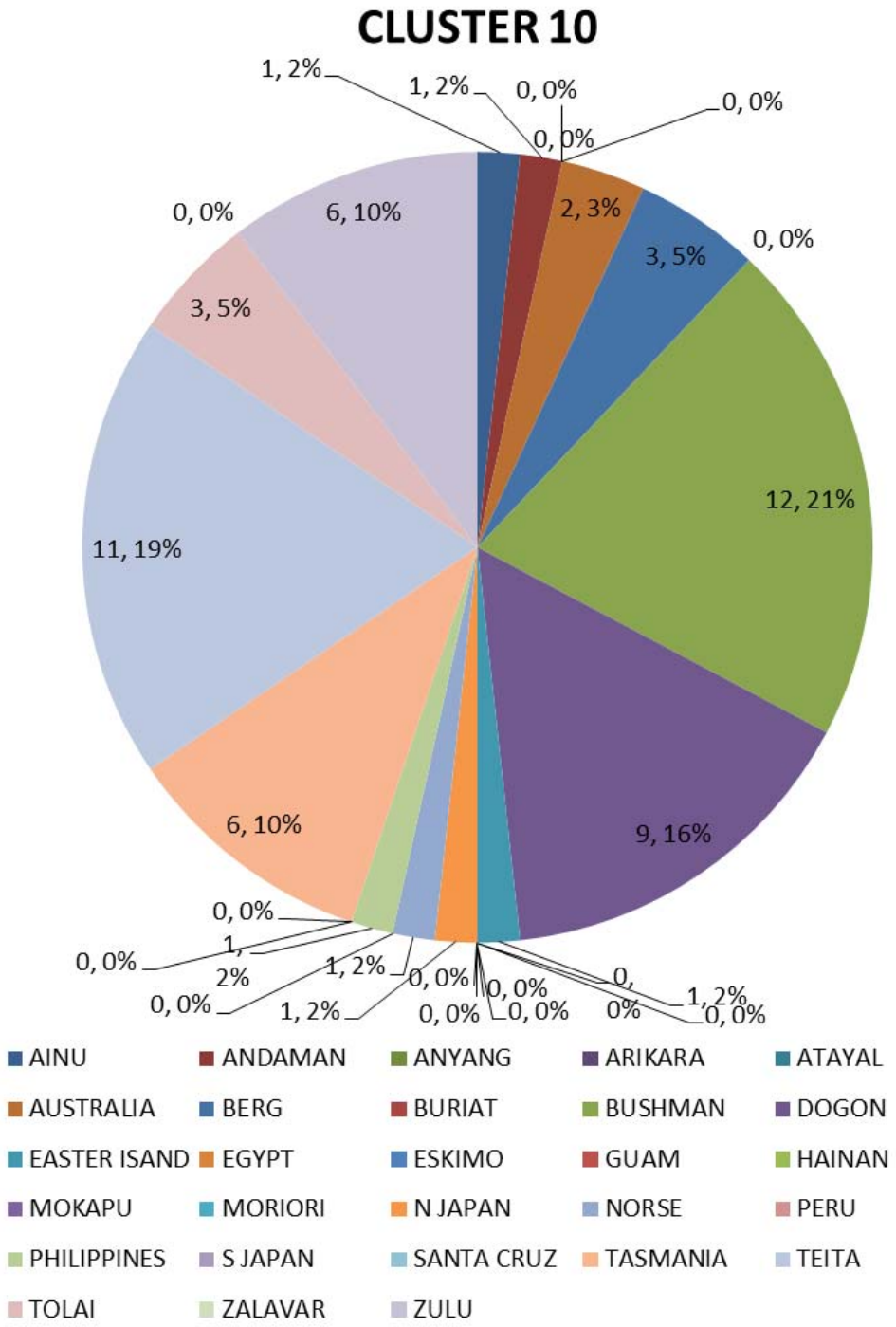


Figure 22: Cluster 10 composition

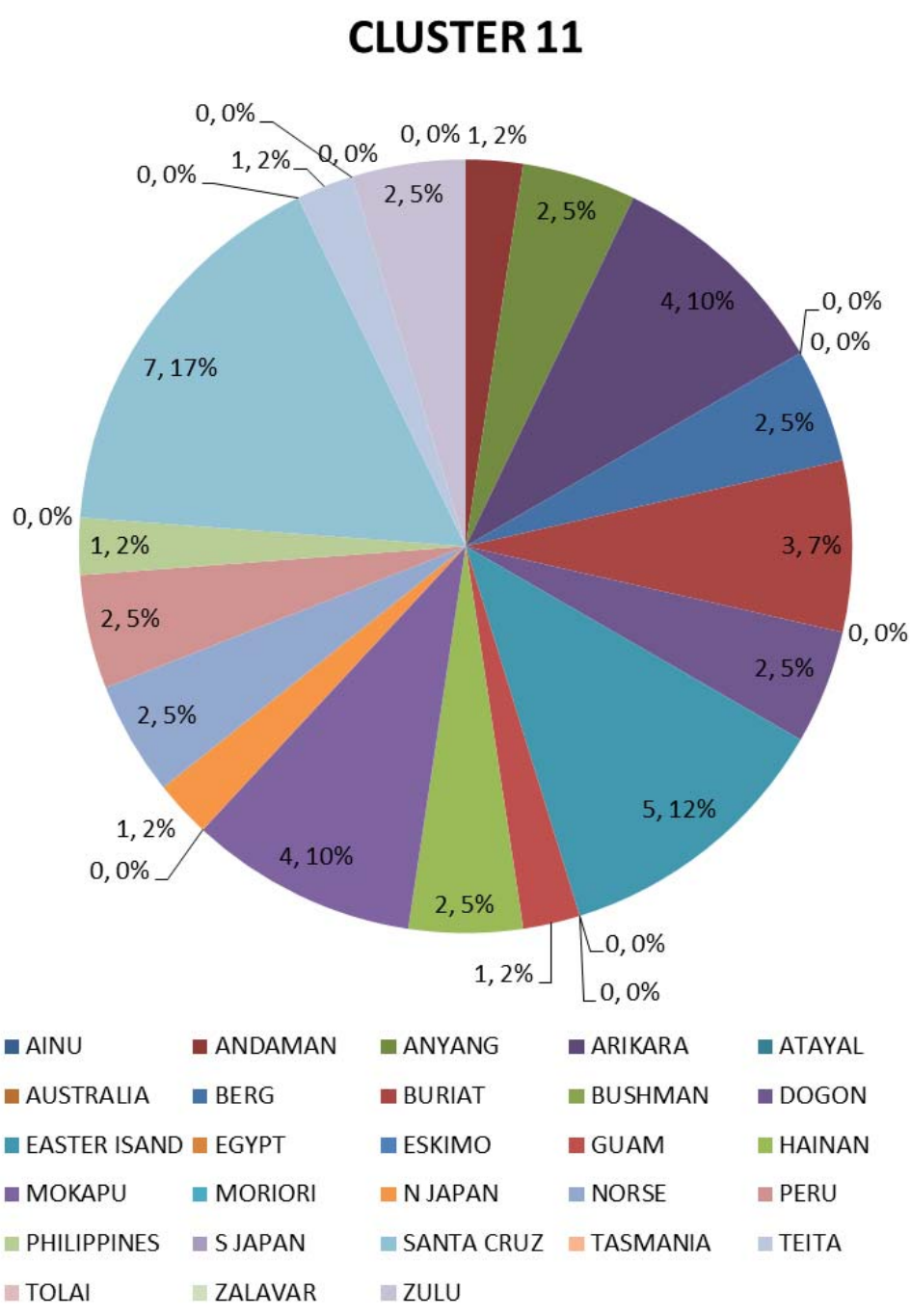


Figure 23: Cluster 11 composition

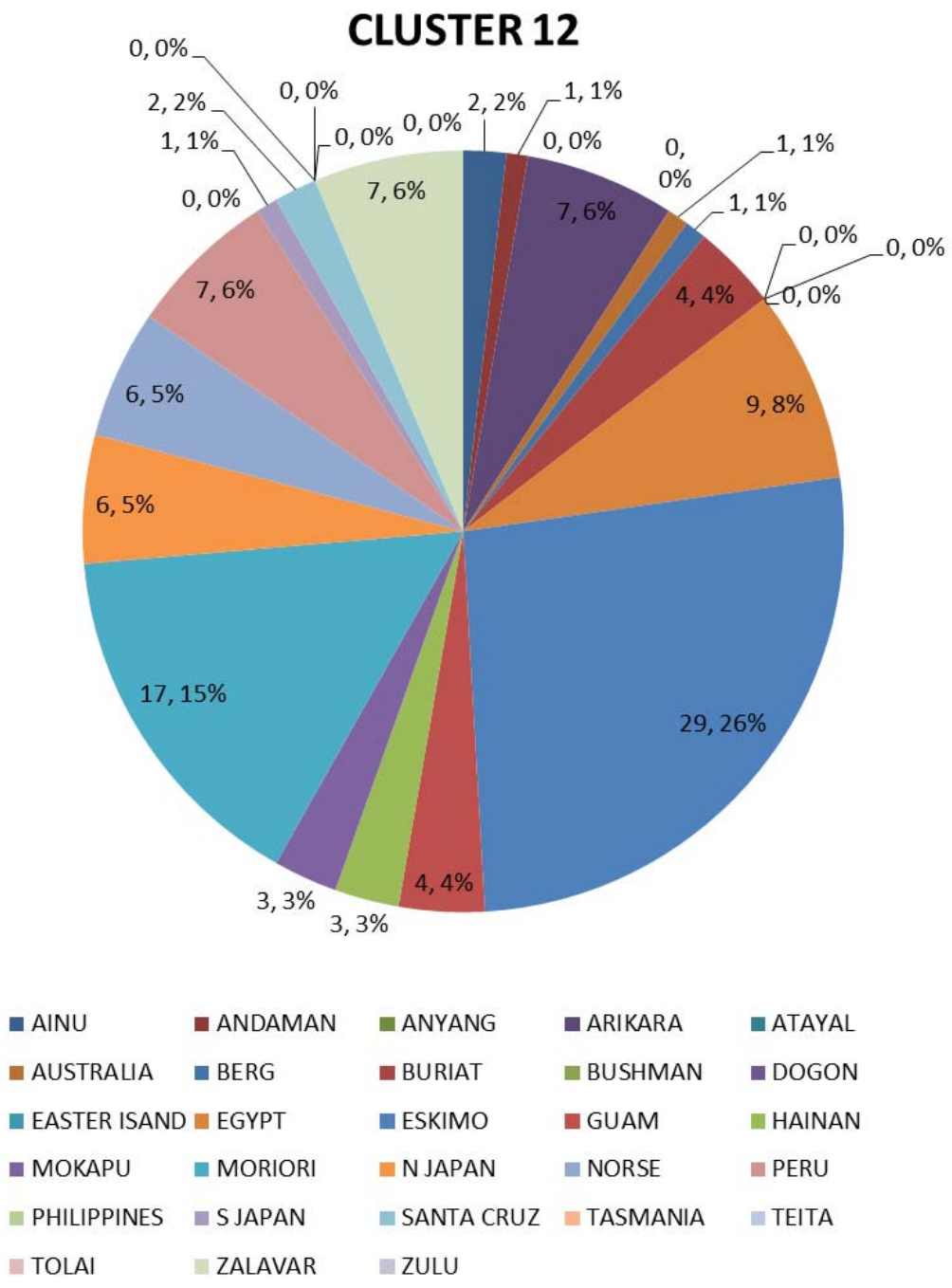


Figure 24: Cluster 12 composition

This same pairing of Eskimo (44%) with Moriori (11%) is also repeated in cluster 13 (Figure 25), but with the inclusion of a significant Buriat component (16%). This configuration, which includes relatively unique and remote groups, is not only suggestive of their shared Asiatic heritage but is also indicative of a trend in craniofacial morphology, e.g., the frontal, facial flatness that is common among these three groups (Howells, 1989, p. 14-15). The strongly "African" pattern of cluster 10 (Figure 22) reemerges here in cluster 14 (Figure 26), creating again a cluster composed of the Bushman (15%), Teita (13%), Dogon (11%), Tasmanian (11%), and Zulu (9%) populations in very similar proportions. The demographic information available for each of the individuals making up these groups was carefully reviewed, and, while sex is only known with certainty for the Zulu, sex-based trends were not detected. There is, therefore, no immediate explanation that can simply account for the split distribution of these groups across clusters 10 (Figure 22) and 14 (Figure 26). Cluster 15 (Figure 27) represents the only truly European component: it includes the Berg of Austria (27%), the Norse of Scandinavia (15%) and the Zalavar of Hungary (9%). While this cluster contains all three members of Howell's regional "Europe" group, it is interesting to note that it did not capture the morphologically "similar" Egyptians from North Africa, as it might be expected since the Egyptians are appended to the Europeans groups in most of the other clustering cases. The relatively limited presence of these European populations in large numbers in other clusters (nowhere else do these three groups appear together among the top five contributors), implies some notable inconsistencies in morphology, which, at least for these three series, undermines any definition of a common "type" (i.e., compare the brachycranial Bergs to the long-skulled Norse). Their frequent distribution across many clusters, in relatively small amounts, may also serve as a signal of inter-regional "overlap." This should not be surprising given their geographic intermediacy: west of Asia, North of Africa and East of the Americans. Cluster 16 (Figure 28) appears to capture a segment of

the southern variation within the New World as it predominantly includes Howells' American Indian series (1989, p. 105, 106) from Santa Cruz (23%), California, and the Yauyos region of Peru (35%).

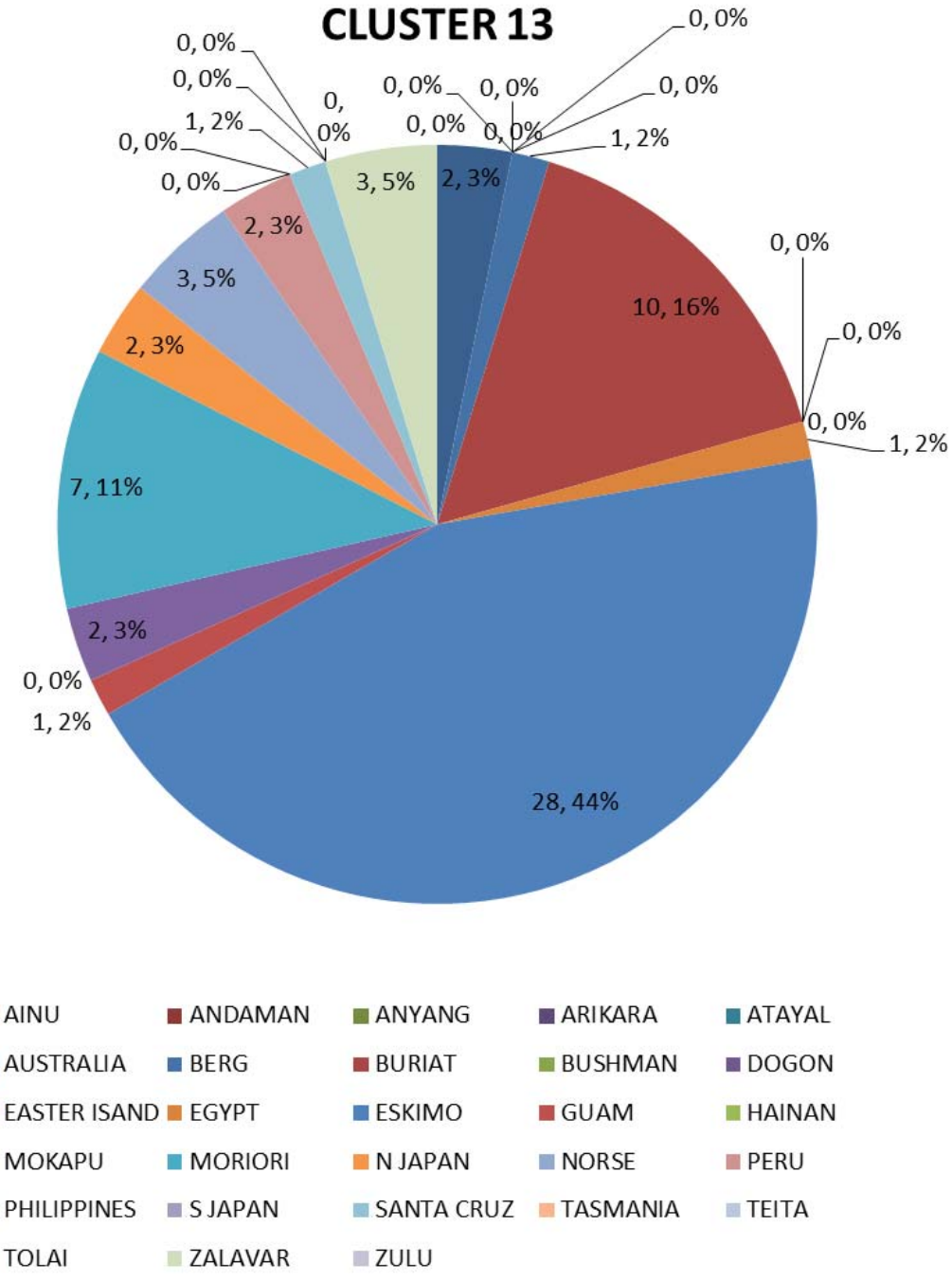


Figure 25: Cluster 13 composition

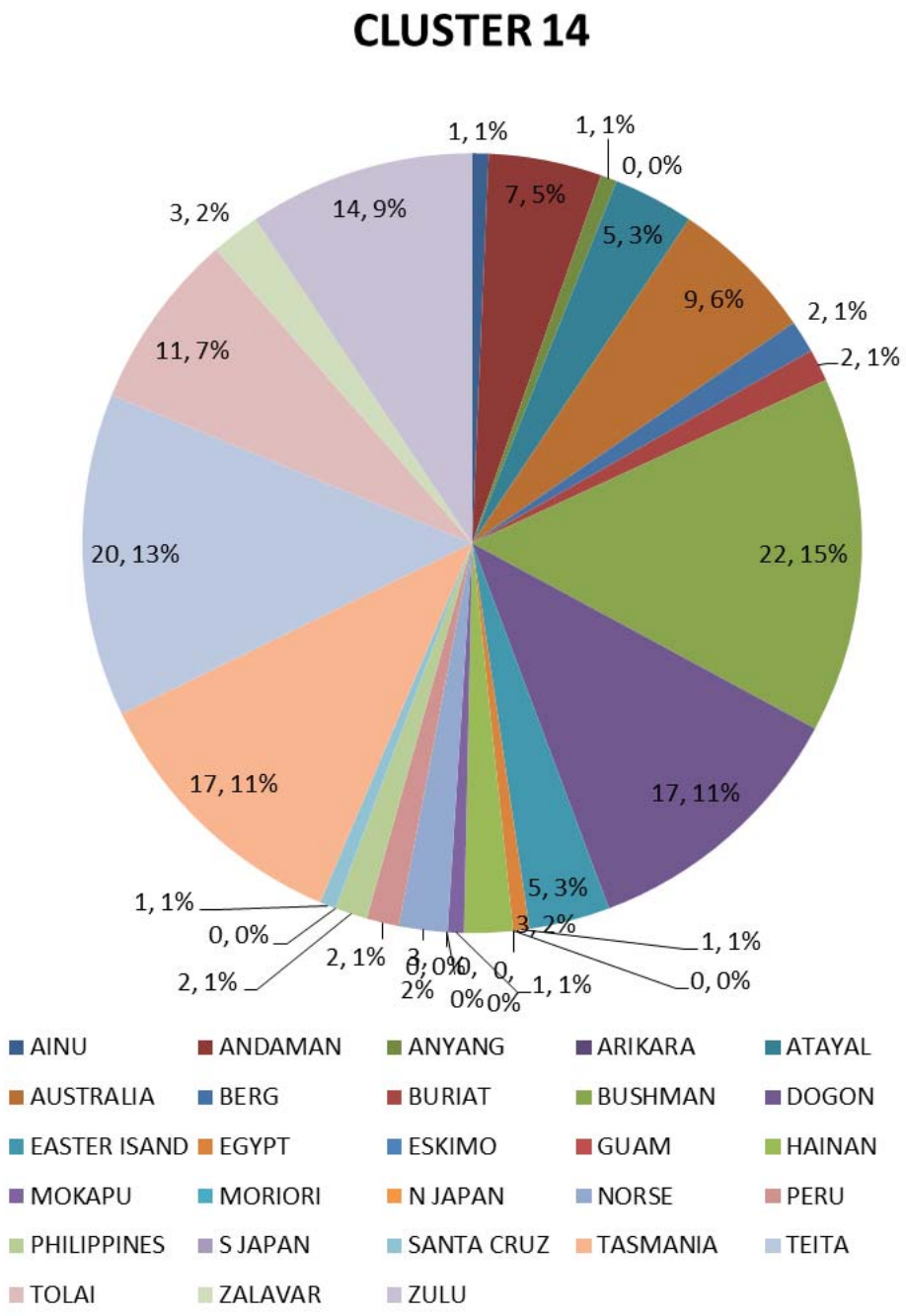


Figure 26: Cluster 14 composition

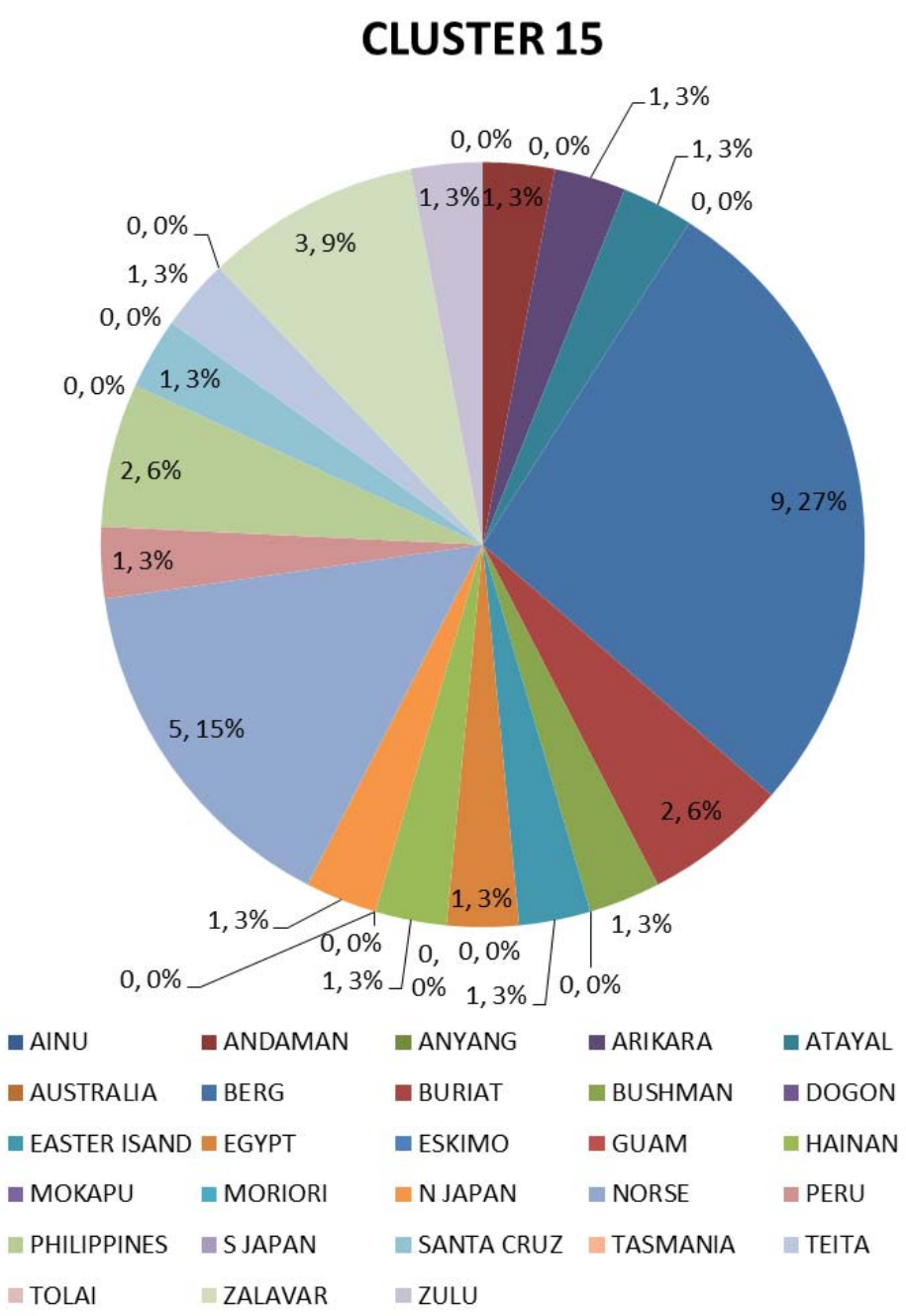


Figure 27: Cluster 15 composition

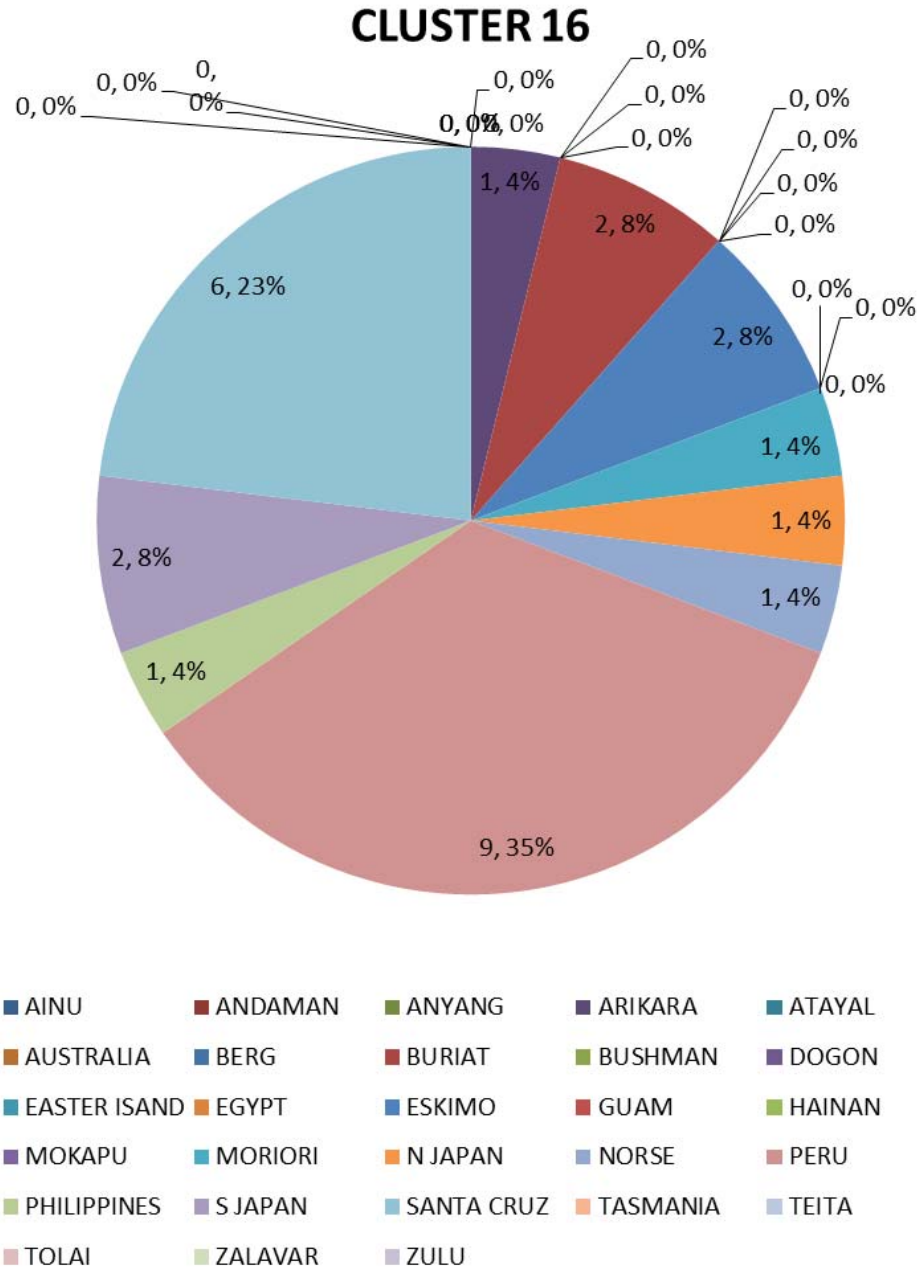


Figure 28: Cluster 16 composition

Since this MCLUST solution considered the full (scaled) Howell's data set, there were 28 potential male and 26 potential female contributing populations, represented by a total of 2504 individuals. Rarely, however, were all of these possible groups included within any one component. Cluster 2 (Figure 14) is clearly the most "mixed" example; clusters 3, 5, and 12 are complex but to a lesser degree and with more obvious "key" contributors (Figure 15, Figure 17, Figure 24). Typically, the clusters were formed, instead, by a few dominant populations. For all of these cases, some geographic, historical, or evolutionary evidence attesting to the intra-cluster relationships can be found in the literature concerned with the morphogenetic links between local, regional and world-wide groups. The fact that population-based trends were clearly identifiable in all of the components, except cluster 2, can be taken as evidence of strong latent structure. Most importantly, these obvious trends suggest that the clusterings given by this solution have anthropological significance.

The second series of pie-charts further clarifies the clustering patterns obtained from the optimal MCLUST model using Howells' (1989) full data set. Here, they are used to discuss "classification" in terms of each population. These graphs display the percent of each population contributing to the makeup of one, some or all of the 16 clusters. When considered collectively, these charts enable visual comparison in the *patterns* of allocations observed among any number or combination of Howells 28 groups. Six distinct groups are clearly identifiable. Their patterns of component membership are most similar. In all cases, inter-population correspondence was found to be the best when the pie-chart combinations were restricted to no more than five series. The groups are presented below in order of likeness, from left to right, based on the degree of similarity in their component membership patterns.

Patterns in allocation were identified among the North Japanese, South Japanese and Guamanians (Figure 29). They share the greatest membership in clusters 2, 3, and 5. They are found in cluster 6 in almost identical proportions (5.6%, 5.5%, 5.9%). Although this contribution is small, it is notable for its consistency; a maximum difference of 0.4% separates each of these three populations. This configuration reflects a morphological similarity between Guam and North-East Asia; reiterating Howells' (1989) findings that, although geographically located in Micronesia, Guamanians "appear to have no particularly close relations to Polynesians and instead share many features seen in Far Easterners" (p. 15). Howells' (1989) Ainu, Andaman, Easter Island, Philippine, and Tolai series (Figure 30) are found together in clusters 2, 5, and 6 in similar proportions. Membership in cluster 3 is greatest for the Ainu and decreases with each subsequent population; the Philippines and the Tolai are not represented in cluster 3. From this pattern, cluster 3 is taken to represent a specifically Asian component; this assumption agrees with Howells' (1989, p. 14-15, Table E-1, p. 188-189) findings that neither the Ainu nor the Andaman are strongly affiliated with Australoids populations (e.g., the Tolai) and that the Philippines are most closely matched using Mahalanobis distances with the Hainan Chinese (2.7) than other East Asian populations represented here, e.g., Ainu (5.4). The Zulu, Dogon, Teita, Bushmen, and Tasmanians (Figure 31) are all linked by their shared membership in, especially, cluster 6 but, not insignificantly, in cluster 5. There are two additional trends of association that underlie these similarities. Memberships in cluster 10 and 14 grow in percentage, proceeding in order from the Zulu to the Bushmen. The Tasmanians are most like the Dogon in their membership proportions overall, except the Tasmanians contribute to cluster 1 in a large amount (15%). This value appears especially high when the values for Howells' core Africa populations are considered, e.g. Zulu (0%), Dogon (1%), Teita (0%). The fact that 6% of the Bushman series contributes to cluster 1 may suggests that this population is intermediate in its biological position between the African and the Australian aboriginal groups.

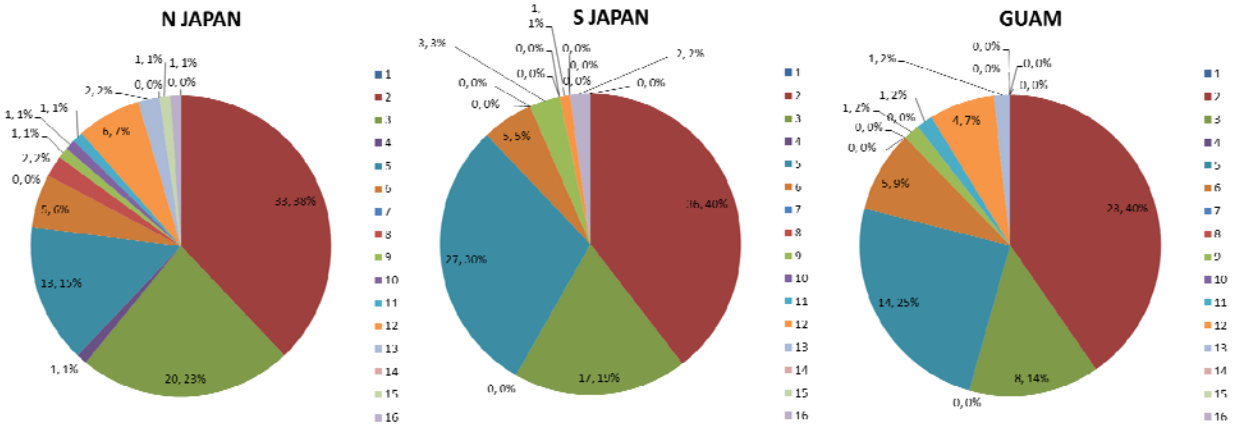


Figure 29: North Japanese, South Japanese, Guam

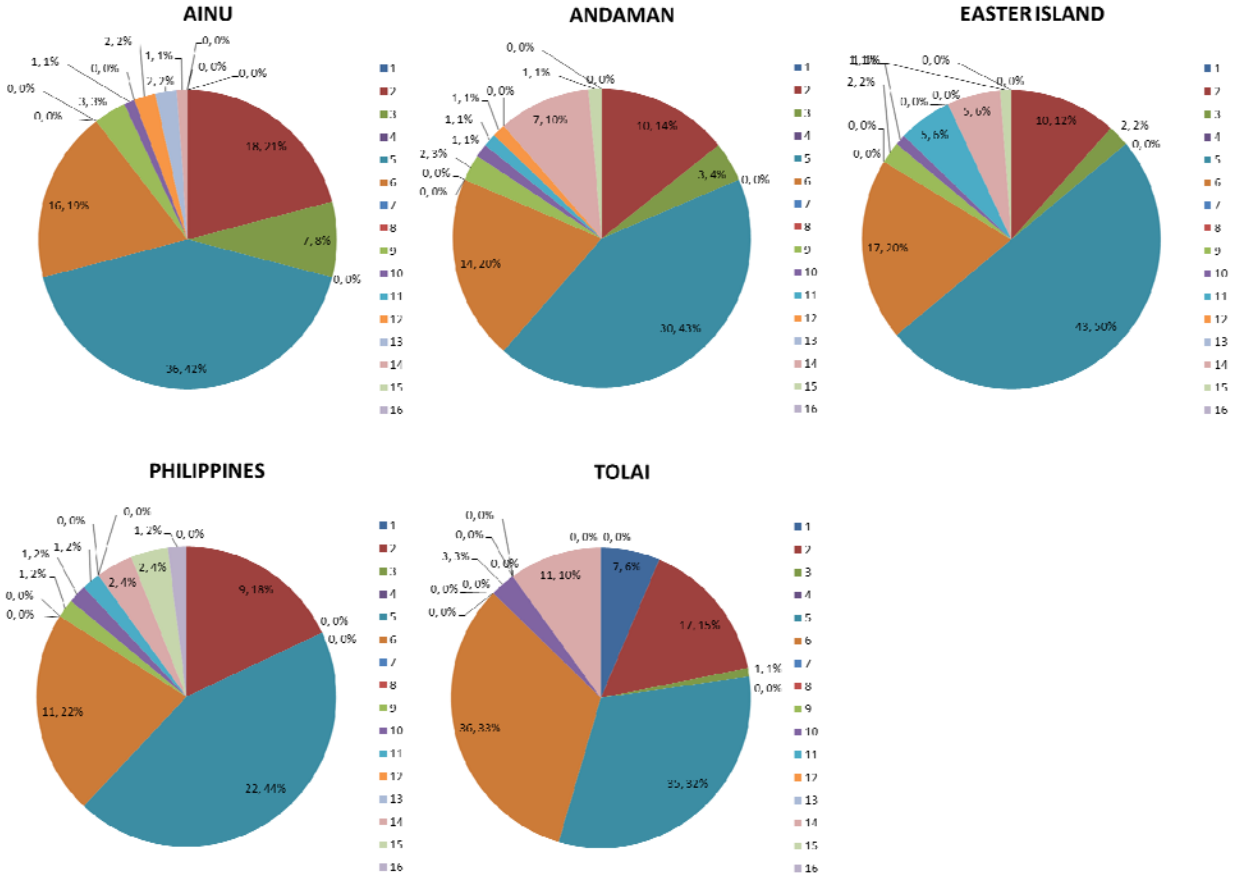


Figure 30: Ainu, Andaman, Easter Island, Philippines, Tolai

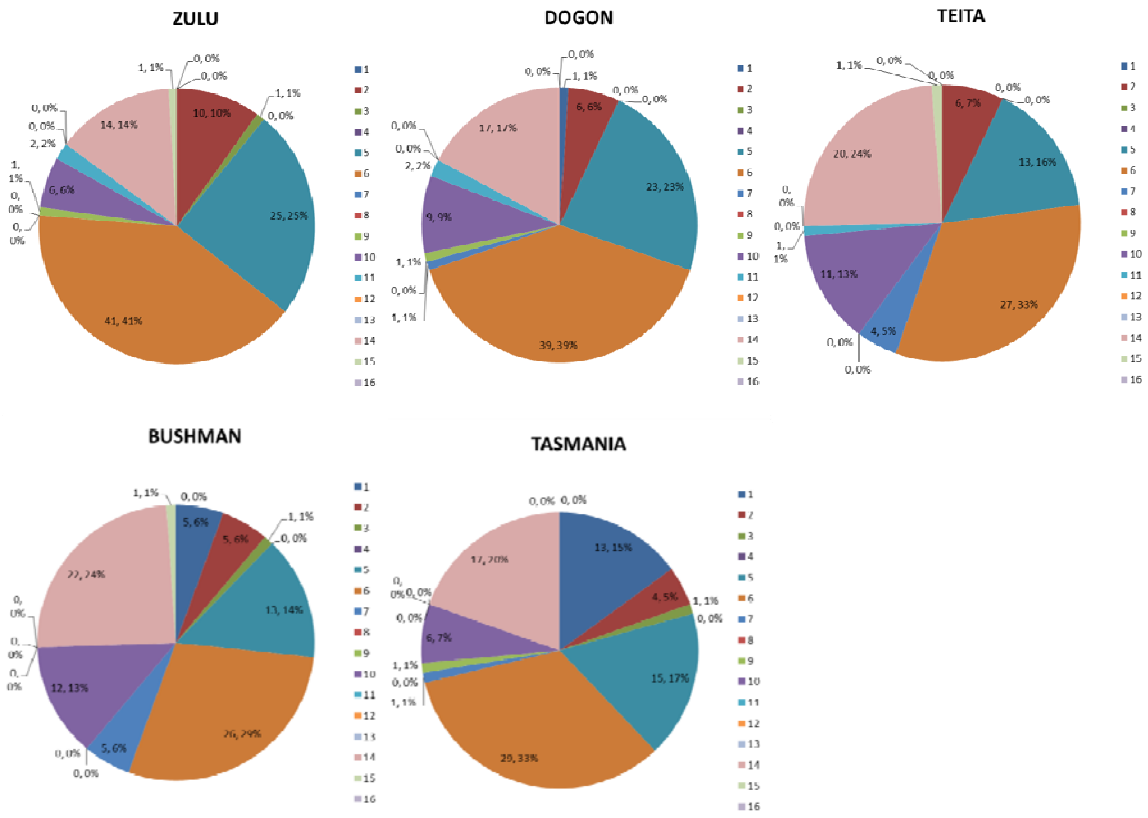


Figure 31: Zulu, Dogon, Teita, Bushman, Tasmania

The Santa Cruz, Peru, Arikara (of the Americas) and the Moriori (of Polynesia) are connected by their notable presence in two key clusters, number 2 and 3 (Figure 32). Other membership patterns are also detectable and serve to delineate the relationships among this series: contributions to cluster 5 decreases from the Santa Cruz to the Moriori, while contributions to cluster 12 increase from Santa Cruz to the Moriori. The Santa Cruz and the Peru series also share membership in cluster 16. While the Buriat and Eskimo are uniquely identified in clusters 4 and 8 respectively, they do show some level of classification affinity with these native groups: group 12 reaches its maximum with the Eskimo, the Eskimo and Buriat share membership with the Moriori in cluster 13. Arguments can be made to append

the Eskimo and Buriat to the Moriori and produce a more broadly distributed collection of indigenous populations of Asian ancestry.

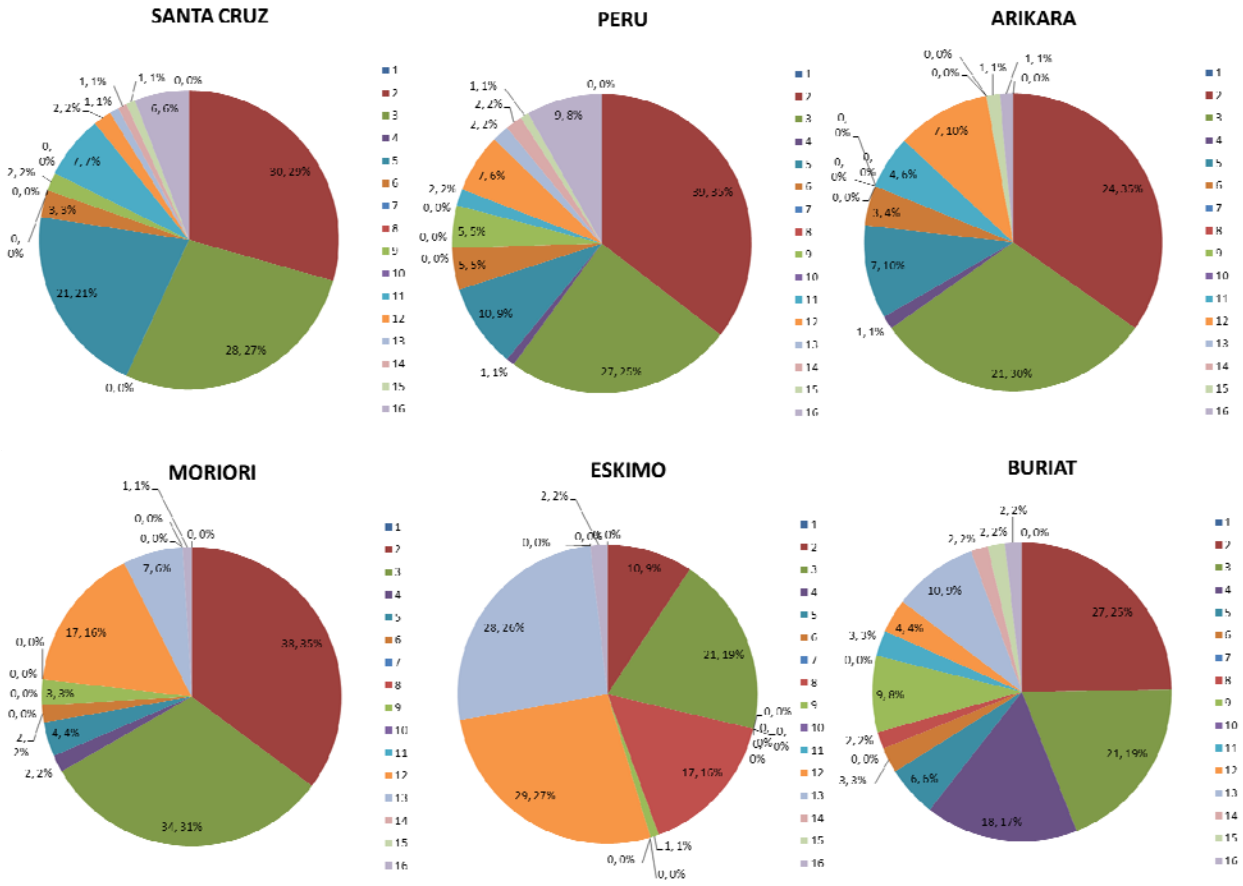


Figure 32: Santa Cruz, Peru, Arikara, Moriori, Eskimo, Buriat

A clearly Caucasian group is represented by Howells's European and Egyptian series (Figure 33). The Berg, Norse, Zalavar and Egyptians are joined by their memberships in cluster 2, which increases from the Berg to the Egyptians, and in cluster 15, which decreases in this same way. The Egyptians are excluded from this group, when the European presence is considered for clusters 3 and 5. The choice of

the relative position of the Egyptians (either before the Berg or after the Zalavar) notably alters the patterns that unite these series, but the relationship between the Norse and Zalavar remains consistent. This suggest, as it might be expect, that the relatively isolated Berg and the "African" Egyptians have complex population-specific morphologies. It is clear, nevertheless, that the Egyptians do not relate well with the other African populations: Egypt is distributed over clusters 2, 3, and 5, while the African series are found mostly in cluster 6 and 14.

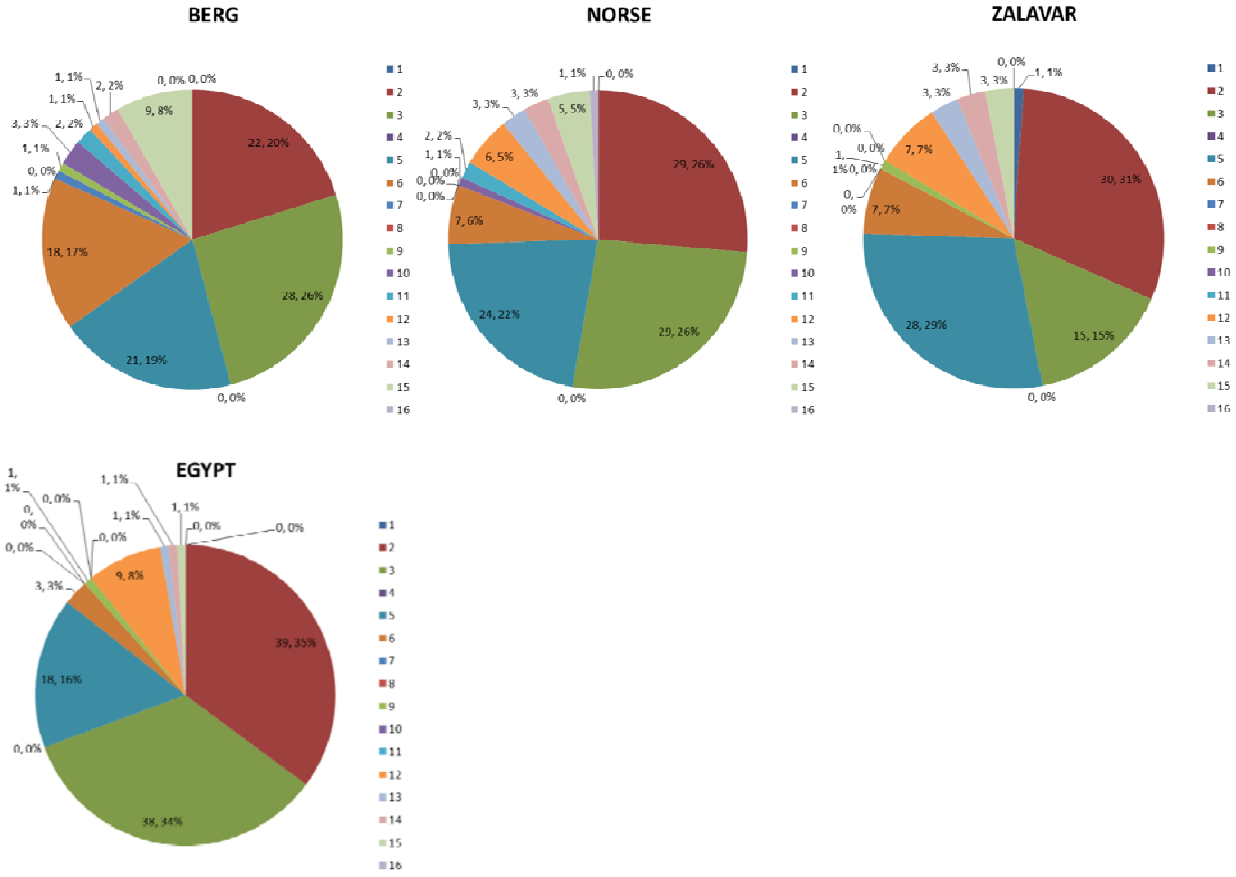


Figure 33: Berg, Norse, Zalavar, Egypt

The Anyang, from mainland China, the Hainan, from South China, and the Atayal, from Taiwan (Figure 34), all share membership in cluster 5 and 6 in similar percents. The proportion of membership increases for cluster 14, proceeding from the Anyang to the Atayal. What is quite interesting is that membership in cluster 14 continues to increase, when moving from the Atayal towards the Africans, Tasmanians and to a lesser degree the Australians. These patterns suggest that the Anyang and Hainan are the most similar of these Asian series (especially considering their shared presence in cluster 3) and that the Atayal may share the morphological signature that links the African and Oceanian groups.

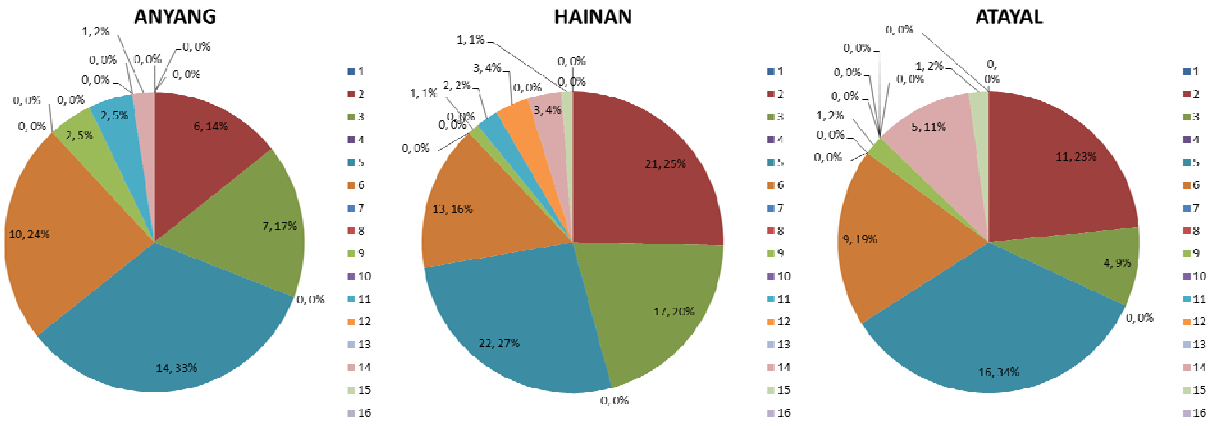


Figure 34: Anyang, Hainan, Atayal

The preceding discussion presents the clustering patterns that were produced by the best MCLUST mixture model, i.e., 16 components with equal covariances. Recall that this model was chosen statistically using BIC-based selection criteria and that the finite mixture clustering procedure did not incorporate any of the *a priori* identifiers associated with the populations and individuals that make up Howells' (1989) full craniometric data set. Despite this intentional exclusion of key metadata on true group memberships (e.g., population name, sampling location, geographic origin, sex), this model-based

clustering analysis was highly effective in detecting structure within the "mixed" population data set. Most classifications were made with acceptably high posterior probabilities of component membership and, overall, this model produced correspondingly low ("implied") error in allocations (i.e., 24% at a threshold of >0.7). Most importantly, it can be seen from the discussion of the per-component and per-population visualizations that it is possible to resolve this statistically significant structure into anthropologically meaningful clusters.

While this 16-component EEE model produced these valid and meaningful results, the clustering patterns are complicated and reflect the choice to accept the "best," but highly complex, model. Even though this model identified 12 fewer groups than the maximum (i.e., 28 known series) included in Howells' (1989) data set, instances of redundancy can be detected in this solution: in several instances, "regional" variation was parsed into nested subsets, which were distributed across several clusters. These findings highlight the known problems associated with selecting for fit over parsimony, and call into question model choice. For this reason, it is important to expand this discussion to contextualize the "best" selected model (16 clusters, EEE) among the other "good" options available. Special consideration is given to the results of the clustering produced under the most *parsimonious model* identified using MCLUST: the "VVV" model with three (3) components. This model was included among the 10 top options (i.e., having the highest BIC values) already reported in Table 8. It should be noted that parsimony in this case is defined in terms of the number of components. If, however, parsimony is redefined in terms of the restriction placed upon the parameterization of the component covariance structure, then this VVV model is the most complex (for the amount of estimation required). Under this latter condition, the chosen "best" 16-cluster EEE solution [re]emerges as the ideal model, now for two

reasons: it has the highest estimated BIC value and it takes the simplest parametric form, by assuming equal covariance matrices for all groups.

While investigating the results of the three component model may be of special interest, in order to determine if the simpler clustering solution produces a better clustering outcome, doing so is problematic for both practical and conceptual reasons. Comparing models which produce clusters in numbers that are different from each other and from the data set's known classifications is not straightforward because the observed proportions differ from the expected: e.g., Howells' data contains 28 true groups, the best model contains 16 clusters, the most parsimonious model contains 3 clusters. Assessing accuracy is also difficult because the maximum number of components over which the posterior probabilities are (free to be) distributed may influence the magnitude of the "greatest" per-individual estimate. Although the posterior probabilities used to determine individual memberships are allowed to vary between 0 to 1 and membership assignments are typically made based on the cluster number associated with the "greatest" of these estimates, the model with 3 clusters will most likely produce posterior probabilities with apparently "higher" "greatest" values than the model with 16 components. This occurs simply because the 3 cluster model is simpler. If, instead, comparisons are made between models with a similar number of components, and, say, different parameterizations this practical concern is lessened: i.e., considering the top three models with 3 components and the highest BIC values (e.g., VVV,3 , 110685.9; VEV,3, 110323.6; EEV,3, 110218.5).

Weighing parsimony against model fit is just as tricky conceptually as it is practically, when the two options differ considerably in component number and parameterization, as they do here (the EEE model with 16 components versus the VVV model with 3 components). Statistically speaking, it has been

established already that these models are reasonably good: both were well-chosen using MCLUST's BIC-based selection method and it has been shown that population [sub]structure can be detected using the 16 component model. There is, nevertheless, an important question that remains to be resolved: will reducing the number of components and changing the restrictions on the covariance structure produce a sufficiently better clustering to warrant changing from the 16 component solution to the three component model? This issue of "improving" the mixture analysis is tied to the question of increasing the clustering's information value. Does the current model help the user to respond to all of the points of concern and adequately address all of the topics of interest? Can the same information be obtained in an equally effective and possibly clearer way using a simpler model? Is it possible to gain further understanding of intergroup relationships by assessing the data from a different perspective (under the favored conditions of parsimony? It seems the most profitable approach to this discussion of "improvement" is to test the reduced 3-component model under simplified terms of analysis that maximize this model's potential. It is useful to take advantage of the fact that this three cluster solution corresponds nicely to the three continent models that have been already tested with success in model-based clustering of genetic data (e.g., Bamshad et al., 2003). It is reasonable, given this precedent, to evaluate how well the three statistical clusters correspond to three macro-groups that are presumed to have clear biogeographic meaning: Asia, Africa, and Europe. In framing the analysis in this way, it is possible to access information that is useful to a discussion of race. In Bolnick's (2008) words, this three-partition, geographic approach is, in fact, considering biological "differentiation among what are essentially races based on continental ancestry" (p.73).

In order to provide a frame of reference for the discussion of these continental relationships, a classification table was generated to display the per-cluster allocation rates for the three component

"VVV" model, as fit with MCLUST (Table 14). It should be reiterated that this three cluster solution, like the previous 16 component analysis, is not predicated upon any information of true group affiliation.

Table 14: Classification table

Cluster Number	***1***	***2***	***3***
Number Of Allocations	452	1117	935
Percent Representation	18.051118	44.608626	37.3402555

To aid in the visualization of the classification patterns associated with this model, a matrix of pair-wise scatter plots Figure 35) was produced to display the allocations of the 2504 individuals over the three clusters, using the first five of the 12 base craniometric variables. A coordinate projection using the first two variables (Figure 36) and a set of random projections (Figure 37) based on three different seeds were also generated for ease of interpretation: each cluster is represented by a colored point cloud, the corresponding component means and standard deviations are indicated, and the ellipse is defined by the covariance matrices of the model.

To evaluate the separation of the 3 clusters and the strength of the individual classifications, this analysis adopted the same standard procedures described in the discussion of the 16 cluster model. The fitted posterior probabilities of group membership were used to (1) calculate the number of misclassifications in each population, (2) arrive at an estimate of the overall error rate in allocation, and (3) to generate the quantiles of uncertainty for this three-component cluster analysis. These finding are reported in Table 15 and Table 16 respectively; Figure 38 and Figure 39 provide graphical displays of the classification uncertainty.

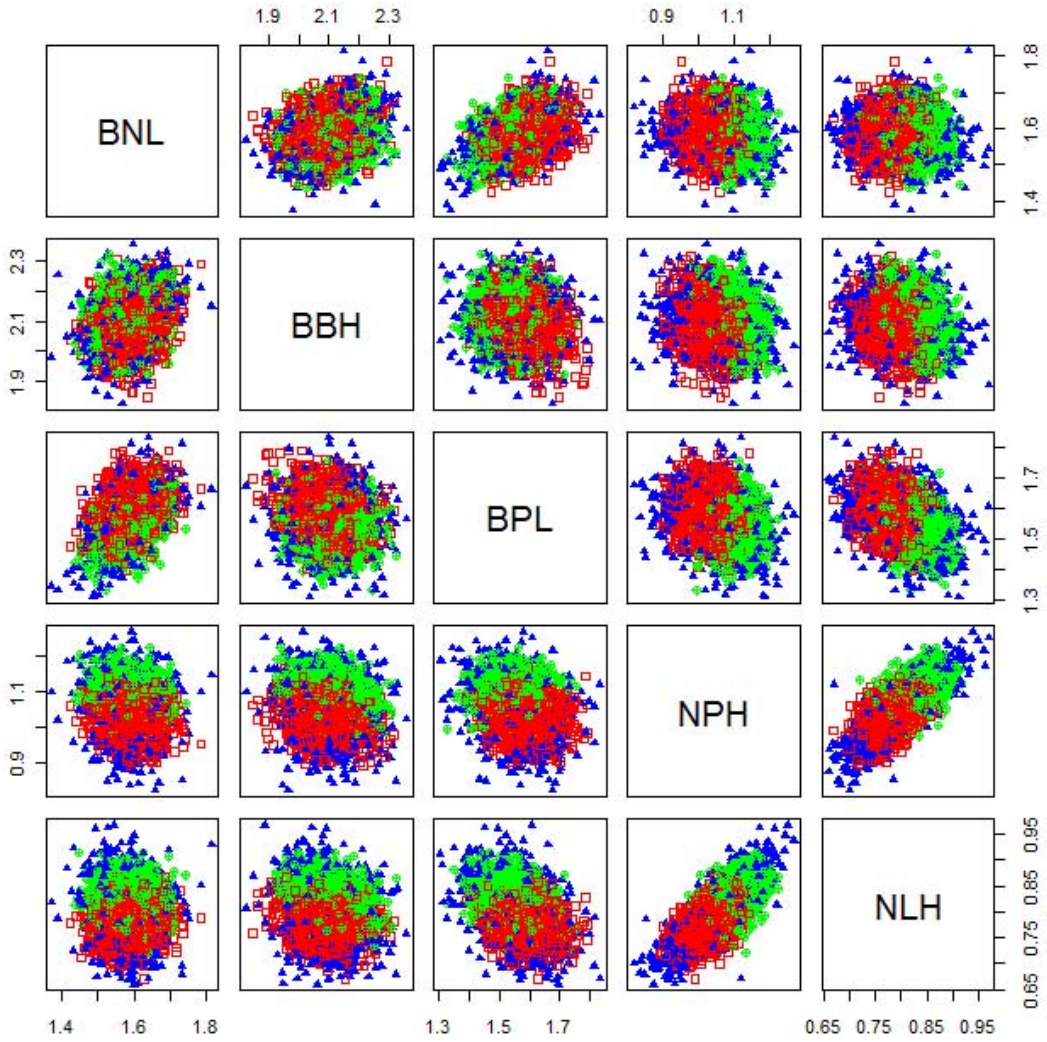


Figure 35: Pair-wise scatterplot matrix showing clusters

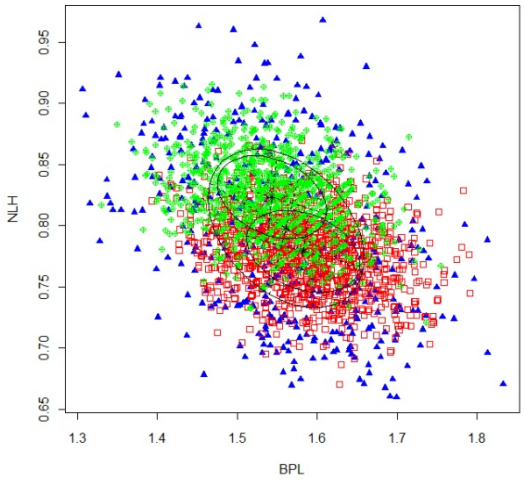


Figure 36: Coordinate projection showing clusters

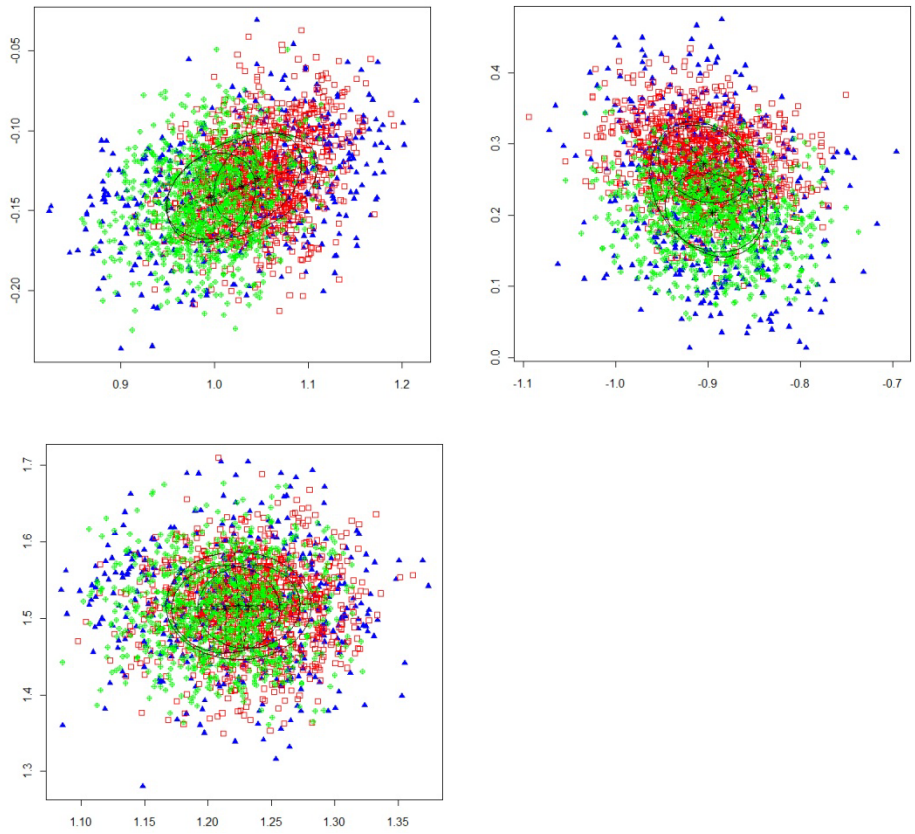


Figure 37: Random seed projection showing clusters

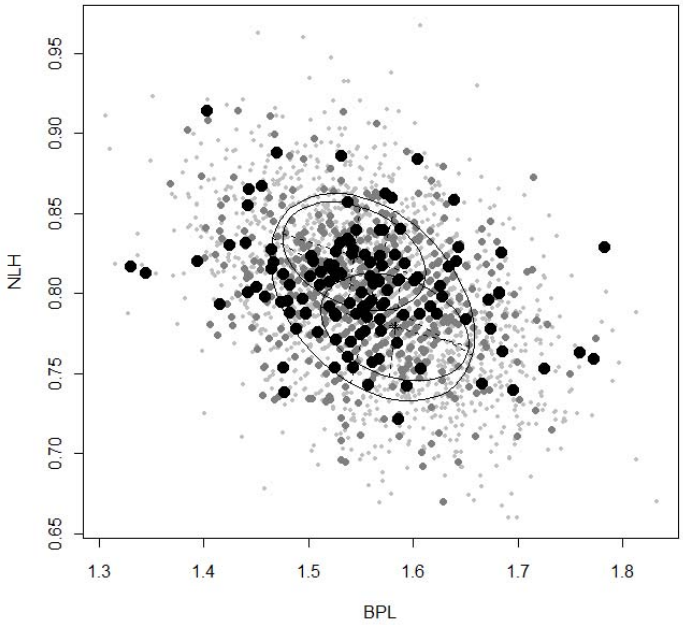


Figure 38: Coordinate projection showing uncertainty

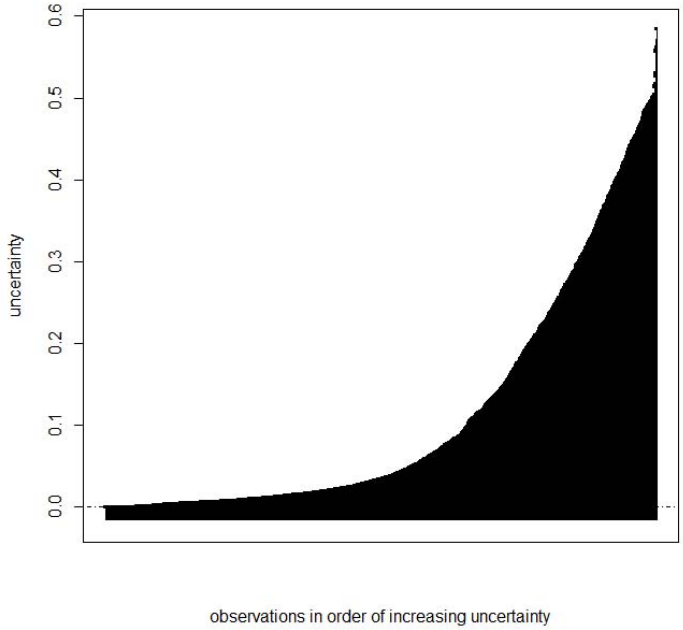


Figure 39: Uncertainty profile

Table 15: Allocation rates

Population (Howells 1989)	*posterior probability > 0.7*	*percent correctly allocated*	*posterior probability → 1*	*percent correctly allocated*	*total number of individuals*
AINU	72	84	40	47	86
ANDAMAN	62	89	44	63	70
ANYANG	33	79	29	69	42
ARIKARA	57	83	40	58	69
ATAYAL	40	85	32	68	47
AUSTRALIA	91	90	75	74	101
BERG	96	88	76	70	109
BURIAT	96	88	77	71	109
BUSHMAN	86	96	69	77	90
DOGON	86	87	74	75	99
EASTER ISLAND	75	87	63	73	86
EGYPT	96	86	61	55	111
ESKIMO	98	91	86	80	108
GUAM	48	84	31	54	57
HAINAN	69	83	52	63	83
MOKAPU	85	85	54	54	100
MORIORI	93	86	73	68	108
N JAPAN	71	82	53	61	87
NORSE	90	82	70	64	110
PERU	98	89	73	66	110
PHILIPPINES	39	78	31	62	50
S JAPAN	72	79	53	58	91
SANTA CRUZ	84	82	59	58	102
TASMANIA	81	93	67	77	87
TEITA	75	90	66	80	83
TOLAI	94	85	72	65	110
ZALAVAR	81	83	58	59	98
ZULU	87	86	72	71	101
Grand Total	*2155*	*86*	*1650*	*66*	*2504*

As expected, there is an improvement in clustering success relative to the reduction in the number of components considered in the mixture model. When memberships were determined using the 3 component VVV model, 86% of the cases were clustered "correctly" at the threshold of > 0.7 and outright classifications were obtained for 66% of the total data set. The implied error rate dropped on average by 20% with the simplified model. It is interesting, however, that, when comparisons were made between the estimates of uncertainty produced by MCLUST for both analyses, the findings associated with the 16 and 3 component models do not appear notably different (see Table 10 and Table 16).

Table 16: Quantiles of uncertainty

Quantile	*0%*	*25%*	*50%*	*75%*	*100%*
Estimate	0	0.008490946	0.033807484	0.175855984	0.583987695

In order to evaluate the 3 cluster model for its relationship to continental ancestry or with the well-known concept of "the three major races " (c.f., Nei & Roychoudhury, 1974; Roychoudhury, 1978; Chakraborty, 1984), it was necessary to first associate each of Howells' populations with the continental group (Asia, Africa, or Europe) that best fit either their geographic location or their evolutionary origin. Collapsing all 28 of Howells' populations into these three biogeographic groupings was intuitive in most cases, especially given the documentation provided for each series by Howells (1989, Appendix A). Some justifiable liberties were taken in these associations; at times, decisions were based on prior knowledge of population origins and affinities as well as the classification patterns revealed by the 16-component cluster analysis. This project chose to (1) include all of the native North American Indian groups (from the Eskimo to the Arikara to the Peruvians), the Siberian Buriats, and the Pacific Islanders within Asia, (2)

assign the Australians, Tasmanians and Tolai to Africa, and (3) classify Egyptians under Europe. In doing so, Asia was significantly expanded, continental Africa was redefined as sub-Saharan Africa plus Australo-Melanesia, and Europe was reimagined along biosocial lines as a Caucasian grouping (c.f., Bamshad et al., 2003).

As visual aids for investigating the patterns of correspondence between continent, ancestry and race, this project produced graphical representations of the percent contribution of each of the three clusters for each of the three biogeographic groups. The percent contribution of each of Howells' 28 populations to the three MCLUST clusters is provided in Table 17.

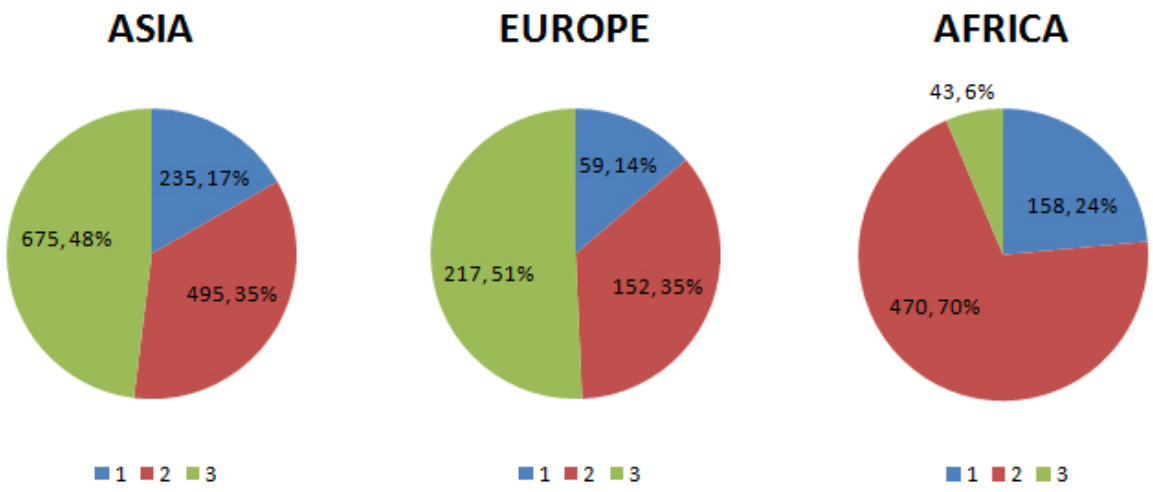


Figure 40: Continental clusters

Table 17: Cluster composition by population contribution

Cluster 1		*Cluster 2*		*Cluster 3*	
BUSHMAN	53.33%	AUSTRALIA	80.20%	MORIORI	81.48%
BURIAT	46.79%	ZULU	79.21%	ARIKARA	72.46%
ESKIMO	46.30%	TOLAI	79.09%	PERU	70.00%
TEITA	33.73%	EASTER ISLAND	79.07%	EGYPT	66.67%
TASMANIA	32.18%	DOGON	75.76%	SANTA CRUZ	60.78%
DOGON	19.19%	PHILIPPINES	72.00%	N JAPAN	58.62%
BERG	16.51%	AINU	65.12%	GUAM	56.14%
MOKAPU	16.00%	TASMANIA	64.37%	S JAPAN	53.85%
NORSE	14.55%	ANDAMAN	64.29%	ESKIMO	53.70%
ANDAMAN	14.29%	TEITA	61.45%	MOKAPU	50.00%
ANYANG	14.29%	ATAYAL	57.45%	ZALAVAR	46.94%
ZALAVAR	13.27%	ANYANG	57.14%	NORSE	46.36%
HAINAN	13.25%	HAINAN	45.78%	BERG	42.20%
ZULU	12.87%	BUSHMAN	44.44%	BURIAT	42.20%
AINU	12.79%	GUAM	42.11%	HAINAN	40.96%
SANTA CRUZ	12.75%	BERG	41.28%	ATAYAL	31.91%
N JAPAN	12.64%	ZALAVAR	39.80%	ANYANG	28.57%
PHILIPPINES	12.00%	S JAPAN	39.56%	AINU	22.09%
PERU	11.82%	NORSE	39.09%	ANDAMAN	21.43%
TOLAI	11.82%	MOKAPU	34.00%	PHILIPPINES	16.00%
EGYPT	10.81%	N JAPAN	28.74%	AUSTRALIA	10.89%
ATAYAL	10.64%	SANTA CRUZ	26.47%	EASTER ISLAND	10.47%
EASTER ISLAND	10.47%	EGYPT	22.52%	TOLAI	9.09%
MORIORI	10.19%	ARIKARA	20.29%	ZULU	7.92%
AUSTRALIA	8.91%	PERU	18.18%	DOGON	5.05%
ARIKARA	7.25%	BURIAT	11.01%	TEITA	4.82%
S JAPAN	6.59%	MORIORI	8.33%	TASMANIA	3.45%
GUAM	1.75%	ESKIMO	0.00%	BUSHMAN	2.22%

It is apparent that there is no "perfect" correspondence between continent and cluster. From this perspective, it can be said that reducing the number of components does not provide, at least in this case, better information. The three-component analysis does, however, provide different information: it

highlights general trends in classification that may be supported by population history and may correspond to world-wide differences in craniofacial morphology. Africa is well separated from Europe and Asia, and Europe and Asia show a great degree of overlap. These patterns are further clarified by the individual allocations, when some of the "continental" populations are compared for their consistency in membership.

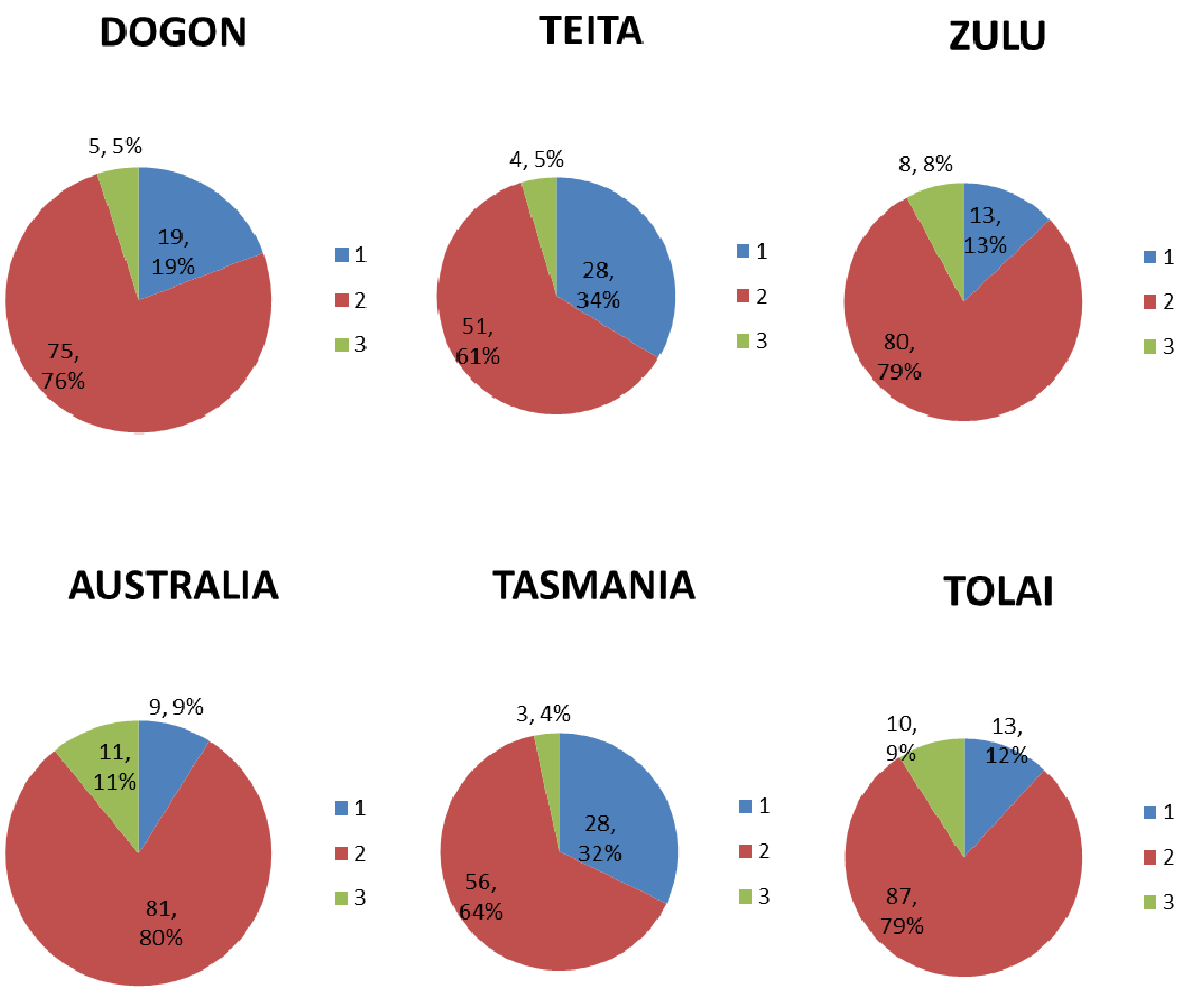


Figure 41: Dogon, Teita, Zulu, Australia, Tasmania, Tolai

It can be seen from Figure 41 that both the sub-Saharan Africans and the Australo-Melanesias share strong membership in cluster2. This pattern not only suggests a characteristic African morphology but also supports the assumed affinity between these two geographic groups.

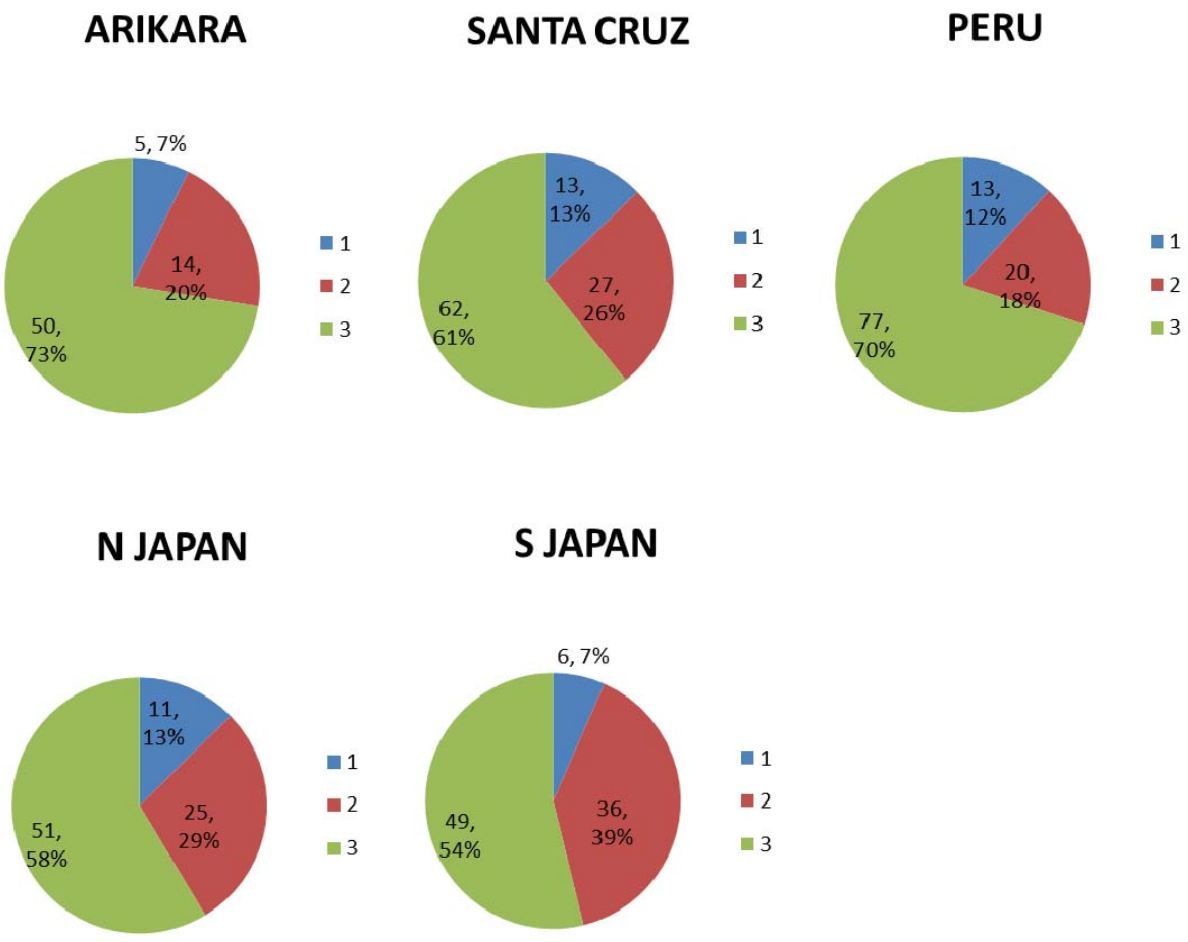


Figure 42: Arikara, Santa Cruz, Peru, North Japan, South Japan

The consistency with which both the East Asian and Native American populations are assigned to cluster 3 both verifies their evolutionary origins and population histories, and suggests an Asian identification for the variation represented by cluster 3 (see Figure 42). This pattern is complicated, however, by the significant "African" contribution to the makeup of otherwise geographically "Asian" groups (e.g. Ainu, Andaman, Anyang, Atayal and Philippine series, Figure 43).

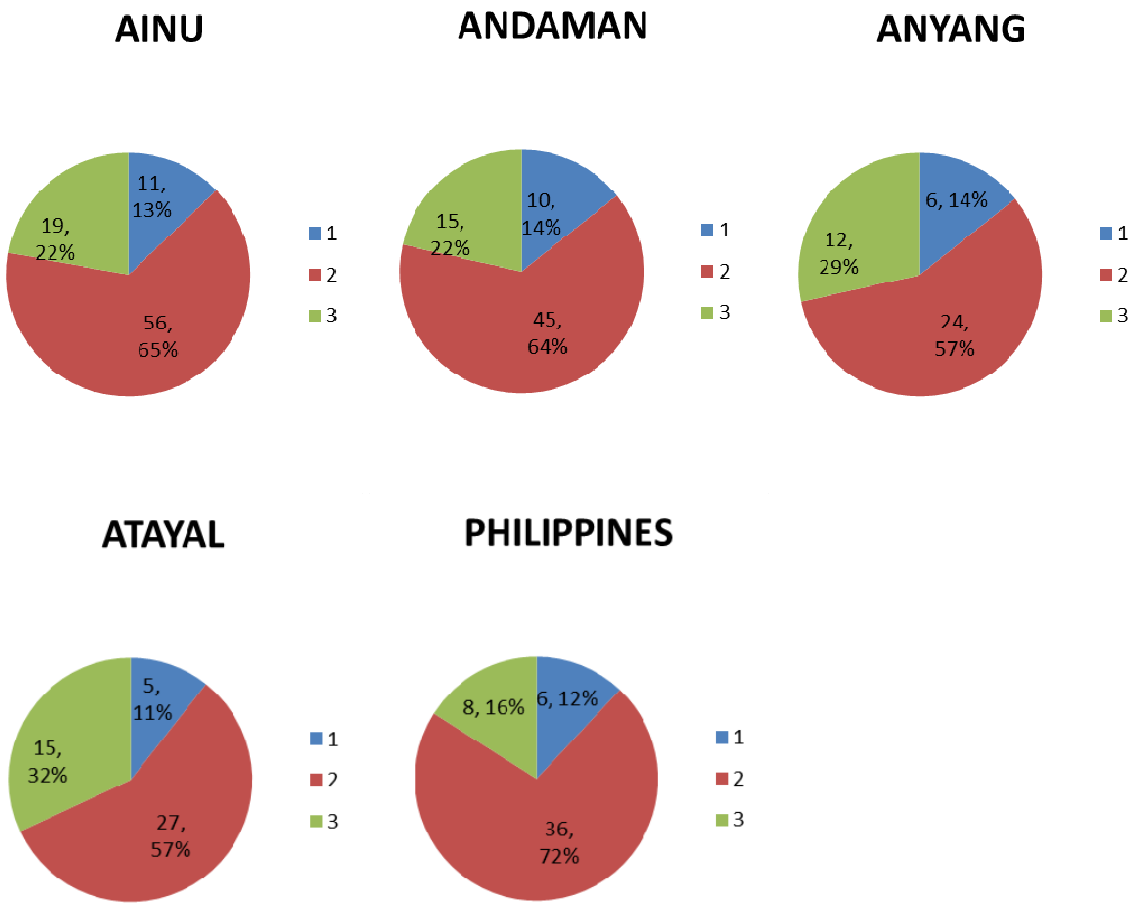


Figure 43: Ainu, Andaman, Anyang, Atayal, Philippines

It is interesting that Howells (1989, p. 13-15) also flagged some of these same populations for their departures from geographic expectations and historical norms. He found, for example, no affinity between the Ainu and the Europeans or Australoids and noted similarities between the Andamanese and Africans but not the Australoids.

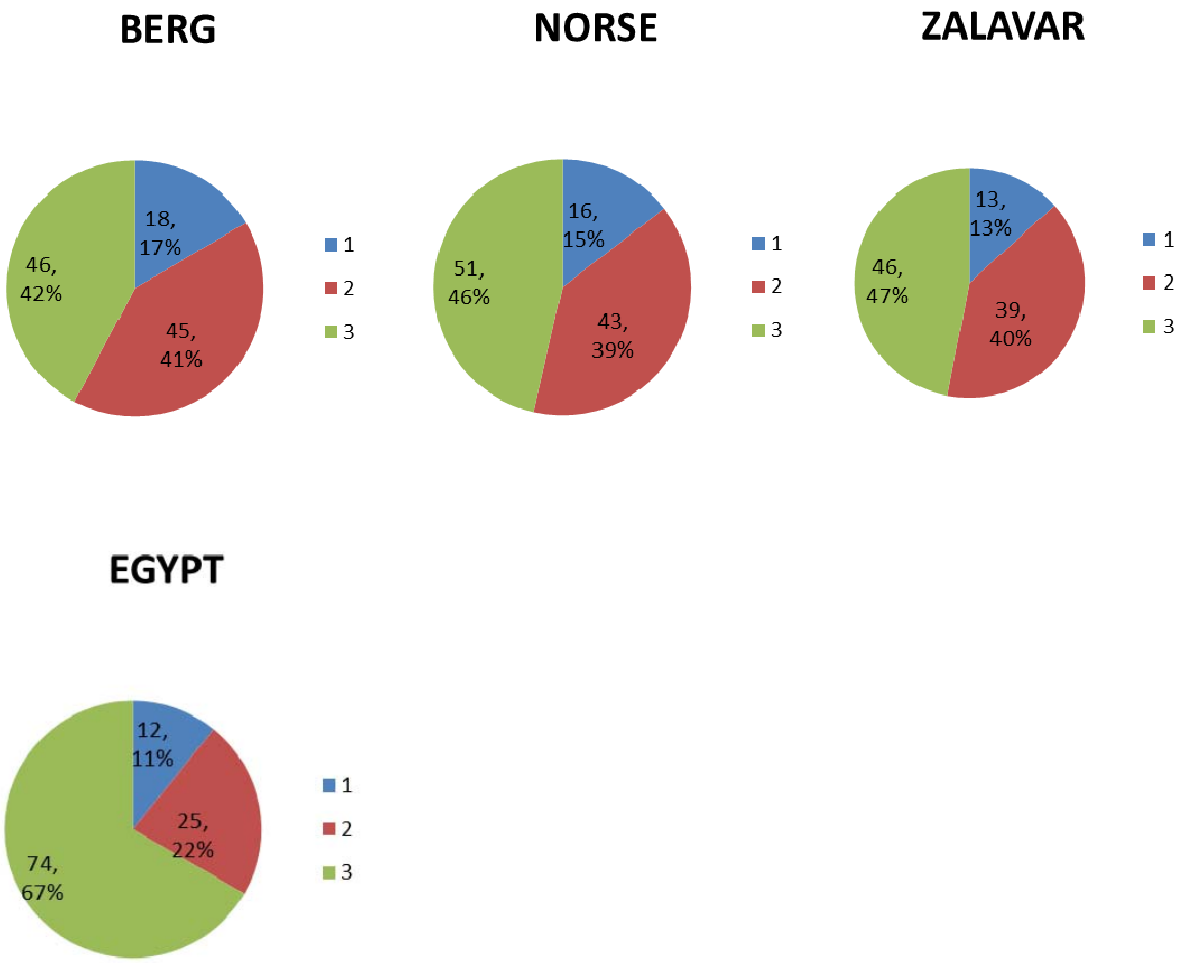


Figure 44: Berg, Norse, Zalavar, Egypt

The European populations (Figure 44) are clearly the most problematic of any continental group: while the Berg, Zalavar, and Norse have similar patterns of representation, they are distributed across both clusters 2 and 3 in close proportions. Egypt stands out among these Europeans series for its large presence in cluster3.

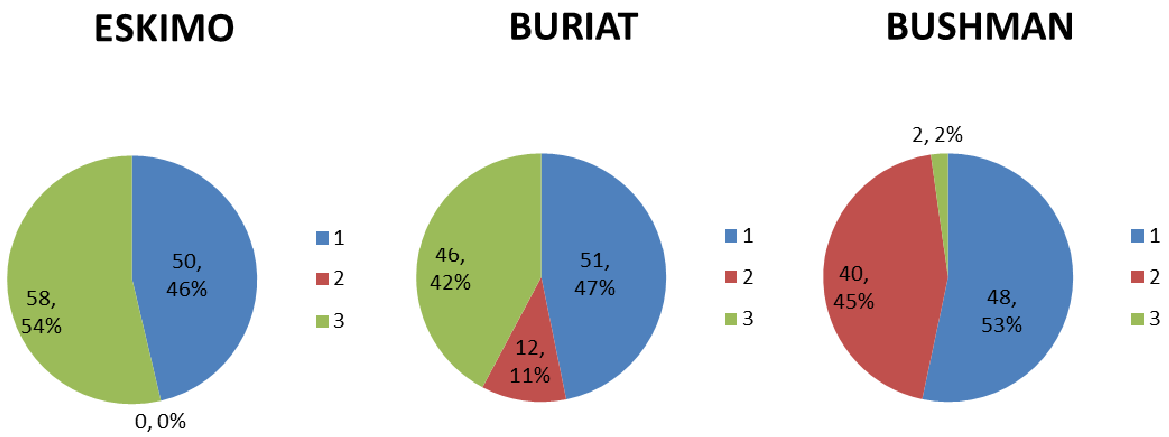


Figure 45: Eskimo, Buriat, Bushman

It appears that cluster 1 is poorly linked to any one of the three continental groups, but there is, nevertheless, a distinctive pattern of association. The greatest single contributors to cluster 1 are Howells' most unique local series, the Eskimo, Buriat and Bushmen (Figure 45), whose morphological patterns are conspicuously divergent from the populations with whom they are presumed to share the greatest affinity.

The considerable degree of overlap between these three clusters may suggest that the most parsimonious solution presents a muddied picture of the patterns of human variation. There are key differences between groups that are lost on account of its overgeneralization. This project's previous concern still remains unresolved: is the best BIC model with many components still better than an

acceptable BIC model with fewer components? This situation is now further complicated by the problem of model identification. While it can be argued that model choice based on the highest BIC or the highest BIC with parsimony requires little user-direction, determining the preferred model between or around these "best" options is less straightforward. Moreover, it is no longer easy to justify the findings or the choice of protocol with arguments invoking fully "unbiased" estimation. How can balance be reached between the best fitting and the most parsimonious mode? And under what criteria should this model be selected? It is reasonable to look to other biological applications of finite mixture methods for the identification of population structure as a guide. In the genetic literature, Rosenberg and colleagues (2002) adopted an approach that involved both statistical and user-directed model selection. Using the STRUCTURE software (Pritchard, Stephens, & Donnelly, 2000), they calculated the probability of their data (X) given each value of G considered (i.e., Pr[|X|G], over the range of g-components) and took the best value(s) of G as the estimate(s) associated with the highest approximation of Pr(X|G). They then chose to further investigate the models with the number of clusters that related most logically to their research goals, given prior expectations for the geographic distribution of human variation. This negotiated approach seems to be in keeping with the opinions that advocate against (or at least question) relying solely upon STRUCTURE's computational method for choosing the number of component in the mixture. Pritchard and coworkers (2000) have remarked that "we see it merely as an ad hoc guide to which models are most consistent with the data, with the main justification being that it seems to give sensible answers in practice" (p. 949). Bolnick (2008, p. 75) has further elaborated on this point, stating that "it is not always clear which of the probabilities associated with a given K should be used when determining the best values for K." While Rosenberg et al., (2002) did not report (or identify for the reader) the model with the most likely number of components, they did present their findings for the model-based clusterings using two to six groups. They focused

especially on their six component analysis, in which they identified a correspondence between the statistical components and biogeographic clusters. Five of these six clusters closely matched major geographic regions: Africa, Eurasia, East Asia, Oceania and America. Following this approach, models with four and five clusters were also evaluated here to determine if, in expanding the three continent schema one which allowed for the separation of Oceania and/or the Americas from Asia, a more satisfying solution is produced.

This project considered specifically a four component VEV model and a five component VVV model, both of which were identified by MCLUST as among the top ten solutions with the largest BIC values (see Table 8). Summary information on the classification accuracy of both these two new models as well as the values for the three and 16 component clustering are presented **in** Table 18**.** As these values indicate, when the number of clusters is low there are only slight changes made to the percent of individuals correctly classified in accordance with the addition or subtraction of components; moreover, the level of model uncertainty in the four and five component models actually exceeds the value calculated for the "optimal," 16 component, clustering solution.

Table 18: Classification success and model uncertainty

Model	*Posterior Probability > 0.7*	*Posterior Probability → 1*	*Quantile of Uncertainty (100%)*
3 VVV	86%	66%	0.583987695
4 VEV	86%	62%	0.66706111
5 VVV	83%	58%	0.70747961
16 EEE	76%	38%	0.6511021

After reviewing in detail these patterns of allocations (by cluster, population, and continental ancestry), it was found that the additional components failed to produce substantially improved results either in terms of their degree of correspondence with predefined geographic partitions (e.g., Asia, Europe, Africa; Africa, Eurasia, East Asia, Oceania and/or America) or their ability to reveal substantially more of the nuanced patterns of affinity between Howells' groups. There were, however, two notable changes to the clustering patterns worth mentioning. When this analysis considered the percent contribution of each cluster to Howells' populations, he fourth component was found to be conspicuously absent among the African populations, including the Bushmen, and it was only very weakly represented among the Australo-Melanesians (i.e., only 3% of the Australians are assigned to cluster 4). Cluster 4 was, instead, represented equally among both the Asian and European series and to a consistent effect: while the number of individuals assigned to cluster 3 (the "Asian" component) did not vary, the number of cases assigned to cluster 2 (the "African" component) was visibly reduced and replaced by cluster 4. When a fifth component was added to the model, this cluster acted as a sub-partition of the Asian variation across all groups except the Eskimo. It is interesting to note that, while all other groups saw increased diversity in their memberships as the number of clusters grew from three to five, the Eskimo remained stable, contributing members to only two clusters (one of which was consistently "Asian"). The strength of the morphological boundaries defining this group may be attributed to their extreme facial flatness and anterior projection, which likely informed Howells' characterization of the Eskimo as "Super-Mongoloid." This association may also reflect the complex negotiation between the pull of evolutionary origins (i.e., Asia) and the selective force of environment (e.g., cold adaptation). When the cluster findings were considered over the range of the analyses presented here (from 3 , 4, 5, and 16 component models), there is a striking degree of similarity in the way in which the populations formed cohesive groups. These patterns of congruence across models with incrementally increasing or

decreasing component numbers, suggest that (1) craniometric variation is hierarchically structured and nested, and that (2) this structure can be mapped onto both to macro-geographic groups (e.g., 3-group model using Asia, Europe and Africa, which represented the major "continental races") as well as smaller scale regional groupings (i.e., the clear sub-Saharan African or Australo-Melanesian subdivisions).

While it has been demonstrated so far that changing the number of components and parameterizations can affect the quality of the clustering solution and utility of the cluster memberships produced, this discussion has been limited to arguments of parsimony and goodness of fit as they apply to models selected using the Bayesian Information Criterion and, more specifically, when following the approach proposed by Fraley and Raftery (1998, 2011) and integrated within the MCLUST package. While BIC has been shown to consistently estimate the number of mixture components under certain regularity conditions (Keribin, 1998, 2000) and numerical tests have demonstrated that it performs well in practice (Fraley & Raftery, 1998), concerns have been voiced over its uncritical acceptance within model-based clustering. Biernacki, Celeux, & Govaert (2000) have argued that the goal of clustering is not the same as the goal of selecting the optimal mixture model and that the integrated likelihood fails to take this difference into account. Baudry et al., (2010) have recently responded in kind, stating that that, while they feel BIC can meet the goal of choosing the number of mixture components for estimating the underlying probability density (c.f., Roeder & Wasserman, 1997), it may fail at meeting the goal of estimating the number of clusters. If this shortcoming is true, then MCLUST's BIC-based approach may not be the most appropriate method for determining the number of clusters for a mixture analysis. The way that the practitioner chooses to arrive at the final clustering solution (i.e., the software used, the algorithms implemented, the measures considered) may have, therefore, a significant effect on the number of components estimated for the mixture, the number of clusters and the corresponding

classifications produced, the amount and clarity of the internal structure that is revealed. All of these factors have bearing on the conclusions that can be ultimately drawn about the distribution of human craniometric variation and the opinions that are, subsequently, formed about population models, like isolation by distance, and concepts, like biogeographic ancestry and biosocial race. With these methodological concerns in mind, is it important to assess the quality of these tests of population structure. To do so, this project adopted a different mixture modeling approach than is standard for MCLUST version 3 (Fraley & Raftery, 2011) and conducted comparative analyses of Howells' (1989) data set using the EMMIX package (McLachlan, 2011; McLachlan, Wang & Peel, 2011).

Using the contributed R package, EMMIX (McLachlan, 2011; McLachlan, Wang & Peel 2011), this project implemented a second model-based clustering of the full skeletal series represented in the Howells (1989) data set. In order to maintain maximum comparability between the two different (software-defined) approaches to finite mixture analysis, this analysis kept to the same conditions under which the previous clusterings were executed with MCLUST: 28 populations (26 for females), 2504 individuals, preselected subset of 12 standard craniometric measurements (see Table 3). Although EMMIX provides the convenient option of modeling mixtures of Gaussian, t, Skew Normal, and Skew t distributions via the EM and ECM algorithms, there is no evidence to suggest that any of the groups represented by all three of this project's data sets have "longer than normal tails or atypical observations" (McLachlan et al., 2011; McLachlan et al., nd, p. 2; McLachlan & Peel, 1998). This project chose, therefore, to fit only a normal multivariate mixture model; a decision which also satisfies the concerns for maintaining consistency between the MCLUST and EMMIX analyses as well as computational tractability. As before, this project elected not to impose prior constraints on the covariance matrices but, instead, allowed for maximum flexibility: since it is not possible to test the optimal parameterization using EMMIX, as it can

be with MCLUST, an "arbitrary" (unequal) covariance structure was predefined for the model. As the multivariate normal mixture is iteratively fit in EMMIX by maximum likelihood estimation via the EM algorithm, it is necessary to determine good starting values. While Fraley and Raftery (2011) have integrated a hierarchical clustering approach in MCLUST, Peel and McLachlan (1999) have argued against hierarchical methods "in situations where the data sets contain a large number of points"for reasons of "space and time" (p. 13). This project chose to initialize classification still using an "automatic approach" (i.e., where starting values are not conditional upon external information supplied by the user), but opting instead to arrive at the initial set of parameter values by using both a "random start" and "k-means partition" method (McLachlan et al., 2011). This approach proceeded by executing both 10 random partitions (sampling 70% of the data set) and 10 k-means cluster analyses (using the full data set). From these results, the clustering solution with the highest log likelihood was identified and its classification information was used to start the EM algorithm (c.f., MacQueen, 1967; Hartigan & Wong, 1979; McLachlan, 2011). In order to determine the optimal number of clusters for the model without using the information theoretic (BIC-based) approach that directed the MCLUST analyses, this project chose to use the likelihood ratio test statistic λ to test for g versus g+1 components. As regularity conditions do not hold for -2 log λ to have its usual asymptotic chi-squared null distribution, a bootstrap procedure was employed here to estimate the corresponding p-value (Aitkin, Anderson and Hinde, 1981; McLachlan & Basford, 1988; McLachlan & Peel, 2000). This project followed the resampling method proposed by McLachlan (1987), whereby bootstrap samples are generated for the normal mixture fitted under the null hypothesis of $g = g_0$ (McLachlan &Peel, 1997; McLachlan & Khan, 2004). When executed in practice, this resampling approach can be based on the MMRESAMP subroutine of McLachlan et al. (1995). It was assumed in the analyses using MCLUST and is continued to be assumed here that the true number of groups in the mixture is unknown. A range of values for the number of g-

components was therefore fit and a bootstrap of the likelihood ratio test (LRT) statistic λ (with 999 replicates) was implemented to assess the effect of additional components on the mixture model at each stage of its growth in component number (i.e., $g=g_0$ versus $g=g_0 + 1$). While the minimum (1) and maximum (30) values for the number of components were specified in advance of running the analysis, this was done under the provision that this incremental testing process will stop when the p-value reaches "insignificance," at the pre-set level of 10% (c.f., Peel & McLachlan, 1999 for the default significance level in EMMIX 1.3). This approach differs considerably from the BIC-based method used in MCLUST by its design and the computational intensity of its procedure. In practice, then, the choice between these two methods may be dictated by sample size, the dimensionality of the data set and the number of components to be tested. It is, however, especially important to recognize the key difference in the philosophies that underlie these approaches. When a bootstrapping of the likelihood ratio statistic is implemented here, it is done so with the intention of identifying the smallest number of components compatible with the given data set and with an interest in obtaining p-values to use as statistical evidence for the number of components to fit. When this minimum "good" number of components is reached, the procedure is terminated. There are no means, therefore, to evaluate the quality of the models that come before or after this threshold of insignificance/significance. This approach is clearly unlike the BIC method, in which the best identified model is chosen over the full range of the specified *g*-components. In this case, the best BIC model is not necessarily the model with the fewest components. Moreover, this optimal BIC model is returned alongside many other, arguably, still "good"models, in ranked order. Using the BIC method, the user is given, first, the flexibility to choose among these additional models, by invoking arguments of, e.g., parsimony, preferred covariance structures, or prior expectations for the given data, and, second, the statistical authority necessary to justify their use, by referencing their individually assigned BIC values. For these reasons of

practice and theory, it is felt here that (1) it is appropriate to conduct a series of clusterings in MCLUST using a subset of the models identified with high BIC values, and that (2) the use of EMMIX should be limited to a single cluster analysis of the data set, which makes use of the single optimal component number, as identified by a significant associated p-value.

For this test of structure using Howells' (1989) full data set, the best initial grouping corresponding to the highest value of likelihood (58001.152) was produced by the random start method, at attempt 2 of 10. The EMMIX clustering algorithm was successfully implemented using this initialization and reliable results were obtained for both the tests of component number and the final classification of Howells' data. The summary criteria for this clustering are reported in Table 19. An insignificant p-value (0.77) was reached for the test of 10 versus 11 components and this result is taken to be a reliable indication of the point at which there is no further improvement in the likelihood. The 10 component solution is accepted, in turn, as the model which realized the selection goal: that is, identifying the model with the smallest number of components compatible with the data set. Additional model selection criteria were computed, for each stage of model testing. These include the corresponding values for the BIC and Akaike's Information Criterion (AIC), "the best known of the information criteria used for determining the number of components" (Steele & Raftery, 2009, p. 6). In order to evaluate the effect of altering the model selection protocol on the final clustering solution in EMMIX, these additional values were produced to represent acceptable but different criteria for model selection. These should aid, therefore, in contextualizing the findings for the optimal number of mixture components (McLachlan & Peel, 2000). Although the utility of the BIC has been discussed already in the context of implementing mixture analysis using the MCLUST package, it is useful to reiterate here the values obtained with MCLUST for the 3 component VVV (110685.856) and 16 components EEE models (110765.65) and provide the

corresponding loglikehoods (56407.22 and 56497.98, respectively) to demonstrate the consistency in their estimation and their relationship to the other information (AIC) and likelihood-based selection criteria (bootstrap estimate of $-2 \log \lambda$). The theoretical justification for the use of the AIC follows that the model with the minimum estimated value should be asymptotically closest in Kullback-Leibler distance to the true model (Akaike, 1983). Unlike the bootstrap of the LRT statistic, which can be fit for a range of components but is, by its nature, subject to stopping rules, the AIC is typically evaluated, like the BIC, over the full range of models specified (here, from a minimum of 1 to a maximum of 30 components). While, in theory, a smaller critical value could be reached among the unrepresented models in the range of g-components that remain untested here (from 11 to 30), it is interesting to observe that the AIC also selects the 10 component model as the best solution. It is worth noting, however, that several studies (e.g., Koehler & Murphree, 1988; Celeux & Soromenho, 1996) have found that for mixtures the AIC overestimates the number of components: when compared to the BIC, which tends to underestimate the true number of clusters. The AIC imposes less of a penalty upon models with larger numbers of parameters, which, instead, biases selection towards models with more mixture components.

Although different in its model specifications, the 10-component solution identified using the EMMIX package produced a pattern comparable in its per-cluster allocations to the 16-component model obtained in MCLUST (see Table 20). While the clusters fluctuate in size, there are no abnormally small clusters suggesting either the isolation of outlying observations or problems with the model fitting algorithm. It is believed, therefore, that these memberships correspond to natural clusters and reflect the underlying structure of Howells' craniometric data. The reliability with which this structure is revealed by the mixture-cluster analysis is represented by the quantiles of uncertainty (in Table 21),

which serve as a measure of the overall quality of this 10-component clustering (Fraley & Raftery, 1998). When the (100%) uncertainty estimates for this 10 component model are compared against the best BIC solution, it is shown that, despite the reduction in complexity that came with the EMMIX solution, the 16-component model has less uncertainty (0.6511021) and, presumably, greater goodness of fit.

Table 19: Summary of EMMIX analysis

Component Number	*Loglikelihood*	*-2log(lambda)*	*AIC*	*BIC*	*p-value*
1	54472.27	-	-108764.53216672339	-108240.22411942878	0.01
2	55773.41	2602.28	-111184.81344456675	-110130.37170500758	0.01
3	56409.14	1271.46	-112274.27596570041	-110689.70053387668	0.01
4	56870.65	923.03	-113015.30117413470	-110900.59205004641	0.01
5	57139.07	536.83	-113370.13610585516	-110725.29328950231	0.01
6	57346.87	415.61	-113603.74582938671	-110428.76932076931	0.01
7	57469.1	244.45	-113666.19267383407	-109961.08247295211	0.04
8	57664.51	390.84	-113875.02826827760	-109639.78437513109	0.01
9	57799.1	269.16	-113962.19092244512	-109196.81333703405	0.01
10	58001.87	405.56	-114185.74722697535	-108890.23594929972	0.01
11	58058.92	114.09	-114117.84101788023	-108292.19604794004	0.77

Table 20: Classification table

Cluster Number	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Number Of Allocations	284	379	86	106	239	259	338	205	152	456
Percent Representation	11.34	15.14	3.43	4.23	9.54	10.34	13.50	8.19	6.07	18.21

Table 21: Quantiles of uncertainty

Quantile	*0%*	*25%*	*50%*	*75%*	*100%*
Estimate	0	0.020775	0.085750	0.245650	0.697900

In order to evaluate this 10 component cluster solution in greater detail, the individual cluster memberships were tabulated by population and sex (Table 22 and Table 23).

Table 22: Cluster memberships by population (females)

Population	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*Grand Total*
Females	130	209	43	53	113	118	110	104	64	212	1156
AINU	1	9	0	1	4	1	1	4	2	15	38
ANDAMAN	4	3	0	1	0	5	0	2	2	18	35
ARIKARA	0	4	0	1	8	3	8	0	1	2	27
ATAYAL	2	2	1	0	2	2	2	1	1	5	18
AUSTRALIA	10	3	4	0	0	1	1	21	2	7	49
BERG	5	10	3	0	0	5	14	0	8	8	53
BURIAT	0	4	0	17	6	1	21	1	3	1	54
BUSHMAN	12	0	8	0	1	2	1	6	13	6	49
DOGON	12	2	3	0	0	2	1	7	7	18	52
EASTER ISLAND	3	1	0	0	1	11	1	2	1	17	37
EGYPT	2	17	0	0	10	7	5	2	0	10	53
ESKIMO	0	8	0	17	26	2	0	2	0	0	55
GUAM	1	10	0	0	3	1	6	1	0	5	27
HAINAN	4	9	0	1	1	10	3	0	1	9	38
MOKAPU	0	12	0	1	10	14	0	1	0	11	49
MORIORI	2	21	0	1	11	3	12	1	0	0	51
N JAPAN	2	8	0	2	4	1	8	1	0	6	32
NORSE	2	11	3	2	9	4	11	4	2	7	55
PERU	1	25	1	7	7	10	1	0	1	2	55
S JAPAN	1	18	0	1	1	3	7	0	0	10	41
SANTA CRUZ	3	24	0	1	1	18	0	0	1	3	51
TASMANIA	7	0	12	0	0	1	0	19	0	3	42
TEITA	22	1	3	0	0	1	1	1	12	9	50
TOLAI	15	0	2	0	0	1	2	19	0	15	54
ZALAVAR	2	7	2	0	7	6	3	2	3	13	45
ZULU	17	0	1	0	1	3	1	7	4	12	46

Table 23: Cluster memberships by population (males)

Population	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*Grand Total*
Males	154	170	43	53	126	141	228	101	88	244	1348
AINU	7	6	1	1	5	2	4	7	0	15	48
ANDAMAN	4	7	0	0	1	7	2	0	2	12	35
ANYANG	6	2	0	0	0	13	5	1	2	13	42
ARIKARA	1	1	0	0	7	4	25	1	0	3	42
ATAYAL	7	4	0	0	0	4	5	0	0	9	29
AUSTRALIA	7	0	3	0	1	0	4	26	6	5	52
BERG	7	9	0	1	1	2	23	1	7	5	56
BURIAT	1	3	0	15	1	1	26	1	6	1	55
BUSHMAN	6	0	13	0	0	0	0	9	11	2	41
DOGON	21	2	2	0	1	3	3	3	6	6	47
EASTER ISLAND	5	0	0	0	1	11	3	0	2	27	49
EGYPT	1	10	0	0	11	11	9	3	3	10	58
ESKIMO	0	0	0	17	30	4	1	1	0	0	53
GUAM	0	9	0	1	1	2	8	0	0	9	30
HAINAN	8	1	0	1	1	8	15	1	2	8	45
MOKAPU	0	15	0	0	9	10	4	1	1	11	51
MORIORI	0	14	0	6	15	2	18	1	0	1	57
N JAPAN	1	10	0	4	9	5	14	0	3	9	55
NORSE	4	11	0	0	8	3	8	1	7	13	55
PERU	3	20	0	1	4	9	7	1	2	8	55
PHILIPPINES	8	3	0	0	1	7	6	1	4	20	50
S JAPAN	4	17	0	0	4	9	6	0	0	10	50
SANTA CRUZ	1	11	0	4	3	13	12	0	1	6	51
TASMANIA	11	1	15	0	1	0	1	6	3	7	45
TEITA	10	0	1	0	1	2	5	3	7	4	33
TOLAI	6	2	6	0	0	1	1	25	4	11	56
ZALAVAR	3	8	1	1	10	5	10	2	3	10	53
ZULU	22	4	1	1	0	3	3	6	6	9	55

Individual posterior probabilities of group membership were assessed to estimate implied classification error: the number of correct allocations by population are included in Table 24. On average, it is shown for this 10 component model that classification success is reasonably high (80% at >0.7 threshold). Even with the change in component number (viewed either as a reduction from 16 or an increase from 3), these "correct" allocation rates are relatively consistent with the previous findings from "good" the MCLUST models (ranging between 76% and 86%, for 3 to 16 components).

The percent estimates reported in Table 24 represent only "implied" rates of success. As such, these percent estimates are determined directly from the mixture analysis and without consideration of true individual memberships. It is helpful in this case to examine the individual cluster compositions and population-specific contributions for patterns of relationship that may hold meaning for discussions of global population [sub]structuring. To maintain consistency between the MCLUST and EMMIX analyses, two series of pie charts were produced here to serve as graphical representations of the 10-component clustering and to facilitate pattern. Using these visualizations as a guide, a summary explanation of the model-based clustering findings for this "EMMIX" 10-component model is provided in the discussion that follows.

In keeping with the earlier treatment of the 16-component clustering and the promising patterns that were uncovered therein, this discussion attempts to identify similar patterns of population affinity using the EMMIX 10-component model. This review operates under the assumption that that within this 10-cluster solution there are patterns that can be objectively organized into meaningful groups of "like" populations, which, in turn, reflect shared biogeographic ancestry or population history.

Table 24: Allocation rates

Population (Howells 1989)	*posterior probability > 0.7*	*percent correctly allocated*	*posterior probability → 1*	*percent correctly allocated*	*total number of individuals*
AINU	65	76	30	35	86
ANDAMAN	57	81	36	51	70
ANYANG	33	79	24	57	42
ARIKARA	55	80	36	52	69
ATAYAL	37	79	22	47	47
AUSTRALIA	85	84	67	66	101
BERG	88	81	64	59	109
BURIAT	95	87	75	69	109
BUSHMAN	75	83	58	64	90
DOGON	74	75	48	48	99
EASTER ISLAND	71	83	48	56	86
EGYPT	88	79	47	42	111
ESKIMO	97	90	83	77	108
GUAM	39	68	25	44	57
HAINAN	69	83	34	41	83
MOKAPU	82	82	55	55	100
MORIORI	86	80	53	49	108
N JAPAN	64	74	43	49	87
NORSE	84	76	50	45	110
PERU	87	79	52	47	110
PHILIPPINES	38	76	21	42	50
S JAPAN	63	69	37	41	91
SANTA CRUZ	79	77	47	46	102
TASMANIA	76	87	64	74	87
TEITA	70	84	53	64	83
TOLAI	94	85	67	61	110
ZALAVAR	76	78	48	49	98
ZULU	78	77	59	58	101
Grand Total	2005	80	1346	54	2504

This discussion begins with the populations presenting the most easily identifiable and consistent patterns of cluster membership: Howells' core African (sub-Saharan), Bushman ("local" African), and Australo-Melanesian series. This "African/Australo-Melanesian" grouping is unified by the percent contributions of each member population to clusters 1, 8 and 10; but it is the presence of cluster 8 that makes these populations identifiably unique. Perhaps more interestingly, is the fact that it is possible to also detect two ordered subsets within this grouping. These second level differences work to divide the Africans from the Australo-Melanesians: (1) Dogon, Zulu, Teita, and (2) Australia, Tolai and Tasmania (Figure 46). The Bushman series, though South African in geographic "origins" but variably sourced by Howells, appears as the mediating group between these two subsets (Figure 47). Cluster 8 is found in greatest proportion among the Australians, Tolai and Tasmanians. Howells' (1989) core African populations are set apart by their smaller proportion of cluster 8 and the larger proportion of cluster 9; while cluster 9 appears to be replaced by cluster 3 in the Australo-Melanesians. It is interesting to observe that the Tasmanians are the single greatest contributor to cluster 3 (31%), across all 28 populations. They are followed in similarly high quantities not by the Australians (8%) or the Tolai (9%), as it might be expected, but by the Bushmen (24%). It is notable that the Bushmen share a high proportion of membership in both cluster 3 (24%) and in cluster 9 (16%). This pattern (Figure 47) may suggest that this complex "local" series occupies some space between these second-level groups and that the clustering algorithm is identifying this instance of morphological overlap. While these conditions of a combined African and Australo-Melanesian membership are seemingly well-defined, it is important to acknowledge that, following this criterion, this composite grouping should also include the Ainu, as almost half of this series was allocated to clusters 8 and 10 (Figure 47). This is clearly problematic. While their cluster allocations recall aspects of the African/Australo-Melanesian relationships, the Ainu have their own pattern of membership. They differ from the African and

Australo-Melanesians by their presence in clusters 2 and 5, which are largely populated by Asian and indigenous populations of presumably Asian origins. While this "atypical" inclusion confounds a neat explanation of the cluster patterns, this association is interesting, in light of Howells' (1989, p. 15) summary statements on the craniofacial morphology of the Ainu: "In many figures, Ainu clearly depart from Europeans and Australoids, further nails in the coffins of two hypotheses of Ainu affiliations."

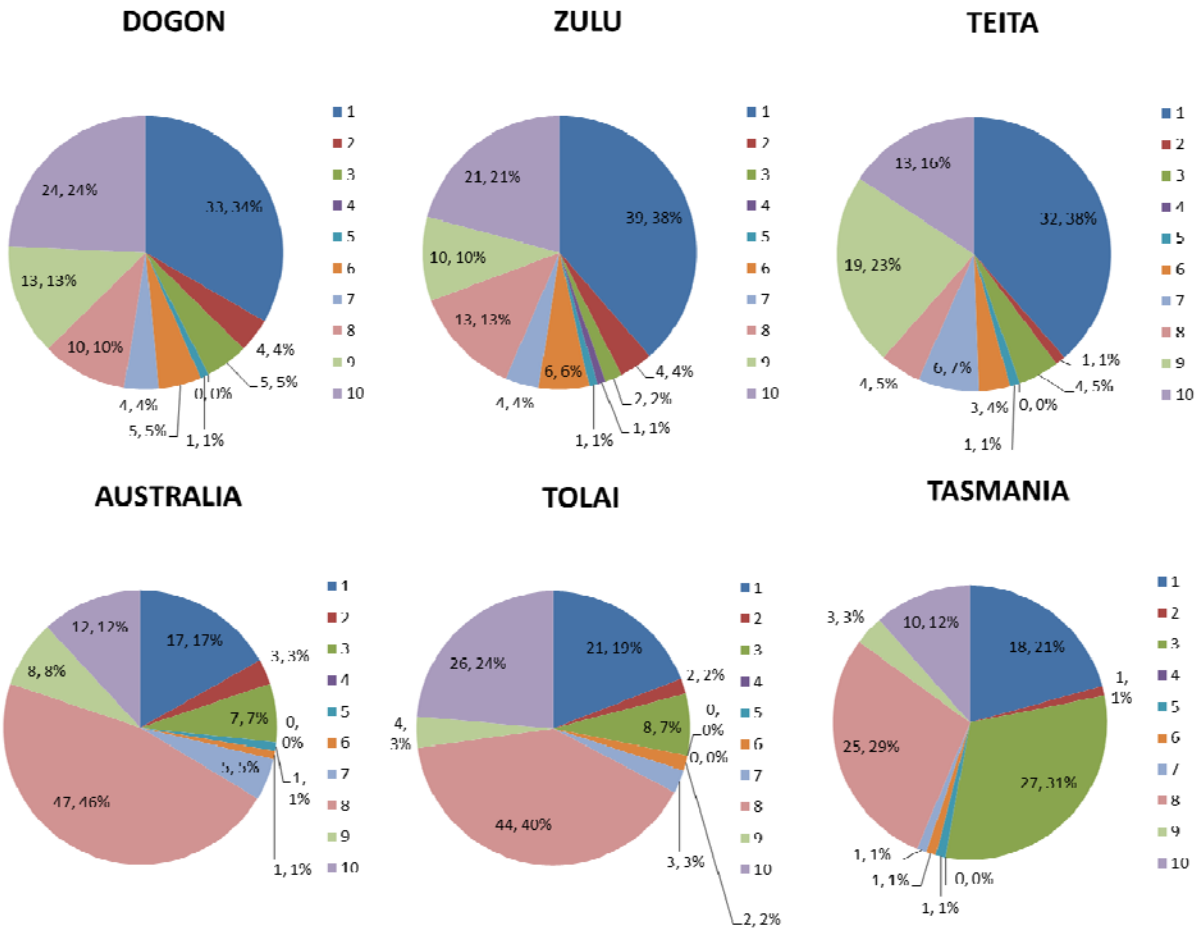


Figure 46: Dogon, Zulu, Teita, Australia, Tolai, Tasmania

The second standout grouping in the EMMIX 10-component analysis includes the Eskimo and Buriats: the most "extreme and isolated" of Howells' series (1989, p. 15). These two "particular" but "like" populations are linked most clearly by cluster 4, in which they appear in almost equal proportions (at a total of 62%). They also share in a similar, but small, amount of cluster 2, which is weighted towards Asian, Island and Native American populations. The major difference in allocation is the large presence of Eskimo in cluster 5 (57 individuals) and Buriat in cluster 7 (47 individuals). It is worth further noting that there are broadly geographic connections between the other populations that make-up these two clusters: the N Japanese, Moriori and Norse are included, while the African and Australo-Melanesians are very weakly represented. These findings reinforce the shared affinity within these special segments and also speak to broad, likely, evolutionary divisions, e.g., African, non-African.

While the relationships between the African and Austro-Melanesians and between the Eskimo and Buriats have been already demonstrated in the discussions of the 16-compoent mixture and were played out again in this EMMIX solution, it is here that the easy correspondence between these two (16 and 10 cluster) models ends. It appears that the reduction in the number of components has come at the expense of interpretive clarity. In this 10-component solution, there are no distinctively Asian, Native American or Polynesians configurations that warrant individualized discussion. Only when these groups are examined as a full set, is it possible to detect similarities (albeit, very broadly-based) in their patterns: there is, for example, some consistency in these groups found in cluster 2 (Figure 49) and cluster 6 (Figure 50), with individual contributions > 15%.

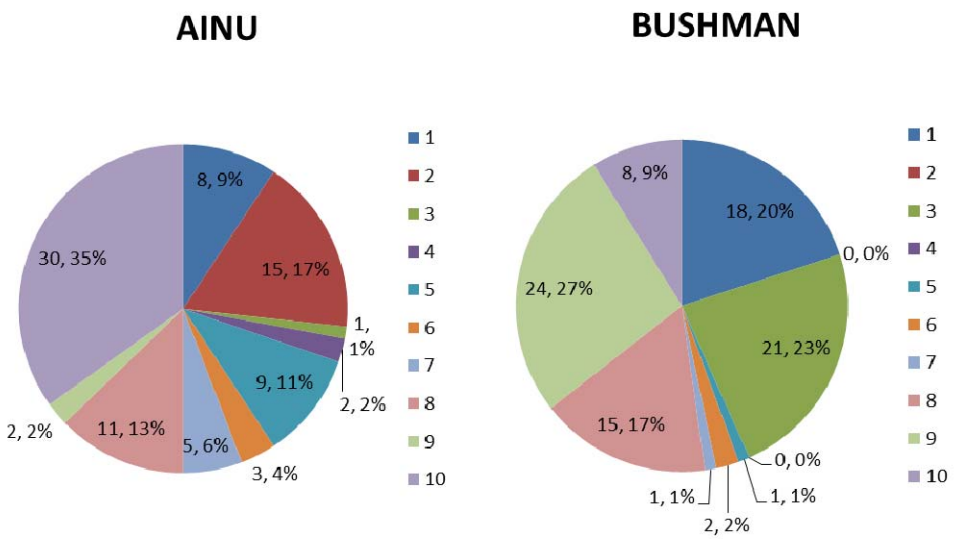


Figure 47: Ainu, Bushman

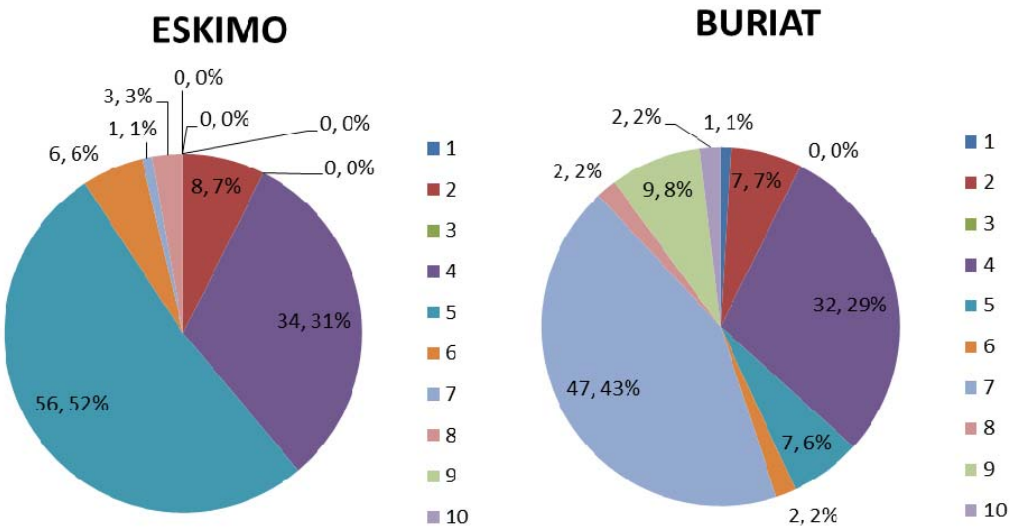


Figure 48: Eskimo, Buriat

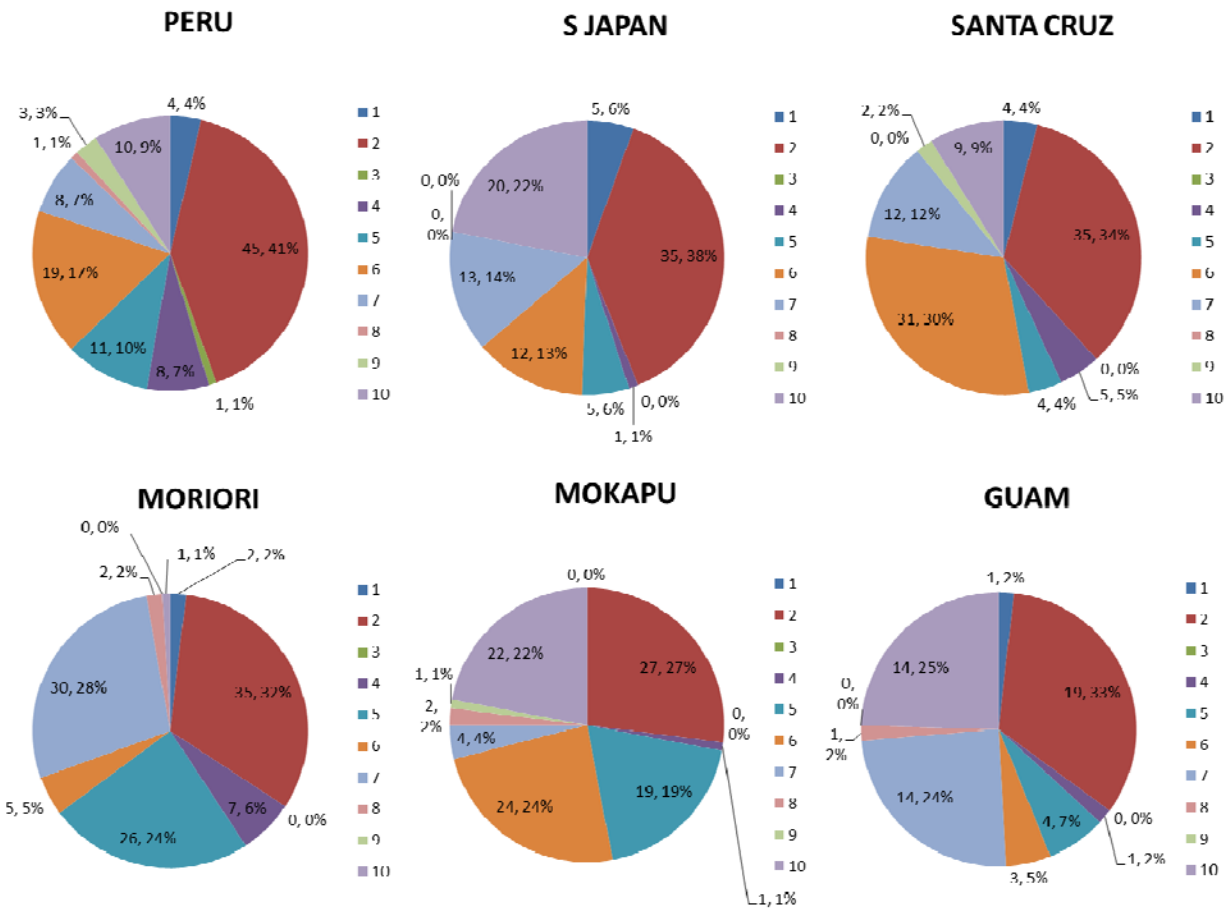


Figure 49: Peru, South Japan, Santa Cruz, Moriori, Mokapu, Guam

This problem of generalization in membership profiles is exacerbated by the fact that Howells' core Northern European groups (including Egypt) are found in many of the same clusters and in proportions equal to these largely "Asian" populations (Figure 51). These widely shared "similarities" produce a homogenizing effect, which in turn suppresses population level discrimination. The limited structure within these populations supports, therefore, a single highly mixed (European, Asian, Native American and Polynesian) but weakly differentiated mega-population.

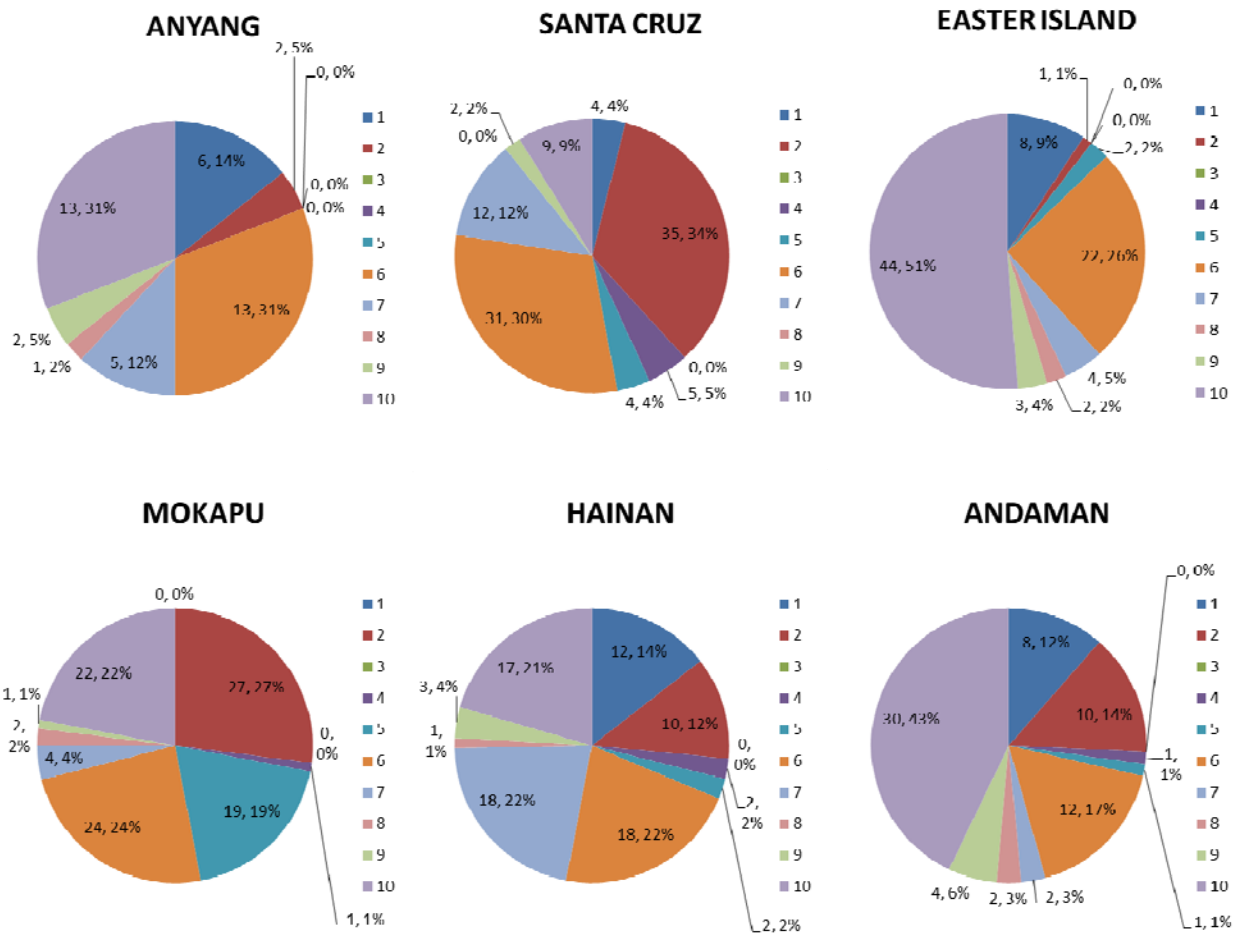


Figure 50: Anyang, Santa Cruz, Easter Island, Mokapu, Hainan, Andaman

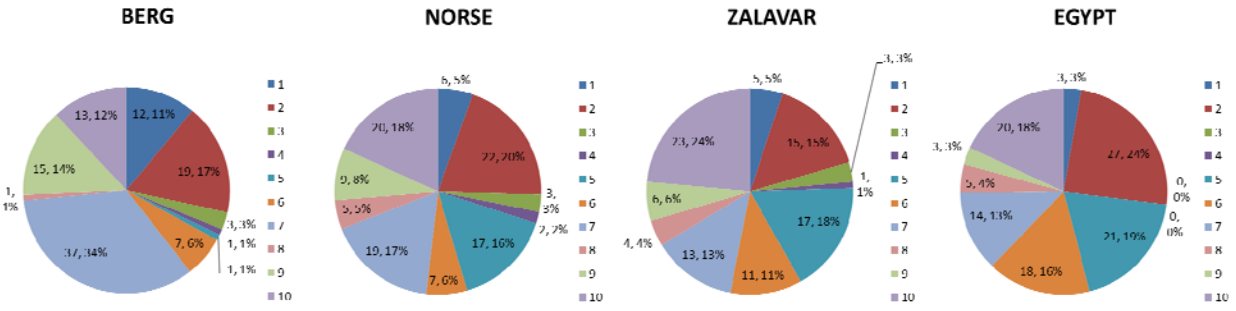


Figure 51: Berg, Norse, Zalavar, Egypt

This distinction between the African/Australo-Melanesian group and all other continental populations as well as the special isolation of the highly atypical Eskimo and Buriat series are supported by the composition of each the 10 clusters. Clusters 1, 3, 8 and 9 (Figure 52) are characterized by African and/or Australo-Melanesian populations: collectively these populations represent at least half of each of these four clusters. The Eskimo and Buriat clearly dominate cluster 4, making up 62% of the total allocations (Figure 53). While they also contribute to clusters 5 and 7 in seemingly large numbers, the Eskimo and Buriat ultimately reflect only a small fraction of the total cluster composition (< 25%) given the sheer size cluster 5 (n = 239) and 7 (n = 338). Clusters 2, 5, 6, 7 and 10 include a mix of Howells' remaining series in varying and largely uninformative degrees.

It is seen here, in a way that it was not for the 16-component solution, that the majority of Howells' populations share key aspects of their cluster memberships with many other groups, and do not otherwise "fit" a specific pattern of similarity (i.e., forming distinctive geographic groupings that define continental or regional clusters). While these findings are discouraging for this project's research questions that focus on identifying relationships at the level of the population, they nevertheless provide support for the utility of finite mixture analysis to respond to the data available and the conditions imposed, and to reveal meaningful cluster information that has information value outside of the model's own "statistical significance." If the picture of relationships presented by the 10-component model is generalized, the EMMIX clustering was most successful at revealing substructuring within the African populations: it enabled the distinction of the African populations from the Australo-Melanesian group and produced a pattern of clusters and classifications that reflected a clear African and non-African divide. These findings correspond well with the evidence that traces the trajectory of human

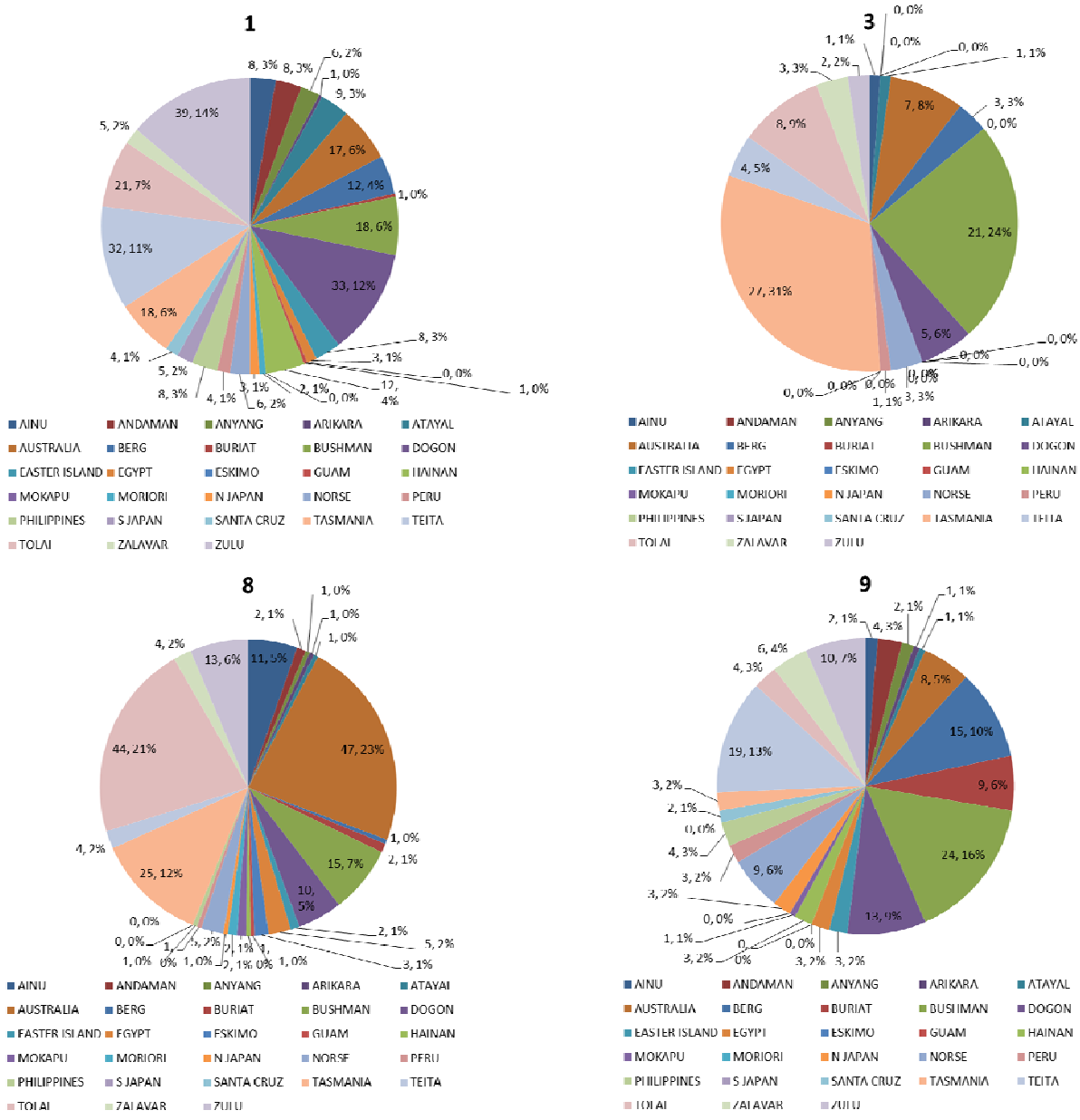


Figure 52: Clusters 1, 3, 8 and 9

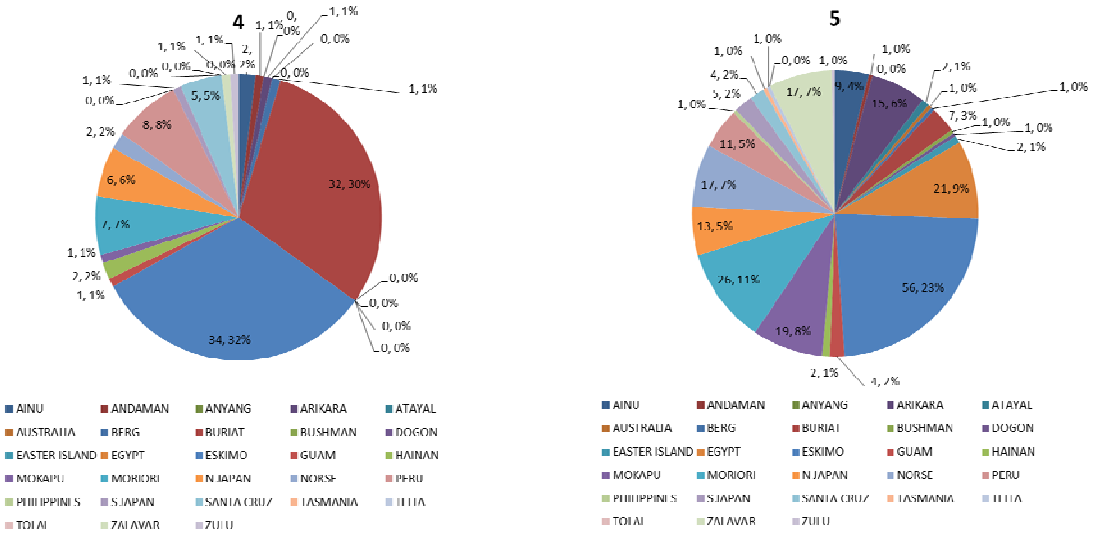


Figure 53: Clusters 4 and 5

evolution out of Africa: (1) African populations are known to exhibit greater genetic diversity than do all other globally distributed populations; (2) molecular evidence suggests that anatomically modern humans emerged first in Africa and later colonized Eurasia and the Americas; and (3) there is a close correspondence in migration and divergence dates between the genetic and the bioarchaeological findings (Jorde, Bamshad & Rogers, 1998; Jorder et al., 2000; Tishkoff & Williams, 2002; Yu et al., 2002; Tishkoff & Verrelli, 2003). These cluster findings demonstrate how the EMMIX mixture methods have effectively teased out the underlying patterns of considerable ancestry using limited craniometric data. Consider, for example, that one the of the earliest modern skeletons recovered outside Africa was found in Australia, dating to 42 kya (Bowler et al., 2003); from this evolutionary perspective, it makes good sense that the clustering algorithm consistently associated African and Australo-Melanesians.

The overall assessment of this EMMIX model suggests that the 10-component solution provides useful information on population relationships. These promising findings, in turn, speak to the potential utility of finite mixture analysis in answering questions of human biological variation. Despite the change in

model selection methodology and component number, the EMMIX analysis of Howells' data produced results that are similar in their statistical properties as those returned by the "good" MCLUST solutions (see Table 8). There are, however, points of weakness in cluster memberships that have affected the ease of interpretation of this 10-cluster solution. These problem points are not as readily apparent (if at all present) in the more complex 16-cluster model. The model choice based on the smallest number of compatible components, while statistically valid, may not in fact be the best approach for craniometric data and the type of structure that this project seeks to reveal. These findings suggest that while model quality may remain consistent between approaches, the particular philosophic underpinnings that support the procedures that choose and define the models may have consequences for the ability of the model-based cluster analysis to resolve the mixture into clusters that are both (1) statistically strong (well separated, with internal cohesion) in the presence of expected "clinal" overlap and (2) the most intuitively useful to population-level studies of globally distrusted variation.

While this analysis has demonstrated that statistical choices, of, in this case, model selection protocol, can influence the outcome of the model-based clustering by imposing conditions that are independent of the data itself, it has equally shown that, despite the variability in the "optimal" component number that may accompany these choices, it is possible to detect clustering trends that remain consistent across the different implementations of the mixture algorithms (e.g., in MCLUST or EMMIX). It has been assumed, thus far, that the patterns observed reflect population structure, i.e., differences in biogeographic ancestry. It is important to also recognize that craniofacial morphology not only encodes population history but also sex. As it has been discussed already, this project has chosen to account for sexual dimorphism in its analyses following a consistent protocol: the sexes are pooled and a size correction is applied (after Darroch & Mosimann, 1985). Although sex is closely tied to size, "correcting"

for sexual dimorphism neither identifies nor resolves the potential effect sex may have on the ways in which the cases are clustered. For this reason, it is advisable to make use of Howells' sex estimates in exploratory analyses using sex-specific, scaled and unscaled data [sub]sets, in order to determine if any of the patterns in allocation are related to trends in male or female cranial shape. These sex-specific morphologies, if present and unaccounted for, may obscure the true population-based structure within a given data set, regardless if an appropriate size correction has been applied.

The test of sex specific morphology begins by first evaluating the most parsimonious MCLUST model with 3 components for obvious male and female contributions within the cluster patterns. Recall that this model was developed using 2504 individuals from all 28 populations, and by pooling the sexes and scaling for size. This simpler model is considered in this case over the 16-component ("optimal") clustering as it can be broadly related to the three major continents. This macro-geographic analysis assumes an acceptable level of specificity when making comparisons between groups by sex since the true sexes are not known or with the same degree of accuracy for all of the samples. When the 3 cluster solution was divided up into female (Figure 54) and male clusters (Figure 55), both the males and female continental patterns were found to be almost identical. These similarities were supported by the exploratory data (EDA) approach that was previously adopted to externally validate this project's model-based protocol and search for potentially problematic trends within the clustering, across several different runs (c.f., Young, Valero-Mora & Friendly, 2006). In this case, it can be seen from simple scatterplots of cluster number by size and sex, using both the MCLUST 16-component model (Figure 56) and the EMMIX 10-component model (Figure 57), that (1) males and females are distributed in expected proportions across all clusters, (2) they occupy different positions along the range of variation within each clusters, and (2) no one cluster is capturing exclusively male or female cases.

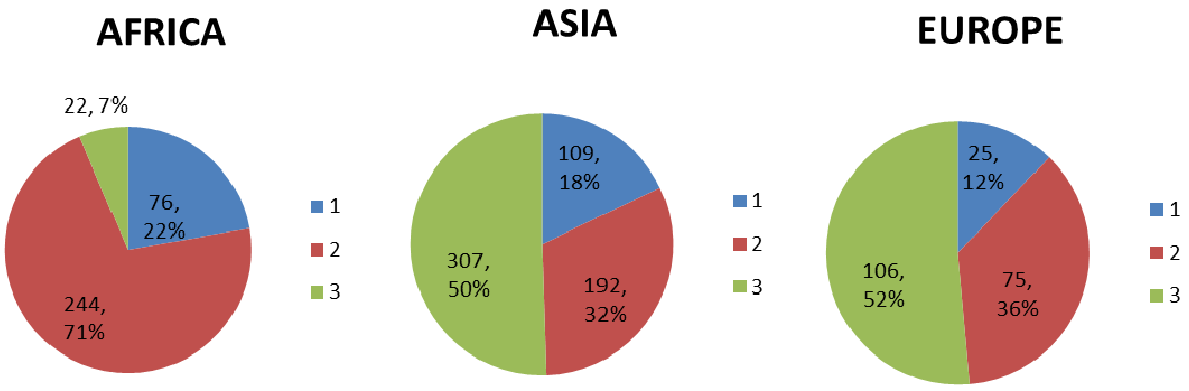


Figure 54: Continental clusters (females)

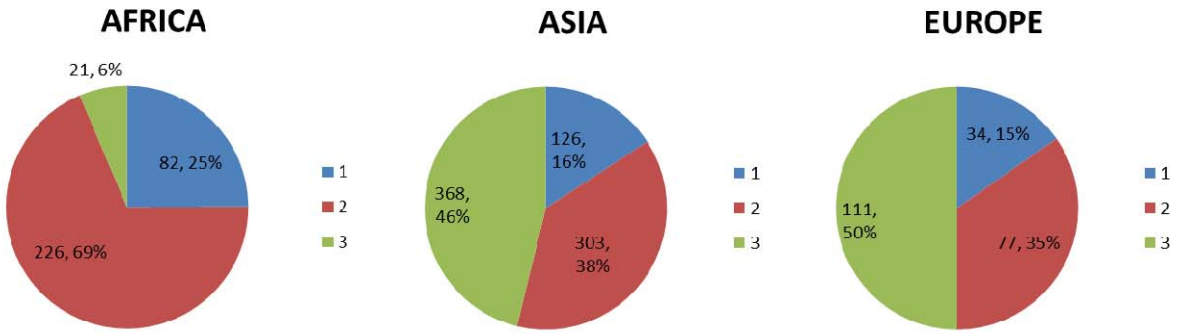


Figure 55: Continental clusters (males)

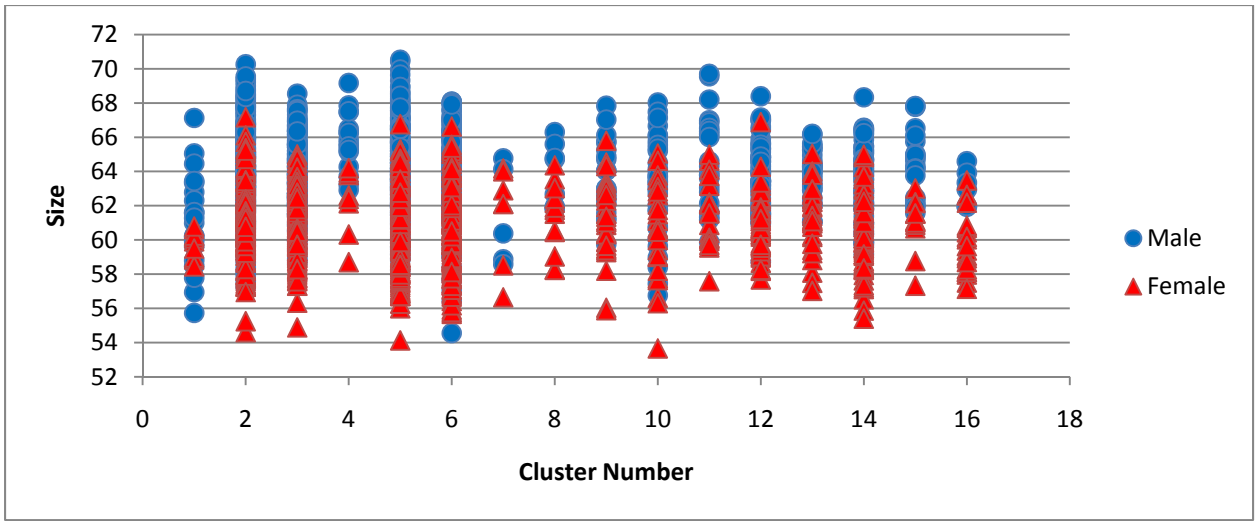


Figure 56: Scatterplot of size and sex (MCLUST G=16)

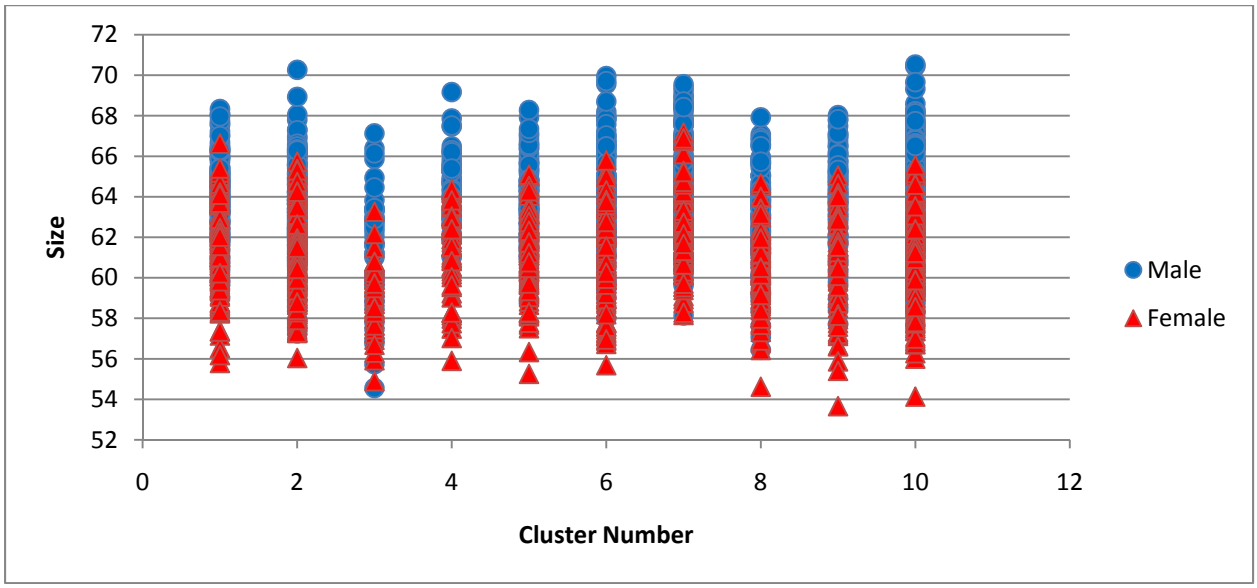


Figure 57: Scatterplot of size and sex (EMMIX G=10)

This consistency in clustering demonstrates that males and females are cohesive in terms of their population-based morphology. These findings not only offer good support for this project's choice to conduct size-corrected tests using pooled sexes, but they are also interesting for the fact that they agree with recent studies of cranial differences at higher taxonomic levels using the more sensitive techniques of geometric morphometry. Fleagle, Gilbert and Baden (2010) conducted craniometric tests of sex differences across primates and with a special interest in human variation. They found for their primates samples in general that the results of female analyses and the pooled-sex analyses looked remarkably similar.

In order to assess independent clustering patterns, separate male (n = 1348, 28 populations) and females tests (n = 156, 26 populations) were conducted, using scaled and unscaled subsets of Howells' data (1989). These MCLUST results are summarized in Table 25 .

Table 25: Summary of sex-specific clustering

Sex	*Treatment*	*Component Number*	*Model Form*	*Maximum BIC value*	*Quantile of Uncertainty (100%)*
Female	scaled	8	EEE	50987.51	0.61644231
Female	unscaled	1	MVN	-71294.85	-
Female	unscaled	3[15]	EEE	-71368.62	0.6214201693
Male	scaled	9	EEE	59206.53	0.62833296
Male	unscaled	4	EEE	-84948.54	0.6290058

[15] Here, the second best model was chosen for the unscaled female data set. It was felt that the increase in information value of the three components model over the model without identifiable group structure (i.e., 1 component) outweighed the slight change in the BIC value.

For the scaled female and male samples, MCLUST selected 8 and 9 component models respectively, using the BIC selection criteria (Figure 58); the coordinate projections of the classifications produced by these two "best" models are presented in Figure 59. As the pooled-sex analysis revealed already, there is clearly a close correspondence between the male and female cluster patterns in size, shape, and overlap (Figure 60, Figure 61). What is notable here is the addition of the 9th component in the male clustering. This imbalance between the male and female cluster solutions is important for several reasons: (1) it indicates that differences in male and female shape exist within an overall similar pattern of morphology, (2) it suggests that males display greater morphological diversity, and (3) since this analysis considered scaled data, it demonstrates that this increased variability in males cannot be adequately explained by size or sexual dimorphism alone. When these findings are taken together, they imply that, even after the effect of sexual dimorphism is "removed" by correcting for "size," males and females continue to differ in shape; but, they do so along the same range of their own population-specific variation, with females representing the most conservative morphology and males presenting the most extreme features. Similar observations have also been made by Fleagle, Gilbert and Baden (2010), who wrote that "a large portion of the diversity in anthropoid crania is driven by the male morphotypes, and it is the male anthropoids that typically display the most extreme scores" (p. 576). The scaled results produced here also suggest that craniofacial morphology more strongly encodes population signatures than sex. These patterns are, again, in line with the primate trends reported by Fleagle and colleagues (2010), from which they have concluded that "thus, the major morphological differences in the shape of primate crania are largely those that have characterized the evolution of different higher-level clades, such as strepsirrhines, anthropoids, cercopithecoids, and hominoids, rather than evolution within these clades" (, p. 570). The unscaled tests conducted here using the same data have conveniently offered corresponding evidence and, in turn, support for the "scaled"findings.

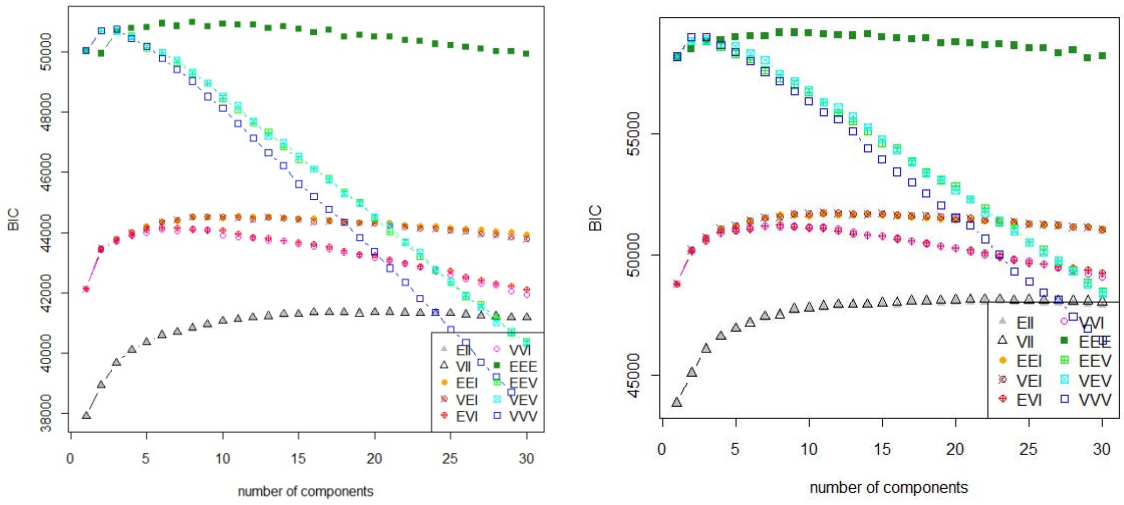


Figure 58: Model selection by BIC (females and males respectively)

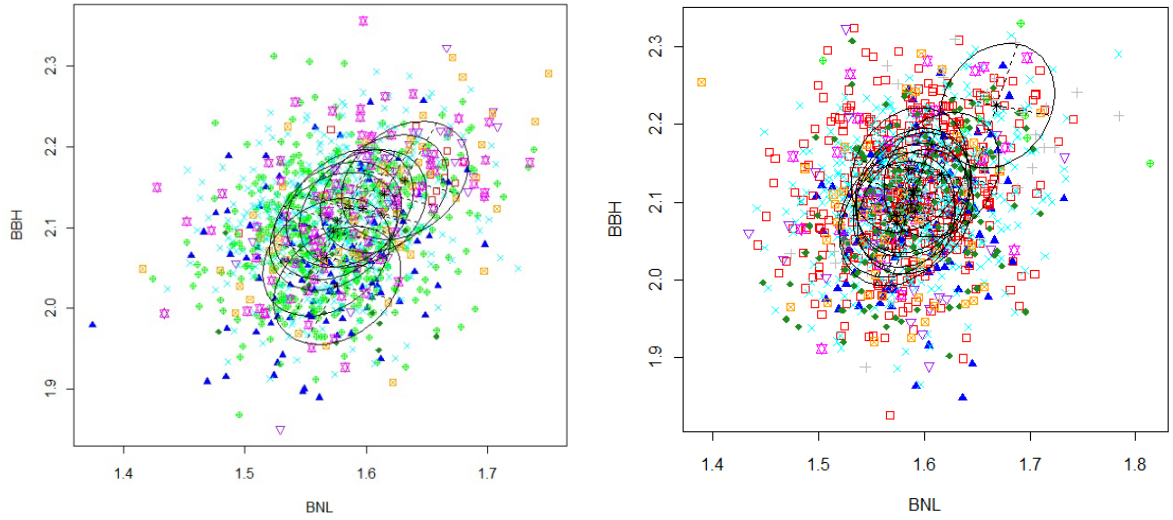


Figure 59: Coordinate projections of the best female and male model respectively

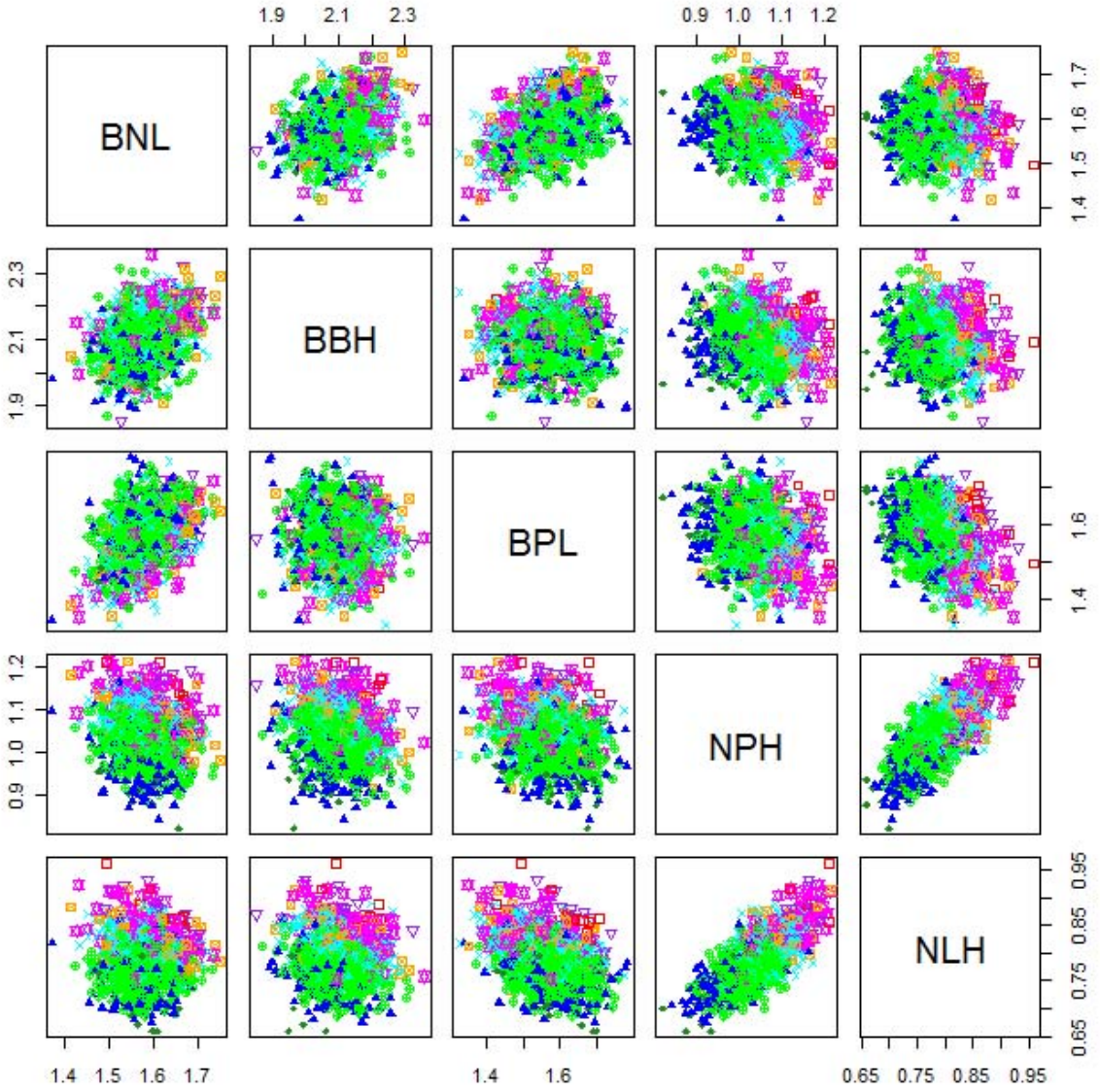


Figure 60: Pair-wise scatterplot matrix showing clusters (females)

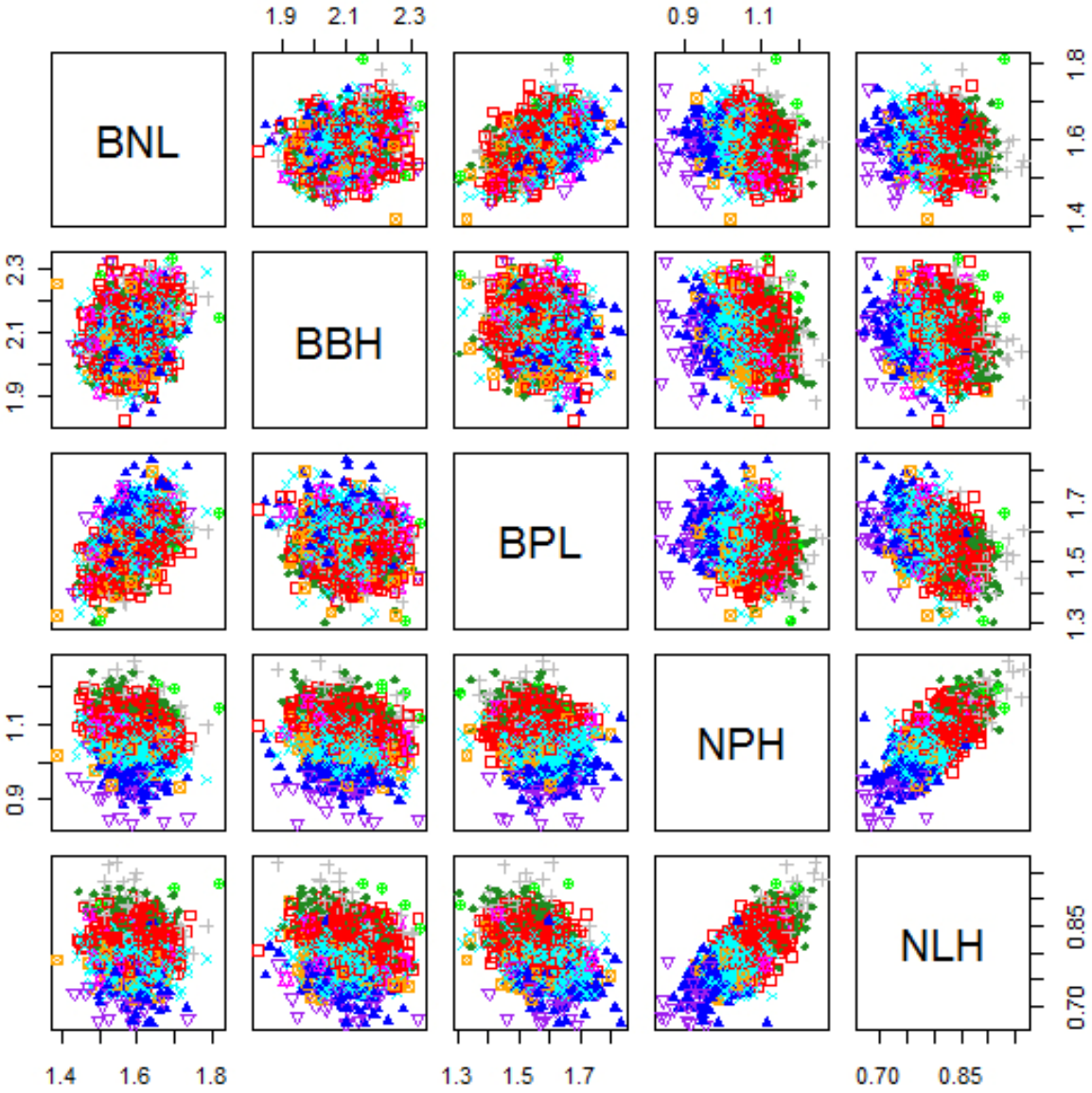


Figure 61: Pair-wise scatterplot matrix showing clusters (males)

For the unscaled female and male samples, MCLUST was used to identify the best BIC models: from these, a 3-component solution was selected for the female data set (the second "best" model) and the 4-component "optimal" model was chosen for the male data set (Table 25). Again, the model options returned by the selection procedure clearly show sex-based differences in the possible component numbers, which, in turn, imply different male and female cluster patterns. It is argued, therefore, that, when size is reintroduced into the mixture analysis, males continue to show greater variability than females and this extra diversity is captured by the inclusion of additional components in the male model (see Figure 62 and Figure 63). Among the males, it is possible to identify a pattern with reasonable clarity: cluster 2 is dominated by Africans, with a smaller Asian component; cluster 4 is dominated by Africans, with a smaller European component; cluster 1 contains a mix of most of the Asians and Europeans; cluster 3 is unique for the strong presence of Eskimo. Among the females, the cluster distinctions are far less refined, with considerable overlap across the clusters: all three clusters have significant Asian components, but cluster 1 is almost exclusively Asian, cluster 2 also contains a mix of Africans and most of the Europeans, and cluster 3 includes select African groups. What is both unique and noteworthy for these unscaled analyses is the obvious homogenizing effect of size: the reduction in component numbers (from 9 to 4 for the males and 8 to 3 [or 1] for the females), and in turn in model complexity, suggests a notable decrease in the quality of the separation of the groups. This weakening of data structure with the addition of size is especially remarkable among the female series. While this project chose to fit the 3 component solution for the unscaled female data set (for better comparability with the male clustering), the maximum BIC estimated for the unscaled female data set corresponds to a multivariate normal model with a single component: that is, the model indicative of a data set lacking any identifiable internal structure, for which the model-based clustering cannot statistically detect meaningful differences between the groups that lie within the mixture.

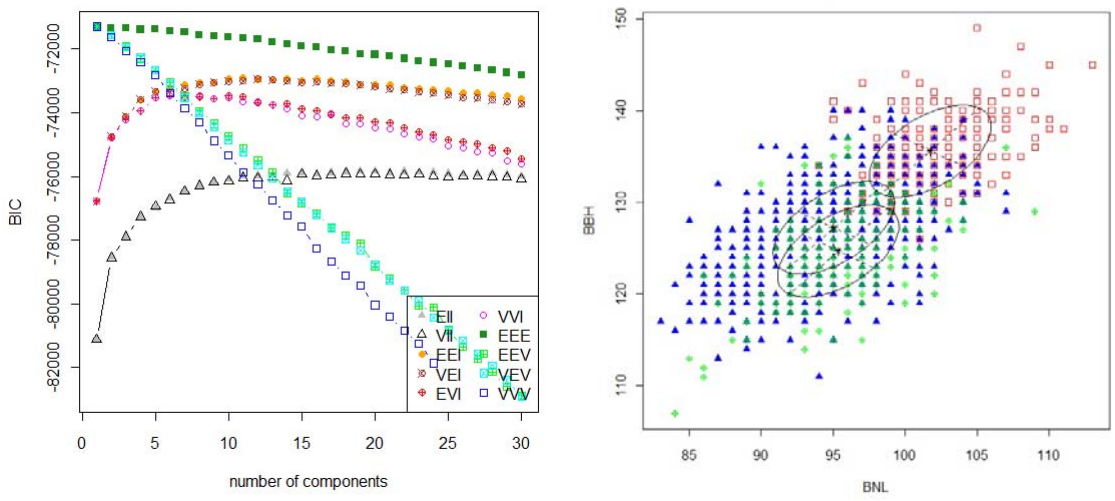


Figure 62: Best female model chosen by BIC

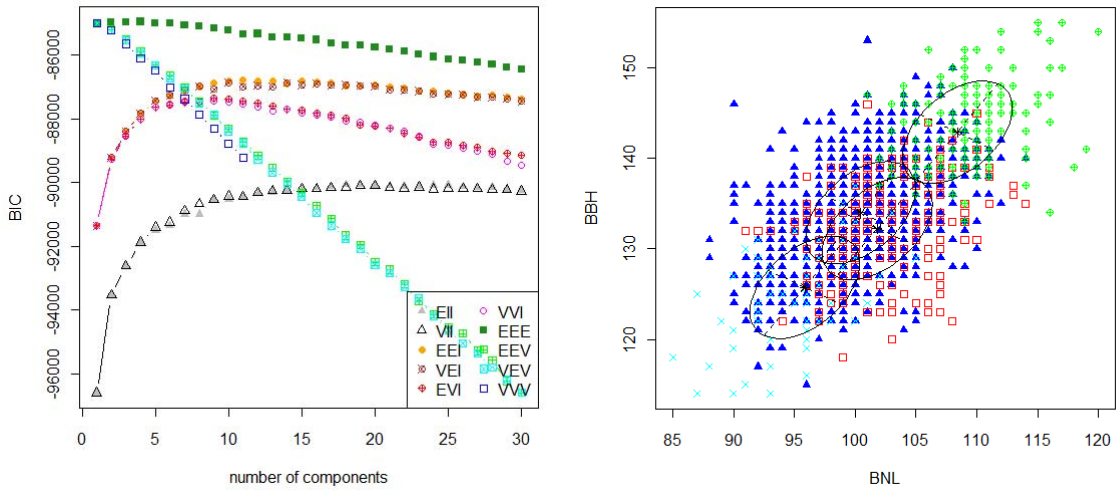


Figure 63: Best male model chosen by BIC

Three key points can be taken from these analyses: (1) there is a clear difference between the cluster patterns obtained using the separate-sex unscaled data and those produced under the size correction (i.e., size tends to mask among population differences, especially among females), (2) there are parallels between the behavior of the female and male clustering under similar conditions (e.g., the introduction of size reduces heterogeneity between populations in both sexes), and (3) the addition or removal of size does not erase the fundamental difference between the female and the male morphological pattern (i.e., males have greater variability, this variability often represents pooled-sex extremes, males define the range of population-specific morphology into which the females fall). These findings call into question the common practice of restricting studies by sex in order to circumvent the methodological issue of correcting for size, as it appears that the presence of size itself may be a far greater confounding factor than any statistical problems that may accompany the application of an appropriate correction technique for sexual dimorphism. This project takes these findings as support for the utility of its chosen size correction method and as justification for the inclusion of both sexes in its analyses: sex adds much needed variability and it does not "bias" the formation of the clusters.

Perhaps the most consistently debated question of methodological protocol, in and out of the field of anthropology, is variable selection. This is expectedly so because the configuration of a data set has the greatest and most immediate repercussions on an analysis. The importance of good variable, like model, selection practices is especially true now, as data storage options and processing power continue to expand and samples sizes grow. As Guyon and Elisseeff (2003) have already observed: "The situation has changed considerably in the past few years...most papers explore domains with hundreds to tens of thousands of variables or features...new techniques are proposed to address these challenging tasks involving many irrelevant and redundant variables and often comparably few training examples" (p.

1157). Following good variable selection procedures has many well-known advantages. In practice, it simplifies data visualization and facilitates data understanding, reduces data collection and storage requirements, and decreases processing and analysis times. In theory, it resolves the problem of dimensionality and improves prediction performance. There are two common approaches to variable selection, each of which has its own shortcomings. It is possible to identify subsets of variables that are "useful" to building a good predictive model; however, with this approach must come the knowledge that culling decisions may lead to the exclusion of potentially (possibly even highly) informative variables. Or, it is equally possible to select (and choose to "filter" by rank) all "relevant" variables, at the risk of redundancy (c.f., Kohavi & John, 1997; Blum & Langley, 1997). While in general, studies on feature selection for pattern recognition and machine learning have argued that "selecting the most relevant variables is usually suboptimal for building a predictor, particularly if the variables are redundant," tests of this position have shown that the addition of seemingly redundant variables can enhance model performance by reducing noise and increasing group separation. When these variables are perfectly correlated they are, however, "truly redundant" and they do not provide additional information and improve model fit or utility (Guyon & Elisseeff, 2003, p. 1158, 1163-1164). These conflicting findings recommend further investigation.

In the context of skeletal biology, many of these studies that choose to employ a large suite of anthropometric variables (often the maximum available) in their analysis of population differences engaged in model building using "relevancy" criteria (c.f., Relethford, 1994, 2001, 2009; Pontikos, 2004; Hubbe & Neves 2007; Elliott 2008; Strauss & Hubbe 2010). Of these studies that focus on craniometrics, the findings reported by Pontikos (2004) are of special interest here. On his anthropology blog, Pontikos has posted a model-based cluster analysis of Howell's full data set (1996) using all 57 metric variables. It

is worthwhile to revisit Pontikos' conclusions, which state that "fourteen separate clusters were identified in the best run, each of which corresponds strongly to either one of the original populations, or to a racial group. It is shown that cranial variation can be used to infer ethno-racial affiliation." This project has chosen, therefore, to [re]evaluate *all* of Howells' data (1989) for evidence of [sub]structuring using model-based clustering: this analysis was implemented with MCLUST and it considered (1) 2486 individuals distributed over 30 populations, (2) pooled sexes, and (3) the full complement of 57 size-corrected metric variables.[16] In addition to its "test" value, this secondary analysis will also enable, perhaps more importantly, a direct comparison between the clustering results produced under the conditions of few and many predictor variables.

This discussion of the 57 variable analysis most profitably begins by, first, a review of the model selection output. Table 26 presents the top ten models identified by MCLUST using the BIC selection criteria: the optimal model is defined as ellipsoidal, with equal covariance matrices (EEE) and 15 components (see Figure 64). Also provided below is an overview of the clustering patterns that define this 15-component model. The raw by-population counts for the number of individuals assigned to each of the 15 clusters are reported in Table 27. These memberships are also graphically displayed using both coordinate (1,5) and random seed projections (Figure 65) and a matrix of pair-wise scatter plots that presents the first five of the 57 variables in the data set (Figure 66).

[16] In this case, data for the N and S Maori (n = 20) from New Zealand was added to the standard Howells data set tested here, as this sample is "of clearly known Polynesian derivation" (Howells 1995, p. 14) and can be reliably used to augment representation in that region and supplement the loss of cases due to missing value deletions.

Table 26: Top 10 models in order

Component Number	*Model Form*	*BIC value*
15	EEE	476917
13	EEE	476874
17	EEE	476860.1
16	EEE	476849
11	EEE	476745.4
14	EEE	476739.2
12	EEE	476722.4
10	EEE	476684.3
18	EEE	476614.8
19	EEE	476467.6

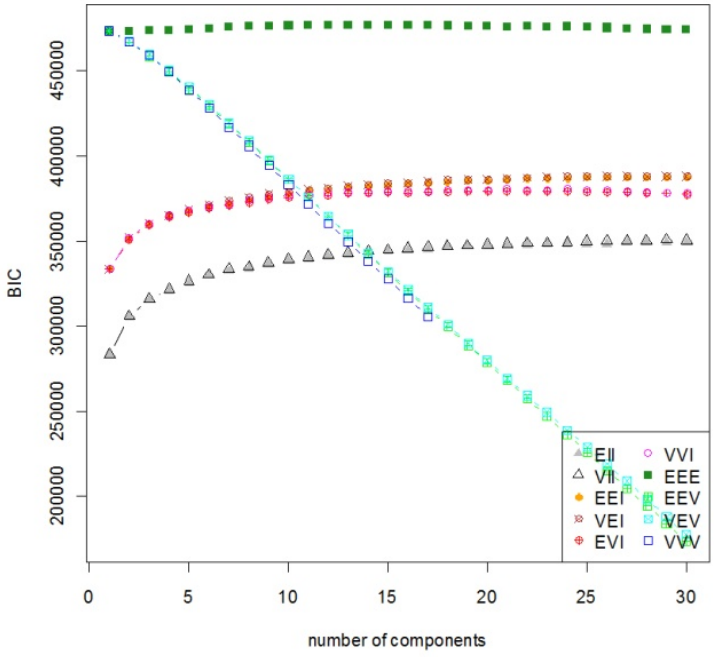


Figure 64: Model selection by BIC

Table 27: Cluster memberships by population (pooled sexes)

Populations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
AINU	0	12	69	0	1	0	0	2	0	1	0	1	0	0	0	86
ANDAMAN	3	4	2	0	0	0	59	2	0	0	0	0	0	0	0	70
ANYANG	1	40	0	0	0	0	0	0	0	1	0	0	0	0	0	42
ARIKARA	0	12	1	33	0	6	0	1	7	8	0	0	0	0	0	68
ATAYAL	0	42	1	0	2	0	0	0	0	1	0	0	1	0	0	47
AUSTRALIA	0	0	0	0	0	0	0	1	0	0	2	0	98	0	0	101
BERG	1	0	0	0	0	94	0	10	4	0	0	0	0	0	0	109
BURIAT	2	1	1	0	0	0	0	0	0	0	0	0	0	104	0	108
BUSHMAN	29	2	5	0	49	0	0	3	0	0	0	0	1	0	0	89
DOGON	2	1	0	0	89	0	0	1	0	0	0	0	0	0	0	93
EASTER ISLAND	0	1	0	0	0	0	0	0	0	2	0	83	0	0	0	86
EGYPT	0	0	1	0	1	0	2	101	0	0	0	0	0	0	0	105
ESKIMO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	107	107
GUAM	0	55	0	0	0	0	0	1	0	1	0	0	0	0	0	57
HAINAN	1	75	1	2	1	1	0	0	0	1	0	0	0	1	0	83
MOKAPU	0	4	0	0	0	0	1	0	0	91	0	4	0	0	0	100
MORIORI	0	1	0	0	0	0	0	0	1	104	0	2	0	0	0	108
N JAPAN	4	79	1	1	2	0	0	0	0	0	0	0	0	0	0	87
N MAORI	0	0	0	0	0	0	0	0	0	8	0	2	0	0	0	10
NORSE	0	2	2	0	0	11	0	92	2	1	0	0	0	0	0	110
PERU	0	1	1	100	0	1	0	0	5	1	0	0	0	0	0	109
PHILIPPINES	0	45	0	1	2	0	0	1	0	1	0	0	0	0	0	50
S JAPAN	3	87	0	0	0	0	0	0	0	0	0	1	0	0	0	91
S MAORI	0	0	0	0	0	0	0	0	0	9	0	1	0	0	0	10
SANTA CRUZ	1	1	0	2	0	0	0	1	96	0	0	0	0	1	0	102
TASMANIA	0	0	0	0	1	1	0	1	0	2	66	1	14	0	0	86
TEITA	1	0	1	0	57	0	0	4	0	0	0	1	2	0	0	66
TOLAI	0	2	0	0	2	0	0	0	0	1	0	1	104	0	0	110
ZALAVAR	0	2	4	2	0	25	0	62	0	0	0	0	2	1	0	98
ZULU	0	4	0	0	88	0	0	2	0	0	0	0	4	0	0	98

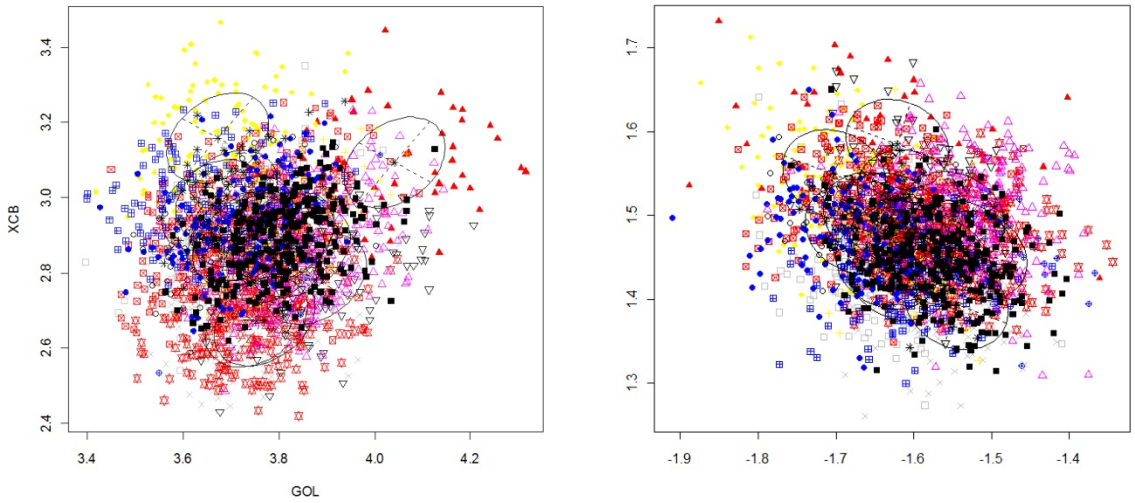


Figure 65: Random seed projection showing clusters

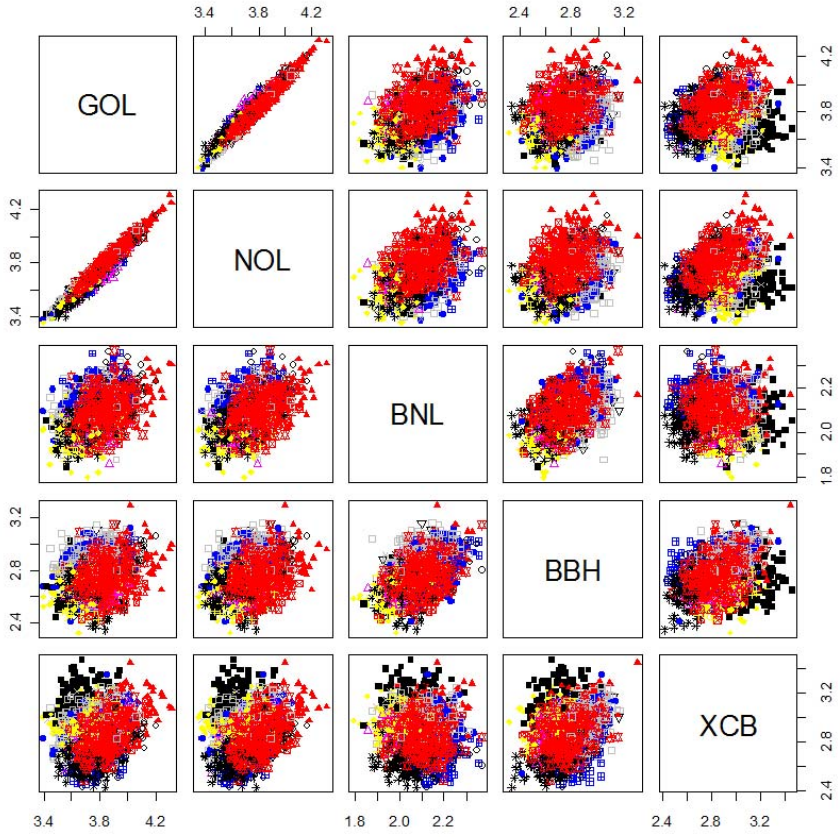


Figure 66: Pair-wise scatterplot matrix showing clusters

It can be seen immediately from these preliminary results that this project was unable to replicate the 14 cluster pattern reported by Pontikos (2004). While Pontikos (2010) has suggested that observed discrepancies between even his own cluster runs may be related to the subsequent addition of the North and South Maori samples (n=20, all males) to Howells' original (1989) data set, i.e., producing analyses based on data sets that differ from each other and the format available for download (Howells, 1996). This explanation is nevertheless viewed, here, with skepticism: the ability of the added Maori cases to effect such a directed change to the cluster solution is highly questionable as these specimens are relatively few in number and Howells' (1995) considered these cases to be "of the same derivation" as the Moriori, who were already included among his core Polynesian series (1989). What is perhaps more interesting and more easily explored, is the fact that the introduction of the 45 additional craniometric variables failed to significantly alter the model chosen using the BIC-base selection criteria, in either the number of components or the parameterization of the variance-covariance matrices: recall that, for the reduced (12) variable subset, MCLUST returned a 16-component EEE model; also note that both the "best" 12 variable MCLUST model (with 16-components) and the "best" 12 variable EMMIX model (with 10-components) are represented here among the top ten models chosen using the full 57 variable data set and identified by their high BIC values (see Table 26). In order to determine if these two 16-componet EEE models are in fact comparable, it is useful to, first, assess classification strength: the implied error rates were calculated by considering the posterior probabilities associated with each allocation (at the thresholds of >0.07 and when $\rightarrow$ 1) and the corresponding uncertainty estimates were generated (c.f., McLachlan & Peel, 2000; Ousley & Jantz, 2005; Fraley & Raftery, 2011). The estimates of correct implied membership (per-population) are provided in Table 28 and the quantiles of uncertainty related to the full clustering of the 57 variable data set are reported in Table 29. The corresponding instances of "misclassification" were visualized with a coordinate projection and the overall quality of

the model's classification was summarized using an "uncertainty profile," for which uncertainty values were plotted in increasing order of magnitude (Figure 67) (Fraley & Raftery, 2011).

Table 28: Allocation rates

Population (Howells 1989)	*posterior probability > 0.7*	percent correctly allocated	*posterior probability → 1*	percent correctly allocated	*total number of individuals*
AINU	79	92	66	77	86
ANDAMAN	68	97	63	90	70
ANYANG	41	98	40	95	42
ARIKARA	53	78	41	60	68
ATAYAL	43	91	39	83	47
AUSTRALIA	99	98	96	95	101
BERG	103	94	90	83	109
BURIAT	106	98	105	97	108
BUSHMAN	87	98	78	88	89
DOGON	90	97	89	96	93
EASTER ISLAND	82	95	78	91	86
EGYPT	103	98	99	94	105
ESKIMO	106	99	103	96	107
GUAM	53	93	51	89	57
HAINAN	79	95	76	92	83
MOKAPU	92	92	82	82	100
MORIORI	100	93	93	86	108
N MAORI	81	93	75	86	87
N JAPAN	7	70	5	50	10
NORSE	101	92	87	79	110
PERU	99	91	91	83	109
PHILIPPINES	45	90	38	76	50
S MAORI	89	98	82	90	91
S JAPAN	10	100	7	70	10
SANTA CRUZ	95	93	92	90	102
TASMANIA	84	98	80	93	86
TEITA	59	89	56	85	66
TOLAI	110	100	101	92	110
ZALAVAR	74	76	46	47	98
ZULU	92	94	80	82	98
Grand Total	*2330*	*94*	*2129*	*86*	*2486*

Table 29: Quantiles of uncertainty

Quantile	0%	25%	50%	75%	100%
Estimate	0	0.0001275200	0.0021617008	0.0303082995	0.6801376391

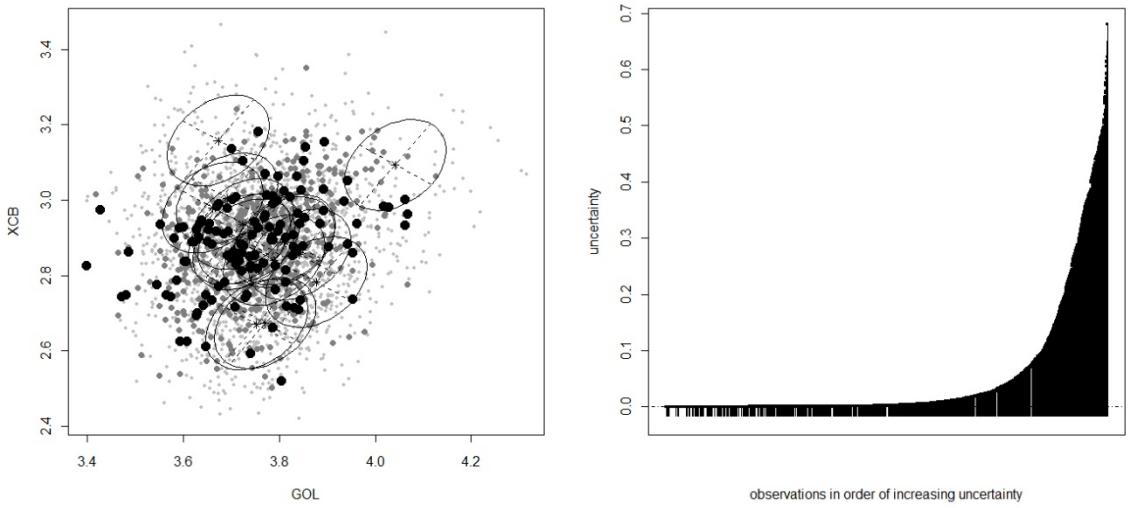


Figure 67: Coordinate projection and data profile showing uncertainty

Statistically, it appears that this 15-component mixture model is highly accurate: this evaluation of its estimated conditional probabilities reveals high levels of classification success (94% and 86%), which, in turn, imply a strong separation between the clusters and a good model fitting. Visually, however, this same quality of clustering is less apparent: the graphical representations display an overlapping cluster pattern and with points of great uncertainty (see Figure 65, Figure 66, and Figure 67). In order to contextualize these findings, the statistical properties of this 15-component solution were compared against the mixture models previously presented. It can be seen from the summary results given in Table 30 that this increase in the number of variables used has contributed to an obvious improvement in outright classification success: the implied error rate for outright allocations has dropped from 62% (for

the 16-component EEE model with 12 variables and 28 populations) to an impressive 14%. Nevertheless, the overall estimates of uncertainty appear to remain consistent across the models with higher complexity (i.e., more components), regardless of variable number.

Table 30: Statistical properties of select mixture analyses

data set; maximum populations	*number of variables*	*model*	*selection criteria*	*posterior probability > 0.7*	*posterior probability → 1*	*quantile of uncertainty (100%)*
Howells 1989; 28	12	3 VVV	BIC	86%	66%	0.583987695
Howells 1989; 28	12	10 arbitrary	-2logλ	80%	54%	0.697900
Howells 1989; 30	57	15 EEE	BIC	94%	86%	0.6801376391
Howells 1989; 28	12	16 EEE	BIC	76%	38%	0.6511021

It is difficult to attach value to an improvement in cluster separation and classification accuracy based upon statistics alone, as the numbers do not reflect the additional estimations that reintroducing these variables require: does this improvement in fit come at the expense of utility? In response to this question, it is reasonable to reconsider this mixture solution in light of the known population memberships provided by Howells (1989, 1995) in order to gain additional practical insight and to assess the *anthropological* worth of the classifications. To this end, a series of graphical representations were generated that allow for the easy evaluation of each population's cluster composition (after Table 27) and, in turn, a better appreciation of the patterns of membership and association produced by this cluster analysis. Since the primary goal for the application finite mixture analysis to Howells' craniometric data set is to test for cryptic population structure within a mixture of globally sampled biogeographic groups, it is appropriate to considered the findings from the 57 variable clustering from a holistic perspective. Populations were collated based on the cluster with which they had the greatest

membership and then each population was ordered on two levels: first by percent membership within the given cluster and, when necessary, by the raw number of individual members. The organization of the cluster findings in this way resulted in the formation of six agglomerative, but distinctive groups. The relationships observed among the members of each of these macro-groups do demonstrate a correspondence with anthropological definitions of biogeographic ancestry and they also permit a general association of to be drawn between the statistically inferred clusters and the major world-wide races. Three additional clusters were shown to correspond almost exclusively to a single population and they bear little to no relationship to the other clusters; these outlying clusters imply, in part, that broad level racial categories poorly capture the full range of global variation.

Because of the greater clarity in these composite groups (which are composed of linked "like" populations) that was gained from the increase in variables used in this model, it is possible to more easily generalize the associations between cluster number and a biogeographic, evolutionary or racial identifier. Cluster 2 represents the East Asian component of variation (Figure 68); capturing, in order of relationship, Howells' (1989) South Japan, Guam, Anyang, North Japan, Hainan, Philippines and Atayal series. While the Ainu are largely allocated to their own unique cluster (number 3), they share closest membership with the populations in cluster 2. This pattern likely reflects both their affinity with the Far East as well as their complex population history. The Caucasian populations are grouped together in cluster 8 (Figure 69), which clearly includes Norse and Zalavar (two of Howells 3 local European series) and the Egyptians. This group is labeled as "Caucasian" rather than European, because of the strong relationship between the Egyptians (North Africans), Scandinavians, and Hungarians - this "link" clearly cannot be attributed to geography. The Berg, although related to the Zalavar and Norse by shared memberships in clusters 8 and 6, is mostly relegated to its own cluster (number 6). While the unique

"Berg" cluster may reflect this population's geographic isolation (with local origins, in an Austrian mountain village), there is an obvious range of distinction: it progresses, in order, from the Berg, to the Zalavar, to the Norse, and it ends with Egyptians (the only population of the Caucasians, without members in cluster 6). If this range is considered in terms of general morphology, it may also correspond to the relative transition in cranial shape from brachycephalic to dolichocephalic. A similar pattern of "core" versus "marginal" groups is also observed for the Polynesian populations, who together make up cluster 10 (Figure 70). While the Moriori, Mokapu, South and North Maori are all clearly linked, the Easter Islanders remains along the periphery of cluster 10, with the majority of these cases assigned to their unique cluster (number 12). It is interesting to note that the Moriori and Maori are assigned almost exclusive membership in the same cluster; while this clear association reinforces Howells' (1995) expectations, given their shared origins, it equally discredits Pontikos' (2010) explanation for the discrepancy in cluster number (14 versus 14 components) between his independent analyses. Cluster 4 includes the Amerindian populations (Figure 71), with the Peru and Santa Cruz representing the extreme of the cluster allocation: the Peruvian series is almost exclusively assigned cluster 4, while the cases from Santa Cruz are largely restricted to cluster 9, with a small percentage in cluster 4. The Arikara occupy an intermediate position along this spectrum of variation; with 50% of these individuals being allocated to cluster 4 and 10% to cluster 9. All of Howells' African series are collected together in cluster 5 (Figure 72), but 33% of the Bushman are assigned membership in their own cluster (number 1). This suggests that within the "African" continental group lies a detectable morphological distinction between the sub-Saharan populations (the Dogon, Zulu and Teita) and the more geographically widespread (or more broadly sampled) San Bushmen. Cluster 13 is composed exclusively of the Australo-Melanesians (Figure 73), with a large part of the Tasmanians sample also assigned to a unique cluster (number 11). The three remaining populations, the Andaman, Buriat and Eskimo each form their own exclusive

clusters (Figure 74). It is interesting to note that five of Howells' seven "local" series are either assigned wholly or in a large degree to their own population-cluster, with a level of certainty and representation that was not achieved in the previous mixture analyses. The heightened sensitivity of this model is also attested to by, for example, the distinction made between the African and Australo-Melanesian populations. While the increased separation between the populations and clusters brought about by the increased number of variable is helpful for improving estimates of classification accuracy, it also results in the loss of information on the ties that bind some of these composite groups together (e.g., severing the connection between the Africans and Australo-Melanesians), which may be highly valuable when the true relationships among groups are not well known.

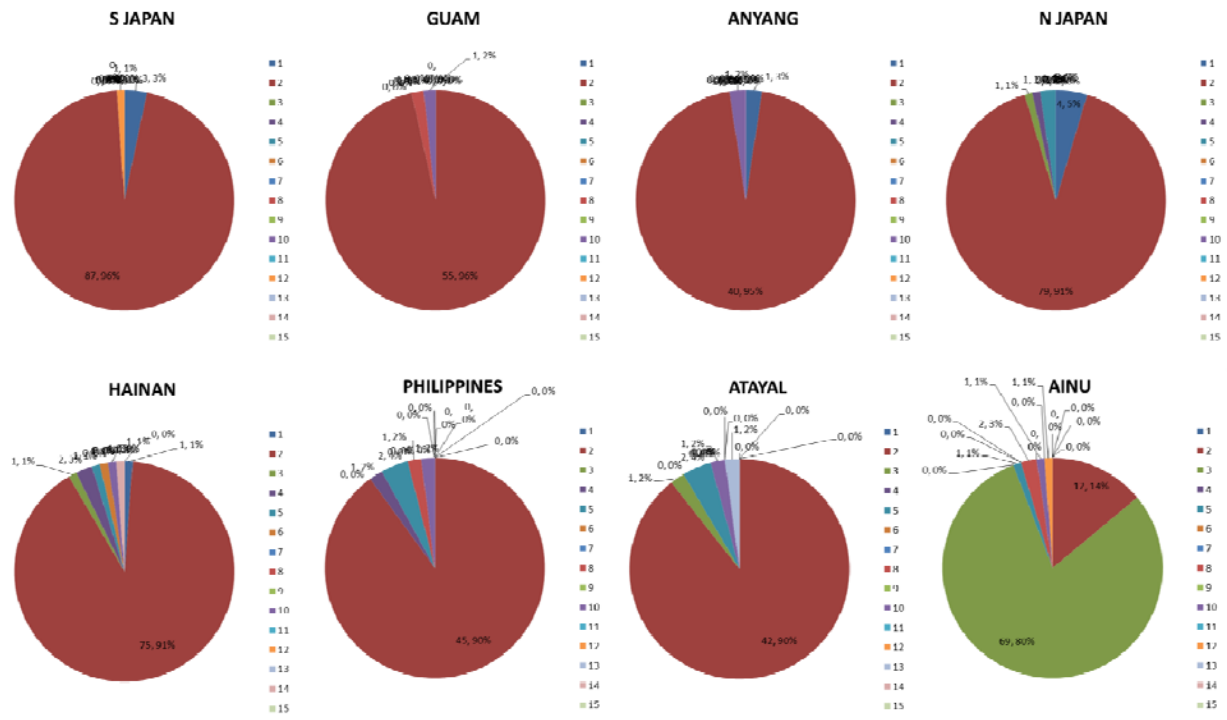


Figure 68: Cluster 2 memberships; Cluster 3 (Ainu)

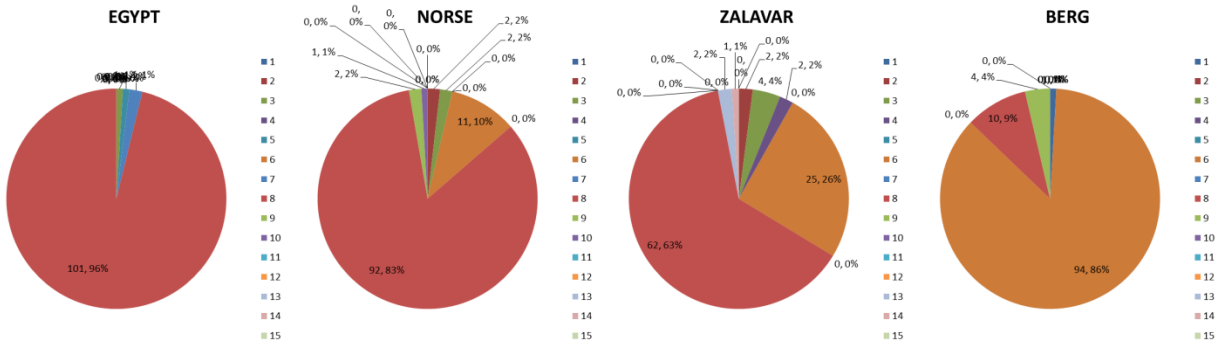


Figure 69: Cluster 8 memberships; Cluster 6 (Berg)

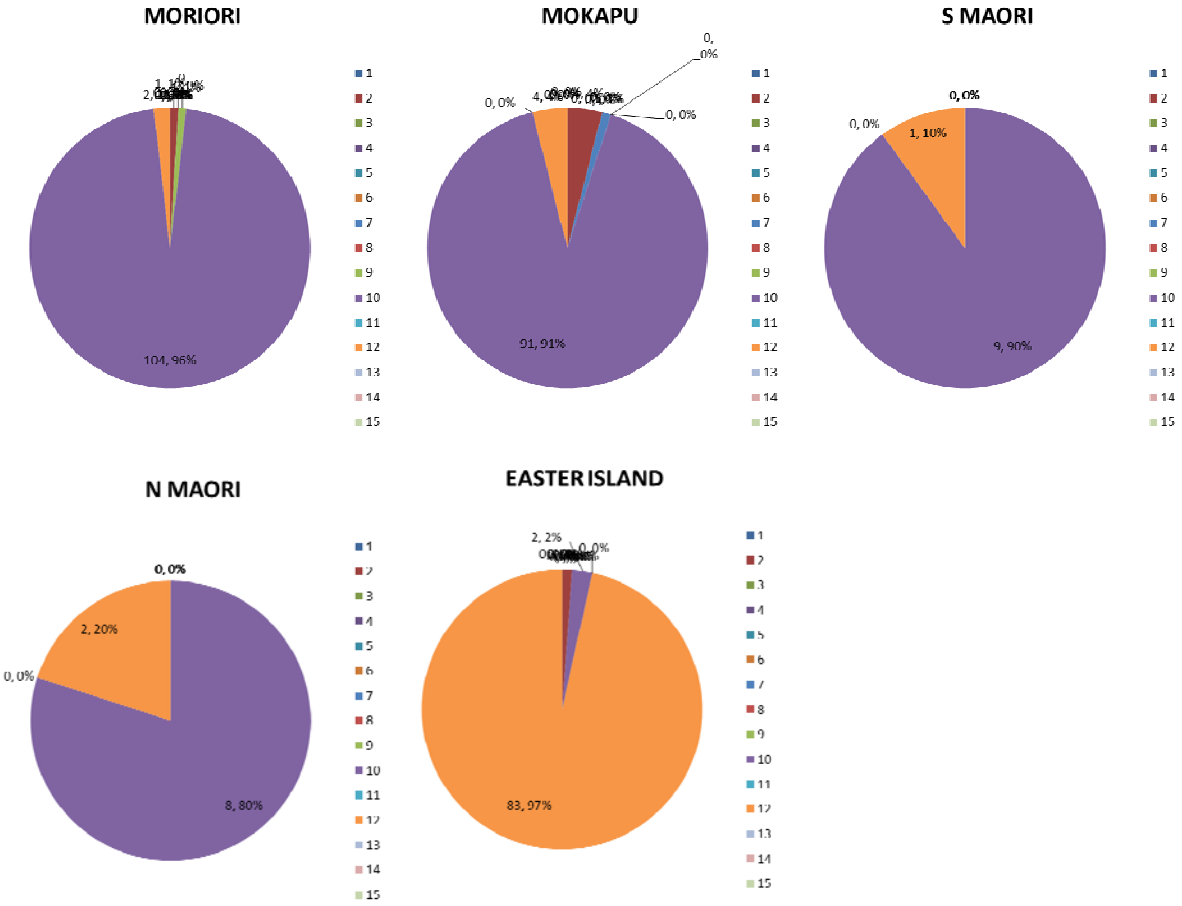


Figure 70: Cluster 10 memberships; Cluster 12 (Easter Island)

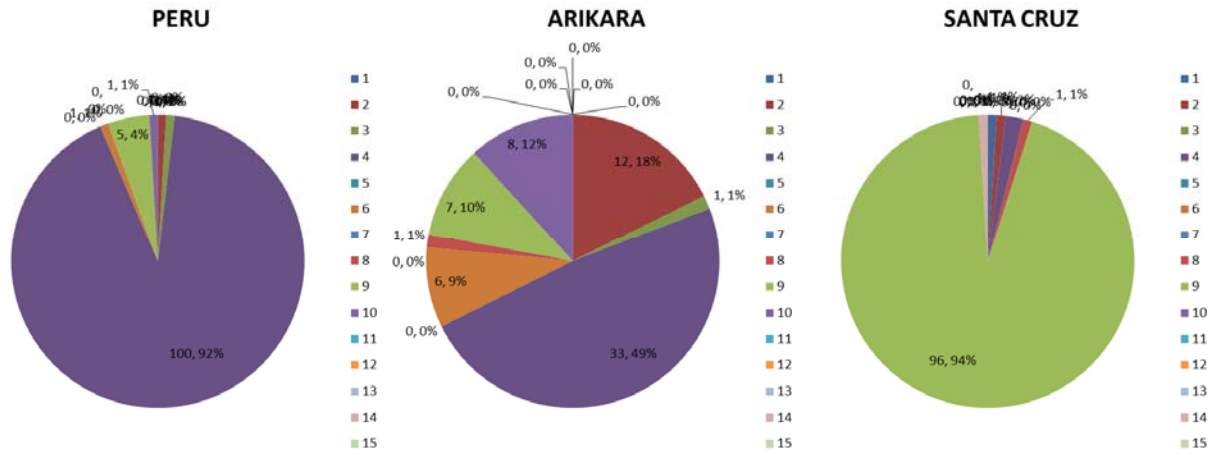


Figure 71: Cluster 4 memberships; Cluster 9 (Santa Cruz)

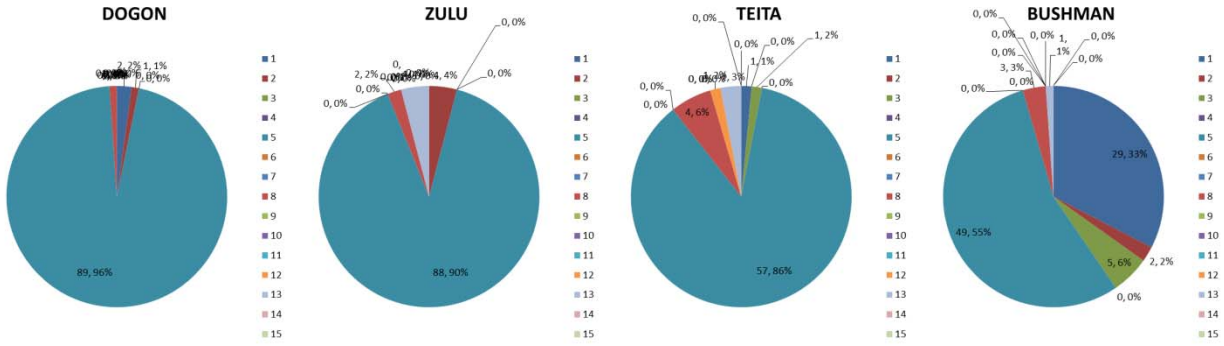


Figure 72: Cluster 5 memberships; Cluster 1 (Bushman)

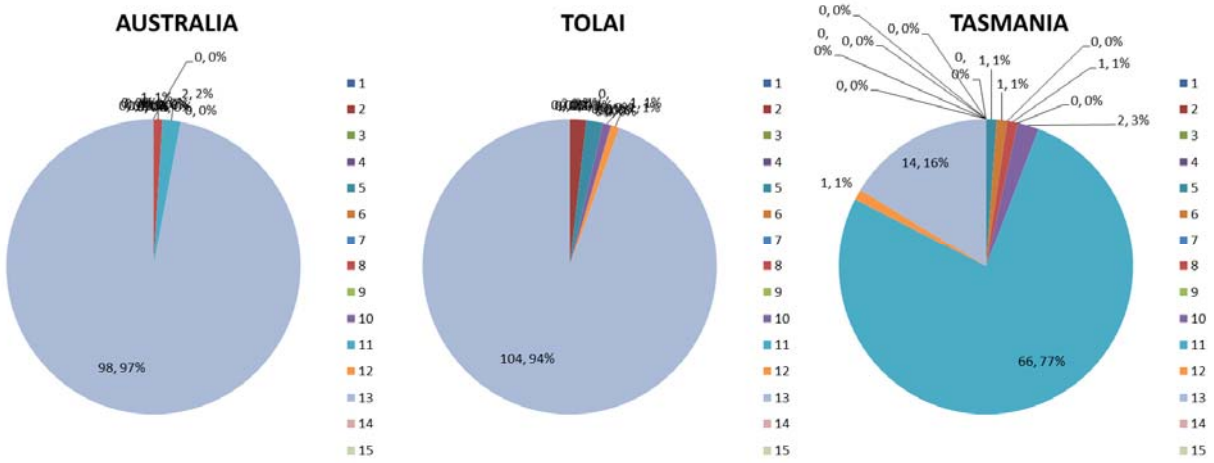


Figure 73: Cluster 13 memberships; Cluster 11 (Tasmania)

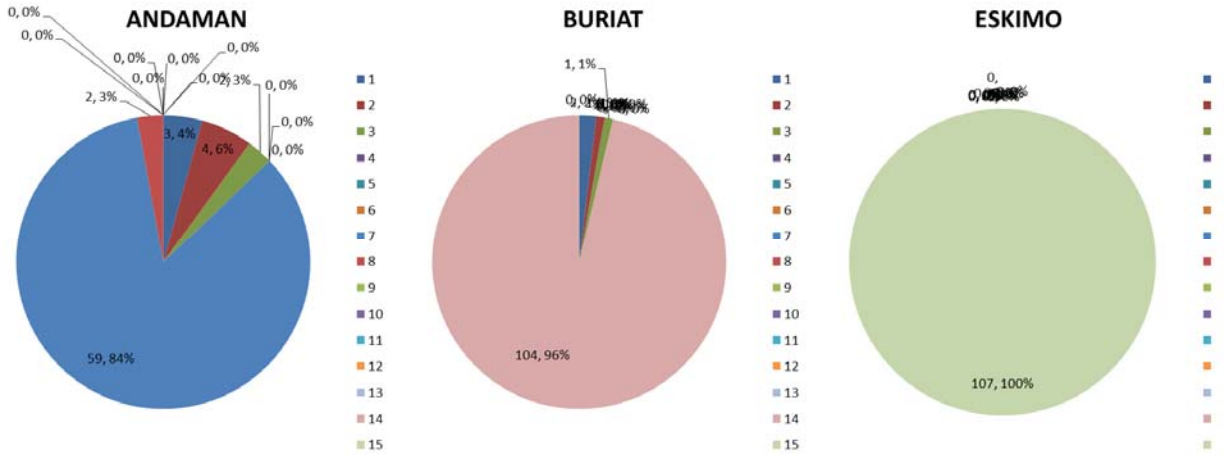


Figure 74: Cluster 7 (Andaman); Cluster 14 (Buriat); Cluster 15 (Eskimo)

When these allocations are collapsed down to a simpler configuration (Figure 75), that considers three "different" continental ancestries, there is an easy association between the true population affiliations (based on Howells' identifiers) and the "major races." Asia is represented by cluster 2, 3, 4, 7, 9, 10 and 12; Africa by clusters 5, 11, 13, 14 and 15; Europe by cluster 6. The only cluster that shows multiple continental groups in significant proportion is cluster 1, which is composed of 67% African (specifically, Bushmen) and 31% Asian populations. The clarity with which cluster-specific trends can be detected and associated with biogeographic proxies (e.g., Howells' population names, regional agglomerations, continental races) offers further support for the conclusions obtained using the simpler 12 variable models: that human craniometric variation is hierarchically structured with nested subsets, so that that this population substructuring or stratification can be identified and interpreted at various demographic levels (e.g., continents, regions, races, tribes, ethnic or local groups).

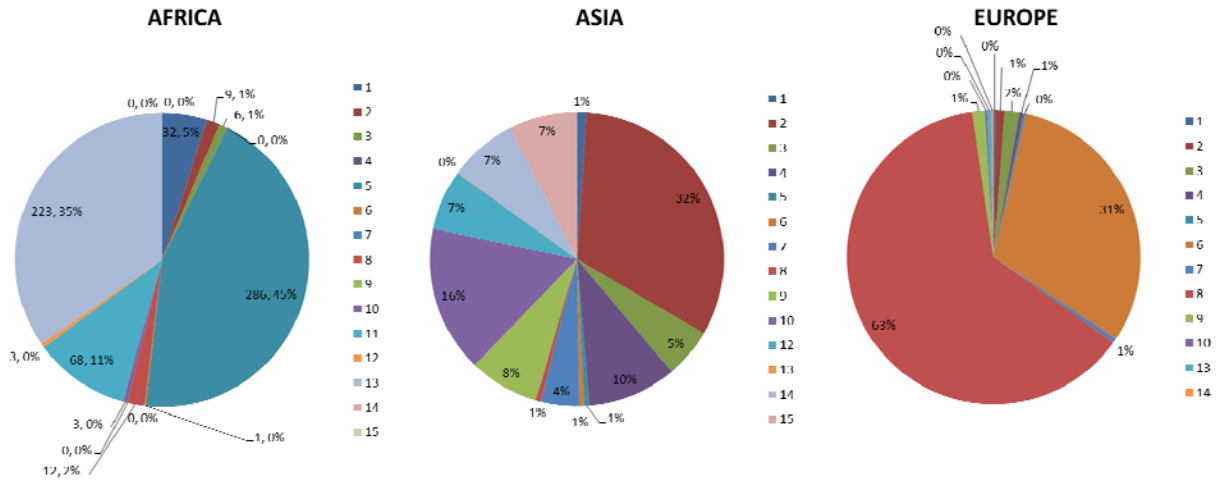


Figure 75: Continental clusters

When further attention is paid to the findings produced by the 15-component EEE model, with 57 variables, Pontikos (2004) is correct in remarking that "[i]t is striking that such population-specific clusters were identified even though the algorithm makes no use of population affiliations, indicating that some populations are craniometrically distinctive, with all their members sharing a high degree of within-group similarity" (np). This analysis offers convincing evidence for the presence of population structure and for meaning within group proxies. It also attests to the ability of finite mixture methods to produce robust clustering results and accurate classifications in the absence of *a priori* identifiers and without user-direction. What these results fail to convincingly demonstrate is that the loss of computational tractability associated with the increase in variable number is warranted: even when the among-group differences are maximized by the increase in variables, the results remain consistent in their trends. Neither the models produced using the 12 variable subset or the full suite of 57 variables returned one-to-one relationships between the cluster number and Howells' 28 (or 30) populations. Similarly, they both show the ability of model-based clustering to resolve of the mixture of craniometric populations into statistically significant clusters with anthropological meaning, and they both corroborate, albeit to different degrees of clarity, the recent arguments put forth by human geneticists for the presence of population substructure, which is not only scientifically detectable but also bears close correspondence to contemporary concepts of biogeographic ancestry and biosocial race (e.g., Rosenberg et al., 2002; Bamshad et al., 2003).

Tests of Structured Variation: The Combined Howells and Hanihara Data Set

A common concern in global population studies is the question of the most appropriate sampling practices. The problems with ascertainment that result from the practical limitations of available data are often unavoidable; for this reason, they are widely excused when adequate steps are taken to correct or account for potential sampling bias. Disagreement emerges most readily when researchers adopt fundamentally different perspectives on what defines "random" sampling and constitutes the most "useful" or "informative" data set. There are two dominant approaches: samples may be obtained from geographically distant groups in order to identify pockets of "different" variation within the full global spectrum, or data may be drawn from samples which are distributed more evenly across space, in order to capture the gradated patterns of diversity, as morphologic variation is spread over global geography and among world-wide populations. It is well known that Howells (1989) data set was collected under the first of these approaches (c.f. Leathers et al., 2002; Relethford, 2009), as he, himself, wrote that "the first 18 populations are meant to serve as a sort of core of model humanity, distributed 3 to a region, to give some kind of central estimate of means and interpopulation variation...the populations were also presumed to be non-extreme in size or shape...these choices bear little relation to the population number of the present world..."(Howells, 1989 p. 2). As such, the sampling strategy that defined the distribution of these groups serves to maximize within-population differences and should, therefore, in theory, also maximize cluster separation when finite mixture analysis is applied. In order to test how the mixture clustering methods respond to different degrees of within-data set diversity and evaluate the consequences of sampling on the detection of global population structure, this project undertook two different tests. It combined craniometric data obtained from Howells' (1996) and Hanihara's (2008) global data sets and subjected this large and very diverse "combined" sample to

standard model-based clustering analysis, using all available male and female populations and 14 preselected variables. The goal of this first test is to evaluate the effect of "improved sampling" and clinal variation on cluster resolution. Howells (1989) craniometric data set was also combined with cases drawn from Howells (1995) TEST data set and this new and highly variable mixture was also subjected to standard model-based clustering analysis. The goal of this second test is to evaluate the behavior of model-based clustering algorithms in the presence of extreme variation and, ultimately, to assess the ability to detect meaningful structure within a mixture that includes highly atypical and problematic morphologies.

While Howells' (1996) data has been extensively preprocessed, Hanihara's (2008) data set contains many incomplete cases: after "duplicate" observations were deleted and the remaining data merged, 1333 cases were identified as suitable candidates for missing value imputation. This combined data set was subjected to multiple sex-specific imputations using the imputeData procedure, implemented with the contributed R package, mix (Schafer 2010). A final complete case data set (n = 14281) was derived following the method described by Steele, Wang and Raftery (2010) and using the corresponding R package, MImix 1.0 (Steele, Wang & Raftery, 2010). Pairs plots were produced for the 14 variables, and partitioned by sex, to display the location of the missing and non-missing values for each imputation. These visualizations were generated using the function, imputePairs, available through the R package, MCLUST (Fraley & Raftery, 2006, 2011). Three example pairs plots are presented here for the female (Figure 76) and male (Figure 77) data imputations: these graphs display the missing values relative to the full data set for variables 1 and 2, 4 and 5, and 11 and 12. These plots provide a convenient visualization of how the missing values are distributed: they attest that they are dispersed randomly throughout the point cloud. This information (which appears consistent across all plots) is used as support for assuming

that this missing data meets the minimum imputation requirement of MAR (missing at random), i.e., where the probability of missingness for a given variable Y_i can depend on other observed variables but not on Y_i itself, after controlling for another variable (Rubin, 1987, 1996). A second data set was also produced, for which missing values were treated by case-wise deletion. This approach yielded a reduced but unbiased sample of 12,948 individuals representing "complete cases".

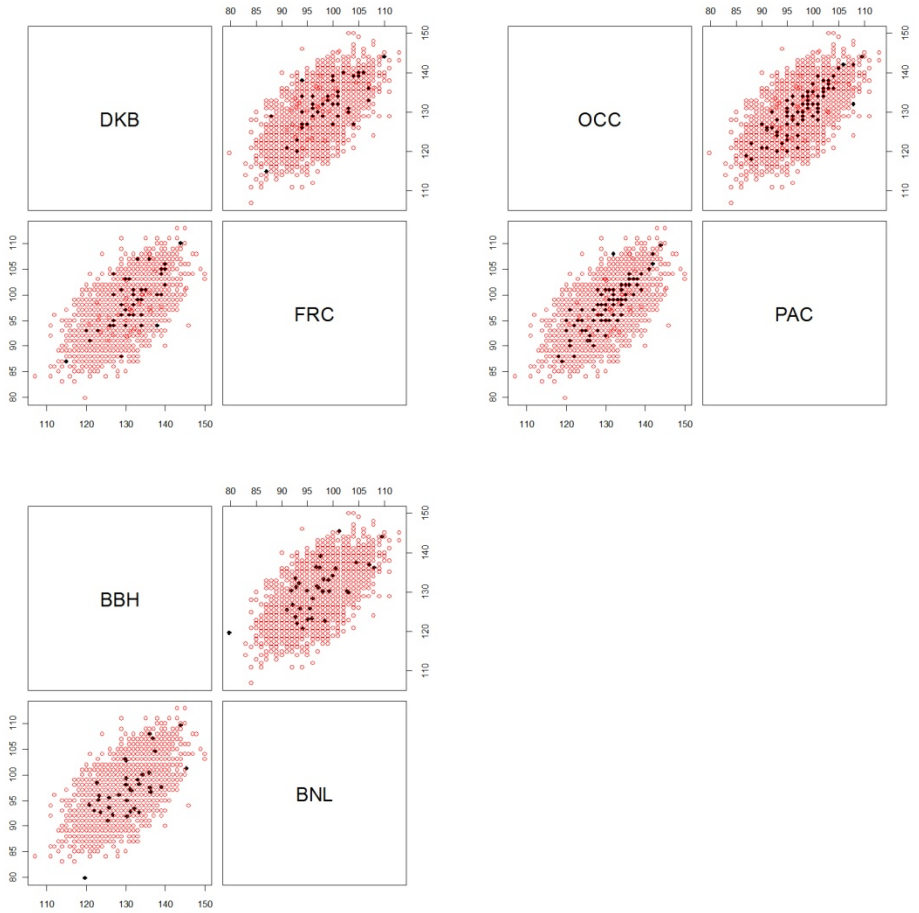


Figure 76: Pairs plots for missing data (females); imputed values in black

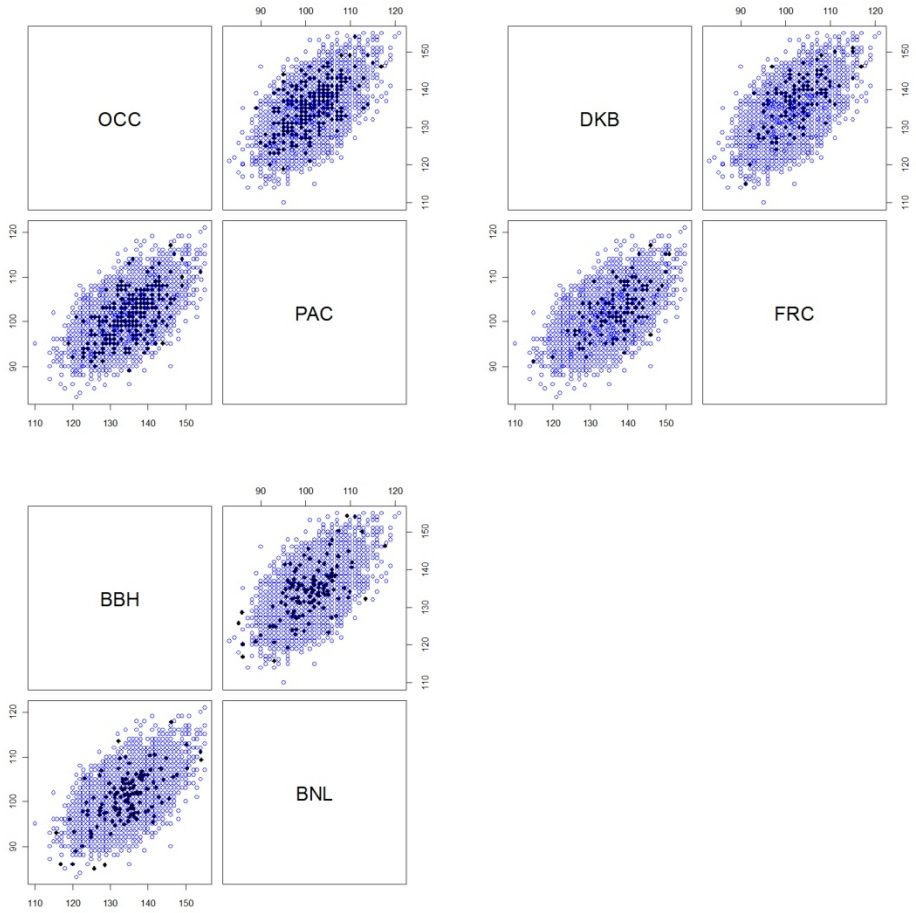


Figure 77: Pairs plots for missing data (males); imputed values in black

While the imputation of missing values provides a convenient way to circumvent the problems associated with conducting statistical analyses on incomplete data sets and the reduction of sample size, these advantages are gained at the loss of some degree of certainty: it is necessary to perform all subsequent tests under the assumption that the estimated "replacement" values represent the "true" (missing) values. For this reason, it is of considerable interested to determine if this imputed data set

provides, in actual practice, better clustering results than the alternative, reduced data set, which was produced using case-wise deletion. Parallel mixture analyses were undertaken for both the imputed (n = 14281) and non-imputed (n = 12948) combined Howells and Hanihara data sets using the MCLUST package (Fraley & Raftery, 2011). As before, a large component range (1 to 30 clusters) was identified for testing and full flexibility in parameterization was specified when executing the automatic model selection method. The top ten models with the highest BIC values generated for both data sets are presented in Table 31. The calculated value of the Bayesian Information Criterion was maximized in both cases for estimates that corresponded to similar model choices (see Figure 78). For the imputed data set, MCLUST returned an 8-component model with unconstrained covariances, and, for the case-deleted data set, it identified a very comparable 7-component model with unconstrained covariances.

In order to identify the model (and data set) with greatest utility, plots were produced for the cluster memberships associated with the optimal 8 and 7-component solutions selected for the imputed and case-deleted data sets, respectively (Figure 79,Figure 80). Classification strength was then assessed by calculating the implied error rates and generating individual uncertainty estimates (c.f., McLachlan & Peel, 2000; Ousley & Jantz, 2005; Fraley & Raftery, 2011). Summary information on classification accuracy is provided in Table 32 and Table 33 and on classification uncertainty, in Table 34 and in Figure 81 and Figure 82.

Table 31: Top 10 models in order

Howells and Hanihara Data Set, *with Imputation*			Howells and Hanihara Data Set, *with Case-Wise Deletion*		
Component Number	Model Form	Estimated BIC	Component Number	Model Form	Estimated BIC
8	VVV	778666.6	7	VVV	705931.6
9	VVV	778524.3	8	VVV	705799.1
11	VVV	778237.2	11	VVV	705535.5
10	VVV	778196	8	VEV	705504
7	VVV	777988.8	10	VVV	705424.8
11	VEV	777482.8	6	VVV	705297.7
9	VEV	777408.4	9	VVV	705227.8
10	VEV	777348.5	7	VEV	704886.9
13	VVV	777334.1	12	VVV	704847.1
12	VEV	777270.6	13	VVV	704588.4

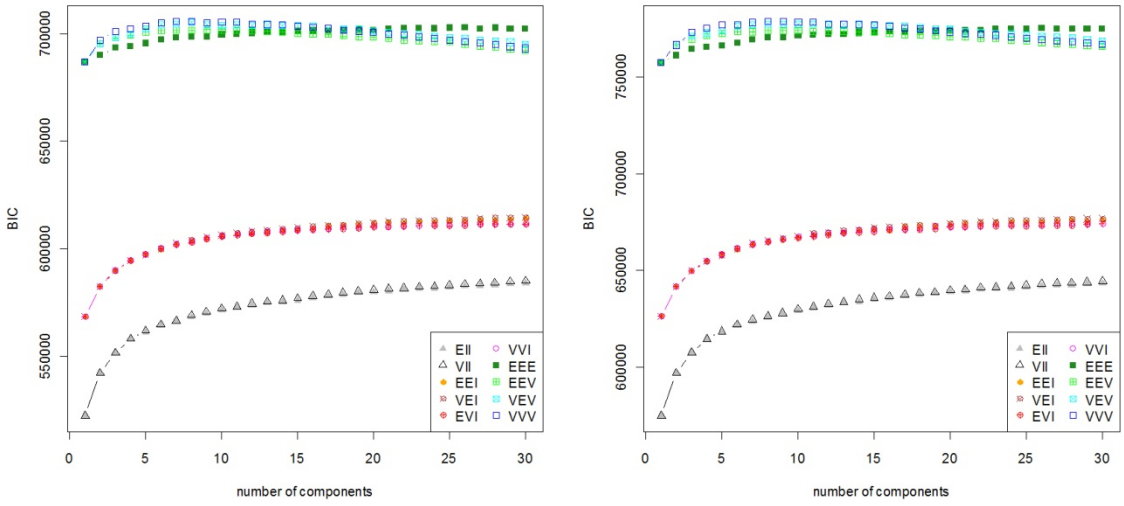


Figure 78: Model selection by BIC for data sets with missing values treated by imputation and by case-wise deletion respectively

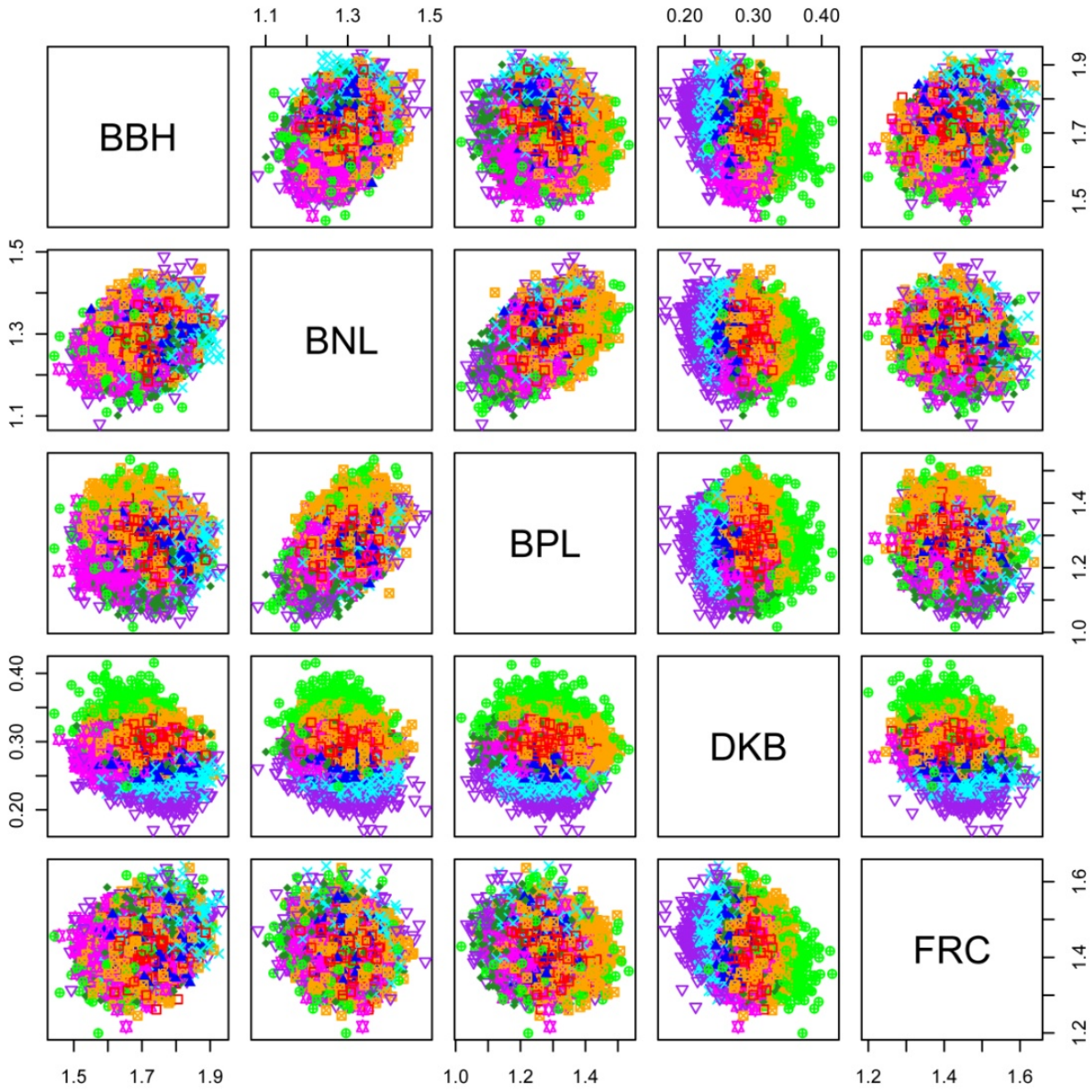


Figure 79: Pair-wise scatterplot matrix showing clusters (imputed data set)

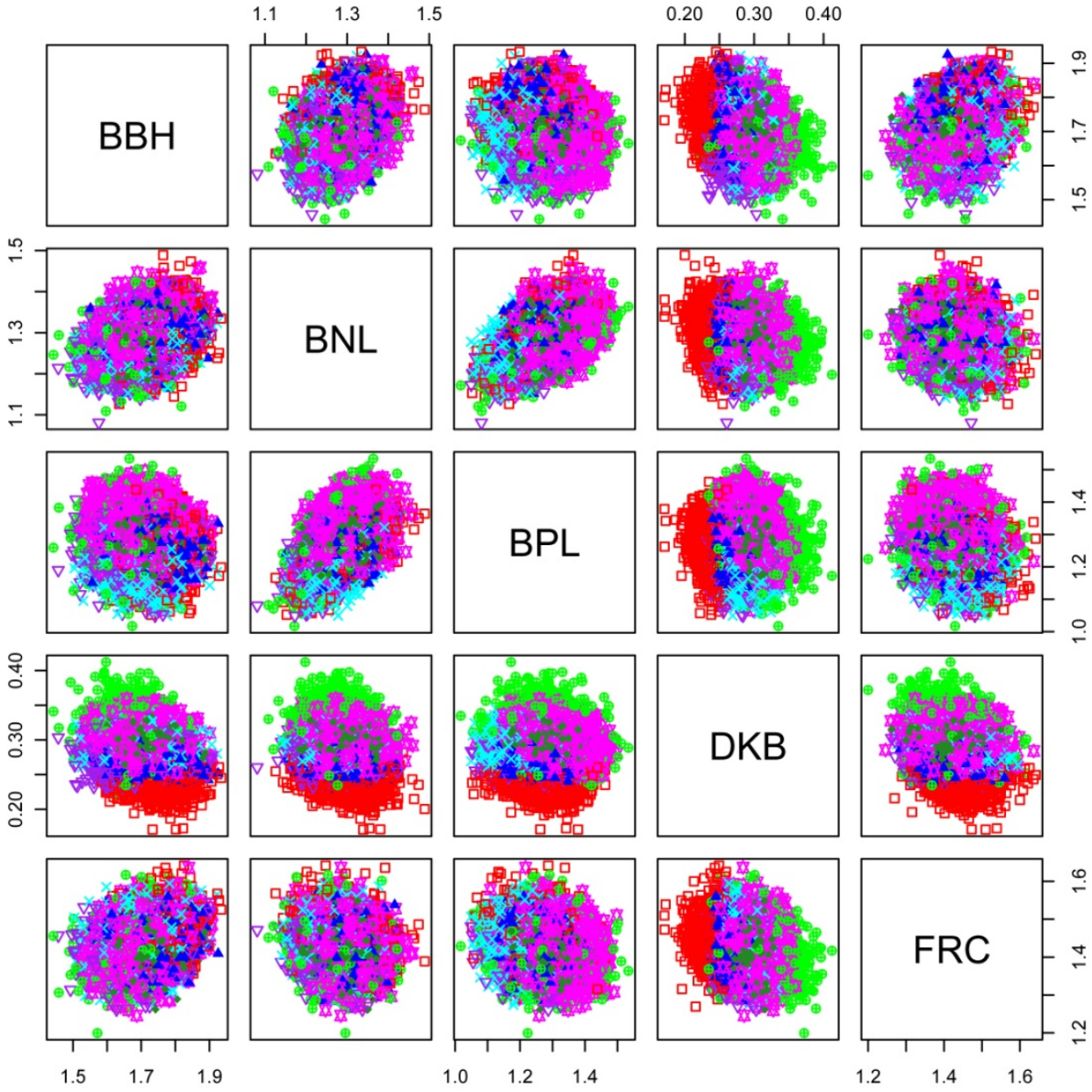


Figure 80: Pair-wise scatterplot matrix showing clusters (case deleted data set)

Table 32: Allocation rates (imputed data set)

Population (Howells 1989, Hanihara2008)	*posterior probability > 0.7*	*percent correctly allocated*	*posterior probability → 1*	*percent correctly allocated*	*total number of individuals*
AL: Aleut	234	79	131	44	297
AT: Australia and Tasmania	411	84	296	61	488
CA: Canada	79	72	49	45	110
CSA: Central and South America	502	67	228	30	753
EA: East Asia	874	69	489	39	1269
ES: Eskimo	761	79	511	53	962
EU: Europe	1039	70	574	39	1476
IS: Indian Subcontinent	351	73	197	41	478
MEL: Melanesia	399	77	265	51	521
MIC: Micronesia	76	68	41	37	111
NA: North Africa	552	71	284	36	779
NEA: Northeast Asia	384	75	231	45	512
PNG: Papua New Guinea	314	75	215	52	417
POL: Polynesia	616	72	355	42	852
SEA: Southeast Asia	656	68	338	35	962
SSA: sub-Saharan Africa	1095	84	770	59	1308
SWA: South and West Asia	223	66	108	32	339
UK: United Kingdom	329	75	213	48	441
U.S.: United States	1486	67	740	34	2206
Total	10381	73	6035	42	14281

Table 33: Allocation rates (case deleted data set)

Population (Howells 1989, Hanihara 2008)	*posterior probability > 0.7*	*percent correctly allocated*	*posterior probability → 1*	*percent correctly allocated*	*total number of individuals*
AL: Aleut	219	80	166	61	273
AT: Australia and Tasmania	377	84	289	64	449
CA: Canada	69	68	44	43	102
CSA: Central and South America	490	69	261	37	712
EA: East Asia	776	71	396	36	1096
ES: Eskimo	688	78	498	57	879
EU: Europe	984	72	577	42	1358
IS: Indian Subcontinent	327	75	191	44	434
MEL: Melanesia	356	81	249	56	442
MIC: Micronesia	75	71	45	43	105
NA: North Africa	519	71	265	36	728
NEA: Northeast Asia	356	76	253	54	467
PNG: Papua New Guinea	280	76	205	56	368
POL: Polynesia	579	73	367	46	798
SEA: Southeast Asia	575	68	302	36	847
SSA: sub-Saharan Africa	1046	85	729	59	1229
SWA: Southwest Asia	193	68	97	34	284
UK: United Kingdom	274	73	194	52	373
U.S.: United States	1364	68	751	37	2004
Total	9547	74	5879	45	12948

Table 34: Quantiles of uncertainty

Missing Data Treatment	*Model*	*0%*	*25%*	*50%*	*75%*	*100%*
Imputation	8, VVV	0	0.04098471	0.1375287	0.3199737	0.7525033
case deletion	7, VVV	0	0.03706082	0.1209636	0.3117271	0.738998

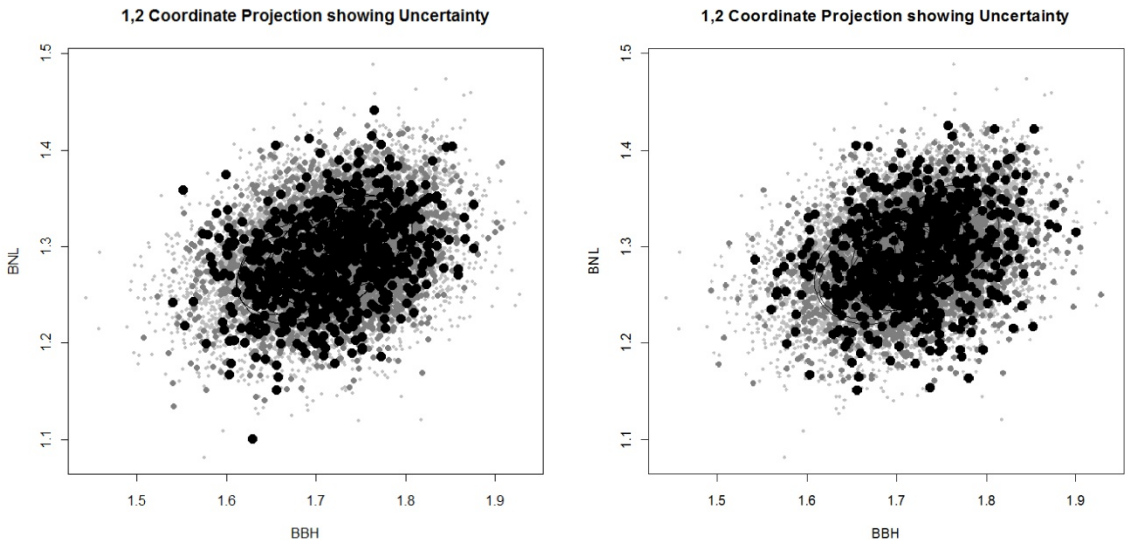


Figure 81: Coordinate projections displaying misclassifications for the imputed and case-deleted data sets respectively

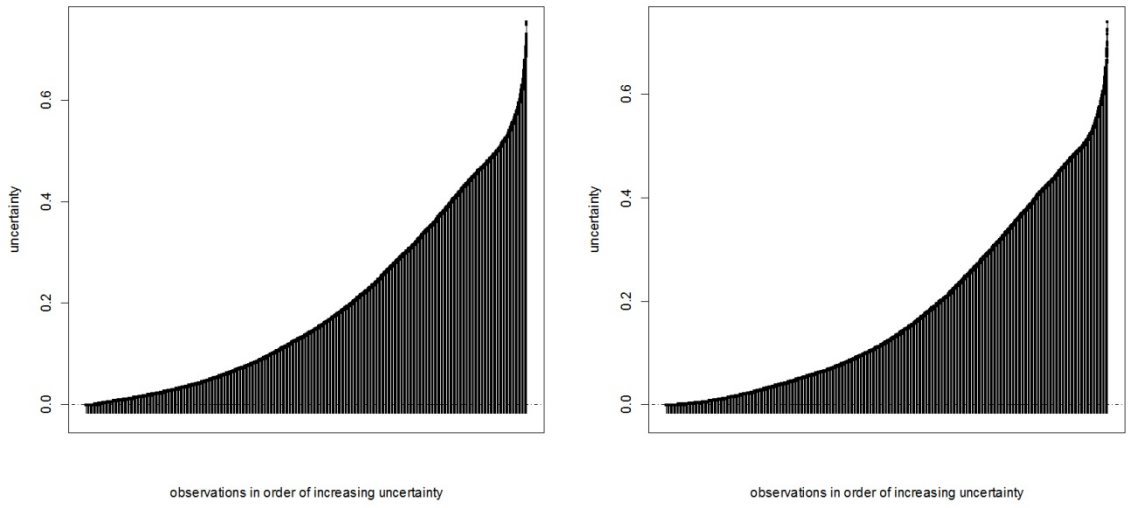


Figure 82: Uncertainty profiles displaying misclassifications for the imputed and case-deleted data sets respectively

As can be seen from these graphical displays, the cluster patterns and the degree of cluster separation do not appear significantly different between these two analyses; moreover, the evidence on clustering quality indicates that, in fact, the 8-compoment VVV model, associated with the imputed data set, provides the weaker of the two clustering solutions. Given these findings, this project has chosen to restrict its use of model-based clustering methods to the combined Howells (1996) and Hanihara (2008) data set for which missing values have been treated by case-wise deletion (n = 12948).

When the findings presented above for the 7-component VVV model are reconsidered, now in terms of the ability of mixture modeling to detect global population structure using a more widely distributed, more diversely populated, and more evenly sampled data set, there are apparent differences in the quality of the results. Overall, it appears that classification success has decreased and uncertainty has increased with these changes to the sample's conditions (Table 35). If, however, this combined sample analysis is compared to the optimal MCLUST (16-component) clustering of Howells (1989) data set alone, a different pattern is shown: there is an improvement in the estimated accuracy of the outright classifications alongside an increase in uncertainty. This change in uncertainty is not unexpected as, in merging these data sets, samples were introduced from areas in which populations overlap; what is interesting is the fact that this uncertainty does not vary in concert with classification accuracy. While these percent estimates of correct allocations are not remarkably high, they are equivalent to those obtained using only Howells 28 populations, for which geographic distance is large and the range of variation represented is limited (i.e., 7 main geographic regions; approximately 50 crania per series). It appears, therefore, that Howells' data collection strategy does not artificially inflate the levels of differentiation among populations by exaggerating group separation in such a way that adversely affects the ability to reliably detect meaningful group structure and obtain realistic classification results (in so

far as the analysis of the combined Howells and Hanihara data sets is an accurate representation of this method's clustering potential).

Table 35: Model statistics for the Howells and Howells and Hanihara data sets

Data Source; Populations or Size	*Number of Variables*	*Model (Component, Parameterization)*	*Selection Criteria*	*Posterior Probability > 0.7*	*Posterior Probability → 1*	*Quantile of Uncertainty (100%)*
Howells 1989; 28	12	3 VVV	BIC	86%	66%	0.583987695
Howells 1989; 28	12	10 arbitrary	-2logλ	80%	54%	0.697900
Howells 1989; 30	57	15 EEE	BIC	94%	86%	0.6801376391
Howells 1989; 28	12	16 EEE	BIC	76%	38%	0.6511021
Howells 1989 and Hanihara 2008; 12948	14	7 VVV	BIC	74%	45%	0.738998

In order to more fully appreciate the import of the objective statistics presented for the 7-component VVV model and see how the introduction of graduated variation into these tests of latent population substructure affects the clarity of the clusters or their anthropological relevance, it is reasonable to consider in detail the patterns of associations made between groups. This project's interest in the combined Howells and Hanihara data set lies in the ability of mixture modeling to reveal global trends when sufficient sampling is used to address clinal variation. For this reason, it is argued here that a simplified naming system can be adopted for use in *a posteriori* evaluations of this combined data set. The total number of the combined population names was reduced to correspond with the 19 regional divisions (see Table 36) defined by Hanihara (2008). The raw classification data associated with this mixture analysis is given in Table 36. These figures provide the number of individuals (by regional division) assigned to each of the seven clusters, after implementing the model-based clustering methods using the MCLUST package (Fraley & Raftery, 2011). In order to bring greater clarity to these results,

graphical representations of the cluster memberships for each regional group were generated. These visualizations facilitate pattern recognition and the detection of both interregional and broad geographic relationships (e.g., Figure 83, Figure 84, Figure 85).

Table 36: Cluster memberships by population

Population	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Grand Total*
AL: Aleut	72	16	4	3	6	8	164	273
AT: Australia and Tasmania	17	9	51	1	229	140	2	449
CA: Canada	29	8	0	5	4	12	44	102
CSA: Central and South America	314	47	5	63	41	106	136	712
EA: East Asia	364	147	21	66	145	265	88	1096
ES: Eskimo	301	437	3	18	10	9	101	879
EU: Europe	186	84	83	703	82	112	108	1358
IS: Indian Subcontinent	95	25	10	43	90	163	8	434
MEL: Melanesia	43	28	8	11	174	171	7	442
MIC: Micronesia	45	12	1	1	8	30	8	105
NA: North Africa	198	49	12	142	118	190	19	728
NEA: Northeast Asia	87	37	23	28	8	23	261	467
PNG: Papua New Guinea	44	11	8	8	140	153	4	368
POL: Polynesia	329	124	14	11	114	131	75	798
SEA: Southeast Asia	196	27	26	65	153	294	86	847
SSA: sub-Saharan Africa	49	15	177	25	523	425	15	1229
SWA: Southwest Asia	86	13	9	77	29	52	18	284
UK: United Kingdom	48	24	16	240	13	16	16	373
U.S.: United States	790	103	25	96	146	326	518	2004
Grand Total	3293	1216	496	1606	2033	2626	1678	12948

The increased sampling density and greater geographic coverage brought to this study by the inclusion of Hanihara's cranial series has complicated the mixture analysis. This introduction of clinal variation appears to have had a homogenizing effect on the overall cluster pattern: this is obvious in the reduction in the number of components from 16 for the best model for Howells' data set to the 7-

components identified for the combined Howells and Hanihara data set. The increased simplicity in the component number is offset, however, by an increased complexity in the covariance structure. The VVV and VEV parameterization dominate the best models identified. In fact, in out of a matrix of 300 BIC estimates (30 component, 10 parameterizations), only two EEE models rank within the top 25. These EEE models are taken to be even more problematic for the fact that they are associated with unreasonably high component numbers, e.g., 26 and 28. These changes in model structure are played out in the clustering solution, in such a way that there is greater overlap in memberships and, in turn, the clusters reflect broader biogeographic trends instead of the population-level or small scale groupings consistently produced for Howells (1989) data alone.

Cluster 4 (Figure 83) appears to capture the Caucasian component of variation, including the UK, Europe, Southwest Asia and North Africa (at values >20%). Cluster 5 (Figure 84) corresponds to the pattern of African and Australo-Melanesian memberships that were observed repeatedly in previous analyses, but now the logical inclusion of Papua New Guinea and the Indian Subcontinent has been made (at values >20%). Cluster 3 appears to represent a subset of this larger grouping, as it contains the sub-Saharan Africans, Australians and Tasmanians in the greatest numbers. Cluster 1 presents itself as the antithesis to Cluster 3, as it takes in all of the regional groups except the sub-Saharan Africans, Australians and Tasmanians.

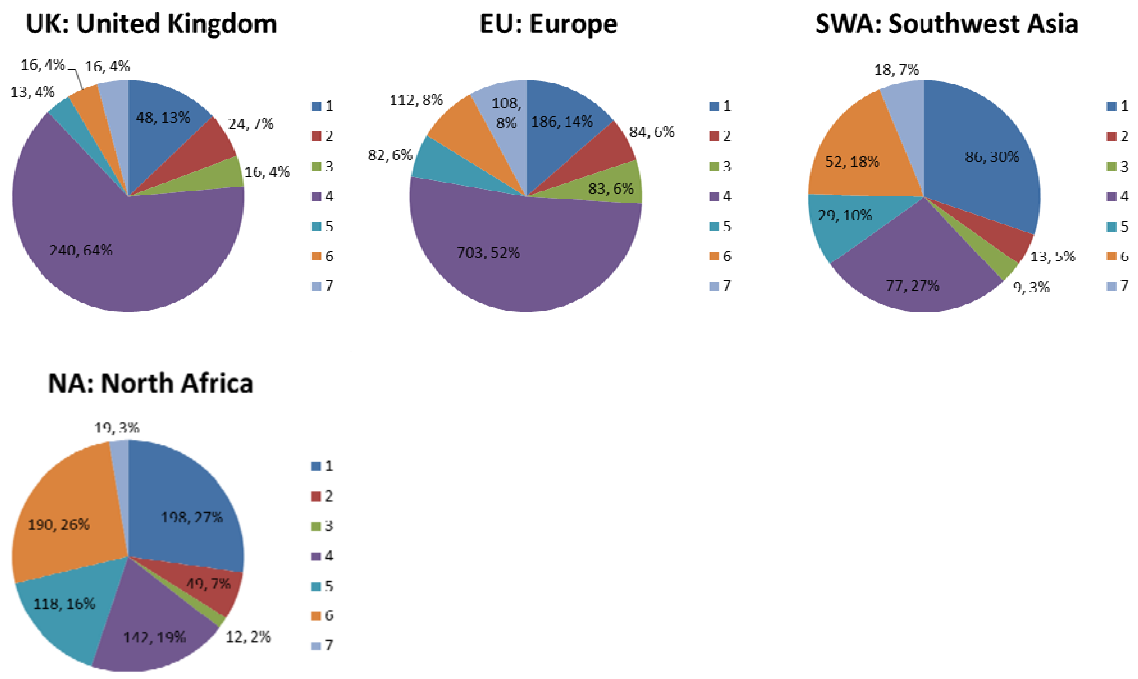


Figure 83: Cluster 4

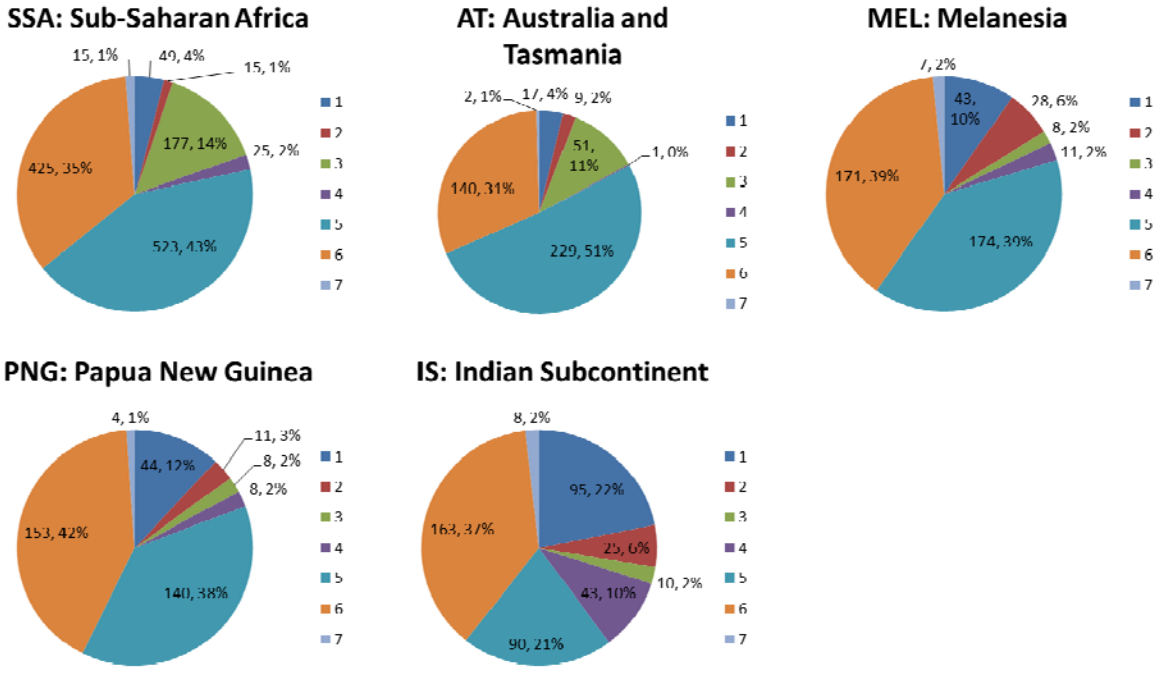


Figure 84: Cluster 5

Cluster 7 (Figure 85) brings together the Native American and Asian groups. In capturing the North East Asian populations specifically, it reflects the evolutionary trajectory of these groups and their "original" relationships. The Eskimo also have a dominant presence in cluster 2. While the Eskimo continue to retain their distinctive membership pattern (split into a few large groups), the inclusion of more similar populations (e.g., the Aleuts of Alaska and the Inuit of the Canadian North) give greater perspective to their unique morphology and their position within the range of cold adapted groups. Cluster 6 appears to capture an opposite morphological trend to the cold adaptations that are likely influencing most of the memberships in cluster 7. Although it is broadly distributed among the 19 groups, there is an obvious absence of northern regions and cold adapted populations, e.g., NE Asia, Eskimo, Europe, UK, and the Aleut. It is assumed, therefore, that memberships in cluster 6 is defined by both Asian origins and southern environmental adaptations. Canada is also included within cluster 6 (at only 12%); this percent contribution most likely reflects the limited populations sampled from Ontario and British Columbia.

To evaluate the repeatability of the findings, and, in turn, assess the strength of this population structure that underlies the expression of these cluster patterns, the combined Howells and Hanihara data set was also subjected to model-based clustering using the EMMIX package (McLachlan, 2011; McLachlan, Wang & Peel, 2011). Mixture modeling was conducted following the same protocol discussed when clustering Howells' (1989) data set with EMMIX. It is important to recall that this package differs in its model selection procedure from the standard MCLUST implementation. Most notably, the optimal number of clusters for the EMMIX mixture model was chosen (1) using the likelihood ratio test statistic λ to test for g versus g+1 components (with a predefined minimum of 1 and maximum of 30) and (2) employing a bootstrap procedure to estimate the corresponding p-value

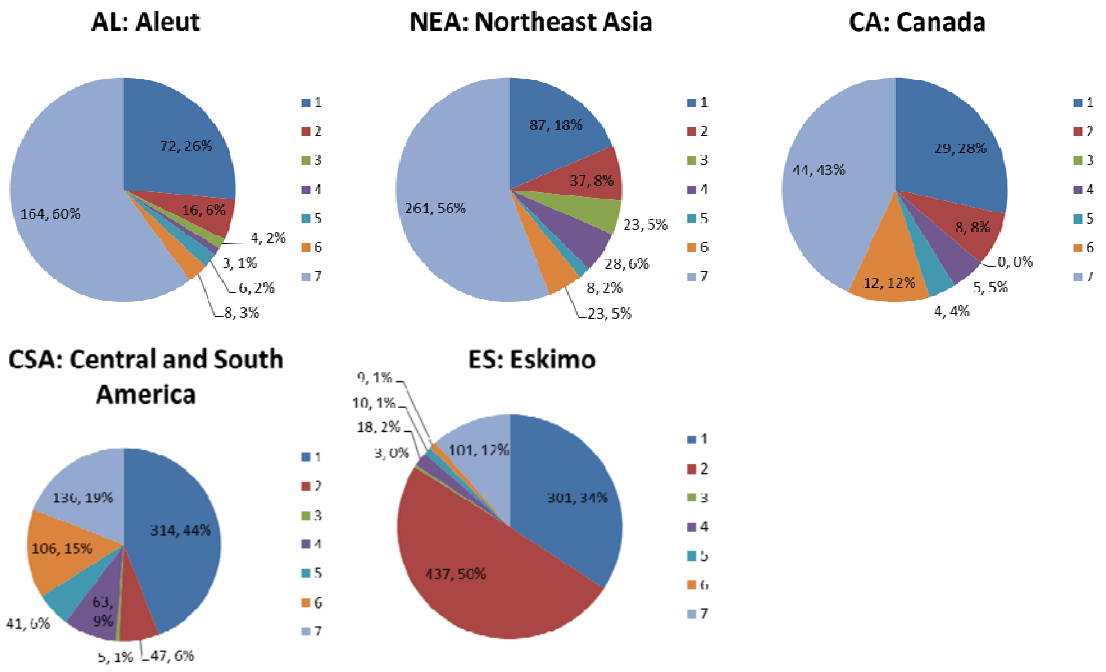


Figure 85: Cluster 7

(McLachlan, 1987; McLachlan & Basford, 1988; McLachlan & Peel, 2000). Given the computational requirements dictated by a data set of this size (n = 12948), Aitken's acceleration was used to speed up the rate of convergence for the EM algorithm in the bootstrap tests by reducing the number of iterations necessary for each fit, i.e., where the likelihood value at convergence is predicted and this estimate is used to calculate the likelihood ratio test statistic (Peel & McLachlan, 1999). The results of this model testing procedure are given Table 37. These include not only the estimates of -2logλ and the corresponding p-values obtained for each successive test of the number of components in the mixture but it also reports, for comparative interest the estimates generated for the AIC and BIC (McLachlan & Peel, 2000). While the maximum number of components was preset at 30 (thereby, considering values of g above and below the "true" number of 19 regional groups), the algorithm encountered problems of

stability of in computation after testing the 11th component. It was decided, therefore, to terminate the process before the p-value reached "insignificance" (Peel & McLachlan, 1999). As it is believed here that (1) the intensive and lengthy computational requirements for assessing the significance of model tests at large component numbers are highly impractical and that (2) the information value of most results produced under models with very many components would be lost to their extreme complexity, this problematic analysis was limited to 20 components during the course of the testing procedure. This range (1-20), although narrower than initially intended, still captures the number of true groups (e.g., 19 +1) as well as the wide array of optimal components numbers identified in previous tests: e.g., 16 component solution, representing the maximum best model for Howells data set returned by MCLUST. For this reason, only the 20-component EMMIX clustering is considered in the discussion that follows.

Table 37: Summary of EMMIX model statistics

Component Number	*Loglikelihood*	*-2log(lambda)*	*AIC*	*BIC*	*p-value*
1	344250.78	-	-688263.55857027497	-687374.78367556061	-
2	349849.78	11197.99	-699221.55040529743	-697436.53191927448	0.01
3	352511.68	5323.80	-704305.35110286076	-701624.08902552922	0.01
4	354277.32	3531.29	-707596.64541659865	-704019.13974795851	0.01
5	355470.28	2385.91	-709742.55128843698	-705268.80202848825	0.01
6	356370.49	1800.43	-711302.98589685129	-705932.99304559396	0.01
7	357038.29	1335.59	-712398.57810235640	-706132.34165979049	0.01
8	357828.03	1579.47	-713738.05010864837	-706575.57007477386	0.01
9	358385.40	1114.75	-714612.80146203924	-706554.07783685613	0.01
10	358865.60	960.39	-715333.19041853177	-706378.22320204007	0.01
11	359396.65	1062.12	-716155.30758973653	-706304.09678193624	0.01
12	359644.28	495.25	-716410.55823048891	-705663.10383138002	0.01
13	360130.00	971.45	-717142.00953845738	-705498.31154803990	0.01
14	360376.38	492.75	-717394.76438750187	-704854.82280577580	0.01
15	360720.22	687.68	-717842.44604324759	-704406.26087021292	0.01
16	361004.73	569.02	-718171.46733913897	-703839.03857479570	0.01
17	361336.17	662.87	-718594.33749791398	-703365.66514226212	0.01
18	361547.61	422.89	-718777.22388490604	-702652.30793794559	0.01
19	361904.59	713.95	-719251.17106232222	-702230.01152405317	0.01
20	362161.15	513.12	-719524.29086906789	-701606.88773949025	0.01

Despite the acknowledged problems in the identification of the "correct" number of components to fit to the combined Howells and Hanihara data set and the complexity (e.g., large cluster number) of the final clustering solution, this analysis produced cluster patterns that are (1) grossly interpretable, given prior expectations for interregional relationships, and (2) comparable to the general trends observed for the (MCLUST) clustering with 7 components. Table 38 reports the model-based classification results obtained using the EMMIX package for the 19 regional groupings, when assuming 20 components and arbitrary covariances.

Table 38: Cluster memberships by population (pooled sexes)

Population	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
AL: Aleut	18	7	1	2	2	3	1	7	36	22
AT: Australia and Tasmania	1	1	96	1	0	13	104	2	6	3
CA: Canada	4	5	1	1	1	3	0	3	0	11
CSA: Central and South America	39	10	7	19	15	49	8	81	16	88
EA: East Asia	54	32	26	24	21	141	42	72	33	27
ES: Eskimo	331	157	35	10	0	35	0	33	3	21
EU: Europe	33	14	12	229	154	43	12	234	23	9
IS: Indian Subcontinent	15	1	37	12	8	31	25	35	0	1
MEL: Melanesia	5	8	79	5	3	26	42	8	1	3
MIC: Micronesia	4	3	5	2	0	19	3	2	3	2
NA: North Africa	26	2	24	46	25	36	28	111	7	7
NEA: Northeast Asia	14	12	4	17	5	11	1	10	105	16
PNG: Papua New Guinea	2	1	72	2	2	17	33	3	2	1
POL: Polynesia	73	31	88	4	5	113	8	16	12	20
SEA: Southeast Asia	5	4	14	12	23	85	45	31	23	13
SSA: sub-Saharan Africa	5	2	81	11	29	18	113	12	11	0
SWA: Southwest Asia	11	1	13	19	7	5	4	51	2	2
UK: United Kingdom	4	2	1	88	47	6	1	89	1	3
U.S.: United States	75	35	56	23	25	90	18	94	46	259

Population	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
AL: Aleut	0	0	1	12	11	112	27	3	0	8
AT: Australia and Tasmania	40	22	76	7	0	1	0	7	49	20
CA: Canada	0	0	5	13	15	17	14	3	0	6
CSA: Central and South America	1	0	36	41	27	40	100	65	8	62
EA: East Asia	15	15	98	76	29	55	96	92	19	129
ES: Eskimo	0	0	4	3	40	76	98	4	0	29
EU: Europe	16	16	32	28	182	61	45	186	10	18
IS: Indian Subcontinent	17	12	76	14	9	2	17	58	25	39
MEL: Melanesia	15	12	104	12	4	3	11	20	56	25
MIC: Micronesia	1	0	12	10	0	2	7	7	2	21
NA: North Africa	22	11	82	19	21	17	64	108	36	36
NEA: Northeast Asia	0	2	3	26	30	149	36	9	2	15
PNG: Papua New Guinea	12	7	88	14	4	1	9	17	42	39
POL: Polynesia	12	4	71	26	10	54	60	14	17	160
SEA: Southeast Asia	8	29	63	162	17	24	44	130	34	81
SSA: sub-Saharan Africa	51	190	134	106	7	2	14	67	347	29
SWA: Southwest Asia	7	3	22	11	25	15	27	36	8	15
UK: United Kingdom	6	3	4	4	35	12	11	52	0	4
U.S.: United States	6	15	116	193	125	190	333	108	24	173

From these memberships, seven clusters of particular interest have been singled out for a detailed discussion. Cluster 4 is shared in greatest amount (>15%) by the European and United Kingdom samples; corresponding, in part, to the larger "Caucasian" grouping associated with MCLUST's cluster 4 (c.f., Figure 83). An even stronger comparison can be made between cluster 4 and cluster 8, which, except for the addition of the Eskimo, includes the four same regions: UK, SW Asian, Europe and North Africa (with values >15%). Cluster 19 is found in greatest amount (>10%) among the sub-Saharan African, Melanesian, Australian and Papua New Guinean groups; cluster 3 captures the Australo-Melanesian subset from the larger cluster 19 and also includes Polynesia. Cluster 20 further expands upon this South East Asian and Western Pacific trend: containing mostly Polynesia, Micronesia, Papua New Guinea, East

Asia and South East Asia (>10%), but also including, to a lesser degree of representation, the Indian Subcontinent, Melanesia, and the Americas. Obvious similarities can be drawn between these pattern and the memberships discussed in reference to MCLUST cluster 5 (see Figure 84). Finally, cluster 16 shares its pattern of membership with MCLUST cluster 7. Again, the inclusion of the Aleut, NE Asians, and Canadian First Peoples, implies relationships based upon evolutionary origins and environmental adaptations. For this 20-component clustering, it can be said that the increased complexity in cluster number has allowed for the logical extraction of related but peripheral regions from the agglomerative groupings that defined of the 7-cluster MCLUST model. For example, the Central and South Americans have been removed from the largely North American grouping (Figure 85) and are now mostly assigned to clusters 10 and 14, and the Eskimo are now distributed across two almost identical clusters, for which they are an especially dominant group (making up almost half of clusters 1 and 2, and contributing 331 and 157 individuals, respectively). These findings suggest, again, that global population structure is detectable, at varying levels, even in the presence of overlapping morphologies or groups (i.e., clinal variation) and under the conditions of increased sampling breadth and density.

Tests of Structured Variation: Howells' "TEST" Data Set

In order to evaluate the effect of extreme variation on the performance of model-based clustering procedures, MCLUST was used to apply mixture analysis to a data set that combines Howells' (1989) "standard"and "TEST" (1995) craniometric samples. This analysis considered, among all of the available "modern"TEST crania, only complete cases, for which reasonable matches could be made with one of

Howells' original 28 populations (i.e., setting the minimum association criteria at geographic or cultural "affinity"). Howells' (1995, p. 39-46) "Prehistoric"or "paleo" specimens were also included, as their age and extreme features make them an ideal morphological out-group. They can be used to investigate the ability of the mixture cluster analysis to detect boundaries between extremes in contemporary human variation and cranial morphologies that lie well outside the normal range of modern populations. This project conducted parallel analyses on this combined data set, using the 12 base variable subset (n = 2730) and the full suite of 57 measurements (n = 2711). The top ten models with the highest calculated BIC values are provided in Table 39. The corresponding BIC graphs are presented in Figure 86, in order to visualize the behavior of the BIC estimates over the preset range of (1 to 30) components and (10) possible parameterizations .

Table 39: Top 10 models in order

Howells (1989) and TEST (1995) Data Set, with 12 variables			*Howells (1989) and TEST (1995) Data Set, with 57 variables*		
Component Number	*Model Form*	*Estimated BIC*	*Component Number*	*Model Form*	*Estimated BIC*
5	VVV	120101.9	21	EEE	504372.3
16	EEE	120092.5	20	EEE	504054.9
13	EEE	120059.3	15	EEE	503691.2
19	EEE	119935.8	22	EEE	503679.5
4	VVV	119935.6	24	EEE	503571.8
17	EEE	119929.3	17	EEE	503516.8
3	VVV	119910.7	18	EEE	503496.8
4	VEV	119884	16	EEE	503492.1
18	EEE	119853.4	26	EEE	503387.8
20	EEE	119810.5	23	EEE	503369.8

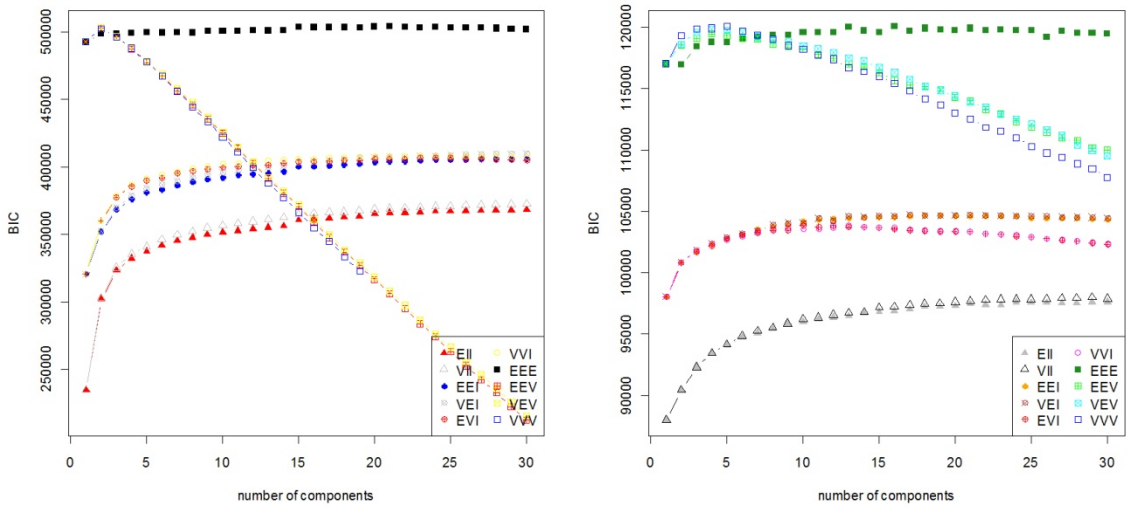


Figure 86: Model selection by BIC for 12 and 57 variables respectively

While the BIC is maximized at a 5-component VVV model for the 12 variables data set, the 16 EEE solution was chosen to be discussed in more detail. While this model is the "second best" of the options identified, it is of special interest because (1) it is identical in its form and component number to the model selected for the previous analysis of Howells (1989) data set without his TEST (1995) specimens (see Table 8 for 12 variable analysis using MCLUST) and (2) it assumes the simplest covariance structure, which, as prior testing of Howells' data has already shown, is likely the most appropriate choice for these among-population analyses. Furthermore, this 16-component EEE model is an equally sound alternative, from an objective statistical perspective, since its BIC estimate is both high and within 10 points of the BIC associated with the optimal 5-component VVV model. A summary of the statistical properties for these models are given in Table 40, and the matching classification information is visualized using coordinate projections of the cluster memberships (Figure 87, see also Figure 11) and their uncertainty (Figure 88, see also Figure 8).

Table 40: Allocation rates and uncertainty by model

Data Source	*Number of Variables*	*Model (Component, Parameterization)*	*Selection Criteria*	*Posterior Probability > 0.7*	*Posterior Probability → 1*	*Quantile of Uncertainty (100%)*
Howells 1989, 1995	12	16 EEE	BIC (preferred)	78%	48%	0.70782484
Howells 1989, 1995	57	21 EEE	BIC (highest)	85%	94%	0.7133545641
Howells 1989 (no TEST data)	12	16 EEE	BIC (highest)	76%	38%	0.6511021

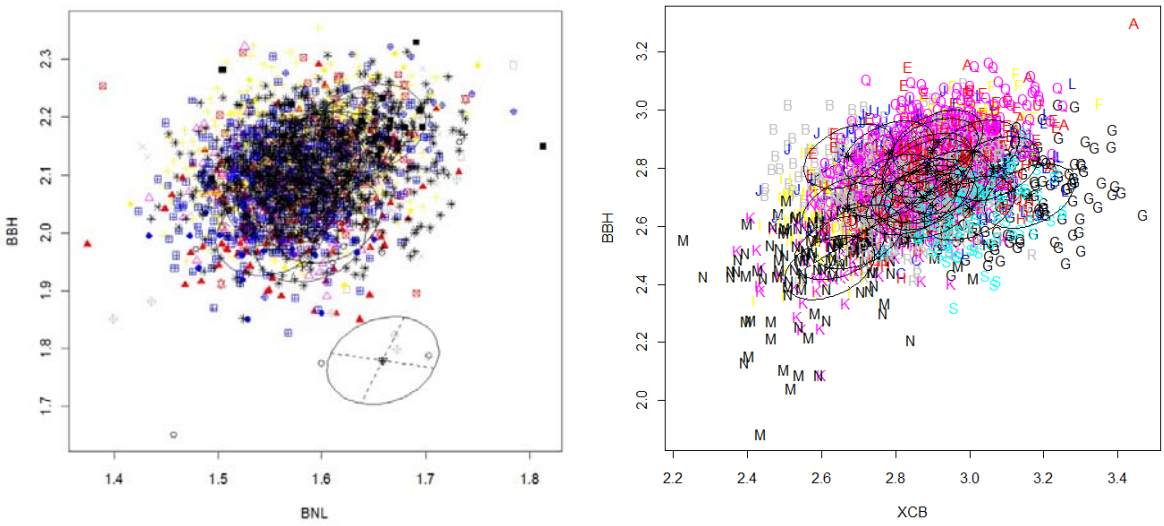


Figure 87: Coordinate projections showing clusters for 12 and 57 variables respectively

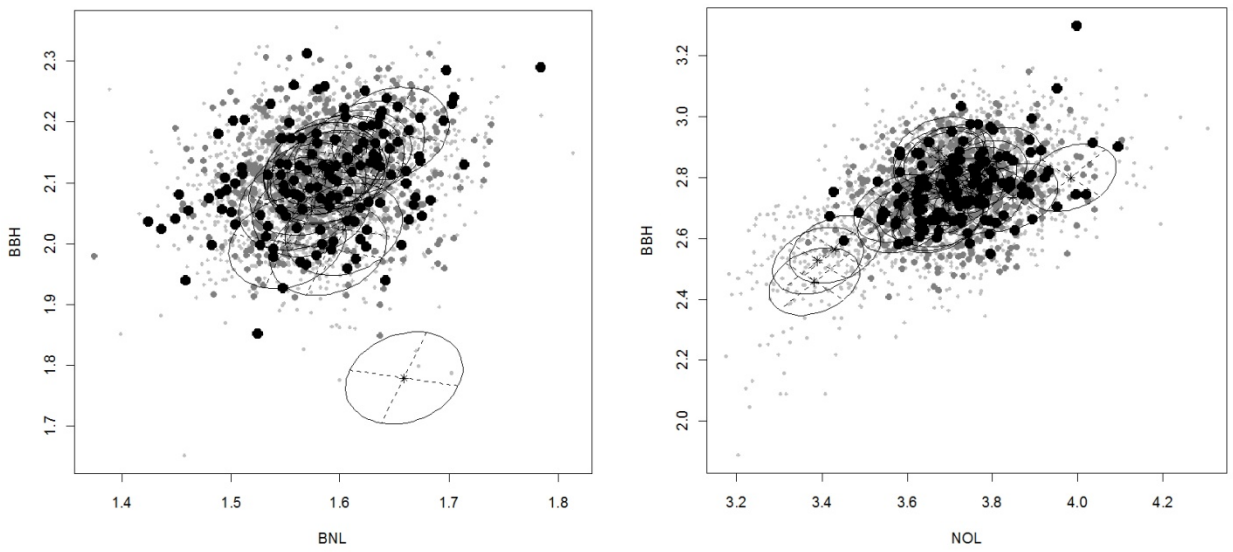


Figure 88: Coordinate projections showing uncertainty for 12 and 57 variables respectively

From these model statistics alone, it may be initially assumed that altering Howells' (1989) standard data set to include a sample of his "TEST" cases (1995) has had little effect on the clustering solution. Upon closer inspection of the classifications, it was found that, in fact, the cluster patterns are different. Perhaps the most obvious inconsistency in results is observed between the two 16-component cluster solutions, in terms of the distribution of cases over the clusters. It is seen here, for the first time, clusters with disproportionately high and low numbers: the largest cluster is populated with 908 cases (33% of the data set) while the smallest cluster includes a single individual (see Table 41; compare with Table 9). This pattern of "lumping," where almost 2/3 of the cases are spread over three major clusters (e.g., 577 individuals in cluster 2, 491 in cluster 5, and 908 in cluster 6), neither corresponds to the true distribution of Howells' (1996) data by population (recall his care to ensure balanced representation for the standard data set) nor to the previous cluster findings.

Table 41: Classification table

Cluster Number	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Number Of Allocations	178	577	28	78	491	908	21	18
Percent Representation	6.52	21.14	1.03	2.86	17.99	33.26	0.77	0.66
Cluster Number	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Number Of Allocations	22	24	62	139	42	63	1	78
Percent Representation	0.81	0.88	2.27	5.09	1.54	2.31	0.04	2.86

This imbalance in cluster assignments is, however, only a reflection of the more fundamental differences in the within-cluster patterns: the 16 clusters produced using Howells' (1989) data set with and without the addition of his "TEST" (1995) sample differ both in the populations represented in each cluster and in the portions in which the contributing populations are present. The ultimate effect of these differences on the quality of the cluster solution and its explanation is as might be expected. When many populations are pooled into a few very large clusters, the clarity with which interpretations can be given to each cluster is understandably diminished. There is also a greater risk of overemphasizing the uniqueness of certain groups by giving too much interpretive value to the patterns that are "apparent" within very small clusters: the practitioner should be wary of situations in which, for example, 80% of Cluster G is composed of population g_1 but Cluster G contains a total of only 10 individuals. When the sample size is large relative to the cluster size, the utility of these findings for population-specific inference is highly questionable. In these instances, it is wise to look instead for trends that govern the overall patterning of the clusters, i.e., when certain populations repeatedly come together in large numbers to form a cluster, and the remaining population-specific memberships in that cluster are small. From this perspective, it can be said that, although the clusters produced from the data set with the TEST samples included are interpretively fuzzy at the level of a single population, there are clearly dominant patterns with information value. Most importantly, these patterns are consistent with the

trends that have been seen several times already. The Africans and Australo-Melanesians group together to make up 66% of cluster 1, 79% of cluster 4, 51 % of cluster 5 and 78% cluster 16; they are also conspicuously absent from cluster 2, which is quite unexpected given its size (n = 577). Howells' European populations (Berg, Norse, Zalavar), along with the Egyptians, represent 59% of cluster 10. Cluster 13 is composed largely of Native American and Asian populations (48%). The variation specific to the "Americas" is more strongly represented in Custer 14, with Peru and Santa Cruz alone contributing 37% of its members. The Siberian Buriat and the Greenland Eskimo group together in cluster 12 (62%) and they are also found, as before, in their own, largely unique but small, clusters, e.g., Buriats in cluster 3 (68%) and Eskimo in cluster 7 (81%). When the same data set is reconsidered but with the use of all 57 of Howells' craniofacial variables, these patterns are remarkably enhanced: cluster separation appears strong and inter-groups differences are well-defined. For the 21-component EEE model, almost every cluster can be related to either a single population or biogeographic grouping. A summary of these memberships is provided in Table 42. This repeated identification of the same agglomerative groupings, despite changes to the methods, models or data sets, further validates argument for detectable and meaningful population substructure. The fact that the degree of clarity with which among-group membership can be revealed is so significantly improved by the increase in the number of variables included in the cluster analysis, suggests that the approach to variable selection protocol may have a much more profound effect on the clustering outcome than a given sampling strategy and it may carry far more weight than acknowledged, thus far, in craniometric literature.

Table 42: Cluster memberships by population

Cluster	*Composition: key % memberships*	*Biogeographic Grouping*
1	95% Andaman	
2	90% Bushman	
3	85% Peru, 8% Arikara	Native American
4	26% Bushman, 21% Zulu, 15% Teita	sub-Saharan African
5	mix (467 cases)	Asian and Micronesian
6	31% Egypt, 32% Norse, 24% Zalavar, 5% Berg	European/White[17]
7	78% Berg, 13% Zalavar	European
8	97% Tasmania, 3% Australia	Australo-Melanesian
9	87% Ainu	
10	89% Easter Island, 3% Mokapu	Polynesian
11	85% Arikara, 4% Peru	Native American
12	83% Moriori, 10% Maori, 3% Mokapu	Polynesian
13	50% Dogon, 37% Zulu, 6% Teita	sub-Saharan African
14	86% Mokapu, 3% Moriori	Polynesian
15	98% Buriat	
16	90% Santa Cruz, 4% Arikara, 3% Peru	Native American
17	47% Tolai, 43% Australia, 6% Tasmania	Australo-Melanesian
18	100% Eskimo	
19	mix (16% Bushman, 6% Paleo Specimens)	
20	mix (35% Peru, 10% Paleo Specimens)	
21	mix (24% Peru, 10% Paleo Specimens)	

Perhaps the more interesting, and, at the same time, troubling, observation to take from these results is the way in which the out-group of paleoanthropological specimens (n = 24) have been handle by the clustering procedure. As it has been observed for Howells' (1989) standard data set, the "blind" clustering methods adopted here have, so far, consistently identified and largely isolated the "extreme" local populations, which Howells (1989) took to be good examples of the maxima in craniofacial shape

[17] White or Caucasian may better capture the variation represented by this group, as it also includes Egyptians sampled from North Africa (Howells, 1989).

occurring along the range of human morphological variation. The level of dissimilarity between the craniofacial characteristic of, for example, the Eskimo, Buriats, and, often, the Bushmen, and Howells's (1989) 18 "core" populations has been so effectively captured by this project's mixture procedures that most of these "extreme" cases have been allocated with high probability into their own clusters. Even when the clustering algorithm was faced with making sense out of the more densely sampled and evenly distributed series from the combined Howells (1989) and Hanihara (2008) data set, it produced, at the very least, clusters with limited memberships that reflected logical groupings (e.g., joining the cold adapted, native Eskimo, Buriat, and Canadian Inuit samples into a single cluster). Based on this prior behavior, it was anticipated that the highly divergent paleoanthropological specimens would be clustered with even greater certainty and clarity. In theory, "Archaic," "Neanderthal," or "Prehistoric" crania should be more cohesively clustered and effectively segregated than any of the series that have been previously analyzed here because the mixture analysis would now be operating at a much greater level of differentiation (e.g., between species or, at least, among highly divergent groups). Even Howells' or Hanihara's most unique cases are only atypical when compared to average modern human craniofacial morphologies: they still reflect the "normal" extremes of craniometric variation, and, most importantly, they have been drawn from members of contemporary (living) populations.

After a careful review of the cluster memberships and classification rates for the "paleo" cases (reported in Table 43 and Table 44), it can be said that these 24 specimens (the original crania and/or their casts) are widely and irregularly distributed (see also Figure 89 and Figure 90). For example, in the cluster solution associated with the 16-component EEE model (with 12 variables), the fossil specimens are included among the large and small clusters, with degrees of classification certainty similar to Howells' (1989) "local" series. Despite the highly identifiable morphologies that characterize each of these crania

as "ancestral" and of evolutionary significance, only one individual case, La Ferrassie I, was assigned to its own unique cluster. Even this single example is problematic as it was a cast of La Ferrassie I, not the real specimen that was specially identified. Howells' (1995) measurements for the original specimen as well as its cast (*Duckworth Lab: ALXVI*) were included in the "out-group"and each of these were treated as separate "La Ferrassie I" individuals. While the original intention behind including data for this cast in additional to measurements taken on the real specimen was to demonstrate repeatability in classification, the mixture analysis, in fact, allocated the original La Ferassie I cranium to cluster 9, and the cast to cluster 15. This pattern of allocation reveals something interesting about the Duckworth cast: this statistical differentiation between the cast and the original specimen calls into question the accuracy of the reproduction as a truthful copy of the original fossil. This finding is arguably of scientific value itself, with further research potential. It is interesting to note that both the cast and original of Skhūl V and the original of La Chapelle -aux-Saints were assigned to the (same) cluster 9, but the McGregor Restoration of the La Chapelle cranium (*Peabody Museum: 59926*) was allocated to cluster 8, alongside Djebel Irhoud 1 and Steinheim. The singular isolation of the Duckworth cast reveals, however, little about La Ferassie I's hallmark features and the distance of its relationship to other ancient and modern human morphologies. While Pontikos (2010) has stated for his analysis of the "famous" Upper Paleolithic skulls that "MCLUST assigns them to very meaningful clusters" (np), the results reported here fail to corroborate his findings (Table 43). For example, cluster 9 is interesting in that it has the greatest concentration of fossils (8 cases; 32% of the cluster) and it tends to attract the "real" crania when casts are considered, it is difficult to make easy associations between its members. The fossil allocated to cluster 9 represent a variety of species, from distant geographic locations, and spread over large time-spans. Kabwe (*Homo heidelbergensis,* from Zambia, 300-125 kya) is clustered with Shanidar 1 (*Homo neanderthalensis*, from Iraq, 45 kya) and with CroMagnon 1 (*Homo sapiens*, France, 30 kya). Moreover,

cluster 9 also contains crania from Howells' modern human series, including the Tasmania (32%), Tolai (14%), Teita (5%), Australia (5%), Zalavar (5%) and Norse (5%). While not identical, this cluster most closely resembles Pontikos' (2010) cluster 6, but it does not support the definitive characterization that he provides: "First of all, notice that cluster #6 is a Neandertal cluster. It includes La Ferassie I, La Chappelle, Skhul V, Shanidar 1, Djebel Irhoud 1. Of course I am aware that there are controversies about some of these skulls, but, MCLUST doesn't seem to have any doubts: they are all placed in cluster #6 with 100% probability and no other skulls have any probability of belonging to this cluster" (np).

If the highly structured 21-component model is looked to for further clarification of these weak "paleo" memberships, it is possible to see, in part, a more cohesive pattern in the classification of these specimens. All 24 cases are distributed over three clusters: cluster 19 contains 21% of the fossils, cluster 20 contains 46% and cluster 21 contains 33%. It is interesting to note that few of the patterns of association are repeated. One might expect that the fossils grouped in the 16-component model would have maintained that grouping when the allocation patterns were simplified, i.e., when reducing the number of cluster with some "paleo" membership from eight (16-component model) to three clusters (21-component model). Perhaps the most notable pattern is the association between Skhūl V, Djebel Irhoud 1,and Djebel Qafzeh 6 in cluster 19, but these classifications are confused by the inclusion of Abri Pataud and, more importantly, by the presence of "recent" crania from several of Howells' (1989) series: e.g., Bushman (16%), Atayal (10%), Eskimo (7%), Mokapu (7%), Moriori (7%) and Santa Cruz (6%).

Table 43: Cluster memberships for the "Prehistoric" specimens

MCLUST model	*16 EEE, 12 Variables*								*5 VVV, 12 Variables*			*21 EEE , 57 Variables*		
Specimen[18]	*1*	*5*	*6*	*8*	*9*	*10*	*15*	*16*	*2*	*4*	*5*	*19*	*20*	*21*
Abri Pataud	1	0	0	0	0	0	0	0	0	1	0	1	0	0
Afalou-bou-Rhummel 5	1	0	0	0	0	0	0	0	0	1	0	0	0	1
Afalou-bou-Rhummel 9	0	0	0	0	0	1	0	0	0	0	1	0	1	0
Australoid (Highdown?)	0	0	0	0	0	0	0	1	0	0	1	0	0	1
Kabwe 1 (Broken Hill)	0	0	0	0	1	0	0	0	0	0	1	0	0	1
Chancelade	0	0	0	0	0	1	0	0	0	0	1	0	1	0
Cro Magnon 1	0	0	0	0	1	0	0	0	0	0	1	0	0	1
Djebel Irhoud 1	0	0	0	1	0	0	0	0	0	0	1	1	0	0
Djebel Qafzeh 6	0	0	0	0	0	0	0	1	0	0	1	1	0	0
Grimaldi	0	0	0	0	0	0	0	1	0	0	1	0	0	1
Keilor	1	0	0	0	0	0	0	0	0	1	0	0	0	1
La Chapelle	0	0	0	1	1	0	0	0	0	0	2	0	1	1
La Ferrassie I	0	0	0	0	1	0	1	0	0	0	2	0	2	0
Markina Gora	0	0	0	0	1	0	0	0	0	0	1	0	1	0
Mladec 1	0	0	1	0	0	0	0	0	1	0	0	0	0	1
Predmost III	0	0	0	0	0	1	0	0	0	0	1	0	1	0
Predmost IV	0	0	0	0	0	0	0	1	0	1	0	0	1	0
Shanidar 1	0	0	0	0	1	0	0	0	0	0	1	0	1	0
Skhūl V	0	0	0	0	2	0	0	0	0	0	2	2	0	0
Steinheim	0	0	0	1	0	0	0	0	0	0	1	0	1	0
Upper Cave	0	1	0	0	0	0	0	0	0	1	0	0	1	0
Grand Total	**3**	**1**	**1**	**3**	**8**	**3**	**1**	**4**	**1**	**5**	**18**	**5**	**11**	**8**

[18] This analysis includes the craniometrics obtained by Howells for both the original specimens and/or casts(s). Each set of measurements was treated as a separate case; duplicate values (cast + specimen) are reported, therefore, for La Chapelle, La Ferrassie I, and Skhūl V.

Table 44: Implied classification rates for select series

TEST Population (Howells 1995)	*Number of Variables*	*Model (Component, Parameterization)*	*Posterior Probability > 0.7*	*Posterior Probability → 1*
Paleo Out-Group	12	16 EEE	88%	63%
Paleo Out-Group	57	21 EEE	100%	100%
Eskimo	12	16 EEE	87%	65%
Eskimo	57	21 EEE	99%	97%
Buriat	12	16 EEE	75%	39%
Buriat	57	21 EEE	97%	95%
Bushman	12	16 EEE	80%	56%
Bushman	57	21 EEE	90%	77%

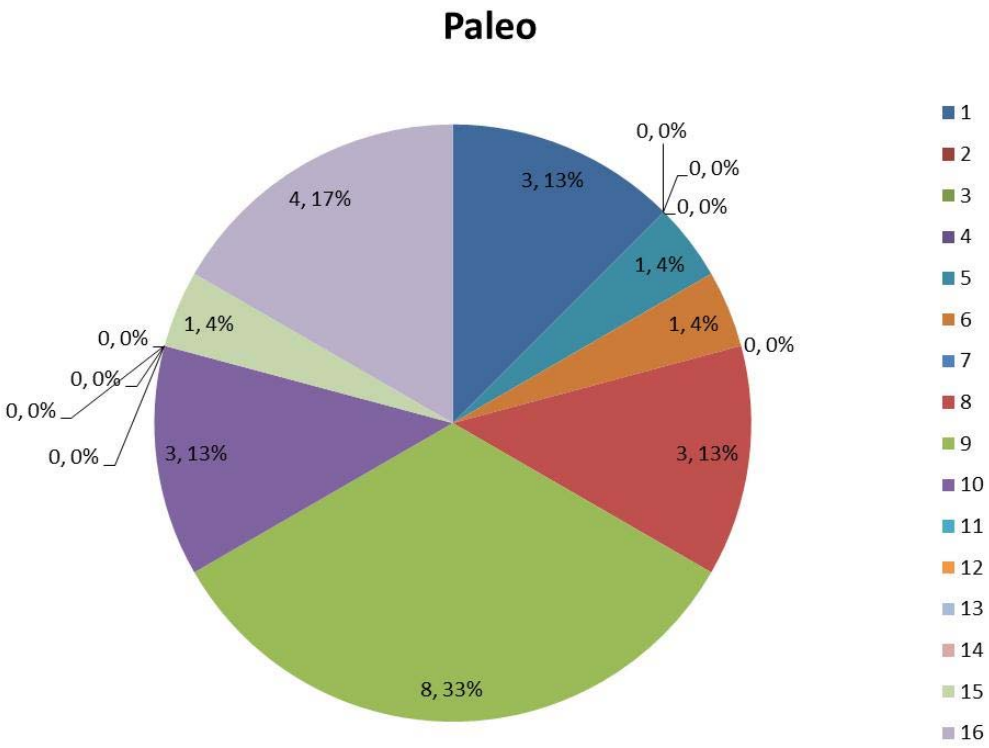


Figure 89: Cluster composition for the "Prehistoric"sample (12 variables)

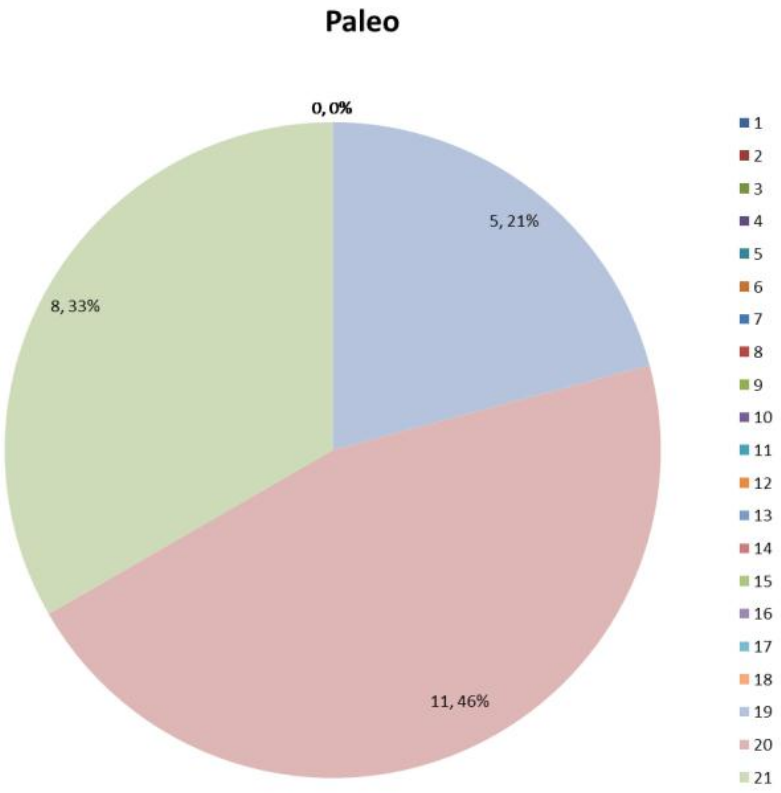


Figure 90: Cluster composition for the "Prehistoric"sample (57 variables)

While the clustering results from this project's tests of the combined sample of Howells' (1989) standard and TEST (1995) data sets do encourage a better appreciation of the influence of variable selection and sampling strategies on the ability of mixture methods to detect strong population structure, they have equally complicated any prior understanding of the effects of extreme variation on the quality of mixture solutions and the intelligibility of the results that these clustering procedures produce. There are obvious questions that remain unresolved. What does it mean when the mixture analysis clearly identifies extremes in normal human morphologies (e.g., isolating the Eskimo as an "outlying" group) but fails on most accounts to capture the exaggerated dissimilarities in shape, either, among the fossils

themselves, or, between the fossils and Howells' (1989) "recent" populations? Why does the model-based clustering procedure not treat these unique paleoanthropological specimens in a way that might be expect or as Pontikos (2010) has reported: that is, isolating the fossils in a cluster that represents an "out-group" or forming clusters that include fossils which are more similar in terms of their evolutionary history (e.g., from the same species, geographic region, time period). This project has attempted to make sense of what these findings may mean methodologically by conducting additional model-based clustering tests. Special emphasis is placed here on observing the behavior of the cluster analysis when the finite mixture methods are applied to the same combined data set, but with the paleoanthropological specimens excluded. Using the BIC selection procedure, MCLUST identified the best model (BIC = 117276.0) as ellipsoidal, with 4 components and unequal covariances (Figure 91). These findings are quite informative: this solution matches well with the cluster results obtained for the more parsimonious models produced for Howells' (1988) standard data set (Figure 92, see Table 18) and it is comparable in its level of uncertainty (Table 46, Figure 93). These similarities demonstrate that (1) the inclusion of samples from Howells TEST data, which represent atypical or problem cases and some extremes in normal human variation, did not sufficiently affect the cluster solution to alter the overall patterns produced, and that (2) the inclusion of the paleoanthropological out-group had a considerable effect on the outcome of the clustering, resulting in weakened cluster separations and obscured patterns.

In order to validate these findings, the project employed mixture based discriminant analysis (after Fraley & Raftery, 2002; c.f., McLachlan, 1992; McLachlan & Peel, 2000), which was implemented with the MCLUST package (Fraley & Raftery, 2011). All of the observations available in Howells' (1989) standard craniometric data set were designated as the required "training cases" and a subset of only

"recent" individuals from his TEST (1995) data set was selected to serve as target "unknowns." The prehistoric specimens were excluded from this target sample as the goal of this analysis is to compare classification success using data sets that represent normal non-extreme and normal extreme variation. In the first application of the mixture based discriminant analysis (MDA), the training data was modeled with a mixture density having a single normal component for each group and, then, the lowest error rate for all of the training data was identified by determining the most appropriate cross-cluster constraints, using both BIC selection and leave-one-out crossvalidation. In a second test, model-based clustering was applied to fit a normal mixture model as a density estimate for each group in the training set, considering these models separately using BIC selection only. The summary results of these analyses are reported in Table 45, and the errors produced from classification of the target data are displayed in Figure 94. While the error rates associated with modeling the training and classifying the target data may appear high, what is important to focus on is the fact that the very conservative crossvalidated error rate (using only the training data) and the target error rate (using a subset of Howell's TEST data set) are almost equal. This agreement between these two estimates indicates that the discrimination model generated from the "known" training data can classify Howells' (1995) collection of "unusual" cases with high morphological variability just as well as it can classify the "typical" cases sampled from Howells (1989) series.

Table 45: Mixture DFA results using Howells' (1989 and 1995) data sets

Training Data, n	*Target Data, n*	*Training Error*	*Target Error*	*CV-Model, Associated Error*	*BIC Estimate, Associated Model*
Howells (1989), 2504	Howells (1995), 202	0.48	0.59	EEE, 0.59	106952.55, EEE

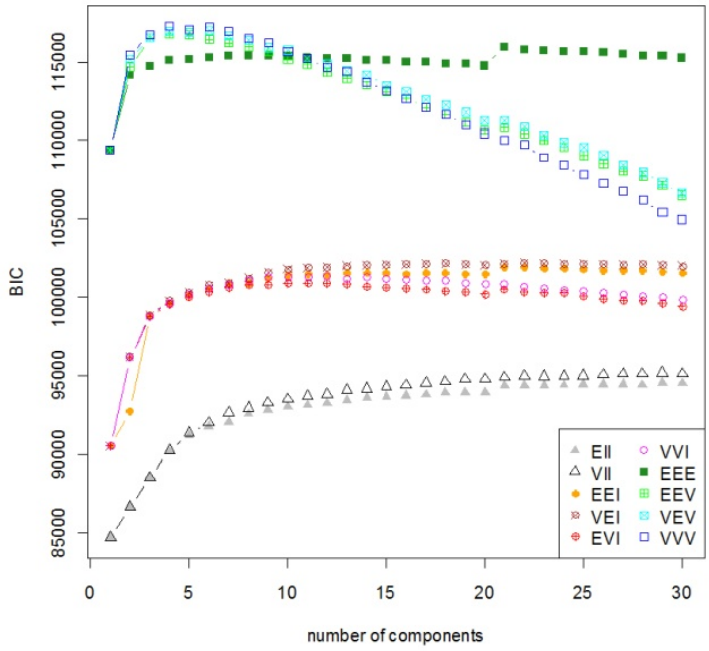


Figure 91: Model selection by BIC ("Paleo"specimens exluded)

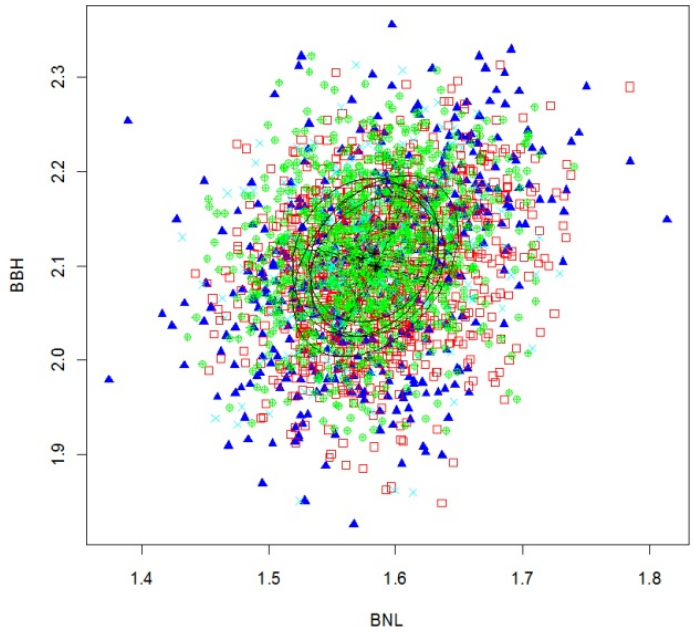


Figure 92: Coordinate projection showing clusters ("Paleo"specimens excluded)

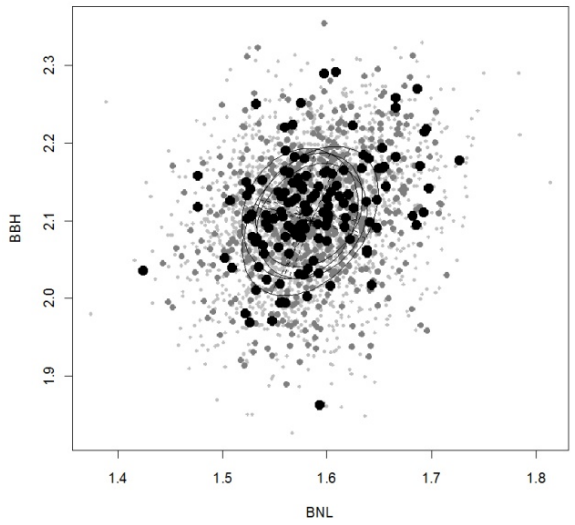


Figure 93: Coordinate projection showing uncertainty ("Paleo" specimens excluded)

Table 46: Quantiles of uncertainty ("Paleo" specimens excluded)

Quantile	*0%*	*25%*	*50%*	*75%*	*100%*
Estimate	0	0.006057065	0.027679293	0.157885070	0.556860994

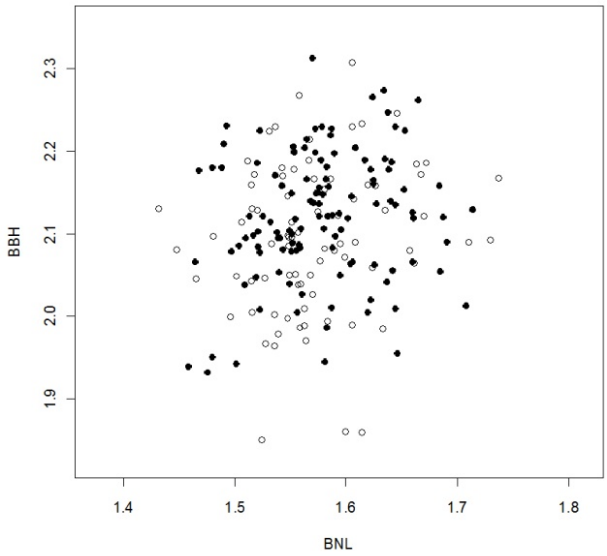


Figure 94: Coordinate projection showing errors (filled circles) for target data using mixture DFA

When the findings of these two, unsupervised and supervised, mixture analyses are taken together, they attest to the confounding effect that the "paleo" samples had on the cluster patterns and membership results. It is reasonable to conclude that the loss of clarity in the mixture analyses of Howells' (1989, 1995) combined data sets is not related to his sampling procedure but is a result of the unexpected problems arising from the decision to conduct cross-species comparisons. This preliminary evidence suggests broader consequences for research design and methodological protocol as it implies that that the behavior of the model-based clustering can be significantly altered by the inclusion of extreme outlying observations, in ways that are unexpected and not easily predicted, especially when cases are truly "unknown" and *a posteriori* assessments cannot be made. These findings clearly reiterate the cautionary statement made already by Konigsberg, Algee-Hewitt and Steadman (2009), that "we cannot recommend indiscriminately turning mclust loose..." (p. 83). It appears, therefore, that considering context is critical when model-building: even though finite mixture clustering is well-known for its ability to operate effectively without the addition of *a priori* population identifiers, this does not mean that it should be employed without giving careful thought to data choices and research goals or without structuring analyses in ways that maximize the utility of this unsupervised clustering approach.

Tests of Race: The Forensic Anthropology Data Bank

The issue of contextualizing mixture analyses gains even greater importance when consideration is given to the problem of revealing structured human variation and teasing out patterned information that is useful for finer levels of population identification. It has been demonstrated, so far, that it is possible to

detect latent population structure within a mixture of craniometric distributions when the analysis is concerned with global variation and it considers many geographically diverse populations. Without the inclusion of external demographic metadata or "proxies" for true relationships, a correspondence was found between statistically inferred clusters and populations. In many cases, natural groupings formed among geographically proximate, evolutionarily or historically related, and biosocially similar populations: the Teita, Dogon, and Zulu consistently clustered together in a sub-Saharan African group, clusters containing Native American populations typically included North and East Asians, and the European Berg, Norse, Zalavar and the North African Egyptians were commonly joined together, demonstrating not just biogeographic (European) affiliations but representing broader Caucasian variation than transcends continental borders. This ability to identify clusters of individuals that have biogeographic and biosocial meaning using world-wide data offers promising support for the question of race in forensic anthropology; however, further refined testing is necessary.

It has been shown that population structure is hierarchical, with nested populations occurring at various levels of stratification. The ability to generalize population relationships in terms of geographic regions, continental ancestries or major races and produce different, but still meaningful results suggests that these changes in scope can be negotiated by the researcher to often profitable ends. This project has yet to demonstrate, however, if this structure can be detected at much finer levels of subdivision, e.g., in the context of the self-declared and census-defined races found within the modern American population-at-large. This project's discussion of "race" in forensic anthropology is framed around the model-based clustering analysis of a large craniometric sample of 918 modern Americans drawn from the Forensic Anthropology Data Bank (FDB). This reference data set is especially suited to this test of finer-grained structure and to advancing the project's broader interest in the meaning and utility of race

because the FDB (1) represents the diversity in self-identification and the variation in morphology expected for the United States and (2) closely approximates the population-at-large from which most North American forensic casework is expected to be drawn.

In a recent discussion on the evidentiary value of estimating sex and reporting race in human identification casework, Konigsberg, Algee-Hewitt and Steadman (2009), have suggested that "forensic anthropologists are 'so good at identifying' races because practicing forensic anthropologists are essentially implicit Bayesians " (p. 86), which is to say, that forensic practitioners often "bring prior information to their cases" (p. 84). This information may include the practical knowledge of recovery location and conditions, as well as experiential knowledge of the commonalities between morphology and individual biosocial identity that underlie the morphoscopic estimation of key demographic parameters, including ancestry and ultimately race in forensic casework (c.f., Rhine,1990, 1993; Hinkes, 1993; Hefner, 2009). The incorporation of this information into statistical estimates of identity is well known and implied, for example, in the directions given for the proper use of the FORDISC software: "the researcher guides the analysis by choosing the populations against which to classify the unknown." (Freid et al., 2005, np).

For this reason, this project is justified in dividing the FDB sample into three test groups, which represent progressively smaller subset of the U.S. population-at-large: (1) all five races, (2) White, Black and Hispanic, and (3) White and Black. This project has chosen these designations because they reflect the recording practices for documented or self-reported race in the Forensic Anthropology Data Bank (Jantz, 2008) and they correspond closely to the race and ethnicity categories defined by the Federal

Office of Management and Budget (OMB) and used by the U.S. Census Bureau (2010) in their summary "race" reports (see Table 47).

Table 47: U.S. Census (2010) and FDB (2008) race statistics

U.S. Census Populations 2010	*Percent Representation*	*FDB Populations*	*Cases*	*Percent Representation*
USA	Total population: 308745538	5 Major Races (missing data deleted)	918	
White persons	72.40%	White	618	63.00%
White persons not Hispanic	63.70%	Black	210	23.00%
Black persons	13%	Hispanic	20	2.00%
Persons of Hispanic or Latino origin	16.30%	Native American	16	1.70%
American Indian and Alaska Native persons	0.90%	East Asian	54	6.00%
Native Hawaiian and Other Pacific Islander	0.20%			
Asian persons	4.80%			

Perhaps most importantly, these race designations correspond to the language used in medico-legal investigations, for which the forensic anthropologist is charged with translating information on the unknown individual's "identity" obtained from the assessment of skeletal biology into terminology that has cultural currency and practical utility in missing persons and human identification casework (Sauer, 1992). It is for this reasons that "Hispanic" is included among the five key test populations, even though the U.S. Census Bureau (2011) assumes that "Hispanics may be of any race" and may be incorporated into other "applicable" race categories. It is acknowledged here that the case may be similar for the FDB data, where individuals of Hispanic origin may have self-identified as Hispanic and/or as some other race, e.g., Black, White, admixed. This project's departure from the Census identification guidelines is supported here with the fact that the most recent Census Report states that only 2.9% of the total

population identified as two or more races. Moreover, this project makes the argument that "Hispanic" ethnicity meets the same criteria as modern American races, when consideration is given to the definition provided by the Census: a composite biosocial group, with complex histories and varying degrees of admixture (U.S. Census Bureau, 2011). Admixture mapping and ancestry testing have revealed, for example, similarly fluctuating levels of stratification among Black Americans and Hispanics. Nei and Livshits (1989) isolated 20% of American Black genetic make-up as White. Shriver and colleagues (1997) identified U.S. Hispanics as 60% European. Recent studies have also shown that these values are heavily influenced by location of individual origin (c.f., Parra et al., 1998; Bryc et al., 2010). While Bryc et al., (2010) have reported for African Americans that "the median proportion of European ancestry is 18.5% (25th–75th percentiles: 11.6–27.7%)," they also noted that there is "very large variation among individuals" (np). Similar degrees of variation have been detected among Hispanics. Approximately 98% of the rural Mexican population of the State of Guerrero has a substantial AmerIndian ancestry, while at the same time cohabiting with a genetically defined sub-population of Mestizos (Bonilla et al, 2005). Mexico City studies also show fluctuations in the percentages of parental combinations, reflecting the diversity of within-city populations and, in turn, the potential variability among the samples selected: Tiburcio et al., (1978) identified 27.6 % AmerIndian, 70.8% European, and 1.4% African parental contributions; Lisker et al., (1988) reported 56.3% Indian, 33 % White, and 10.7% Black genes among individuals from the state of Puebla; and Cerda-Flores et al., (2002) documented, for their sample of Mestizo populations, parental percentages of 37-49% Amerindian, 50-60% Spanish-European, and 1-3% African. It is also suspected here that there may be similar instances of overlap between other closely related group, like East Asians and Pacific Islanders or Native Americans and Native Hawaiians. This fluidity in biosocial identification should be taken into account when considering the strength of the clustering results for the FDB data by race category.

This project begins its discussion of race in America by reporting the results obtained for the model-based clustering of a mixture of FDB craniometric data (n=918) for five populations that represent biosocial races: White, Black, Hispanic, East Asian, and Native American. The top ten models identified by MCLUST using the BIC selection method (Fraley & Raftery, 1998) are presented in Table 48 and the behavior of the BIC over the range of models tested (30 components, 10 parameterizations) is graphically displayed in Figure 95. For the purpose of method comparison, the optimal cluster number identified using EMMIX, following McLachlan's (1987) resampling procedure, is also included here. The best MCLUST model is discussed here in greater detail. The summary statistics for the implied classification success and overall quality of this 7-component EEE model are presented in Table 49 and Table 50. The corresponding visualizations of the cluster membership patterns and classification uncertainty are given in Figure 96, Figure 97, and Figure 98 (c.f., McLachlan & Peel, 2000; Ousley & Jantz, 2005; Fraley & Raftery, 2011).

Table 48: Model selection 5 races

FDB Data Set (2008): 5 races and 12 variables			
Component Number	*Model Form*	*Selection Statistic*	*Estimate*
7	EEE	BIC	38724.54
6	EEE	BIC	38694.26
5	EEE	BIC	38662.9
4	EEE	BIC	38650.57
9	EEE	BIC	38631.02
8	EEE	BIC	38622.46
10	EEE	BIC	38593.65
13	EEE	BIC	38577.84
11	EEE	BIC	38545.3
14	EEE	BIC	38531.44
5	arbitrary	-2logλ	262.01, p-value 0.01

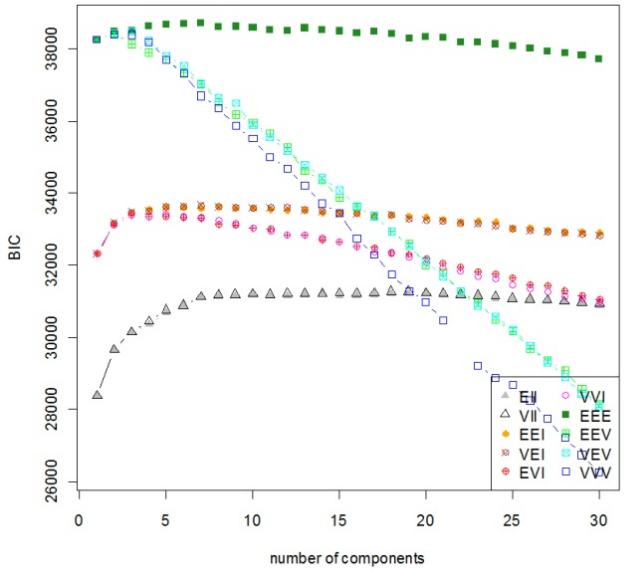


Figure 95: Model selection by BIC

Table 49: Allocation rates

FDB 5 Races: 7 clusters	*posterior probability > 0.7*	percent correctly allocated	*posterior probability → 1*	percent correctly allocated	*total number of individuals*
Black	175	83	130	62	210
East Asian	41	76	29	54	54
Hispanic	18	90	11	55	20
Native American	11	69	6	38	16
White	507	82	343	56	618
Grand Total	752	82	519	57	918

Table 50: Quantiles of uncertainty

Quantile	*0%*	*25%*	*50%*	*75%*	*100%*
Estimate	0	0.02134887	0.07170227	0.22671006	0.62585639

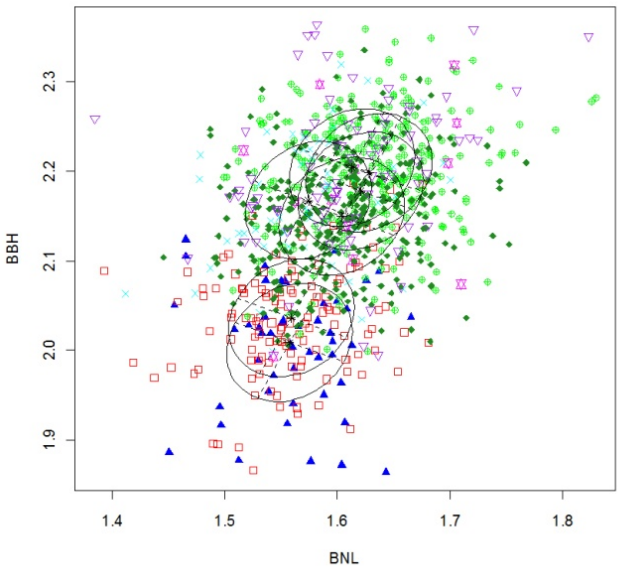


Figure 96: Coordinate projection showing clusters

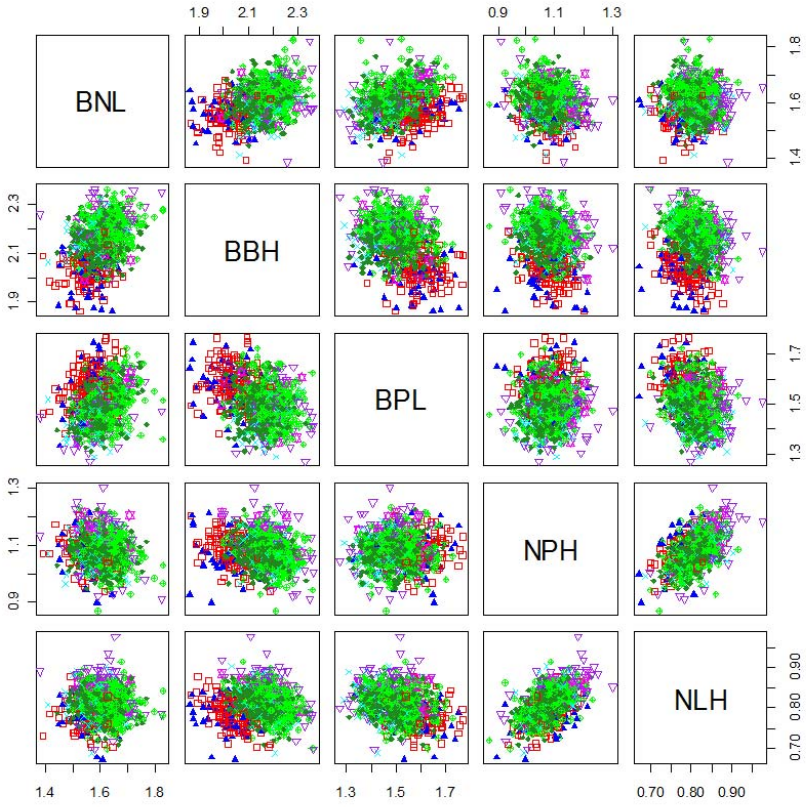


Figure 97: Pair-wise scatterplot matrix showing clusters

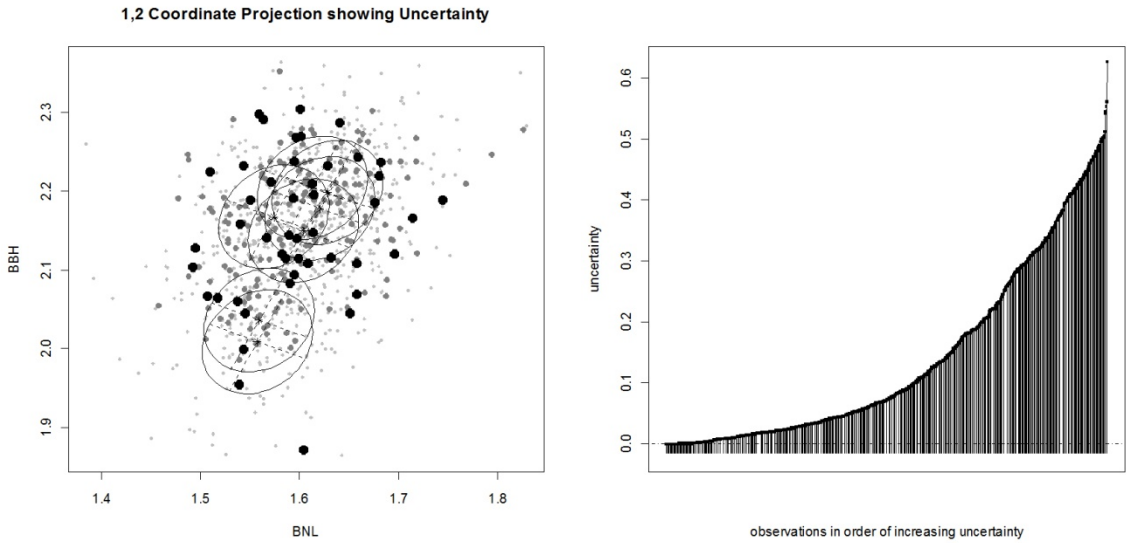


Figure 98: Coordinate projection and data profile showing uncertainty

While there is no direct correspondence between the number of components and the number of true groups, the reasonably high posterior-probabilities (and in turn, implied rates of classification success) suggest strong underlying structure. Similarly, while no one cluster accounts for the variation present within a single race, as the per-population allocations are spread across the 7 clusters (see Table 51), when these classifications are more simply displayed, it is possible to detect patterns that suggest meaningful trends of association and admixture. Pie charts representing the cluster membership of each race-based group are provided in Figure 99.

Table 51: Cluster memberships by race

Race	1	2	3	4	5	6	7	Total Cases
Black	46	106	14	3	0	39	2	210
East Asian	1	5	12	2	0	29	5	54
Hispanic	0	1	7	0	0	9	3	20
Native American	0	2	3	4	1	6	0	16
White	2	12	236	42	11	246	69	618

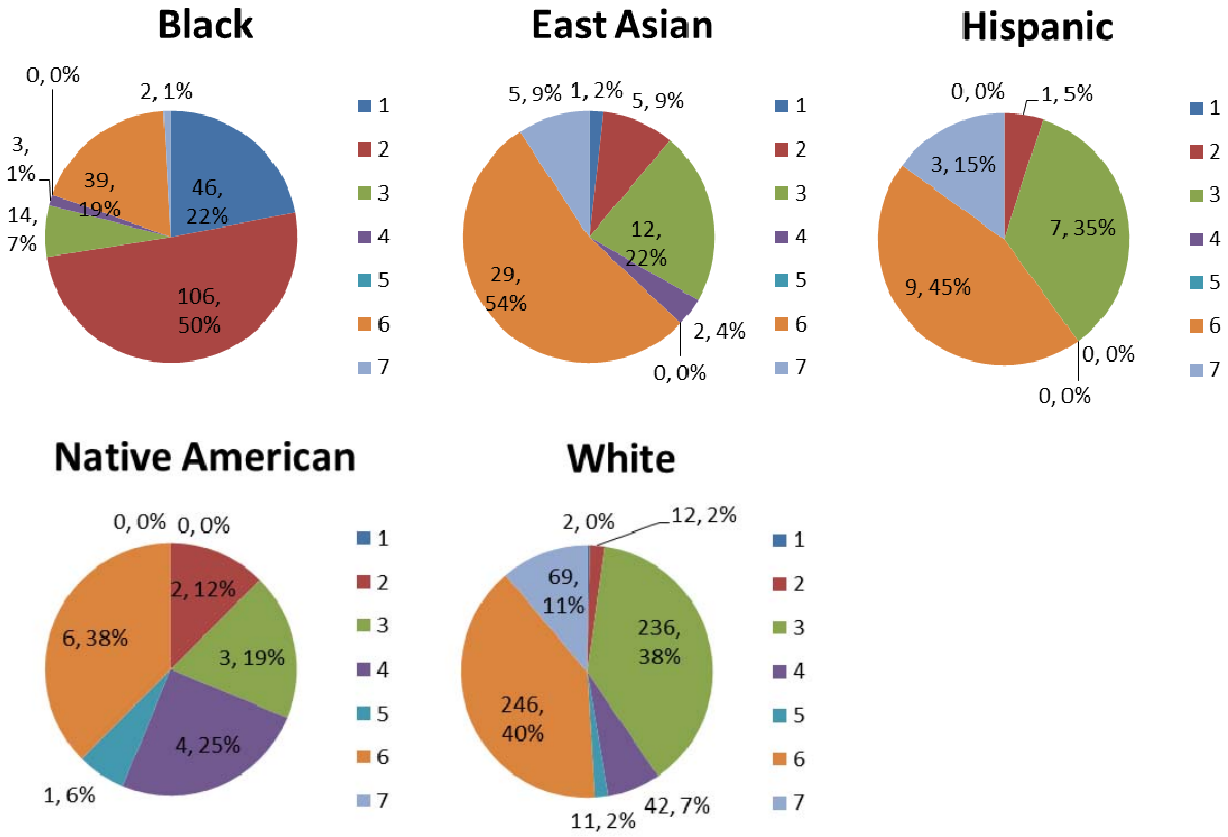


Figure 99: Black, East Asian, Hispanic, Native American, White

From the population-specific pie charts, it can be suggested, first, that clusters 1 and 2 represent the Black component of variation. It is posited that cluster 2 may correspond more closely to the morphological features associated with African origins as it is shared among all groups. Given the known history of admixture among West African slaves and their descendents with Native American populations, it should not be surprising to find shared membership between these two groups (Parra et al., 1998). The presence of cluster 2 within the Asian group may represent the same type of evolutionary patterns that were observed in the global studies, where certain African, East Asian and Oceanic groups shared morphological signatures. The low level of Black contribution to the composition of the Hispanic

group also agrees with what is known about the data under analysis. The majority of the cases making up the FDB Hispanic category represent individuals from the American Southwest and Mexico, whose morphology and genetic ancestry show close ties to Native Americans and Europeans, e.g., with estimated ancestral proportions of 63% and 34% respectively (c.f., Bonilla et al., 2004; Spradley et al., 2008). It is not surprising, then, that they also differ considerably in their morphology from Afro-Caribbean Hispanics, like Cubans, who most often classify as Black (Ross et al., 2004,) and have been shown to be primarily of West African genetic ancestry, e.g., 87.9% (c.f. Parra, Kittles & Shriver, 2004; Benn-Torres et al., 2008). It is suspected here that the Black group is separated largely over 2 clusters on account of the well-documented incidence of secular change in craniofacial morphology among American Blacks (Jantz & Meadows Jantz, 2000; Jantz, 2001; Wescott & Jantz, 2005), where the smaller cluster 1 represents a unique and recent African American trend. It is also possible to identify a corresponding 2 cluster pattern in the American Whites, which may be interpreted in similar terms: cluster 6 represents the shared evolutionary component and cluster 7 captures relatively recent developments in craniofacial shape. Ousley, Jantz and Freid (2009) have also considered secular change as a viable explanation for the extra splitting patterns observed in their k-means cluster solution of FDB data.

It is notable that the Hispanic group shares 95% of its cluster membership with American Whites. This close correspondence between Hispanics and Whites agrees with the expectations of the U.S. Census (2011): it is common for members of Hispanic ethnicity to self-report as White. This relationship may also speak to the relatively recent formal acceptance of "Hispanic" as a documented biosocial identity. It only first appeared in 1980 as an option on the U.S. Census questionnaire and even still in 2000 there

remained considerable ambiguity in its meaning and confusion over its quantification, as the following OMB (1997) directives attest:

> People of Hispanic origin may be of any race and should answer the question on race by marking one or more race categories shown on the questionnaire, including White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, and Some Other Race. Hispanics are asked to indicate their origin in the question on Hispanic origin, not in the question on race, because in the federal statistical system ethnic origin is considered to be a separate concept from race. (U.S. Census Bureau Population Division, 2011, np).

Finally, cluster 3 is taken here to account for the Asian components of variation, whose presence among the White, Native American and Hispanic groups can be explained by their shared evolutionary history; cluster 4 may serve to distinguish the American Indian morphology within these groups.

This project also evaluated the membership patterns among these five races produced under the other "good" MCLUST models: as the number of clusters was increased and decreased, the patterns remained consistent (e.g., see Figure 100 for the 5-component clustering). Blacks and Whites were distributed over two clusters each. Hispanics were almost always assigned to a predominantly White cluster. Native Americans and East Asians were commonly joined in the same clusters, shared obvious membership with the White group, and were more irregularly assigned to predominantly Black clusters. What is most clearly visible regardless of the "level" of the analysis (e.g., few or many cluster solutions) is a Black and White divide, where each of these groups represent one extreme of the range of the total (five race) variation. This pattern is supported by the fact that the Black and White (including "White" Hispanic) groups have the lowest inferred classification errors and the highest incidence of outright clustering (i.e.,

posterior probabilities ≈ 1.0). The Black group alone occupies the extremes of the uncertainty values, which not only attests to the well-known fact that African populations exhibit the greatest amount of within population diversity but also shows how average patterns of craniometric variation is useful for providing evidence of past population history (e.g., presumably, here, West African origins) despite significantly changed environments (c.f., Parra, Kittles & Shriver, 2004).

For comparison, this project also considered the best EMMIX model to assess how reliably these trends could be detected, under different methodological and modeling conditions. It was found that the EMMIX 5-component model, which corresponds directly to the true number of groups, failed to provide an equally strong or nuanced pattern of memberships. In general, it produced a more even distribution of groups across the clusters (e.g, the Black cases appear clearly split among three clusters, in almost identical numbers) and there was a concomitant loss of population-specific information. In contrast, however, the White and Hispanic relationship was strongly maintained and the Black and White observations were still "differently" allocated. This dissimilarity in clusters patterns between MCLUST and EMMIX solutions may be in part due to the choice of covariance structure: MCLUST returned, in all top ten models, an equal variance (EEE) parameterization, while arbitrary covariances were specified for the EMMIX analysis. Following the recommendations of Peel and McLachlan (1999), this analysis used unconstrained variances because the true relationships among the sampled observations are (or are taken to be) unknown. This imposed "freedom" may have, in fact, produced a weakened the EMMIX model.

The estimates of the inferred classification success of the chosen 7-component MCLUST model (Table 49) demonstrate, for this five race data set, that the Native American and East Asian groups have the

greatest level of membership uncertainty, and in turn, the highest rate of misclassified observation. From a purely statistical perspective, the inclusion of these more weakly differentiated populations may only serve to introduce noise into the "best" model, significantly affect the performance of the chosen model, and, ultimately, detract, from the overall information value of the multi-group analysis. It is of interest, therefore, to reassess this clustering solution with these "problem" groups removed. To this end, model-based clustering was applied to a reduced sample that included only Black, Hispanic and White cases from the FDB (n = 848). The top ten models identified by MCLUST using the BIC selection method (Fraley & Raftery, 1998) as well as the best EMMIX model are reported in Table 52; the behavior of the BIC (Figure 107) over the range of models tested (30 components, 10 parameterizations) is also graphically displayed.

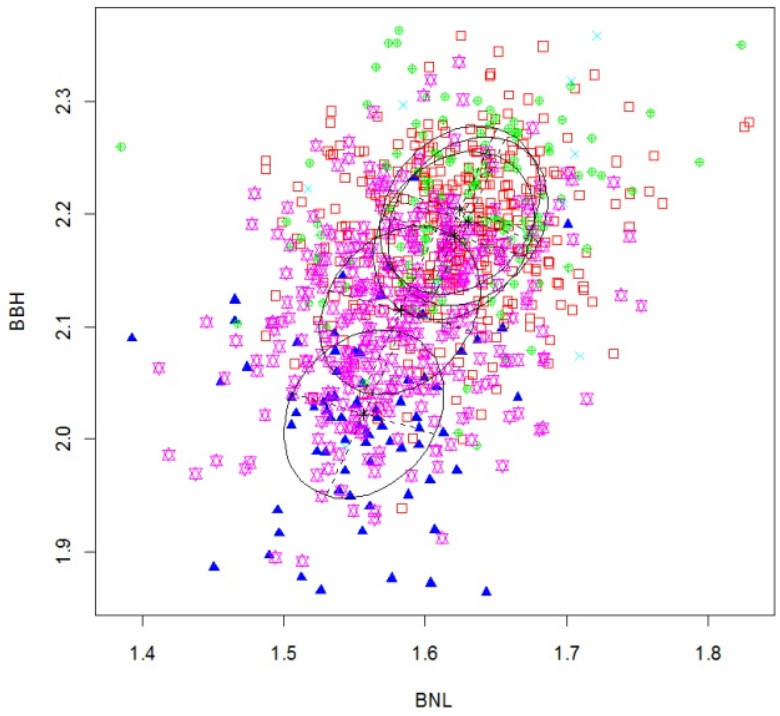


Figure 100: Coordinate projection showing clusters

Table 52: Top 10 models in order and EMMIX model

FDB Data Set (2008): Black Hispanic and White races, 12 variables			
Component Number	*Model Form*	*Selection Statistic*	*Estimate*
8	EEE	BIC	35609.89
10	EEE	BIC	35589.83
6	EEE	BIC	35586.77
9	EEE	BIC	35574.35
7	EEE	BIC	35566.21
11	EEE	BIC	35530.59
5	EEE	BIC	35530.27
4	EEE	BIC	35495.56
12	EEE	BIC	35463
13	EEE	BIC	35398.45
4	arbitrary	-2logλ	361.06, p-value 0.01

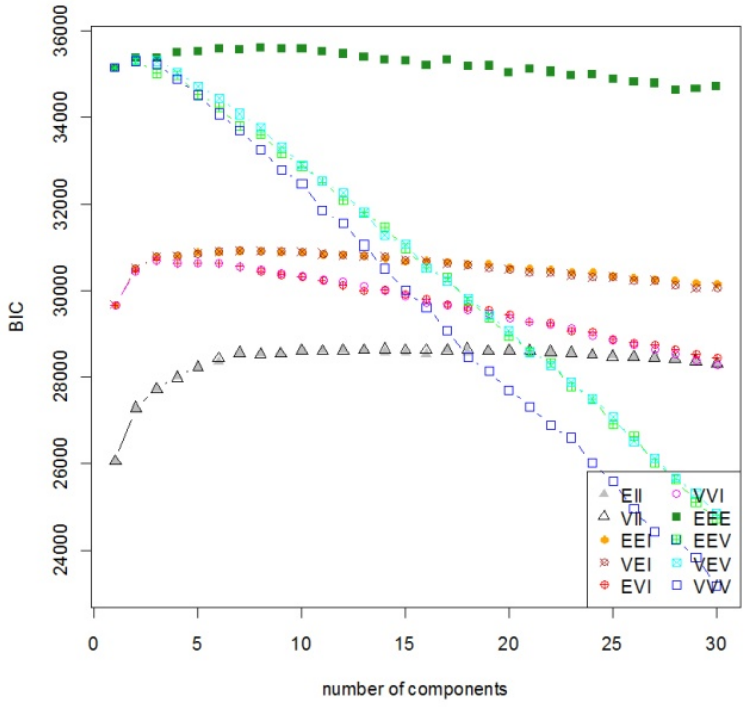


Figure 101: Model selection by BIC

The summary statistic on classification success and overall model quality are reported in Table 53 and Table 54 for the best identified MCLUST solution: 8-components with equal covariances. The corresponding cluster membership patterns and classification uncertainty are also displayed in Figure 102, Figure 103, and Figure 104 (c.f., McLachlan & Peel, 2000; Ousley & Jantz, 2005; Fraley and Raftery, 2011).

Table 53: Allocation rates

FDB 3 Races: 8-cluster	*posterior probability > 0.7*	*percent correctly allocated*	*posterior probability → 1*	*percent correctly allocated*	*total number of individuals*
Black	170	81	135	64	210
Hispanic	17	85	13	65	20
White	565	91	468	76	618
Grand Total	752	89	616	73	848

Table 54: Quantiles of uncertainty

Quantile	*0%*	*25%*	*50%*	*75%*	*100%*
Estimate	4.440892e-16	3.341778e-03	2.013955e-02	1.207219e-01	6.165571e-01

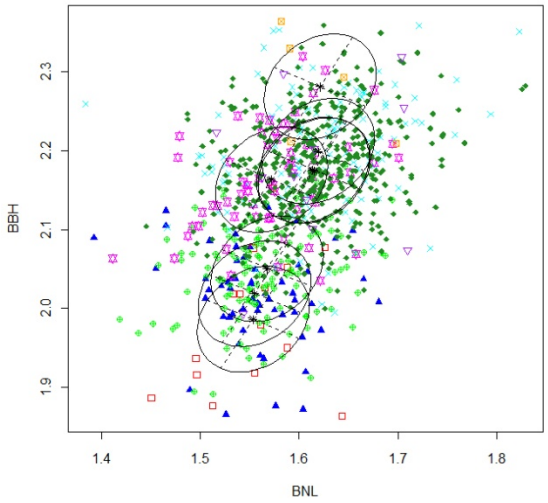


Figure 102: Coordinate projection showing clusters

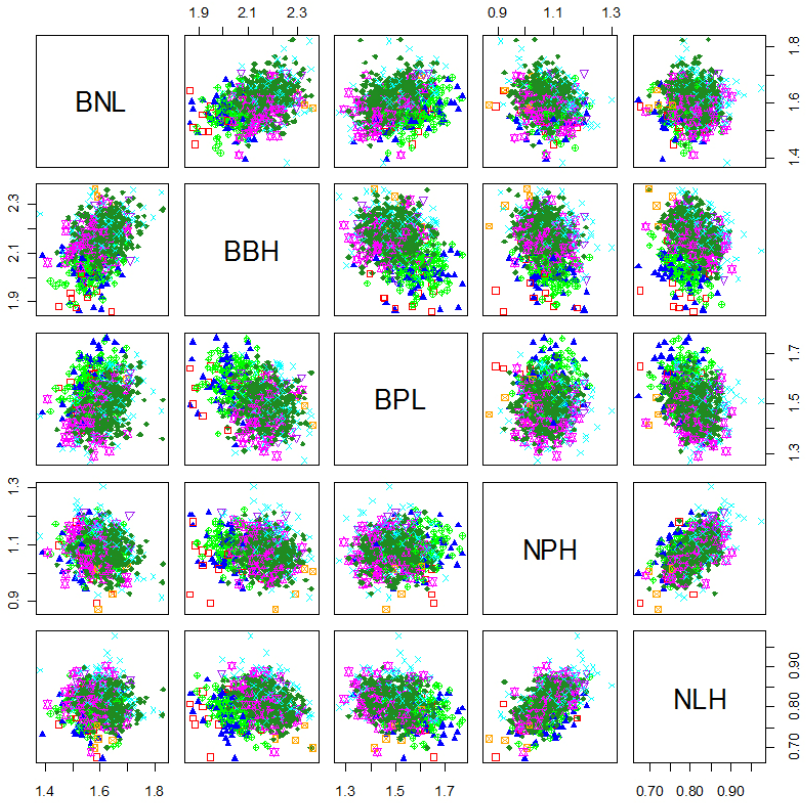


Figure 103: Pair-wise scatterplot matrix showing clusters

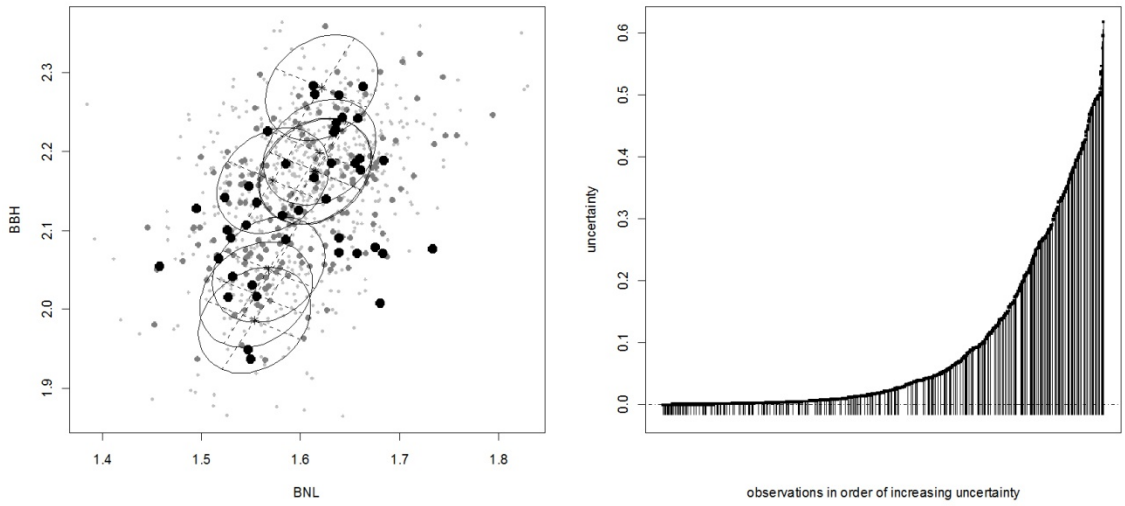


Figure 104: Coordinate projection and data profile showing uncertainty

It is interesting to note that the chosen MCLUST model exceeds the true number of groups. If the best model choice is considered in isolation (Table 53), it can be shown that, despite an attempt to reduce noise and, in turn, simplify the cluster solution, the reduction in populations has brought about, instead, an increase in model complexity and a decrease in cluster clarity: the component number grew from 7 to 8, but the model retained its EEE parameterization. If this discussion is expanded to also review the measures of model fit (e.g., classification success and uncertainty; Table 53 and
Table 54), error is relatively low even though the component-number is high. This suggests that the structure that underlies this complex clustering is strong; it can be expected that, in turn, the patterns revealed by this analysis should carry some interpretable meaning. To determine if this meaning is of biological relevance and informative of race, pie charts were produced to display group-specific variation in terms of cluster make up and to permit an easier assessment of relationships between the races as well as the magnitude of their separation. As Figure 105 shows, the removal of the East Asian and Native American samples did not sufficiently simplify the clustering solution, so as to produce a direct relationship between cluster number and race. Nevertheless, there are clear patterns of relationship that reinforce the previous findings. There is some overlap between the Black and White groups, as one would expect for the FDB's "African American" samples given the documented history of admixture and recent patterns of gene exchange in the United States (versus *within*-Africa, [c.f. Collins-Schramm et al., 2002; Bryc et al., 2010; Stefflova et al., 2011]). The Black and White groups are, however, not symmetrical in their cluster membership: the Black sample is more like the White sample, than the White sample is like the Black (i.e., 14% of Blacks are allocated into cluster 3, while only 3% of Whites are assigned to cluster 3; cluster 4 is exclusively White and Hispanic). It is notable that these findings correspond with the genetic expectations that document similarly directed gene exchange, where mean African admixture among White Americans is comparably low (e.g., roughly 0.7% African

and 99.3% European admixture, [c.f. Shriver et al., 2003; Goncalves et al., 2007]). Each of these two groups also demonstrates their own, largely unique component of variation, e.g., 26% of the Black FDB data is contained within cluster 1, which is to be expected given the broad evolutionary (spatial and temporal) divergence between this group's European and African ancestries.

The Hispanic sample closely resembles the White FDB sample in its cluster makeup, but, unlike the White or Black groups, it is the only population without a unique or an exclusively "Hispanic" cluster. This is a likely effect of the composite biosocial make up of "Hispanic" as a mixture of already predefined groups or races (c.f., Census, 2011). It also reflects current genetic data: Parra, Kittles and Shriver (2004) found, for example, that the Hispanic individuals have both European (62.7 ±2.1%) and Indigenous American ancestry (34.1 ±1.5%), with a relatively small proportion of West African ancestry (3.2 ±1.5%). The strength of the data structure that enables the division of these groups is better demonstrated and the clarity with which these population specific patterns can be identified are further enhanced when the population make up of each cluster is evaluated. In Figure 106, it is shown that cluster 1, 2, and 3 contain the American Black samples and clusters 4 through 8 contain the American White samples; the individuals identified as Hispanic are largely allocated to the White clusters (15 cases in cluster 4 and 6 versus only 5 cases in clusters 1 and 3; with cluster 3 also having the largest White component of the Black clusters). This mixture analysis is able, therefore, to effectively differentiate between Black and White morphology, albeit over several clusters. The subsetting of this variation into multiple Black and White clusters may reflect the complex patterns of admixture and genetic sources that characterize the American population-at-large and which complicate the use of singular terminology (e.g., simply Black or White). This point was made very recently by Stefflova and colleagues (2011) for genetic analyses, who wrote that,

> We have confirmed that there is a genetic thread connecting Africa and the Americas, where each colonial system supplied their colonies in the Americas with slaves from African colonies they controlled or that were available for them at the time. This historical connection is reflected in different relative contributions from populations of W/WC/SW/SE Africa to geographically distinct Africa-derived populations of the Americas, adding to the complexity of genomic ancestry in groups ostensibly united by the same demographic label (np).

This project has shown so far that, regardless of the number of populations included without the mixture analysis (using the FDB or Howells or even the combined Howells and Hanihara data sets), the MCLUST and EMMIX clustering methods identify a clear Black and White or African and non-African divide. In order to further explore the relationship between the two groups which appear to be driving all of these cluster solutions, FDB sample was reduced so as to consider data for American Whites and Blacks only, in the final iteration of these tests of fine-grained population structure. In Table 55, the top ten models identified by MCLUST using the BIC selection method (Fraley & Raftery, 1998) as well as the best EMMIX model are presented, for purposes of comparison. A graphical representation of the behavior of the BIC over the range of models tested (30 components, 10 parameterizations) is given in Figure 107. Despite this extreme reduction in the populations included in this mixture clustering, the BIC is again maximized by a model with many more groups than represented in the data: 7-components and equal covariance matrices. It is notable that this solution corresponds to the same number of clusters proposed for the initial FDB analysis, which considered the full five races of interest (Black, East Asian, Hispanic, Native American and White).

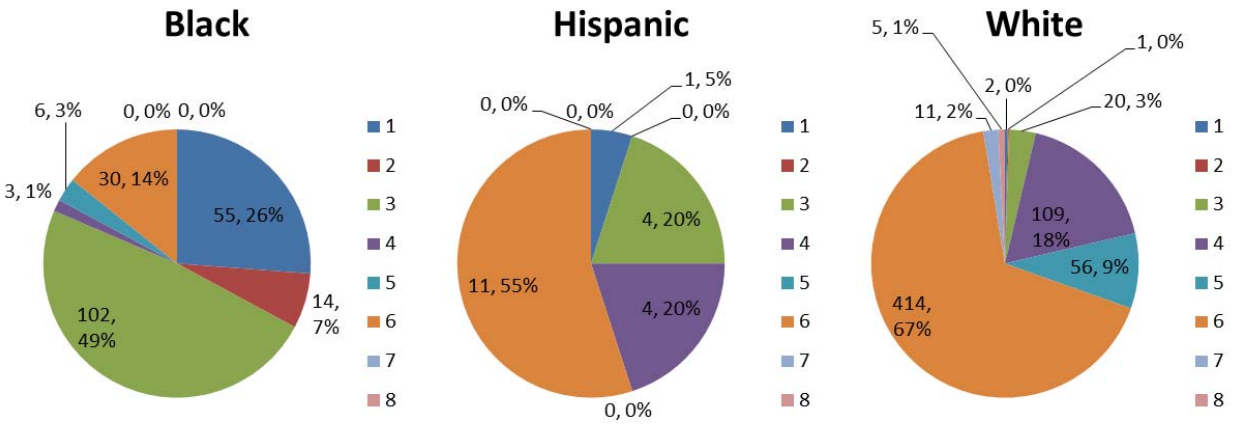


Figure 105: Black, Hispanic, White

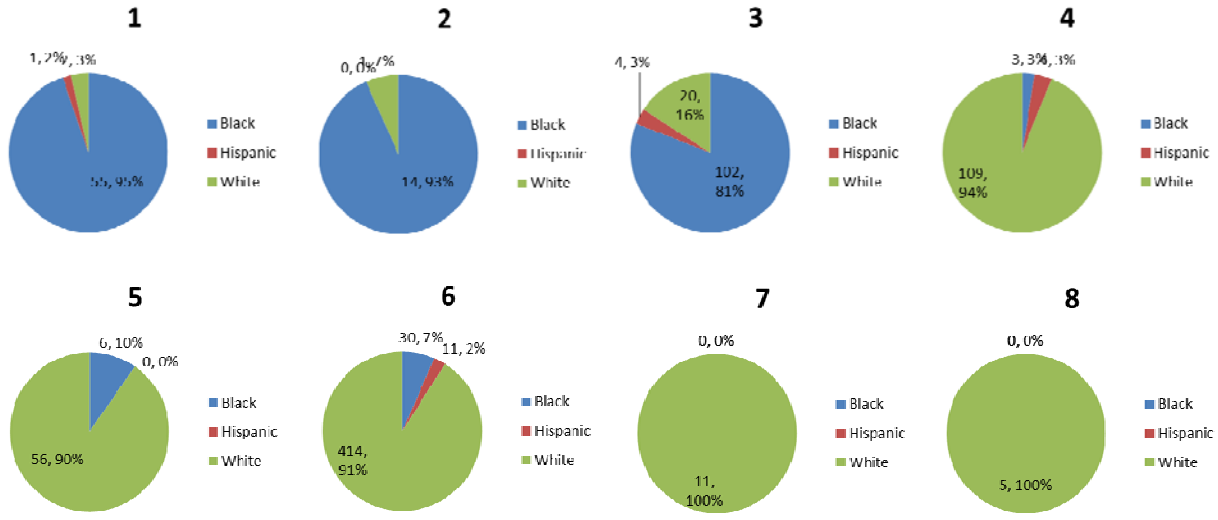


Figure 106: Cluster composition by race

Table 55: Top 10 models in order and EMMIX model

FDB Data Set (2008): *Black and White races, 12 variables*			
Component Number	Model Form	Selection Statistic	Estimate
7	EEE	BIC	34720.42
8	EEE	BIC	34703.38
6	EEE	BIC	34645.38
5	EEE	BIC	34639.25
10	EEE	BIC	34635.71
11	EEE	BIC	34634.11
12	EEE	BIC	34614.02
9	EEE	BIC	34606.26
13	EEE	BIC	34575.98
4	EEE	BIC	34541.71
4	arbitrary	-2logλ	371.75, p-value 0.01

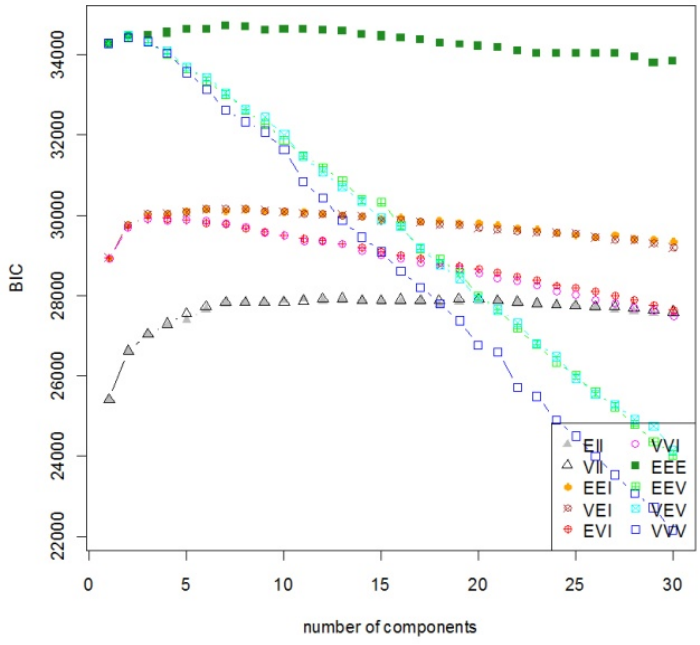


Figure 107: Model selection by BIC

While the optimal MCLUST model is complex, given its large component number relative to the true number of races considered, it is among the models with the fewest components: in fact, the true solution ranks only 13th among the EEE models, suggesting that solutions with low component number (i.e., more parsimonious) may not adequately capture the degree of structured variation within this mixture. For this reason, the 7-component EEE model is adopted for further scrutiny. Table 56 and Table 57 provide the associated summary statistics on classification success and overall model quality for this 7-cluster solution. The cluster membership patterns and classification uncertainty are also displayed graphically in Figure 108, Figure 109, and Figure 110 (c.f., McLachlan & Peel, 2000; Ousley & Jantz, 2005; Fraley & Raftery, 2011).

Table 56: Allocation rates

FDB 2 Races: 7-cluster	*posterior probability > 0.7*	*percent correctly allocated*	*posterior probability → 1*	*percent correctly allocated*	*total number of individuals*
Black	177	84	129	61	210
White	570	92	483	78	618
Grand Total	747	90	612	74	828

Table 57: Quantiles of uncertainty

Quantile	*0%*	*25%*	*50%*	*75%*	*100%*
Estimate	0	0.002118342	0.015644170	0.105810234	0.633808573

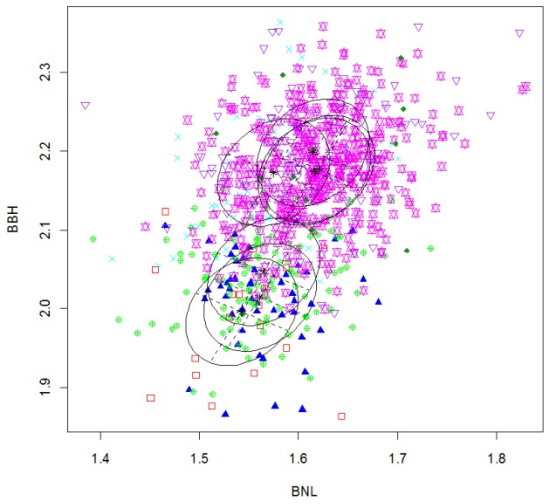


Figure 108: Coordinate projection showing clusters

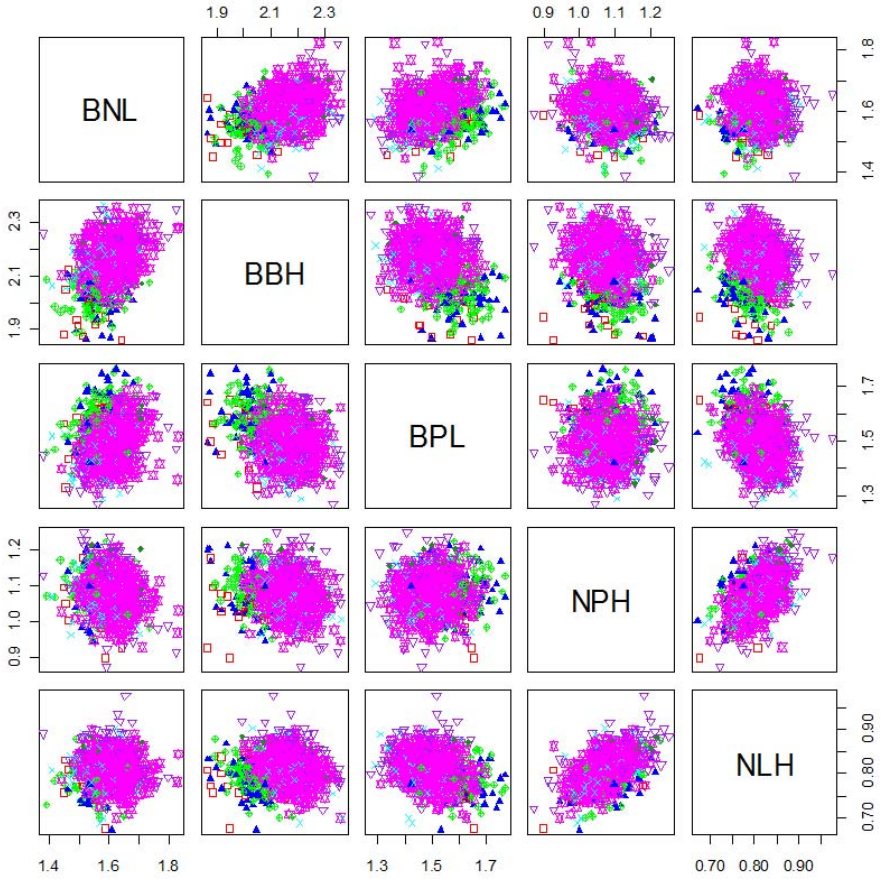


Figure 109: Pair-wise scatterplot matrix showing clusters

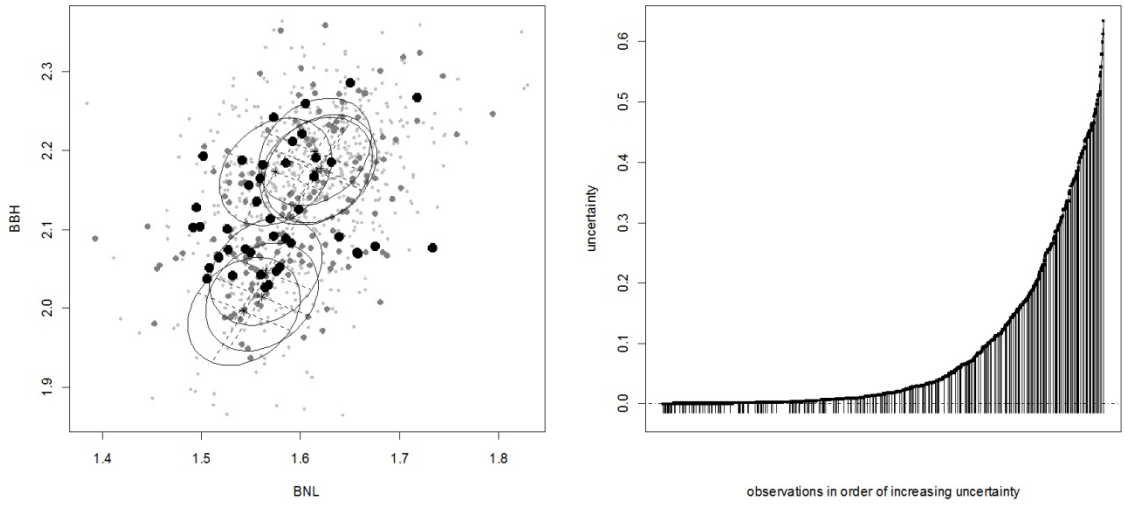


Figure 110: Coordinate projection and data profile showing uncertainty

What is of particular interest here is that, despite the reduction in the number of true races and the retention of the large number of clusters, a pattern nevertheless emerges when clusters are assessed in terms of their composition. As Figure 111 shows, clusters 1, 2 and 3, clearly capture the Black component of the mixture's variation, while clusters 4, 5, 6 and 7 represent the variation present among the White cases. The configuration of these clusters should come as no surprise: it is essentially a repetition of the same pattern of Black versus White clusters that has emerged throughout all of the analyses, at all level of hierarchy. This organization is made clearer when the number of cases assigned to each cluster is considered. Most of the White individuals (434 cases) are grouped together in cluster 5, while the Black individuals are more evenly dispersed across clusters 1 and 3 and, to a lesser degree, cluster 5. It appears, therefore, that this complex 7-component model could be easily collapsed into a 3-component model, consisting of 1 White cluster and 2 Black clusters. As before, consistent trends can be detected: (1) the Black population is largely allocated into 2 groups, which may be suggestive of within population divergence related secular change, (2) the greatest within-population diversity is found in the

Black population, (3) the percent contribution of White variation across all of the "Black" clusters (average of 11%) corresponds in general to recently reported estimates of Black-White admixture for the American population (c.f., Bryc et al., 2010). It is likely that, in the absence of secular change, a reliable two-group solution could be arrived at, which would permit highly accurate discrimination between both races. This assumption is supported by the classification patterns apparent in the more parsimonious MCLUST and EMMIX models (see Table 55). Upon closer inspection of the MCLUST model with four components (presented in Figure 112), a divide can be seen between the Black and White cases (represented by the red-blue and turquoise-green symbols). This consistency in patterning across a variety of the models demonstrates that the progressive reduction in components yields an almost identical picture of the differences between the two underlying groups. This, ultimately, attests to the strength of the data structure through all degrees of stratification or level of analysis: from world-wide continental to within-population (among "American") biosocial races.

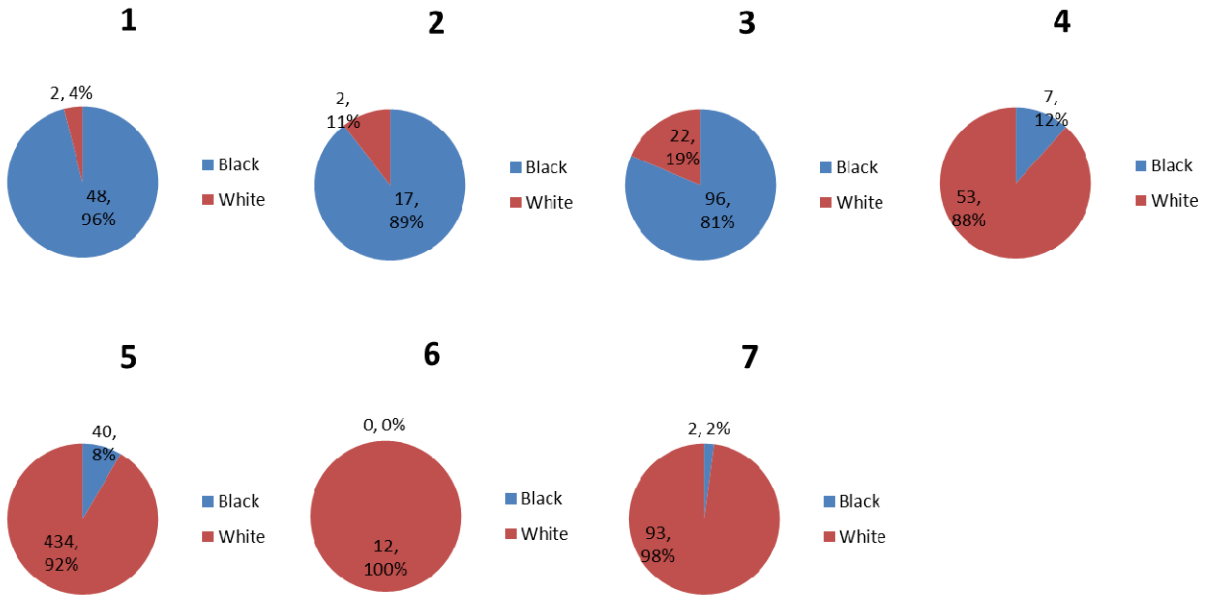


Figure 111: Cluster composition by race

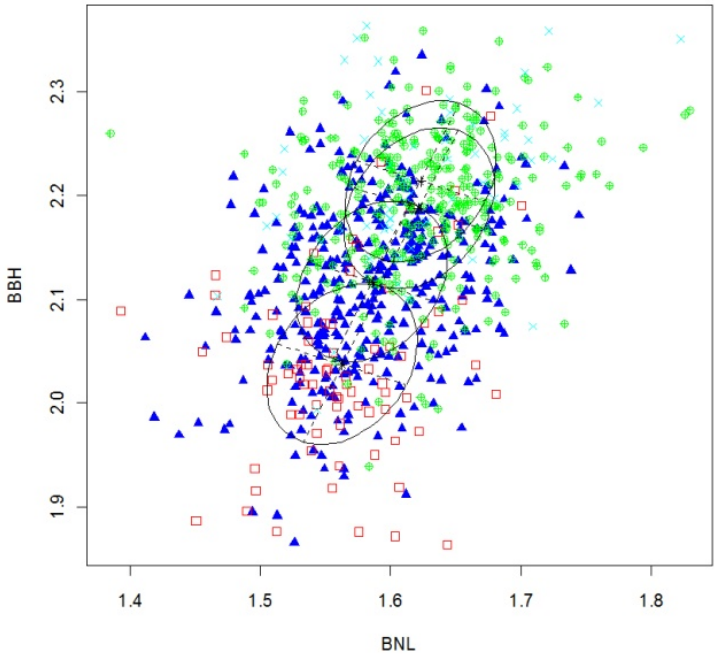


Figure 112: Coordinate projection showing clusters

While cluster results have been obtained that suggest stable trends despite differences in time, across different geographic ranges, between sexes and through many levels of hierarchy (local, regional, population, continental, global structure), the confidence with which meaningful conclusions can be drawn from these patterns of relationship is limited by concerns for the effect of sex and secular change on the composition of the clusters and the strength of their separations. Understanding the consequences of these factors on the cluster solutions may also offer valuable help in explaining the patterns that the mixture analyses have produced. As key demographic information is available for the individuals sampled from Forensic Anthropology Data Bank (Jantz, 2008), it is possible to further evaluate these cluster findings in a way that was not heretofore possible, given the limited case-specific metadata provided for the skeletal series assembled by Howells (1996) and Hanihara (2008): for most cases, sex has been "estimated" and birth year is not known or has not been reported.

The effect of sex on the cluster solution is a critical concern both for methodological and interpretive reasons. The current analyses has been structured in a way that controls for sexual dimorphism and "size" by imposing a highly effective and simple correction method. By extension, the choice was made here to pool the sexes, under the assumption that the suite of craniometrics representing males and females from a particular population will be more similar to each other than to same-sex members from other populations. In order to validate this overall approach, it is important to determine if sex is a driving factor in the model-based clusterings produced here; if so, a reassessment of this project's methodological decisions may be necessary. Recognizing if and how sex-specific morphology can affect individual classification is important for the larger discussion of the quantitative estimation of ancestry or race in forensic anthropology. As several studies have reported that membership assignments differ when sex is altered in discriminant function analyses and, specifically, when using the FORDISC software (e.g., Belcher, Williams & Armelagos, 2002; Williams, Belcher & Armelagos, 2005; Campbell & Armelagos, 2007), it is essential to independently assess the effect of "knowing" sex with some degree of statistical certainty on the quality of the classification results. This is also true in more general terms, as the determination of sex is a common first step in any forensic analysis, since the conventional approach to building the biological profile requires the adoption of sex-specific standards or the use of estimation methods that are conditional upon known sex (e.g., age-at-death and stature estimation). Controlling for sex or being aware of its effect can allow for more thoughtful commentary on problem allocations and enable a better answer for the obvious question: are misclassifications the result of poorly used methods, inappropriate test samples or improper priors, or sex biases? It should be acknowledged, however, that, despite considerable methodological emphasis, in judicial contexts and for large scale analyses (global tests of structure) sex differences may be of little evidentiary value and

knowing the "correct" sex may be of equally little statistical concern. As Konigsberg, Algee-Hewitt and Steadman (2009) have written, "sex of the individual is only of much use if the sex ratio is very biased for the population at large and the identification is for an individual of the lesser represented sex" (p. 84).

To determine the effect of sex and assess this project's methodological choice to "remove" size, simple exploratory scatter plots of cluster, sex, and size were produced for the FDB data. The size variable was also reintroduced into a mixture analysis of, first, all available FDB race groups (Figure 113) and, then, only American Blacks and Whites, for whom the largest sample sizes are available (Figure 114). In both cases, it was found that sex was evenly distributed: no one cluster represented a single sex and each cluster was well balanced, with members of populations, not sexes, clustering together. It was noted, however, morphological size did alter some of the allocations, although overall patterns were reasonably consistent, i.e., the extreme morphologies remained isolated as extreme, while the more moderate shapes retained their average positions despite some differences in size. These findings offer good justification for this project's choice to pool the sexes and adopt a conservative size correction approach; most importantly they validate the population-specific conclusions drawn from the observed cluster relationships, as in no case does sex appear to drive the analysis and, in turn, confound the interpretation of the mixture solution. These results clearly call into question the validity of the criticisms levied against the utility of discriminant function analysis, based upon arguments that invoke concerns over sex identification. The misclassification problem is not likely related directly to sex, but it may be instead a reflection of the user's poor choice of test data (given the known reference populations in the FORDISC database) and unbalanced sex ratios.

As the importance of secular change has been well documented in modern Americans (Angel, 1976, 1982; Jantz & Meadows Jantz, 2000,2001; Wescott & Jantz, 2005) and is being increasingly demonstrated among populations from in other geographic regions, including Africa and its Diaspora (Cameron et al., 1990; Spradley, 2006), Europe (Buretić-Tomljanović et al., 2004; Buretić-Tomljanović, Ostojić & Kapović, 2006; Weisensee & Jantz, 2011), Japan (Kouchi & Koizumi, 1992; Kouchi, 2000), and Mexico (Little et al., 2006), this project also addresses the effect of this short-term change over time on the resolution of the finite mixture analysis. This discussion is limited to the FDB data, and, specifically, the Black and White samples as (1) evidence of secular change has not been adequately investigated among other American groups, including Hispanics, Asian-Americans, and Native Americans, and as (2) the sample sizes presently available for these additional groups are small. As it can be seen from the scatter plot of the FDB Black and White samples (Figure 115), clusters 4 and 5 contain only White individuals born after 1920. These findings validate the mixture clusterings produced here by giving a credible reason for the subdivision of the FDB groups into multiple within-population components, e.g., where the White sample is distributed across clusters 4 through 7. Although these patterns do not agree exactly with the explanations given by Ousley, Jantz and Freid (2009) for the within-group partitions produced by their k-means cluster analysis using a different sample of FDB data, when taken together, these results do demonstrate the consistent influence of birth year on cluster membership. These findings strongly suggest that secular change should be considered carefully when conducting craniometric analyses using modern American samples.

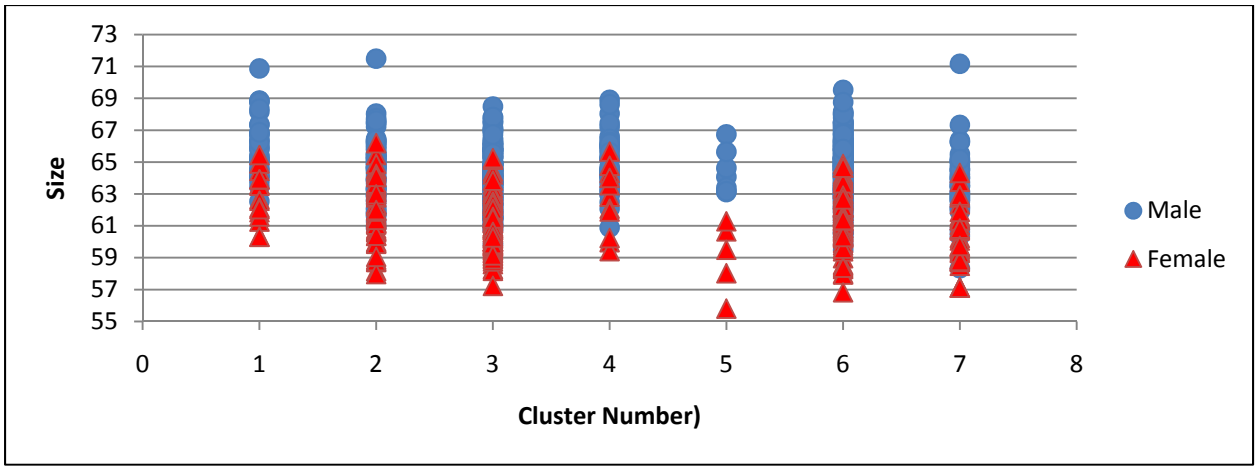


Figure 113: Scatterplot by size and sex (all races)

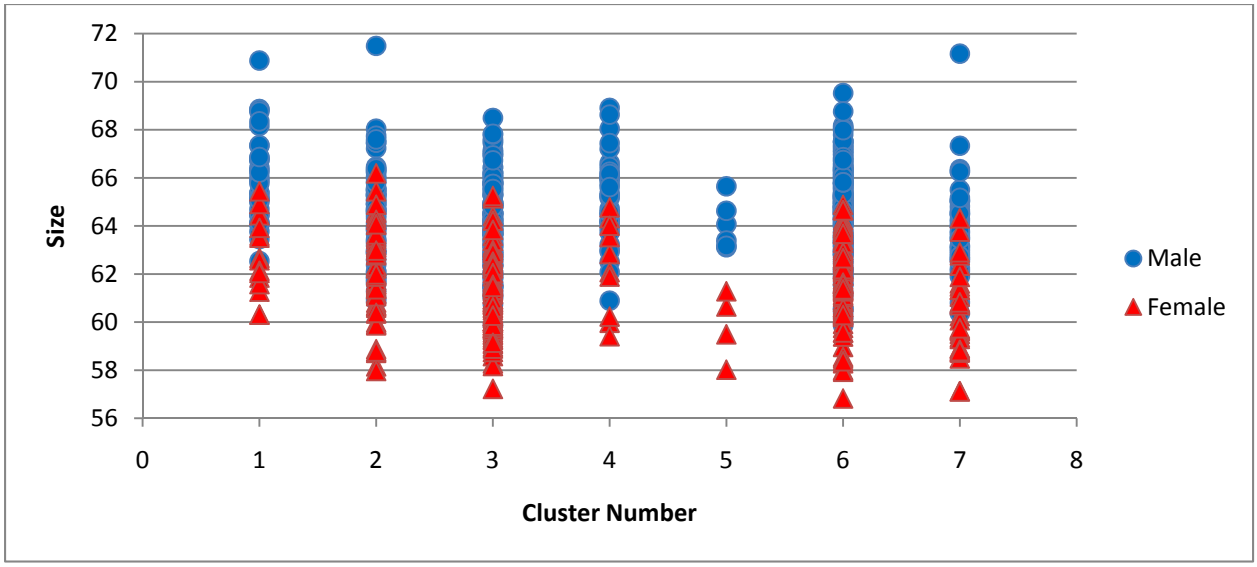


Figure 114: Scatterplot by size and sex (American Blacks and Whites)

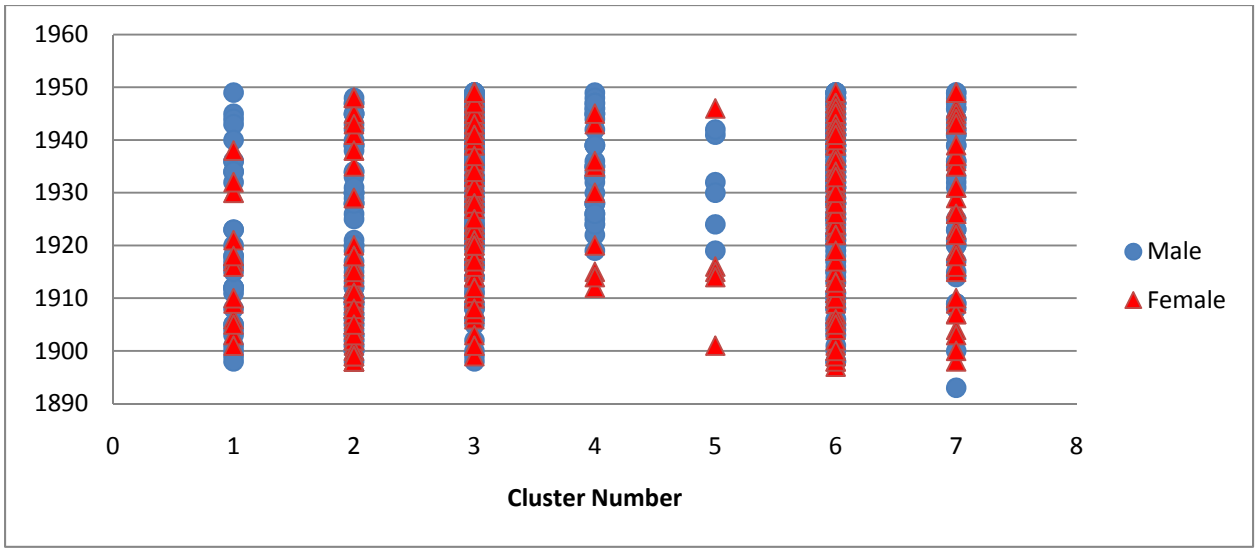


Figure 115: Scatterplot by birth year and sex (American Blacks and Whites)

Much of the ongoing debate surrounding the estimation of ancestry and the role of race in forensic anthropology is centered on the utility of the software package FORDISC, with special emphasis placed upon the limitations of its reference samples and the problems with its supervised classification method (c.f., Brown & Armelagos, 2001; Armelagos & Van Gerven, 2003; Williams, Belcher & Armelagos. 2005; Albanese & Saunders, 2006). As the Forensic Anthropology Data Bank serves as the source for the modern American forensic data used in FORDISC (Ousley & Jantz, 2005) and Howells' (1989) craniometric data set has seen considerable use in tests of FORDISC and its statistical approach (e.g., Hubbe & Neves, 2007; Elliott, 2008; Elliott & Collard, 2009), these two data sets are combined here to expand the modern "race-based" sample and, in turn, to address the significance of racial identification in terms of the broader demographic profile of the American population-at-large (c.f., Konigsberg, Algee-Hewitt & Steadman, 2009, for a discussion of evidentiary significance using the same data sources and comparable samples). This project considers, specifically, the correspondence between the six

major races defined by the Federal Office of Management and Budget (OMB) and the U.S. Census Bureau (2010) and the craniometrically inferred clusters obtained from a finite mixture analysis of the combined FDB and Howells data set. The goal is to determine if biological variation can be intelligibly and meaningfully linked to social identity in the absence of *a priori* information on individual memberships and without employing the user-directed techniques that have inspired arguments against classification, based on presumptions of forced allocations. A translation of Howells' (1989) 28 populations and the five FDB (2008) races into the racial categories employed by the U.S. Census Bureau is provided in Table 58. The individual sampling locations are also reported in order to provide context for the assigned "races" in terms of biogeographic ancestry. While these continental designations may seem obvious, it is important to clarify, for example, that the "FDB Black" group represents African-Americans and does not correspond equally to continental Africans, Afro-Caribbeans, or Indigenous populations whom are known to differ genetically in their parental contributions and admixture proportions and for whom pigmentation is a weakly informative proxy of ancestry (c.f., Parra, Kittles & Shriver, 2004). These geographic associations follow, in most cases, the naming conventions of the United Nations Statistics Division (2011) to ensure methodological transparency, especially in the presence of questionable classifications. Under these guidelines, for example, the Papua New Guineans are differentiated from the other Oceanic populations. Howells' Tolai were grouped with the Pacific Islanders and his Australian and Tasmanian series were assigned to the "Other" category (after Konigsberg, Algee-Hewitt & Steadman, 2009), despite the fact that the global population clusterings consistently associated the Australo-Melanesian populations with the Africans. It was felt that, in this context, it is more appropriate to preference the classification standards that represent bureaucratic race over prior expectations for affinity. Careful consideration was also given to the findings of the U.S. Census Bureau (2011) on racial self-identification by Hispanics: using this data, the decision was made to

classify the FDB Hispanics as White. The FDB sample also includes small set of admixed individuals who are documented as White-Hispanic and White-Black: for the purpose of this analysis, the White-Hispanic cases were reclassified as White in accordance with Census protocol and, in recognition of the realities of biosocial membership criteria and the persistence of the "one drop rule," any admixed individuals with some reported Black ancestry were reclassified as Black (c.f., Ho et al., 2010).

Finite mixture analysis was applied, using the MCLUST package, to the combined Howells and FDB sample (n = 3426) representing the six races recognized by the U.S. Census Bureau (2010, 2011). The top 10 MCLUST models obtained using the full combined data set and the standard suite of 12 variables are presented in Table 59; the corresponding behavior of the estimated BIC values, over a maximum of 30 components and 10 parameterizations, is displayed in Figure 116. In Table 61 and Table 62, the associated summary statistics on classification success and overall model quality are reported for the top BIC model: a four cluster solution with unequal covariance matrices. Table 60 provides the sample breakdown by sex and U.S. Census race as well as the per-cluster allocations; the corresponding cluster membership patterns and classification uncertainty are also displayed in Figure 108, Figure 109, and Figure 110 (c.f., McLachlan & Peel, 2000; Ousley & Jantz, 2005; Fraley & Raftery, 2011).

Table 58: Populations by U.S. Census and geographic labels

Howells (1989) and FDB (Jantz 2008) Populations	*OMB Terminology, Census 2011*	*Continental Regions*
Ainu	Asian	Asia
Andaman	Asian	Asia
Anyang	Asian	Asia
Arikara	American Indian	Americas
Atayal	Asian	Asia
Austrailia	Other	Australia
Berg	White	Europe
Buriat	Asian	Asia
Bushmen	Black	Africa
Dogon	Black	Africa
Easter Island	Pacific Island	Pacific Islands
Egypt	White	Europe
Eskimo	American Indian	Americas
FDB American Indian	American Indian	Americas
FDB Asian	Asian	Americas
FDB Black	Black	Americas
FDB Hispanic	White	Americas
FDB White	White	Americas
Guam	Pacific Island	Pacific Islands
Hainan	Asian	Asia
Mokapu	Pacific Island	Pacific Islands
Moriori	Pacific Island	Pacific Islands
Norse	White	Europe
North Japan	Asian	Asia
Peru	American Indian	Americas
Philippines	Asian	Asia
Santa Cruz	American Indian	Americas
South Japan	Asian	Asia
Tasmania	Other	Australia
Teita	Black	Africa
Tolai	Pacific Island	Pacific Islands
Zalavar	White	Europe
Zulu	Black	Africa

Table 59: Top 10 models in order

FDB and Howells Data Set: census races, 12 variables			
Component Number	*Model Form*	*Selection Statistic*	*Estimate*
4	VVV	BIC	148925.3
18	EEE	BIC	148649.2
14	EEE	BIC	148626.5
5	VVV	BIC	148620.8
6	VVV	BIC	148563.5
24	EEE	BIC	148544.3
5	VEV	BIC	148539.5
3	VVV	BIC	148539.3
13	EEE	BIC	148523.2
23	EEE	BIC	148501.1

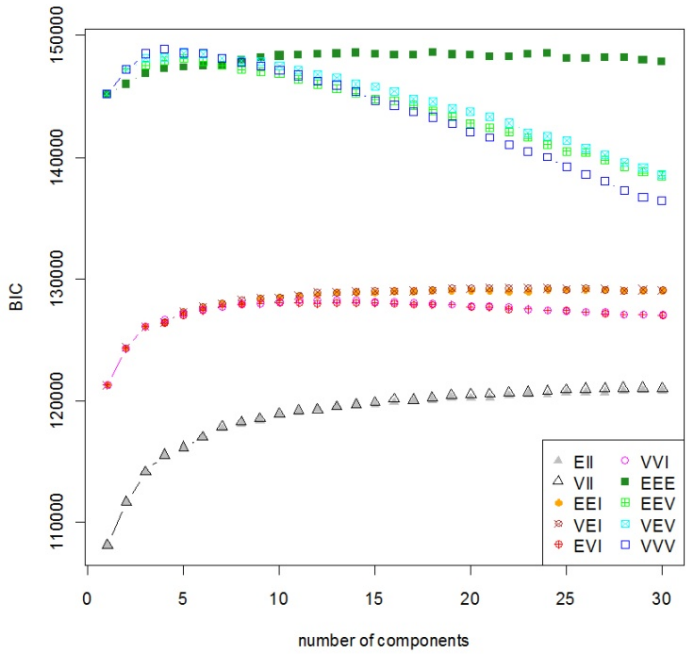


Figure 116: Model selection by BIC

Table 60: Cluster membership by race and sex

Race	*1*	*2*	*3*	*4*	*Grand Total*
Admixed		1		1	2
M		1		1	2
American Indian	73	14	114	204	405
F	28	7	56	103	194
M	45	7	58	101	211
Asian	302	65	112	240	719
F	96	26	47	88	257
M	206	39	65	152	462
Black	271	219	14	79	583
F	136	101	7	37	281
M	135	118	7	42	302
Other	95	77	4	12	188
F	47	35	3	6	91
M	48	42	1	6	97
Pacific Island	190	60	52	159	461
F	95	22	28	73	218
M	95	38	24	86	243
White	160	100	295	513	1068
F	73	38	113	209	433
M	87	62	182	304	635

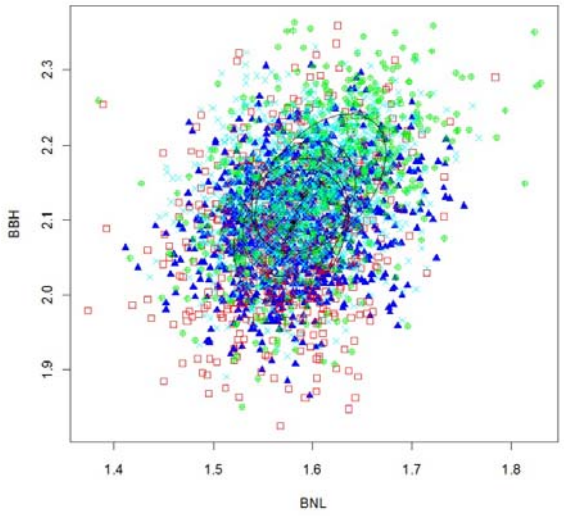


Figure 117: Coordinate projection showing clusters

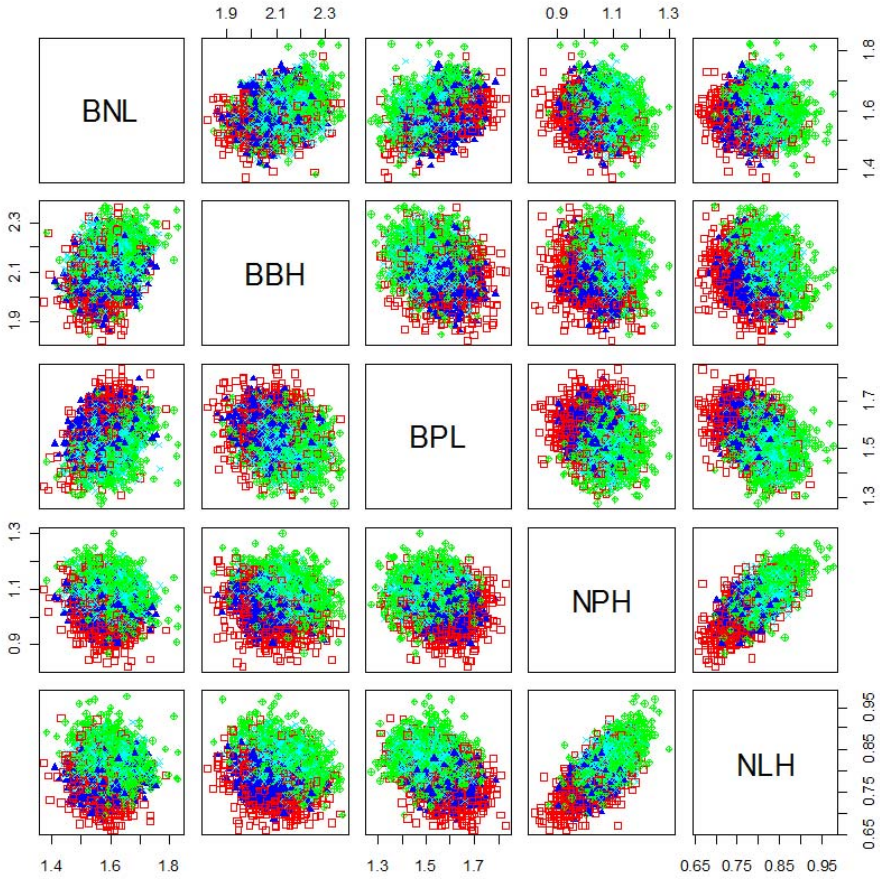


Figure 118: Pairwise scatterplot matrix showing clusters

Table 61: Allocation rates

FDB and Howells Combined Data: 4-clusters	*posterior probability > 0.7*	*percent correctly allocated*	*posterior probability → 1*	*percent correctly allocated*	*total number of individuals*
Admixed	2	100	1	50	2
American Indian	324	80	221	55	405
Asian	566	79	392	55	719
Black	484	83	379	65	583
Other	155	82	115	61	188
Pacific Island	377	82	246	53	461
White	850	80	611	57	1068
Grand Total	*2758*	*81*	*1965*	*57*	*3426*

Table 62: Quantiles of uncertainty

Quantile	*0%*	*25%*	*50%*	*75%*	*100%*
Estimate	0	0.01740344	0.06548544	0.23931628	0.69887600

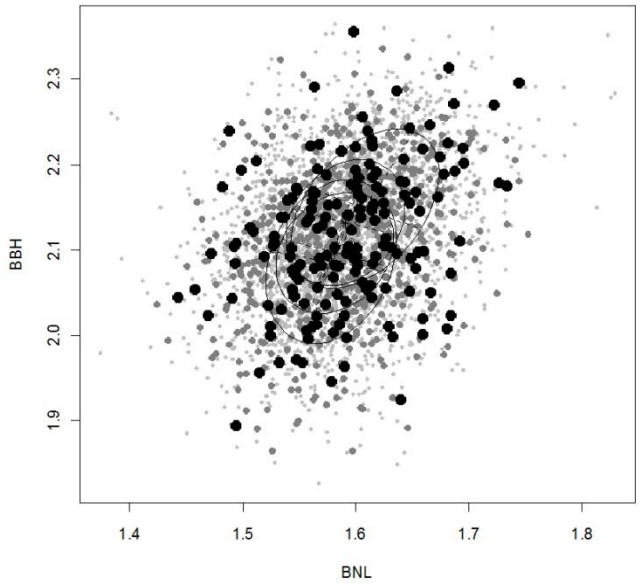


Figure 119: Coordinate projection showing uncertainty

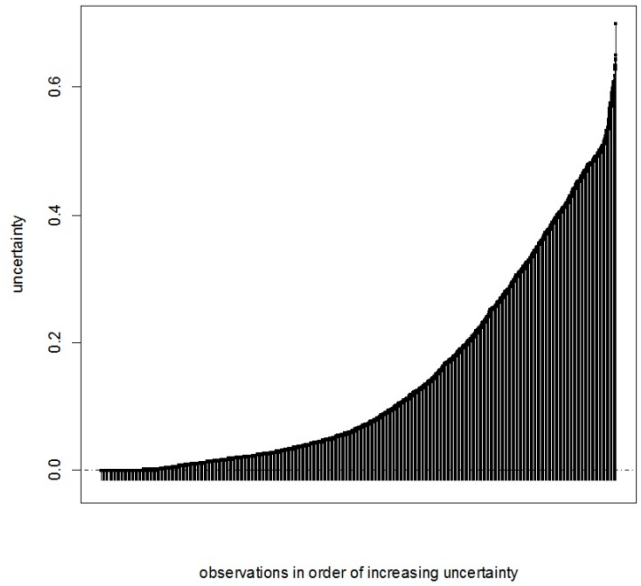


Figure 120: Data profile showing uncertainty

When the inferred success rates for this complex analysis (i.e., many original, presumably morphologically different, groups, and unconstrained variances) are contextualized, it can be seen that , despite the increase in sample size and breadth of biogeographic variation, the percent correct estimates (82% and 57%) are similar to those obtained for the previous analysis of the craniometric data available for all five of the FDB's race-based groups, e.g., 82% and 52% (see Table 49). Interestingly, the patterns of membership have also remained consistent. From the composition of the 4 clusters (Figure 121), it can be seen, first, that this mixture analysis has resolved the total variation into two major groups that are heavily weighted by Black (clusters 1 and 2) and White (clusters 3 and 4) samples. Clusters 1 and 2 contain the African and African-American Black, Australian and Tasmanian ("Other"), and Tolai ("Pacific island") groups in the largest numbers, as well as the Asian samples which had previously shown some affiliation with the Africans and Australo-Melanesians (e.g., Andaman, from the Indian Subcontinent). Clusters 3 and 4 capture most of the European, North African, Hispanic, and American White groups, while also accounting for the non-Black component of the Native American, East Asian, and Pacific Island variation. The breakdown of each Census-based race into its constituent components, in Figure 122, perhaps shows best how these patterns conform to expectations for among-group similarities. Given the allocations of membership that have been observed already for both Howells' and Hanihara's data, these classifications should not be surprising. What is especially notable is the fact that this patterned population structure supports a meaningful and useful correlation between biogeographic ancestry and biosocial identifiers, including the naming convention for American races and the newly emerging ethnicities (e.g., Hispanic). These findings imply, therefore, that (1) there is information contained within the biological data alone that is informative of population history and relatable to modern American social identities and that (2) this information can be accessed without employing user-guided methods (e.g., supervised classification, like discriminant function analysis) and,

in turn, without comparative data (e.g., "priors" or "training sets") or *a priori* identifiers. The utility of discriminant function analysis and its implementation in software such as FORDISC to human identification should not be dismissed on the grounds of "forced allocation" when used properly, i.e., as Ousley & Jantz, 2005, advise. Moreover, forensic anthropologists should not be censured for their "race making" and perpetuation of "typological thinking." The underlying craniometric patterns present within the "modern" data tested here support the reality of race as a viable proxy for population-specific variation and attest to its utility in medico-legal casework, for which skeletal biology must be translated into personal identity in a way that is relevant and accessible to law enforcement and lay communities.

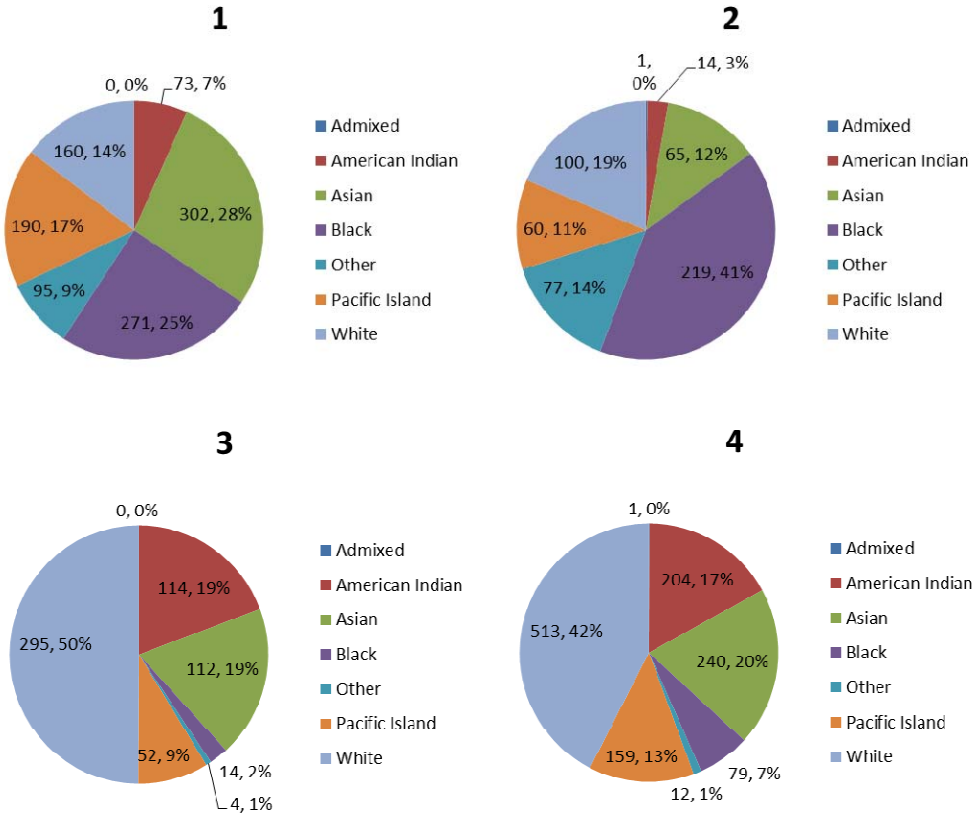


Figure 121: Cluster composition by race

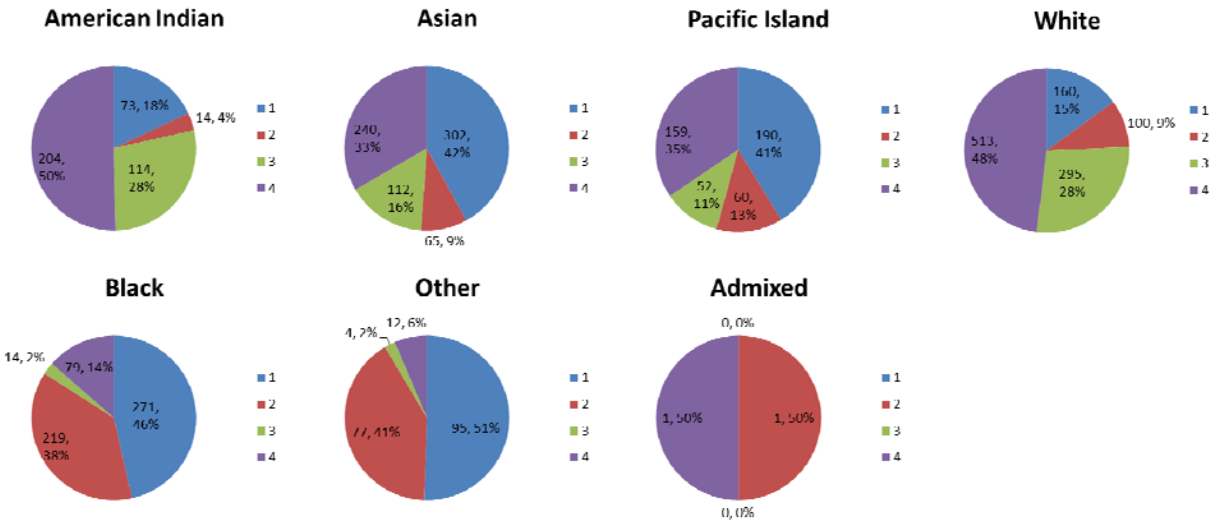


Figure 122: American Indian, Asian, Pacific Island, White, Black, Other, Admixed

Validation Tests: Supervised Classification and Information Visualization

In many of the common practical applications of model-based clustering, true group structure for a given test data set is unknown and the resultant clustering must be evaluated solely upon the quality of the model's statistical properties. This project is fortunate to have the information available on the "known" memberships in order to assess the mixture clustering results using *post hoc* methods. From a broad disciplinary perspective, this freedom affords an active researcher the opportunity to improve the success of future projects by adjusting the conditions of the current research design in light of the valuable information that is learned from *a posteriori* assessments of the mixture findings. On a smaller scale, and of special importance here, it provides this project, in particular, with the necessary information that allows it to speak authoritatively on both the utility of the finite mixture approach to

skeletal biology and the reliability of the conclusions reached. The available data on individual population membership or race [self]identification has been employed for use in supervised classification analyses and with force-directed visualization methods as a means to objectively assess the statistical validity of the cluster findings and, in turn, reaffirm the support that they provide for the reality of biogeographic ancestry and the utility of biosocial race.

As a first step in validating the presence of population-specific structure within the craniometric data set of interest here, this project has chosen to subject the combined FDB and Howells data sets to discriminant function analysis via model-based clustering and traditional linear discriminant analysis (LDA), implemented in R using the MCLUST (Fraley & Raftery, 2011) and MASS (Venables & Ripley, 2002) packages respectively. This analysis has been purposefully restricted to these two sources of craniometric data as they provide a large but manageable sample size (n = 3426) and because they mirror the reference database that underlies the execution of linear discriminant function analysis in the FORDISC software package (Ousley & Jantz, 2005). A training sample of 3402 cases was defined using the same combined data set and 12 linear variables that guided the previous analysis (see Table 60), while a test subset of 24 target individuals was randomly drawn from the Forensic Anthropology Data Bank (Jantz, 2008). This test data set represents all five available races and both sexes and it includes four cases of admixture: the breakdown of this target sample is provided in Table 64. Parallel analyses using a simplified data set that contains only Black and White cases from the FDB were also conducted. This second analysis employed a significantly smaller training sample (n = 201) and a proportionally larger target sample (n = 621). This data was also subjected to a partial classification mixture analysis using the EMMIX software (Peel & McLachlan, 1999; McLachlan, Peel, Basford & Adams, nd), in which approximately 25% of the available cases were taken to be a priori known. For all of these analyses,

equal priors were specified; that is, each test assumes equal probability of membership in any of the groups included within the given analysis, i.e., $1/\pi_g$. The default setting in most discriminant function analyses executed in R uses "proportional" priors based on group size. In this case, using proportional priors would require the belief that the sample size ratios for a given data set truthfully replicate the proportions of each race in the most relevant population-at-large, i.e., across the United States, given this project's interest in U.S. Census races and its use of the FDB. While this project has formally adopted the most conservative (equal prior) approach, comparative tests were conducted using proportional priors and it was found that the error rates did not significantly differ. For the mixture-based DFA, using the full combined Howells and FDB data set, training error rates changed by less than a 10th of a percentage point. This consistency in estimation may be attributed the inclusion of the much larger Howells' data set (representing 2504 of the total 3402 cases) which was carefully constructed to ensure balanced sex and sample sizes (Howells, 1989), as well as the robust fitting abilities of mixture-based DFA, as finite mixture modeling has been shown to capture larger percentages of smaller groups when group sizes are unequal (Holden & Kelley, 2010). A notable (10%) drop in classification accuracy did occur, however, when proportional priors were specified for the standard LDA using the combined FDB-Howells training set and when subjecting it to leave-one-out crossvalidation. This increase in error attests to the well-documented sensitivity of LDA to sample conditions, including unequal and small group sizes, in a variety of practical applications and test environments (e.g., Laitine 2007, Holden & Kelley, 2010; Way et al., 2010).

The fundamental goal in applying DFA to the combined FDB and Howells sample is to provide objective evidence from "supervised testing" that supports the previous identifications of latent data structure and its relevance to anthropological questions on population-specific morphologies and biosocial race.

This project chose to apply a method that is not only familiar to anthropologists and has been traditionally used in quantitative analyses of group differences craniofacial shape, but also has contributed to the human identification toolkit and has received considerable attention for its utility when estimating ancestry and predicting race from skeletal remains (e.g., Giles & Elliot, 1962, 1963; Birkby, 1966; Giles, 1966; Snow et al., 1979; Ousley & Jantz, 2005; Freid et al., 2005; Konigsberg, Algee-Hewitt & Steadman, 2009). This project also evaluates classification accuracy using two methodologically different approaches to discriminant function analysis, in order to identify the most effective technique, address concerns over the limitations of classic LDA, and provide recommendations for future research using craniometric data. This project tests specifically the model-based clustering discriminant analysis (MDA) and linear discriminant function analysis (LDA) methods as statistically defined by Fraley and Raftery (2002) and Venables and Ripley (2002), respectively. It has chosen not to test the efficacy of quadratic discriminant function analysis (QDA) as there is a clear tendency in most anthropological studies to assume the simpler case of homoscedasticity; moreover, QDA is not an available discriminant function option in FORDISC (Ousley & Jantz, 2005), which serves as the key software for implemented DFA in forensic casework contexts.

As discriminant analysis methods are often probabilistic, it is typically assumed that (1) observations in each group are generated by a distribution specific to that group, (2) a model can be fit to each group in the training set, and that (3) data should be assigned to the group associated with the model that has the highest posterior probability (McLachlan, 1992; McLachlan & Peel, 2000). The type of discriminant function model and the method of fitting can, nevertheless, vary among probabilistic discriminant function methods. As a result, methodologically different approaches to DFA exist and the user's choice among these may have important philosophical implications as well as practical consequences for the

classification outcome. When the discriminant function model is a normal mixture fitted by model-based clustering, the supervised classification procedure is known as finite mixture or model-based DA (Fraley & Raftery 2002). Within this class of analysis, there are two methodological approaches, both of which were adopted here for the purpose of comparison. This project, first, modeled the training data with a mixture density having a single normal component for each group and it then chose the model (parameterization) that produced the lowest error rate for all of the training data. It determined the most appropriate cross-cluster constraint for the covariance matrices in two ways: using BIC-based selection and leave-one-out crossvalidation. Although in both of these analyses (using the combined FDB and Howells, and the FDB Black and White training data sets), the BIC and crossvalidation methods chose the same "EEE" (equal covariance) model, this will not always be the case for other data sets. The models selected and, in turn, the results produced by the discriminant analysis, can be, and are frequently, different. While crossvalidation techniques are common protocol in model selection literature (Stone, 1977; Kohavi, 1995), they are more computationally intensive than selecting the best model using the Bayesian Information Criterion. When, for example, the crossvalidation method is implemented via the EM algorithm in MCLUST, one training observation is left out at a time, each model is then fit using the m-step, the observation that was "left out" is then classified using the e-step, and the error rate output for each model represents the fraction of the "left-out" cases that were correctly classified by the model fit to the remaining sample (Fraley & Raftery, 2011). Past experience suggests that, when data sets are large, the less algorithmically demanding BIC selection method offers considerable advantages as it is an equally statistically robust but more efficient option; Fraley and Raftery (2003) found the error rates to only be slightly higher (increase of 2%) when using BIC for their example, Lansing Woods, data set. For this project's second MD analysis, model-based clustering was used to fit a normal mixture model as a density estimate for each group in the training set; in this case, it

was specified that a model be chosen for each class of the training data separately using the BIC selection method available in MCLUST (Fraley & Raftery, 2011). In both of these approaches to mixture discriminant analysis, the parameterizations of the discriminant models are afforded ultimate freedom. This modeling flexibility should enable the best fit and produce optimal (i.e., more realistic or truthful) classification results given the unique properties of the data under analysis. Although discriminant function analysis, as a general class of techniques, can be seen as a specialized case of finite mixture analysis, the mixture DA approach adopted here differs fundamentally from the traditional application of linear discriminant function analysis (LDA), where the same covariance structure is assumed for all groups and this "common" covariance matrix is used for calculating the Mahalanobis distances between cases and groups. Furthermore, this version of MDA is not directly comparable to even the more complex quadratic discriminant function analysis (QDA), where the covariance structure is taken to be unequal and the Mahalanobis distances are calculated using the individual group covariance matrices. What is notable about MDA is not only that it is statistically robust and mathematically satisfying, given its probabilistic foundation, but also that it is highly flexible: in the MCLUST implementation, the user employs well established statistical decision-validation methods to identify the best fitting model out of several possible parameterizations (Fraley & Raftery, 2002, 2011). In order to compare these mixture-based and traditional DF analyses, the multivariate normal based linear discriminant function method was implemented here with the lda function in the MASS package (Venables & Ripley, 2002; see also Ripley, 1996) and classifications were produced for both the combined FDB and Howells and FDB Black and White data sets. Finally, in order to address the issue of small training sample size (a problem that is not uncommon in anthropology, for which appropriate skeletal reference samples are often limited) the FDB data set was simplified. The revised dataset, which now contained only Black and White cases, was subjected to partial classification analysis using the EMMIX software (Peel & McLachlan, 1999;

McLachlan, Peel, Basford & Adams, nd). In this analysis, the true classifications for a small number of randomly determined data points (201 of the total 828 cases) were pre-specified. As these predefined or *a priori* "known" cases retain their classification throughout the fitting process, this method can be seen as a compromise between unsupervised mixture analysis and fully directed discriminant function analysis (McLachlan & Peel, 2000).

From the classification results obtained under these three different modeling conditions, this project produced the corresponding training and target error rates. When appropriate, it derived conservative error rates using leave-one-out crossvalidation in order to adhere to standard practice and to provide the most realistic evidence of the model's predictive ability, and, ultimately, the associated method's utility. Table 63 provides the summary results for these three different approaches to directed classification analysis. Table 64 presents the classification results and overall rate of correct allocations for test data set, consisting of 24 target cases randomly sampled from the Forensic Anthropology Data Bank. These test cases were not resampled from the original training date set but represent a unbiased, independent hold-out subset (after Fielding, 2007). It can be seen clearly from these results that the flexible mixture based approaches, both the MDF and even the partial classification method, yield better rates of classification accuracy overall than the traditional LDA and produce remarkably good results when the original training data set contains many groups in small numbers: compare, for example, the mixture DFA training error rate of 49.4% to the corresponding LDA estimate of 55.2%. While these findings agree with Konigsberg, Algee-Hewitt and Steadman (2009), who chose QDA as the most appropriate approach for craniometric data in their analysis of an example human identification case, they are, nevertheless, at odds with the standard methods applied by most forensic practitioners, who rely heavily upon the use of the FORDISC software - the current version of which performs only LDA

(Ousley & Jantz, 2005). This project suggests that further testing and method refinement is warranted for craniometric-based classification and that incorporating finite mixture approaches into the biological and forensic anthropologists' statistical toolkit may offer improvements to the quantitative estimation of ancestry.

Returning to previous discussions of detectable population structure and its translation into the language of biosocial race, the mixture DFA findings provide credible evidence for the classification of individual cases of forensic significance (i.e., the randomly drawn FDB target sample, n =24) with some acceptable degree of statistical certainty when population-specific identification are of interest (see Table 63 and Table 64). This project obtained 46% accuracy, but this figure is likely highly conservative as it accounts for the classification of 4 admixed individuals for whom "true" or comparably mixed reference populations were not included within the training data set and, so, only best-fit allocations could be made. What is important to note about these classifications is that the "errors" in assignment correspond, in general, to the observations with high levels of classification uncertainty; identifying this trend is key to the generalization of this approach to new data and especially actual casework, for which true memberships are unknown. The fact that the high uncertainty values have clear population referents, are not relatable to some other biological factor, like sex, and are not a random artifact of a poorly fit model or the inappropriate use of the statistical procedure, suggests that they can be taken as a reliable signal of problematic classification. What is even more interesting is that, when cases are misallocated, they are, for the most part, assigned to populations with which they have known biological affinity (e.g., with geographic, evolutionary, or historical links) or for which the mixture cluster analyses have already demonstrated some level of relationship. For example, a FDB case documented as American Indian was identified as an Arikara, a FDB Black case was assigned to the Zulu, a FDB East

Asian case is allocated to the Polynesian Mokapu, and a FDB Hispanic case was classified with the Peruvians. The "biological" logic that underlies these classification results appears, therefore, to contradict the arguments made against the utility of discriminant function analysis on the basis of random classifications, and, specifically, its failure to meet expectations for population-level allocations based on affinity (c.f., Williams, Belcher & Armelagos, 2005).

This project's conclusions are even further supported when consideration is given to the classification results not at the level of the population, but using the modern American terminology of race. As it can be seen in Table 64, when the target cases are assigned based on the six races considered by the 2010 U.S. Census, the estimated rate of correct classification greatly increased, reaching 71% accuracy for MDFA (surpassing the LDA estimate by almost 10%). In several cases, when allocations were similar in affinity but not accurate by population standards, the simplification of the population categories into races allowed for the misallocations to be reversed: e.g., the previously misidentified FDB American Indian was correctly reclassified, changing from Arikara to American Indian. When misclassifications do persist, it appears that, not only are they logical, but they are also quite consistent with the population-specific classifications. For example, the FDB Black individual, who was previously misidentified as Philippino, is now misidentified as a Pacific Islander, and the FDB Hispanic case, which was previously allocated to Peru is now similarly identified as American Indian. It is especially interesting to note that for the admixed cases, neither set of individuals were allocated to the group that would have been expected: the White/Black cases consistently classified as White (and not Black), and the White/Hispanic as Asian (and not White or American Indian). While the repetition of these associations among the misclassified cases may, simply represent an overlap in cranial metrics (especially when uncertainty values are high), they may also reflect some information on personal identity that was not captured in

the individual's choice of self-identification or the standard reporting procedures. It is in instances such as these that morphoscopic and morphometric analysis should be considered in concert, and it may be found that the "error" in classification actually corroborates other lines of evidence, improves the practical utility of the forensic case report and, ultimately, increases the success with which the law enforcement community may identify good possible matches for the unidentified remains. These findings imply, therefore, that the hotly contested methods of supervised classification and the contentious use of biosocial identifiers, like race, do have statistical validity and practical utility - they are valuable for both understanding patterns of population structuring and identification of individual memberships.

Table 63: Summary results of supervised cluster analyses

Method	*Training Data, n*	*Target Data, n*	*Training Error*	*Target Error*	*CV-Model, Associated Error*	*BIC Estimate, Associated Model*
Mixture DFA (MCLUST package)	FDB and Howells, 3402	FDB, 24	49.4%	45.8%	EEE, 55.2%	142438.7, EEE
Mixture DFA (MCLUST package)	FDB Black and White Races, 207	FDB Black and White Races, 621	8.2%	10.8%	EEE, 7.7%	8261.434, EEE
Method	*Training Data, n*	*Target Data, n*	*Training Data Percent Correct (LOOCV)*		*Target Data Percent Correct*	
LDA (MASS package)	FDB and Howells, 3402	FDB, 24	44.8%		38.9%	
LDA (MASS package)	FDB Black and White Races, 207	FDB Black and White Races, 621	90.3%		91%	
Method	*Training Data, n*	*Target Data, n*	*Overall Rate of Correct Allocation*		*BIC Estimate, Associated Model*	
Partial Classification (EMMIX software)	FDB Black and White Races, 207	FDB Black and White Races, 621	94.6%		-34137.97, 2 components with arbitrary covariances	

Table 64: Test data classifications (equal priors)

FDB Sex	*FDB Race*	*MCLUST DFA Classification*	*Classification Uncertainty*	*Census Race*	*LDA Classification*	*Census Race*
Male	American Indian	Arikara	0.201275	American Indian	Arikara	American Indian
Male	American Indian	Anyang	0.749452	American Indian	American Indian	American Indian
Female	Black	Zulu	0.699231	Black	Peru	American Indian
Female	Black	Andaman	0.637564	Asian	Black	Black
Male	Black	Berg	0.702273	White	Hispanic	White
Male	Black	Teita	0.061697	Black	Teita	Black
Male	Black	Black	0.022471	Black	Black	Black
Male	Black	Philippines	0.79574	Pacific Island	Philippines	Pacific Islands
Male	East Asian	Mokapu	0.215943	Pacific Island	Guam	Pacific Islands
Male	East Asian	East Asian	0.042945	Asian	East Asian	Asian
Male	East Asian	Buriat	0.351563	Asian	Hainan	Asian
Male	Hispanic	Peru	0.53303	American Indian	Hispanic	White
Male	Hispanic	White	0.203058	White	White	White
Female	White	White	0.005961	White	White	White
Female	White	White	0.499512	White	Andaman	Asian
Male	White	White	0.007776	White	White	White
Male	White	White	0.274423	White	White	White
Male	White	White	0.199661	White	White	White
Male	White	White	0.602212	White	White	White
Male	White	White	0.026151	White	White	White
Male	White/Black	Berg	0.243409	White	N Japan	Asian
Male	White/Black	Egypt	0.359883	White	Norse	White
Female	White/Hispanic	Ainu	0.498204	Asian	Ainu	Asian
Female	White/Hispanic	S Japan	0.680782	Asian	Hainan	Asian
Percent Correctly Classified		46%		71%	46%	62.5%

Information visualization techniques are often very useful, if not essential, to the interpretation of patterns within large data sets. Standard approaches to simplifying high dimensional data typically involve some form of "dimension reduction," which allows for complex data to be more easily visualized as a collection of points in two-dimensional, or sometimes three-dimensional, space. This simplification is often achieved by applying principal component analysis (Jolliffe, 2002),

multidimensional scaling (Kruskal & Wish 1978), or non-linear dimensionality reduction methods, like LLE or Isomap (Roweis & Saul, 2000; Tenenbaum , de Silva & Langford, 2000). Often these methods are used in concert with clustering algorithms in order to produce more intelligible solutions or clearer displays of patterns that have been otherwise detected, which can aid in validating findings obtained from less widely accessible techniques. The gain in interpretability that these method offer is, however, often offset by the statistical assumptions and computational constraints that they impose upon the data and its analysis. For this reason, this project is interested, instead, in identifying an approach that allows for the simplification of the analysis in a way that not only enables easy interpretation of the data, but also retains the information on the complex latent population structure that has been already revealed by the mixture clustering algorithms.

As the second step in validating the presence of population-specific structure within the craniometric data, this project has chosen, therefore, to adopt the graphical methods of computational geometry. This approach has proven especially useful to the areas of pattern recognition, robotics, computer graphics and operations research; all of which share some relationship to the methods and goals of cluster analyses. This project focuses specifically upon the use of graphical visualizations as a means to design, analyze and interpret algorithms and their output in complex geometric applications (c.f., Lee and Preparata, 1984; Preparata & Shamos, 1985; Eldesbrunner, 1987). Here, it relies specifically upon the use of force directed graph drawing (Di Battista et al., 1999, Kauffmann & Wagner 2001) and mosaic plotting techniques in order to produce Voronoi diagrams (Aurenhammer, 1991). These "maps" allow the incorporation of empirical distance measures based upon on the craniometric data, the clustering results of the mixture analyses, and the background knowledge of the true group memberships into

graphical displays that convey the information of group (population or race) relationships in a visual, highly accessible and easily interpretable way.

The force directed approach is especially appropriate for studies concerned with nested structure and dynamic, interconnected populations as it follows the theory that "many systems of scientific or practical interest are decomposable into subsystems with strong internal and relatively weak external interactions" and that these [sub]systems have geometric analogues (Noack, 2009, p.1; c.f. Simon, 1962). It operates under the expectation that that (1) these systems can be modeled as "networks," with their elements represented by vertices and their interactions by edges, and that (2) the related subsystems can be each identified as unique "communities," which are definable individually by sets of vertices composed of densely packed internal connections with limited connections to the other parts of the network. These networks are represented by "layouts," which position the vertices in a metric space, and by "clusterings," which divide the vertex into subsets. The structure of the "community" is "naturally" revealed by densely connected vertices that are grouped together in similar space and in the same cluster and by sparsely connected vertices that are separated by distance and allocated into different clusters (Noack, 2009). The ultimate result is, therefore, a two-dimensional map that summarizes the patterns of relationships among all observations: each case is represented by a polygon, with its edges touching its nearest neighbors. Many of these neighbors come together organically to form a group, whose composite boundary joins it to other similar groups: the smaller the polygons the more densely packed are the groupings. The algorithms that implement this force-directed method treat the task of graph-making as a straight-edge drawing problem, not unlike spectral (Koren, Carmel & Horen, 2003) or high-dimensional embedding methods (Harel & Koren, 2004). In this case, however, they model the data by simulate a physical system of bodies with forces acting between them (Hu,

2006), often drawing upon the analogy of springs between neighboring vertices of the graph (Fruchterman & Reigold, 1991). These algorithms locate the best placement of the bodies by minimizing the energy of the system, following some "cooling schedule" (Kamada & Kawai, 1989).

Although force directed drawing has been historically used in computer science applications (e.g., Quinn & Breuer, 1979), this approach can be translated to the craniometric data by drawing upon the well-established practice of using anthropometric distance matrices as a means to differentiate between groups and to estimate individual relationships in paleoanthropological, modern human, and, specifically, forensic contexts (e.g., van Vark, Bilsborough & Henke, 1992; Howells, 1995; Ousley & Jantz 2005). Force directed mapping algorithms typically receive "system" information is in the form of a relational data set. From this input, a graph can be extracted, $G = (V, E)$, with, $|V| = n$, vertices and, $|E| = m$, edges (Gansner et al., 2009). The extension of this form to a population-based application using craniometrics is straightforward: let the set of vertices, V, represent the "nodes" or observations in the sampled data, e.g., the crania, and the set of edges, E, represent the relationship between pairs of observations, e.g., the pair-wise distances between crania. Here, the general case of the Minkowski distance is used as the measure of similarity as it works well in practical tests of force directed plotting and it is known to provide "a concise, parametric distance function that generalizes many of the distance functions used in the literature" (Groenen, Kaymak & van Rosmalen, 2006, p.3). Evaluating its utility as an alternative to, for example, Mahalanobis distance (Mahalanobis, 1936), the standard choice for traditional cranial morphometrics, is of practical interest as it is being increasingly incorporated into newer, flexible clustering algorithms (Groenen & Jajuga, 2001). The "dist" function was implemented via the R "stats" package (Mardia, Kent & Bibby, 1979; Becker, Chambers & Wilks, 1988; Borg & Groenen, 1997), to compute Minkowski distance matrices for three of the test data sets, using the base suite of 12

craniometric variables (Table 3). The three test data sets chosen were Howells's (1989) series of 28 populations (n = 2504), the full FDB data set (Jantz 2008), which includes all five available races (n =918), and a subset of this FDB sample (Jantz 2008), which contains only the Black and White cases (n = 828). These samples were previously subjected to finite mixture analysis using the MCLUST and EMMIX packages and the resulting findings are reported above in the corresponding discussions of the model-based cluster analyses.

In order to produce the geometric features required for force directed plotting, an Excel (Microsoft Office 2010) macro was applied to the distance matrices to define the "nodes" (creating entities) and "edges" (creating relations) following pre-set threshold and cutoff rules: the macro specified a maximum distance value of 0.3 between neighbors and the number of possible neighbors was limited to 5. These decisions were based on practical tests of different thresholds and cutoff values from which rules were developed that (1) are reasonable biologically, (2) follow the trends in the literature on force-directed methods, (3) maximize the interpretability of the data and (4) optimize the readability of the final graph. The open source graphing software contained within the information visualization framework *Tulip* 3.5.0 (2007) was used to implement the built-in GEM algorithm (after Frick, Ludwig & Mehldau, 1994) and produce two -dimensional force directed graphs based on the node and (distance-weighted) edge input. The X,Y,Z coordinate data that supported the structure of these visualizations was extracted and the X,Y points were used to build Voronoi mosaic diagrams in R. These diagrams were constructed using calls to the series of "voronoi" functions included in the Delaunay triangulation package, "tripack" (Renka & Gebhardt 2009, c.f., Renka, 1996). As the diagrams that result from this proceedure graphically display the relationships between individuals based purely upon craniometric distance, they can be taken as biologically truthful maps of the distribution of human craniofacial variation and used as templates onto

which can be laid externally derived "grouping" information that is in need of validation. First, each polygon was colored in accordance with the *a priori* known identifiers (e.g., Howells' population names, FDB or Census races, or geographic information) and, then, on a second identical mosaic map, the same polygons were re-colored based on the cluster memberships obtained from the MLCUST and EMMIX mixture analyses. The polygon colors correspond to the standard R color chart (Figure 134). For each test data set, only the best model was considered from those produced under the BIC selection criteria (Fraley & Raftery, 1998) and McLachlan's (1987) bootstrap of the likelihood ratio test statistic. These clustering solutions include the MCLUST 16 component and EMMIX 10 component models for Howells' (1989) data set, the MCLUST 7 and EMMIX 5 component models for the full FDB sample, and the MCLUST 7 and EMMIX 4 component models for the Black and White FDB subset.

Figure 123 through Figure 133, presents the known versus estimated maps for all three data sets. As can be seen for all cases but one, there is a clear relationship among the purely biologically defined craniometric distance mosaic maps (the base diagrams), the color overlays representing documented identity (population, race) or geography and the color overlays corresponding to the clusters produced using mixture-modeling (MCLUST and EMMIX). If the map of Howells' data by true sex is also considered (Figure 125), consistent disagreement between the male-female distribution and the cluster distribution can be seen, regardless of model or component number. This lack of correspondence implies that sex is not a driving force in the cluster analyses (i.e., cases group by population not by sex) and offers further support for the decision to pool the sexes. Out of the several maps produced, there is only one instance in which the cluster solutions poorly match the true population patterns: the EMMIX 5 cluster solution for the full FDB data set (Figure 130). It is interesting that this model corresponds in number of components to the true number of races and represents the more parsimonious of the other good (MCLUST) solutions. This disagreement may arise, in part from the model selection method of EMMIX,

which chooses the smallest number of components that fits the data under analysis: it does not search out more complex, but potentially, better fitting models. This discrepancy among visualizations may also reflect the sensitivity of the force directed mapping method to the initial placement of the nodes. The fact that the mapping algorithm provides in all other instances highly tractable visualizations suggests, however, that the weakness lies within the mixture model selection method.

If the visualizations are considered collectively, it becomes apparent that these maps attest to the presence of meaningful group (population or race) structure in both the globally sampled and modern forensic craniometric data sets and, furthermore, validate the conclusions drawn from the independent assessments of the patterns of per-cluster allocations. Importantly, they do so in a way that is wholly different from the initial analyses in approach but equally mathematically grounded and algorithmically powerful. Furthermore, force directed methods have the benefit of ease of use: they can be implemented with a reasonable degree of statistical and computational difficulty, they are widely generalized to many data types and modeling questions, they can be quickly and simply interpreted despite the high level data that they contain, and they are visually satisfying for their clear two-dimensional display and use of color-coding. The emerging techniques and resources of computational geometry may offer, therefore, a variety of potentially useful tools for expanding anthropological analyses, whose relevance will only increase as data complexity grows with the shifts in research towards the study of human variation using more analytically demanding geometric morphometric and molecular methods.

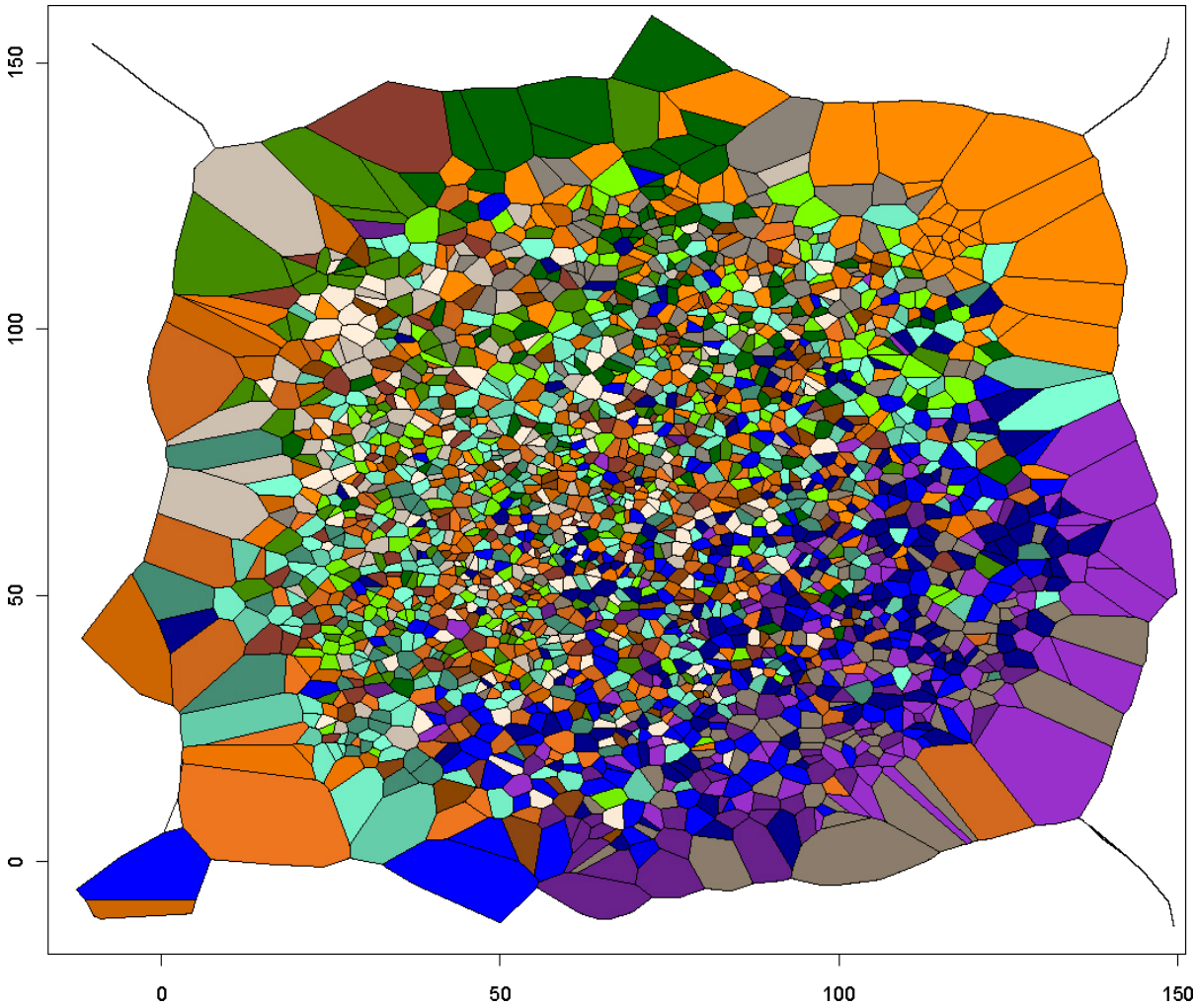


Figure 123: Mosaic plot of Howells' data set (populations by color shade; U.S. Census races by color range)

Specific Populations: AINU=chocolate, ANDAMAN=chocolate2, ANYANG=chocolate3, ARIKARA=chartreuse, ATAYAL=chocolate4, AUSTRALIA=darkorchid, BERG=aquamarine, BURIAT=darkorange, BUSHMAN=blue, DOGON=blue2, EASTER ISLAND=antiquewhite, EGYPT=aquamarine2, ESKIMO=chartreuse2, GUAM=antiquewhite1, HAINAN=darkorange2, MOKAPU=antiquewhite3, MORIORI=antiquewhite4, N JAPAN=darkorange3, NORSE=aquamarine3, PERU=chartreuse4, PHILIPPINES=darkorange4, S JAPAN=coral4, SANTA CRUZ=darkgreen, TASMANIA=darkorchid4, TEITA=blue4, TOLAI=bisque4, ZALAVAR=aquamarine4, ZULU=darkblue

Census Group Colors: Black=blue, American=green, Asian=brown/orange, Other=purple (Tasmania, Australia), White=aquamarine, Pacific Island=gray

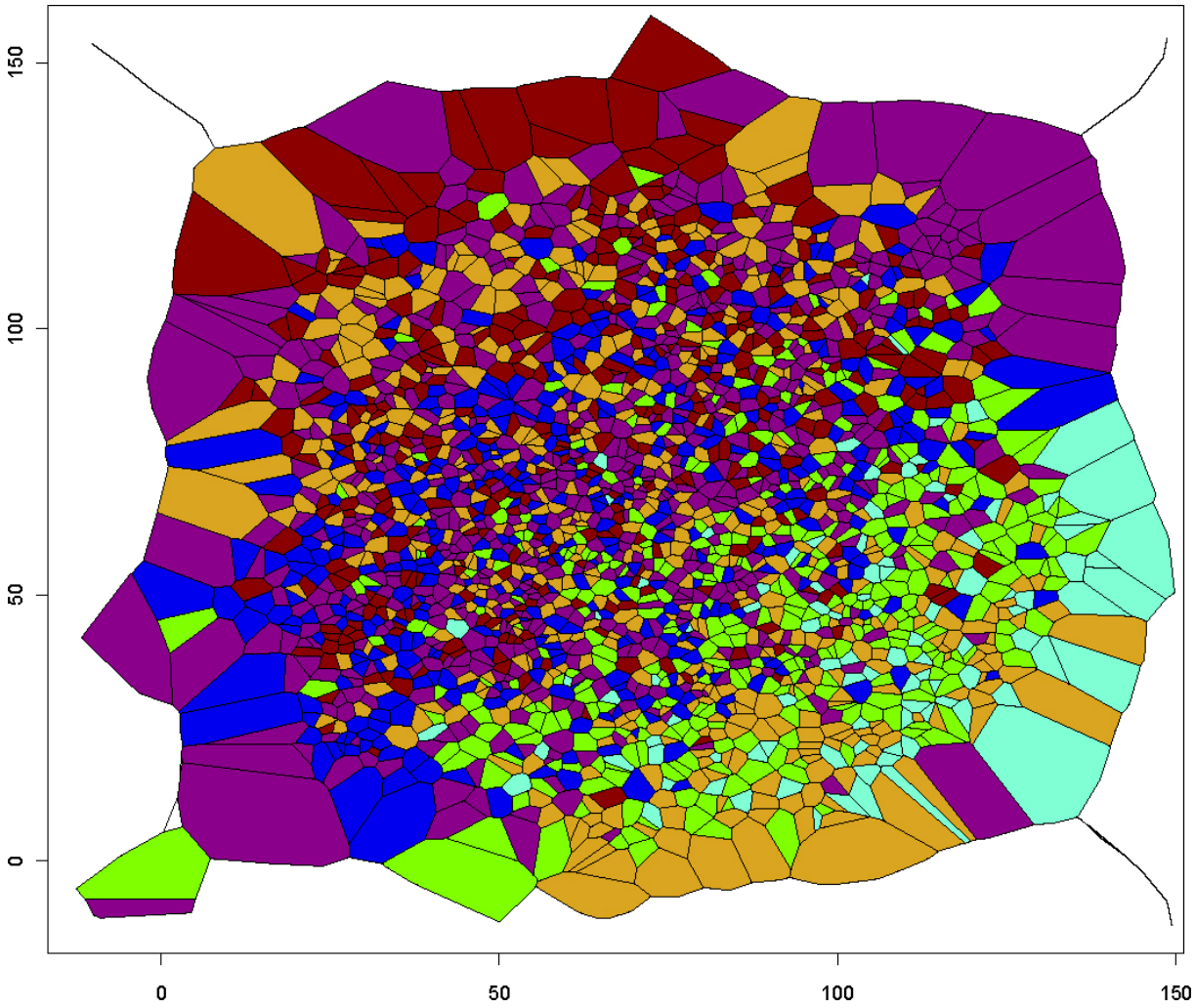


Figure 124: Mosaic plot of Howells' data set (continents by color shade)

Continents: EUROPE =blue2, ASIA=darkmagenta, AMERICAS=darkred, OCEANIA=goldenrod, AUSTRALIA=aquamarine, AFRICA=chartreuse1

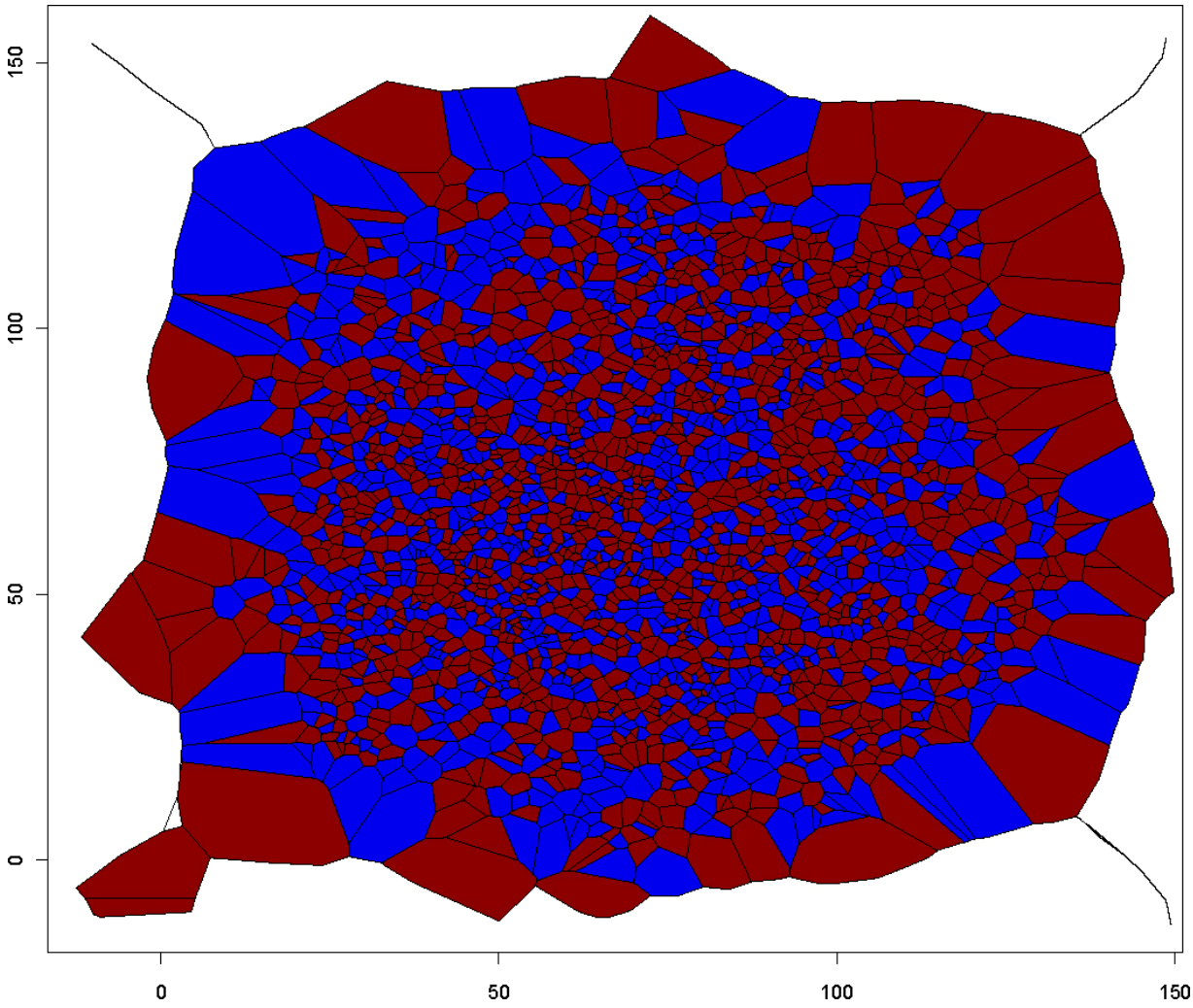


Figure 125: Mosaic plot of Howells' data set (sex by color shade)

Sex: Male=darkred, Female=blue2

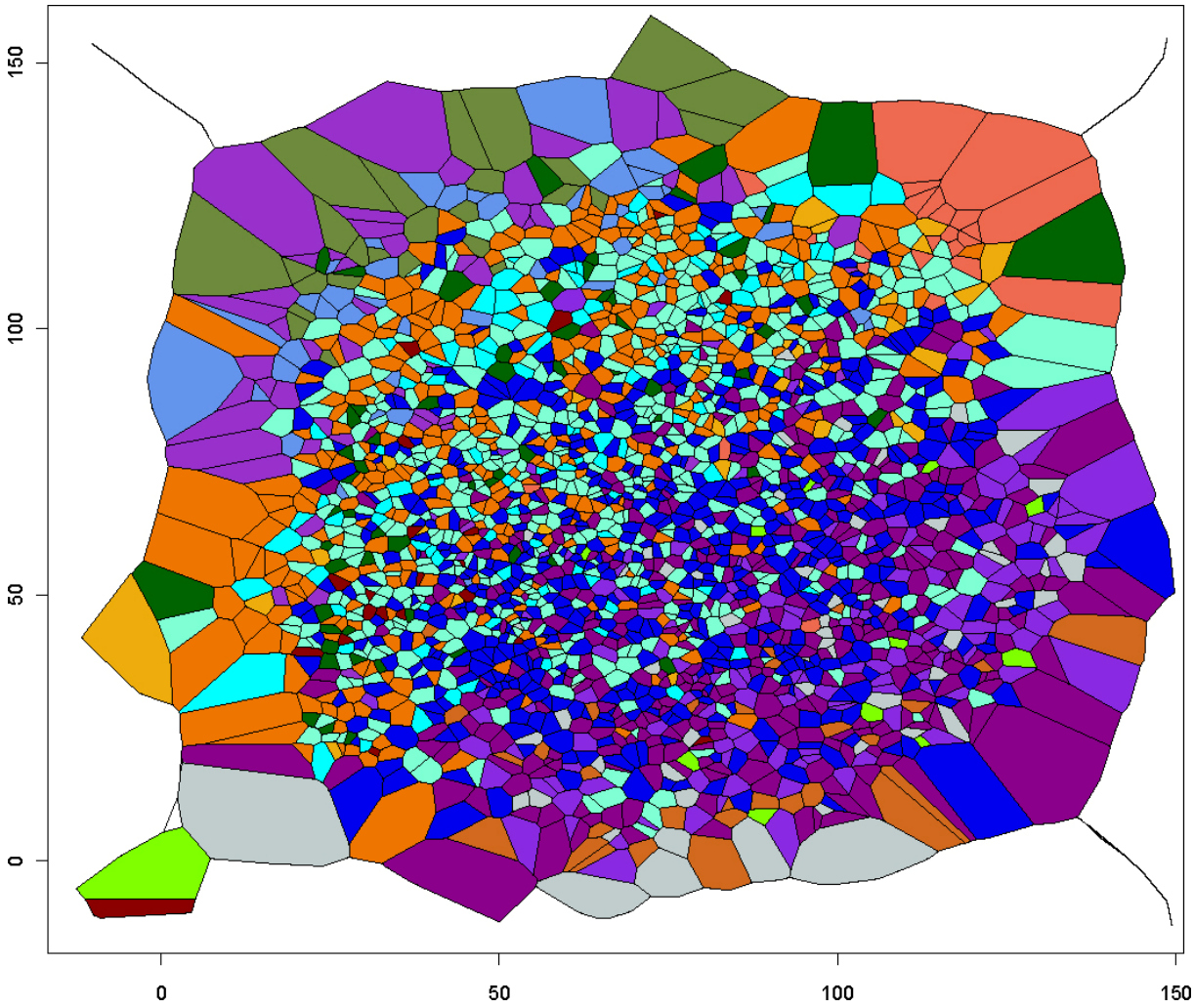


Figure 126: Mosaic plot of Howells' data set (MCLUST 16-component model; cluster by color shade)

MCLUST Clusters: Group 1=chocolate, Group 2=aquamarine, Group 3=darkorange2, Group 4=coral2, Group 5=blue2, Group 6=darkmagenta, Group 7=chartreuse, Group 8=darkred, Group 9=darkorchid, Group 10=azure3, Group 11=cornflowerblue, Group 12=cyan, Group 13=darkgreen, Group 14=blueviolet, Group 15=darkgoldenrod2, Group 16=darkolivegreen4

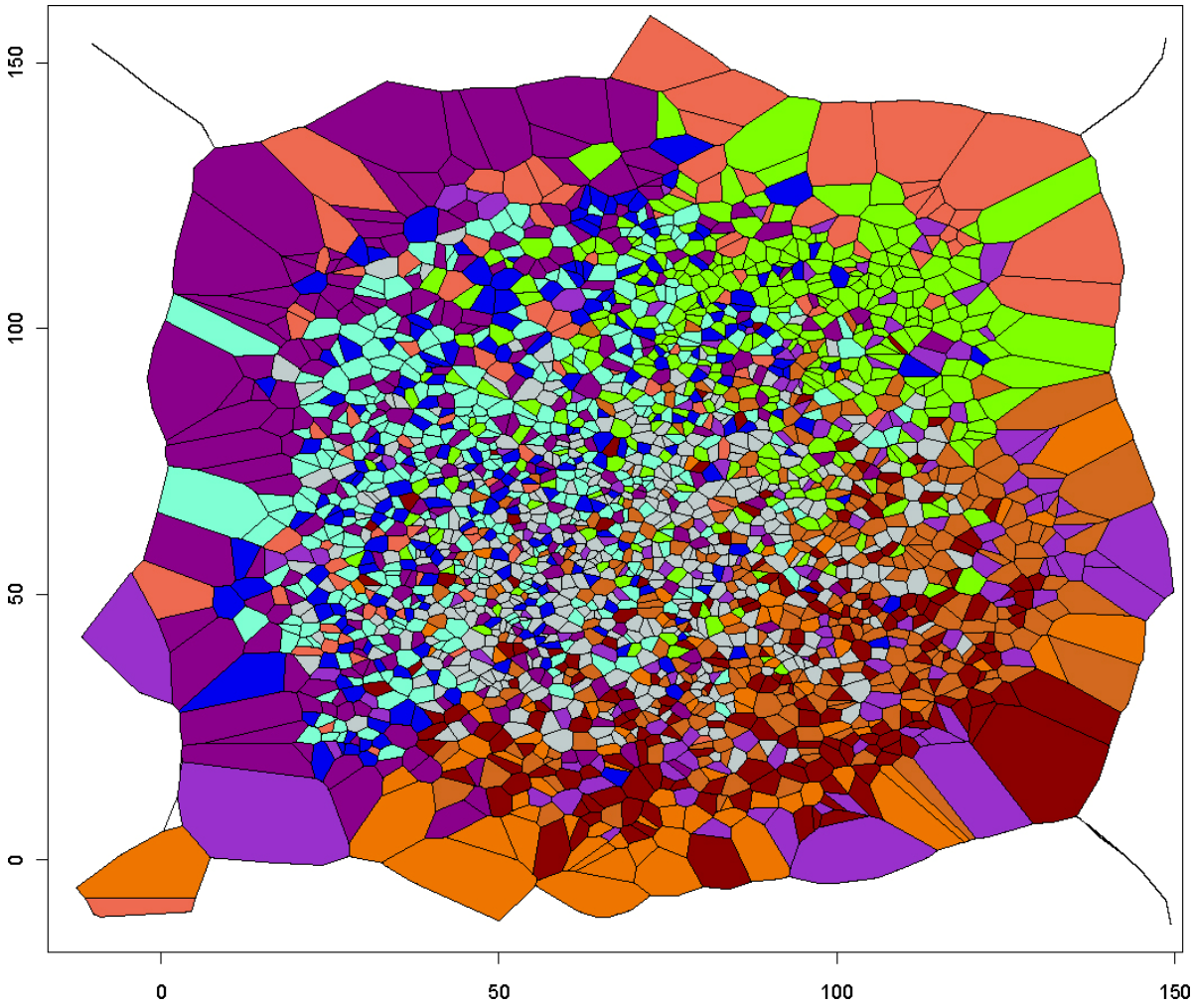


Figure 127: Mosaic plot of Howells' data set (EMMIX 10-component model; cluster by color shade)

EMMIX Clusters: Group 1=chocolate, Group 2=aquamarine, Group 3=darkorange2, Group 4=coral2, Group 5=blue2, Group 6=darkmagenta, Group 7=chartreuse, Group 8=darkred ,Group 9=darkorchid, Group 10=azure3

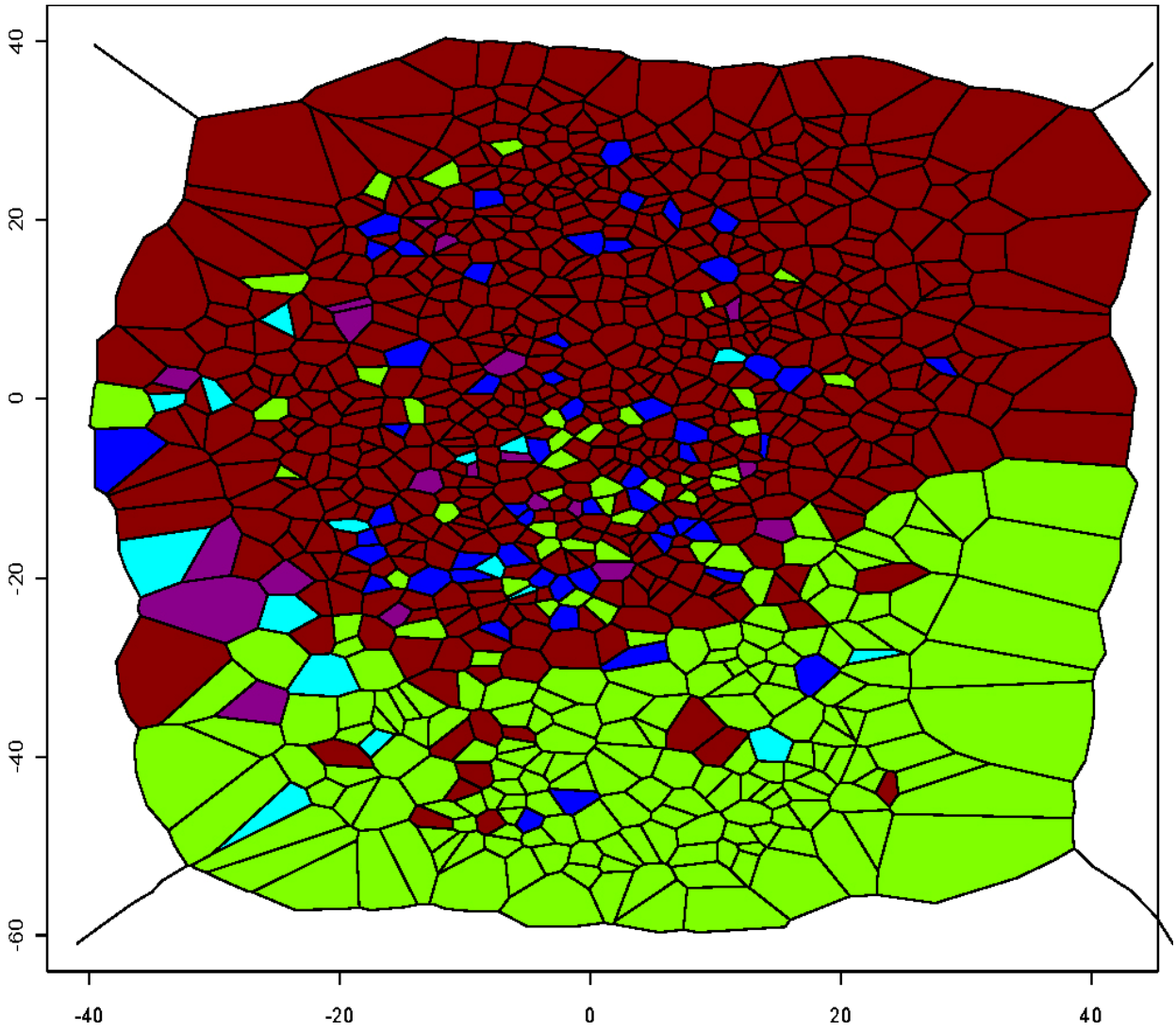


Figure 128: Mosaic plot of FDB data set (5 races; races by color shade)

FDB Races: American Indian=cyan, Black=chartreuse, East Asian=blue, Hispanic=darkmagenta, White=darkred

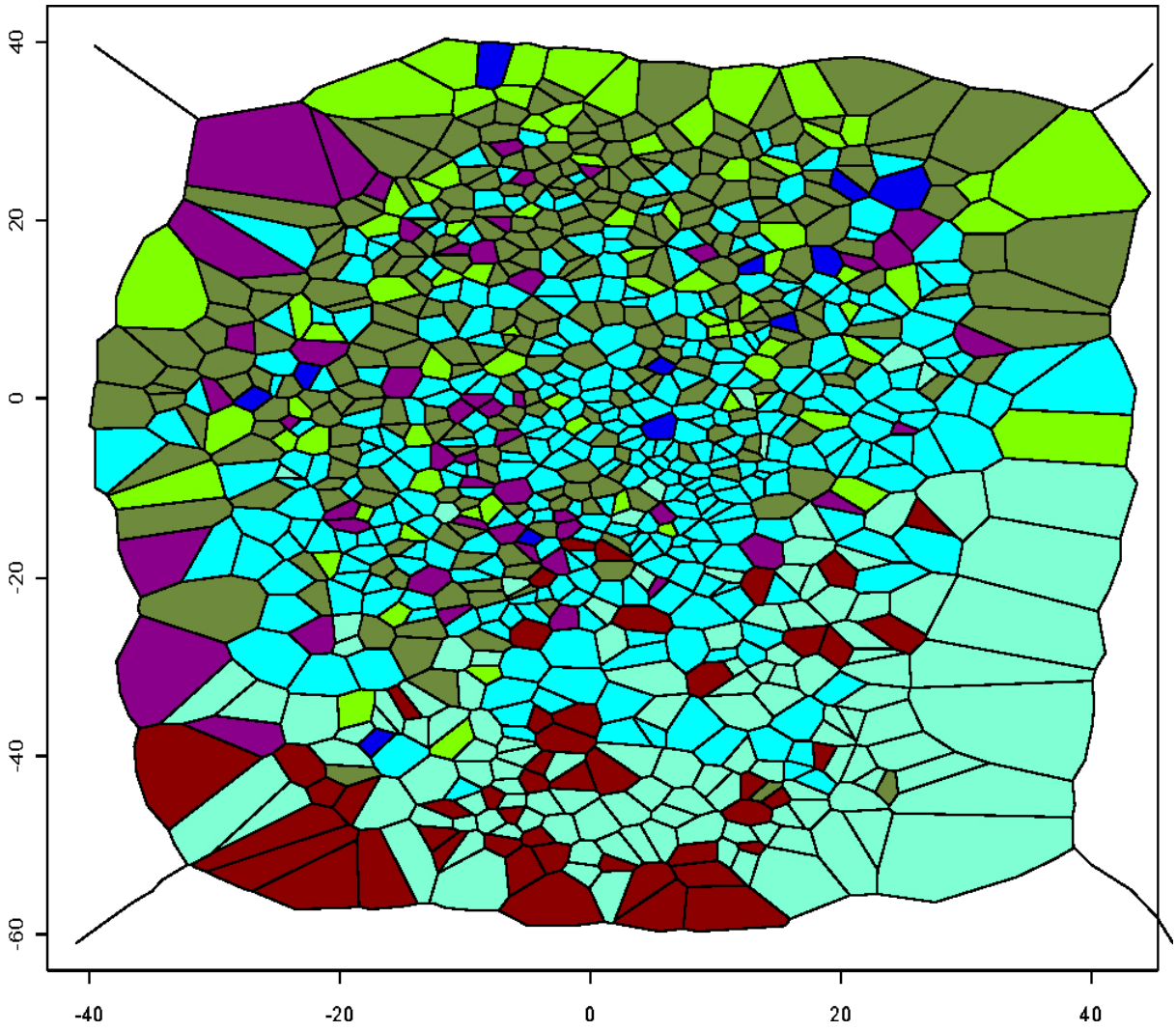


Figure 129: Mosaic plot of FDB data set (MCLUST 7-component model; cluster by color shade)

MCLUST Clusters: Group 1=darkred, Group 2=aquamarine, Group 3=darkolivegreen4, Group 4=darkmagenta, Group 5=blue2, Group 6=cyan, Group 7=chartreuse

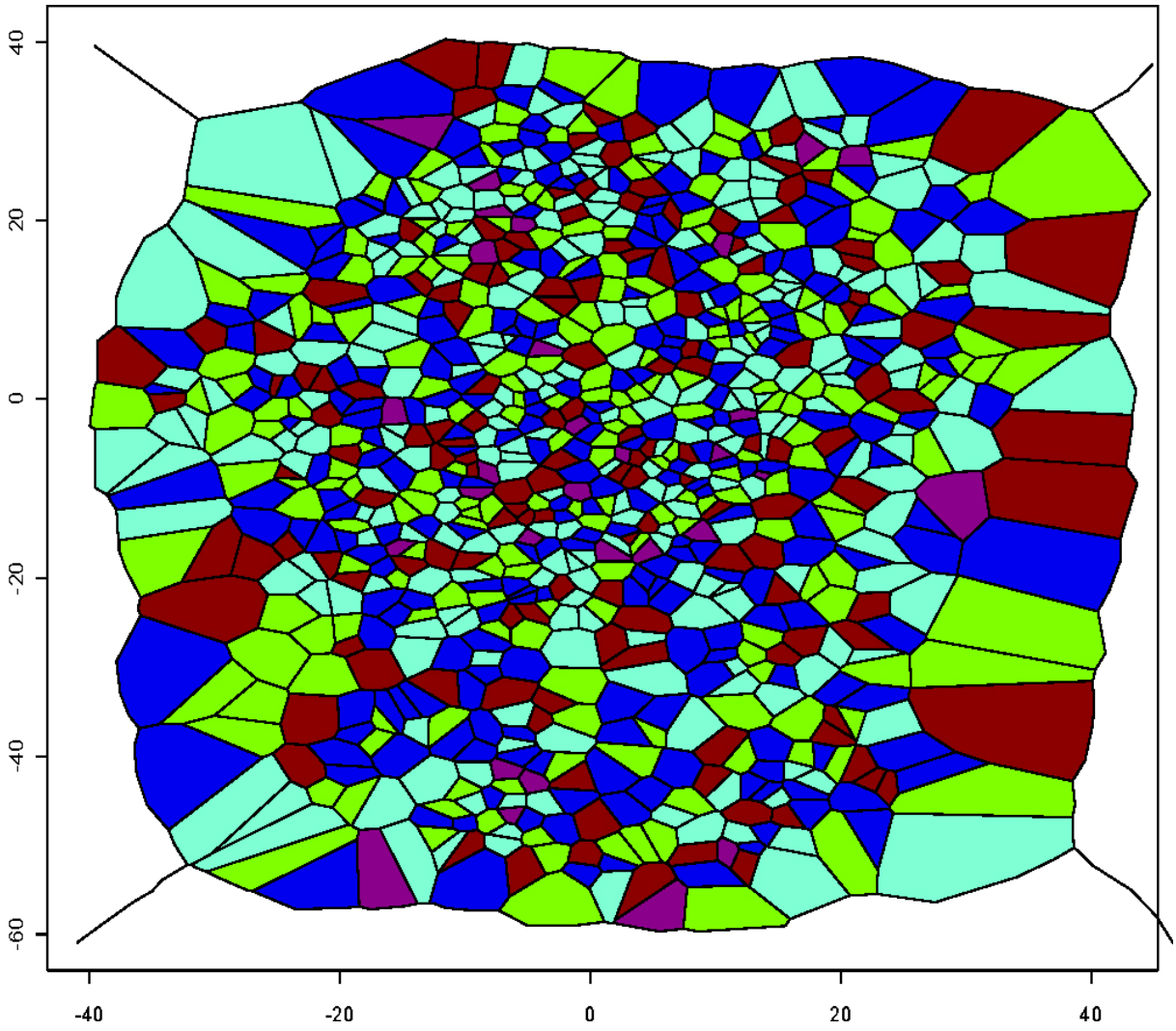


Figure 130: Mosaic plot of FDB data set (EMMIX 5-component model; cluster by color shade)

EMMIX Clusters: Group 1=darkred, Group 2=aquamarine, Group 3=chartreuse, Group 4=darkmagenta, Group 5=blue2

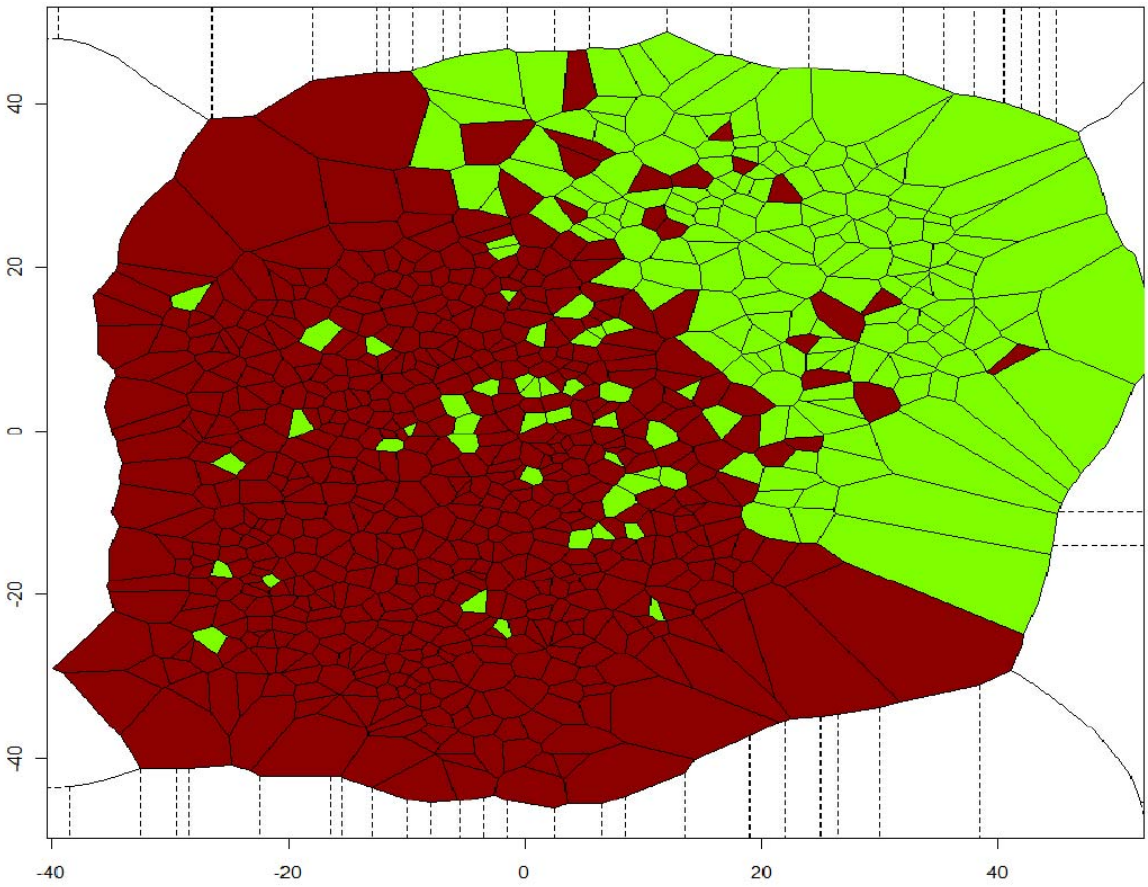


Figure 131: Mosaic plot of FDB data set (American Blacks and Whites; races by color shade)

FDB Black and White Races: White=darkred, Black=chartreuse

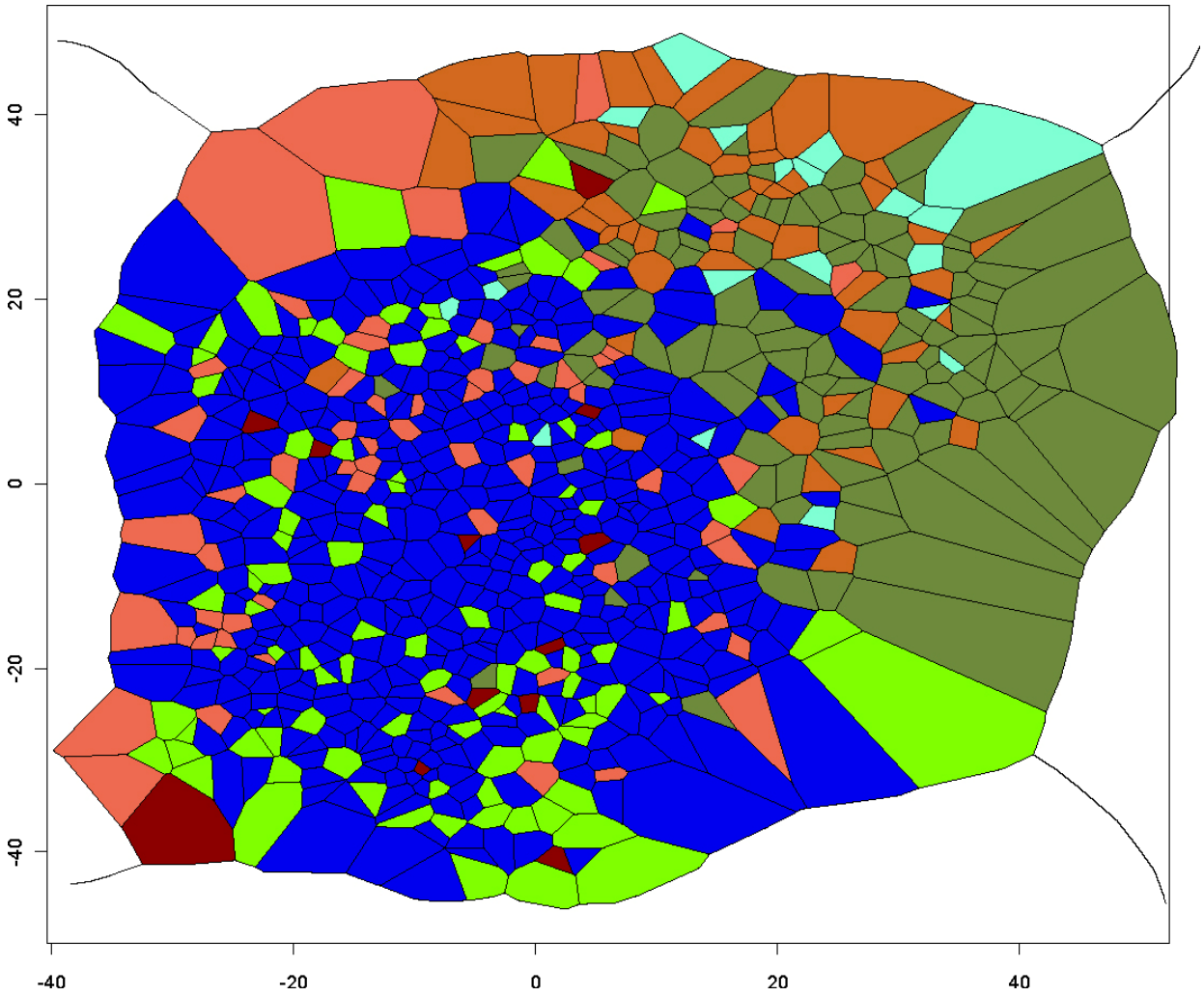


Figure 132: Mosaic plot of FDB data set of American Blacks and Whites (MCLUST 7-component model; cluster by color shade)

MCLUST Clusters: Group 1=chocolate, Group 2=aquamarine, Group 3=darkolivegreen4, Group 4=coral2, Group 5=blue2, Group 6=darkred, Group 7=chartreuse

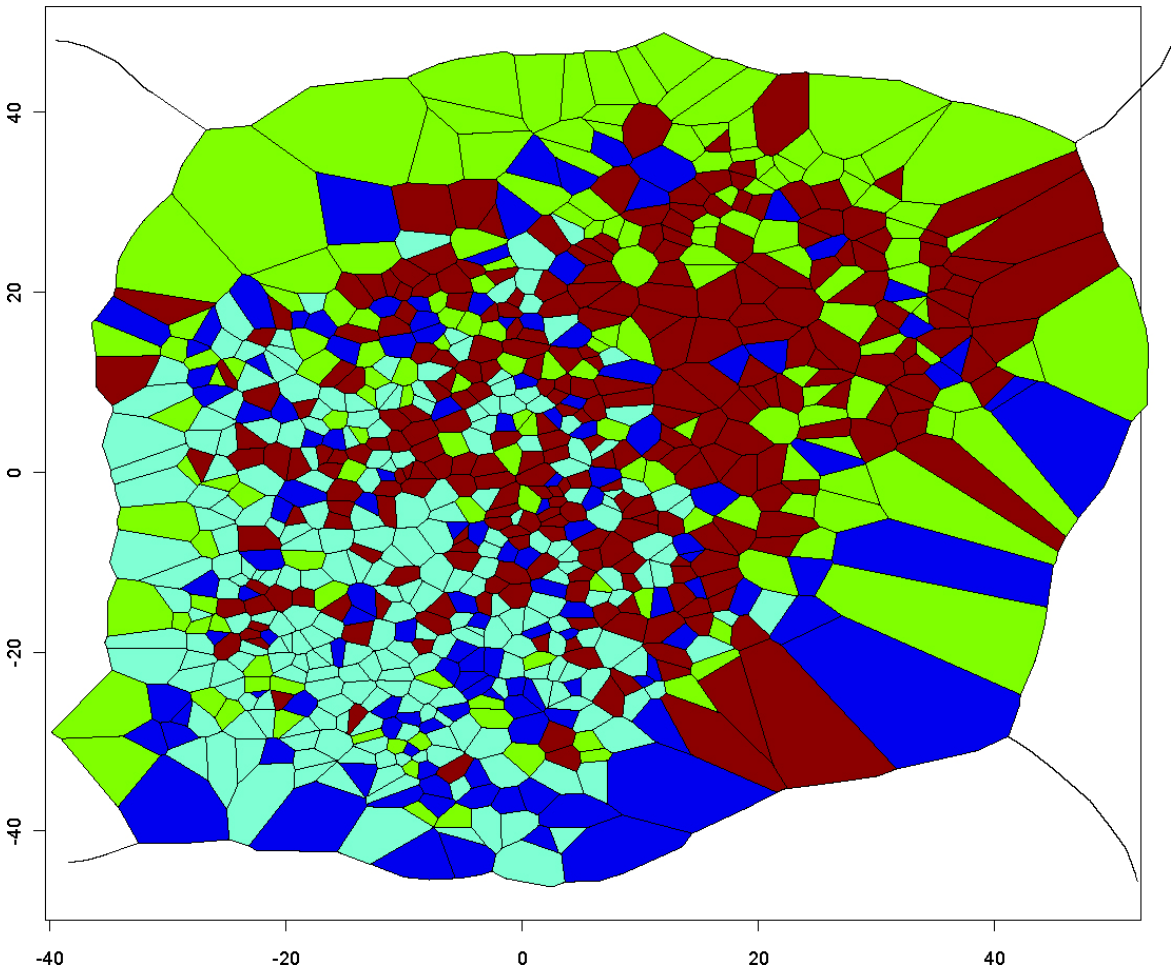


Figure 133: Mosaic plot of FDB data set of American Blacks and Whites (EMMIX 4-component model; cluster by color shade)

EMMIX Clusters: Group 1=darkred, Group 2=aquamarine, Group 3=chartreuse, Group 4=blue2

1	white	#FFFFFF	255	255	255
2	aliceblue	#F0F8FF	240	248	255
3	antiquewhite	#FAEBD7	250	235	215
4	antiquewhite1	#FFEFDB	255	239	219
5	antiquewhite2	#EEDFCC	238	223	204
6	antiquewhite3	#CDC0B0	205	192	176
7	antiquewhite4	#8B8378	139	131	120
8	aquamarine	#7FFFD4	127	255	212
9	aquamarine1	#7FFFD4	127	255	212
10	aquamarine2	#76EEC6	118	238	198
11	aquamarine3	#66CDAA	102	205	170
12	aquamarine4	#458B74	69	139	116
13	azure	#F0FFFF	240	255	255
14	azure1	#F0FFFF	240	255	255
15	azure2	#E0EEEE	224	238	238
16	azure3	#C1CDCD	193	205	205
17	azure4	#838B8B	131	139	139
18	beige	#F5F5DC	245	245	220
19	bisque	#FFE4C4	255	228	196
20	bisque1	#FFE4C4	255	228	196
21	bisque2	#EED5B7	238	213	183
22	bisque3	#CDB79E	205	183	158
23	bisque4	#8B7D6B	139	125	107
24	black	#000000	0	0	0
25	blanchedalmond	#FFEBCD	255	235	205
26	blue	#0000FF	0	0	255
27	blue1	#0000FF	0	0	255
28	blue2	#0000EE	0	0	238
29	blue3	#0000CD	0	0	205
30	blue4	#00008B	0	0	139
31	blueviolet	#8A2BE2	138	43	226
32	brown	#A52A2A	165	42	42
33	brown1	#FF4040	255	64	64
34	brown2	#EE3B3B	238	59	59
35	brown3	#CD3333	205	51	51
36	brown4	#8B2323	139	35	35
37	burlywood	#DEB887	222	184	135
38	burlywood1	#FFD39B	255	211	155
39	burlywood2	#EEC591	238	197	145
40	burlywood3	#CDAA7D	205	170	125
41	burlywood4	#8B7355	139	115	85
42	cadetblue	#5F9EA0	95	158	160
43	cadetblue1	#98F5FF	152	245	255
44	cadetblue2	#8EE5EE	142	229	238
45	cadetblue3	#7AC5CD	122	197	205
46	cadetblue4	#53868B	83	134	139
47	chartreuse	#7FFF00	127	255	0
48	chartreuse1	#7FFF00	127	255	0
49	chartreuse2	#76EE00	118	238	0
50	chartreuse3	#66CD00	102	205	0
51	chartreuse4	#458B00	69	139	0
52	chocolate	#D2691E	210	105	30
53	chocolate1	#FF7F24	255	127	36
54	chocolate2	#EE7621	238	118	33
55	chocolate3	#CD661D	205	102	29
56	chocolate4	#8B4513	139	69	19
57	coral	#FF7F50	255	127	80
58	coral1	#FF7256	255	114	86
59	coral2	#EE6A50	238	106	80
60	coral3	#CD5B45	205	91	69
61	coral4	#8B3E2F	139	62	47
62	cornflowerblue	#6495ED	100	149	237
63	cornsilk	#FFF8DC	255	248	220
64	cornsilk1	#FFF8DC	255	248	220
65	cornsilk2	#EEE8CD	238	232	205
66	cornsilk3	#CDC8B1	205	200	177
67	cornsilk4	#8B8878	139	136	120
68	cyan	#00FFFF	0	255	255
69	cyan1	#00FFFF	0	255	255
70	cyan2	#00EEEE	0	238	238
71	cyan3	#00CDCD	0	205	205
72	cyan4	#008B8B	0	139	139
73	darkblue	#00008B	0	0	139
74	darkcyan	#008B8B	0	139	139
75	darkgoldenrod	#B8860B	184	134	11
76	darkgoldenrod1	#FFB90F	255	185	15
77	darkgoldenrod2	#EEAD0E	238	173	14
78	darkgoldenrod3	#CD950C	205	149	12
79	darkgoldenrod4	#8B6508	139	101	8
80	darkgray	#A9A9A9	169	169	169
81	darkgreen	#006400	0	100	0
82	darkgrey	#A9A9A9	169	169	169
83	darkkhaki	#BDB76B	189	183	107
84	darkmagenta	#8B008B	139	0	139
85	darkolivegreen	#556B2F	85	107	47
86	darkolivegreen1	#CAFF70	202	255	112
87	darkolivegreen2	#BCEE68	188	238	104
88	darkolivegreen3	#A2CD5A	162	205	90
89	darkolivegreen4	#6E8B3D	110	139	61
90	darkorange	#FF8C00	255	140	0
91	darkorange1	#FF7F00	255	127	0
92	darkorange2	#EE7600	238	118	0
93	darkorange3	#CD6600	205	102	0
94	darkorange4	#8B4500	139	69	0
95	darkorchid	#9932CC	153	50	204
96	darkorchid1	#BF3EFF	191	62	255
97	darkorchid2	#B23AEE	178	58	238
98	darkorchid3	#9A32CD	154	50	205
99	darkorchid4	#68228B	104	34	139
100	darkred	#8B0000	139	0	0

Figure 134: R Color Chart

Conclusion: A New [Re]view of Race and Human Variation

"One of the most astonishing features of the contemporary discussion of race is the fact that anthropology, the science that deals with human biological and cultural variation, has managed to be marginalized. [...] The anthropological message has become transformed over the past generation into a vacuous tenet of political correctness. Alongside it is a folk caricature of anthropological idealism: that because races are culturally constructed, everyone is the same as everyone else. If anything, of course, the opposite is true - everyone is different. At issue is the general pattern that structures the natural variation in our species." (Marks, 1995, p. 143)

The introductory chapters of this project review the broadly-based intellectual history of the debates surrounding race. The intention for their inclusion is to provide an overview of how race is conceptualized by members of different scholarly communities, from theoreticians in the physical or natural sciences, to biomedical clinicians and forensic practitioners, to thinkers in the social sciences and philosophy. Within these preliminary discussions is an examination of the social implications of research into contemporary human population variation. This project, first, provides a historical review of the "concept of race," tracing its socio-political evolution with changing times and finding its place within the movements in anthropological theory and practice throughout the 20th century. It also explores the present impact that emerging molecular technologies have had on scholarly interest, policy-making, media presentation, and the public discourse at large. It emphasizes the tension between the contemporary revival of race, through morphogenetic research interested in revealing personal identity (genotyping for commercial ancestry testing, individualized medicine, and forensic identification), and the simultaneous rejection of the scientific reality of race in service to the advancement of human unity.

This focus on a global human identity came with the advent of the Human Genome Project (2001), which has inspired a revolution in how variation is both conceptualized and operationalized, by shifting population studies towards research that emphasizes the fundamental (genetic) similarity among all peoples. In a second discussion, this project focuses on the changing trends in the reception and perception of race concepts among disciplines concerned with the relationship of race, ethnicity, and ancestry to human biology, in physical anthropology and genetics, and its contribution to the meaningful interpretation of both globally and locally distributed variation for use in applied (clinical, medico-legal identification) contexts, which rely heavily upon population-based practices. This overview of race was necessary to contextualize the new approaches adopted here, as well as the theoretical findings that are reported and the methodological recommendations that are made. The major emphasis of this project is placed upon the identification of structured patterns of human biological variation: the issue which Marks recognized already a decade and a half ago as the key point of concern for the, then, current debate over race in anthropological discourse. It is also what Edgar and Hunley (2009) have recently revisited as "the fundamental difference in opinion" (p. 1) in how contemporary biological anthropologists interpret and, ultimately, explain human variation. This question of structured variation, and, by extension, concerns for classification and human identification, still remain among the most critical loci of disagreement in the discourse on human biology that have yet to be fully reconciled in anthropology today.

Out of broadly based reviews of the trends in human biological literature and from dialogues shared among experts in the fields of anthropology and genetics alike, three core views on the question of structured variation have emerged: (1) human biological diversity is clinally distributed (Livingstone, 1962; Brown & Armelagos, 2001; Serre & Pääbo, 2004; Handley et al., 2007), given that only small

amount of the total neutral genetic variation is known to be apportioned among human groups (Lewontin, 1972), (2) neutral (genetic and phenotypic) variation conforms to a pattern of isolation by distance, whereby a smooth increase in biological differentiation occurs with the increase in geographic distance between populations (Relethford, 2004; Prugnolle, Manica & Balloux, 2005; Manica, Prugnolle & Balloux, 2005), and (3) structured human variation reflects both the complex trajectory of modern human evolution and the more recent patterns of population interaction, so that significant past events, like population divergence, bottlenecks and founder effects, and migration with gene exchange or admixture, are reflected in the present expression of quantitative morphogenetic traits (Rosenberg et al., 2002; Risch et al., 2002; Edwards, 2003; Manica et al., 2007; Ousley, Jantz & Freid 2009). Tied to the first two of these three positions is the common assumption that, since there are no geographic discontinuities or boundaries that delineate unique biological entities within the global distribution of human diversity, there is no biological basis for defining distinctive human groups that correspond to the largely cultural concepts of race, ethnicity or ancestry. In contrast, the last point-of-view permits a more flexible interpretation of human variation, by leaving open the possibility that patterns of interaction can be retained as morphogenetic signals of population history, which are identifiable at the level of the regional or local group or in the individual and which may be informative of biogeographic ancestry or biosocial race. Although each of these three core perspectives emphasizes a suite of particular ideas, when the subject of population structure is approached from a holistic perspective, it can be shown that these views are not necessarily mutually exclusive.

It is proposed here that there are three basic questions that can be effectively investigated using craniometric data. These questions, when answered, can contribute to a more integrative understanding of how human variation is distributed. For this reason, they underlie the analyses that

have defined this project. First, is it possible to detect latent population structure using quantitative traits, in the absence of reference populations and without the inclusion of *a priori* known group identifiers? If so, how is this structure organized and does this organization have internal statistical validity as well as external, biological or anthropological, meaning? Finally, what does this organization reveal about the reality and utility of the personal classifiers, like population, biogeographic ancestry or biosocial race, which are fundamental to establishing individual identity? Throughout this project's discussions of theory, method and analysis, it has been argued that the model-based clustering and classification methods of finite mixture analysis (McLachlan, 1992; McLachlan & Peel, 2000) as well as the forced-directed graphing techniques of computational geometry (Aurenhammer, 1991; Di Battista et al., 1999, Kauffmann & Wagner, 2001) can provide the robust and flexible statistical tools necessary to respond to these questions collectively and, in turn, produce reliable results that speak to each of these three key perspectives on human variation.

This project has adopted a multi-part statistical analysis that has applied these finite mixture, model-based clustering methods to three craniometric data sets drawn from three well-documented sources of human craniofacial measurements: Howells' (1996) 28 population and TEST craniometric series, Hanihara's (2008) broadly sampled, global craniometric data set, and a modern American craniometric sample extracted from the University of Tennessee's Forensic Anthropology Data Bank (Jantz 2008). These widely-known and reliable resources have collectively allowed for this project's implementation of large-scale tests of "population structure" at a variety of partitions (e.g., at continental, regional, local levels, through time and between the sexes) and to differing investigative goals (i.e., revealing globally distributed patterns or producing forensic classifications) using a robust set of mixture clustering, supervised model-based classification, and force-directed graphing techniques. These data sets were

also combined with demographic information obtained from the United States Census Bureau (2010, 2011), for the purpose of addressing concerns over forensic estimation procedures. This added information has served as a means to contextualize the craniometric samples within the range of racial diversity represented in the American population-at-large, as these vital statistics allow for the approximation of the distributional patterns expected for contemporary medico-legal casework in North America. Despite the diversity in samples used, the variety of methods applied, and the different conditions under which the corresponding series of tests were executed, a similar set of significant findings were consistently returned. This project has demonstrated, across all analyses, that a latent population structure underlies the mixture of craniometric data. This structure can be clearly detected using unsupervised model-based clustering methods: (1) the statistically inferred clusters are formed from the craniometric variables alone, without the inclusion of any "known" data on true affiliation (e.g., ancestry, race, population designations) and (2) the individual allocations into these clusters correspond to expected patterns of memberships based on distance, human history, and, most importantly, continental, regional, and local (including admixed) levels of ancestry or race-base variation. This structure was also successfully validated in *post hoc* analyses using supervised mixture classification and traditional discriminant function analyses as well as the force-directed plotting techniques that integrate *a priori* group identifiers and distance measures with the cluster findings.

The patterned distribution of the craniometric variation captured by these finite mixture clustering methods is, first, consistent with the prevailing theories on the trajectory of modern human evolution, and supports, specifically, the model of a recent African origin and subsequent dispersal out of Africa. It also corresponds to the expectations for clinal variation and isolation by distance (IBD) models that imply graduated variation in neutral traits over geographic distance and expect instances of overlap in

character distributions. Perhaps, of greatest interest, is the fact that this structured craniofacial variation can also be resolved into clusters that meet acceptable levels of statistical certainty, are informative of population-specific morphologies, and reveal biological affinities that agree with other external evidence on population histories. What is more, these clusters attest to the relationship between human craniometric variation and the contemporary concepts of biogeographic ancestry and biosocial race, as they are employed in, for example, the contexts of demographic data collection, clinical practice, and medico-legal casework. These findings not only show, therefore, that craniometric variation is geographically structured along traditional theoretical lines but that these models need not be adopted to the exclusion of emerging research. These innovative genetic studies support the description of human diversity as discontinuous or "clustered" and are actively revising the discussion of the scientific value of race in the study of human variation (c.f., Rosenberg et al., 2002, 2005; Bamshad et al., 2003; Turakulov & Easteal ,2003; Jorde & Wooding, 2004; Tang et al., 2005; Lao et al., 2006; Wooding et al., 2006). By extension, they can give credence to the well-established, but still contested, practice of using craniometric variables to identify populations and to classify individuals (e.g., Howells, 1973, 1989, 1995; Freid et al., 1995). Evidence that simultaneously supports population theories based on neutral expectations, and also corroborates the utility of clustered biological variables, like craniofacial metrics and multi-locus genotypic data, in revealing population history and speaking to contemporary race-relations may hold significant import for how the fields of human biology continue to understand race and human variation. In this regard, it is worthwhile to recall how, throughout the short but contentious history of the debate over the scientific value of biological differences between human populations, a central tenet to the explanation of human variation has been regularly invoked: that is, only a small proportion of the human genetic variation observed on a global scale is known to account for differences between groups (Lewontin, 1972). These findings have been consistently reiterated in

studies testing the degree of within and among population variation using selectively neutral traits (Nei and Roychoudhury, 1972; Latter, 1980; Barbujani et al., 1997; Jorde et al., 2000;, Watkins et al., 2003; International HapMap Consortium, 2005; Rosenberg et al., 2005) and the FST values estimated from craniometric data have been shown to be comparable in their magnitude and range of variation to the standard values produced using both classic genetic and microsatellite DNA markers (Relethford, 1994, 2002; Roseman, 2004). These findings of low levels of inter-population variation (i.e., 10%–15%) have been extended from the discussions on the significance of human population variation to, specifically, the disputes over the scientific use of race in genetics and anthropology. In a document issued in reply to OMB Directive 15, the American Anthropological Association (1997) responded accordingly, writing that ''data also show that any two individuals within a particular population are as different genetically as any two people selected from any two populations in the world" (p. 1), and a similar statement was later included among the educational materials distributed by the Human Genome Project (2001, p. 812). The new findings on population structure bring into question the limited importance given to the apportionment of diversity among groups. In turn, they demand a review of how mainstream researchers have come to assign value to population-informative traits for the quantitative study of human variation and a fresh look at the utility of clustering and classification solutions in practical identification contexts, using racial, ethnic, and population-based proxies (Bamshad et al., 2004; Witherspoon et al., 2007).

The origins and patterns of dispersal of anatomically modern human (AMH) have long been the focus of a considerable, ongoing debate (Stringer & Andrews, 1988; Wolpoff, 1989; Lahr, 1994; Wolpoff, Hawkes & Caspari, 2000; Wolpoff et al., 2001; Stringer, 2002; Brauer, Collard & Stringer, 2004; Mellars, 2006, 2006; Trinkhaus, 2005; Tattersal, 2009). Much of the recent interest in the spread of modern humans

has been centered on the potential contribution of archaic hominids to the modern human gene pool and the isolation of evidence attesting to specific events of Neanderthal-AMH admixture (Currat & Excoffier, 2004; Serre et al., 2004; Weaver & Roseman, 2005; Green et al., 2006, 2008, 2010; Noonan et al., 2006; Plagnol et al., 2006; Herrera et al., 2009). As population-genetic questions have been at the forefront of these discussion, molecular studies have actively pursued many lines of evidence that involve a wide range of research into the global distribution of human variation (Harpending & Rogers, 2000; Cavalli-Sforza & Feldman, 2003; Garrigan & Hammer, 2006; Relethford, 2008; Weaver & Roseman, 2008). Within this context of analysis, genetic research has revealed three notable geographic trends, which have shown to aid significantly in the characterization of human genetic diversity (Handley et al., 2007; DeGiorgio, Jakobsson & Rosenberg 2009). It has been observed that, as a function of geographic distance from Africa, (1) the level of genetic variation or heterozygosity decreases linearly (Prugnolle, Manica & Balloux, 2005; Ramachandran et al., 2005; Liu et al., 2006), (2) linkage disequilibrium (LD) increases linearly (Jorde, 2000; Weiss & Clark 2002; Jakobsson et al., 2008), and (3) the slope of the ancestral allele frequency spectrum decreases, so that derived alleles are present in higher frequencies in populations that are located farther from Africa (Li et al., 2008). These three patterns are taken by many as "compelling evidence" (Liu et al., 2006, p. 230) for the hypothesis that all modern humans originated from a single location and support the conditions dictated by the "recent African origins" or "Out-of-Africa" models (Cann, Stoneking & Wilson, 1987; Templeton, 2002; Tishkoff & Verrelli, 2003; Tishkoff & Gonder, 2006). As Hanley et al., (2007) write, indeed from this perspective, this "Unique Origin (or 'Out-of-Africa') model is generally accepted nowadays" (p. 432), over the alternative explanations given by the Multiregional Evolution model (*sensu* Wolpoff, 1989; Wolpoff, Hawks & Caspari, 2000). Very recently, further refinements have been made to the theory of single origins. Genetic differentiation has been shown to decrease smoothly with genetic distance over landmasses

that spread out from East Africa, making this region, in particular, the likely location for source for the emergence anatomically modern humans and the starting point for the colonization routes that ultimately led to world-wide dispersal of modern humans. Strong empirical support has also been provided for models that consider a progressive loss of genetic diversity through serial founder events and that have the freedom to incorporate the possibility of multiple migrations and, most notably, gene exchange with archaic populations (Satta & Takahata, 2002; Prugnolle Manica & Balloux, 2005; Liu et al., 2006; Manica et al., 2007; DeGiorgio, Jakobsson & Rosenberg, 2009; Deshpande et al., 2009; Hunley, Healy & Long 2009).

Over the last few decades, several authors have attempted to extract information on modern human origins using the craniofacial and dental variation expressed in world-wide populations (Howells, 1989, 1995; Lahr, 1994; Lahr & Foley, 1994; Relethford, 1994; Relethford & Harpending, 1994; Hanihara, 1996, 2009, 2010). Increasingly, research concerned with within and among group phenotypic variation has employed the genetic relationship (**R**) matrix (Relethford & Blangero, 1990) to produce estimates of genetic distance and genetic differentiation (e.g., Relethford , Crawford & Blangero, 1997; Relethford, 2001). The **R** matrix method, in particular, has allowed researchers in craniofacial variation access to factors like relative effective population size, gene flow, and drift from craniometric data and the opportunity to make more direct comparisons with similar but genetically derived information (e.g., Relethford, 1994, 2002). These comparisons operate under the assumption that, while natural selection has exerted some influence over certain craniofacial dimensions (Roseman, 2004; Harvati & Weaver, 2006; Roseman and Weaver, 2007; Relethford 2010) since the heritability of metric traits is known to fluctuate among cranial regions (Carson, 2006; Martínez-Abadías, Esparza & Sjøvold, 2009), there is still, overall, a close correspondence between craniometric distances and the neutral model of evolution.

The effects of interactions between natural selection, genetic drift, and gene flow do not erase or obscure underlying patterns of population structure and, so, craniometric variation can be used to success as a proxy for neutral variation for a variety of questions of evolutionary, archaeological and contemporary population importance (c.f., Sparks & Jantz, 2002; Weaver, Roseman & Stringer, 2007; Relethford, 2007; Logan & Jantz, 2010).

Within this framework of retained signals, a decrease in within-population diversity with distance from Africa has been detected using genetic and craniometric traits. More specifically, it has been shown that this morphogenetic diversity is higher in sub-Saharan Africa than in any other geographic region. Manica and colleagues (2005) have observed, in particular, that "once geography has been accounted for, there is very little variance left to explain in Eurasia"(p. 370). The distributional pattern has been interpreted as a "genetic signature" indicative of a larger long-term effective population size in sub-Saharan Africa (Relethford, 2001). For skeletal biology, these craniometric findings are increasingly viewed as further support for the hypothesis of a single origin of anatomically modern humans in Africa, followed by a global expansion out of Africa (Manica et al., 2007; Hanihara, 2008; von Cramon-Taubadel & Lycett, 2008). While the project presented here did not set out to address modern human origins and does not adopt the standard methodological approaches that allow for explicit testing of the population processes associated with the appearance and spread of modern human species (c.f., Wall, 2000), evolutionary explanations can nevertheless be given for its model-based clustering results.

As this project has shown thorough its results for several model based clustering tests of world-wide population structure, the mixture algorithms employed here have produced interpretable clusters that capture, with remarkable regularity, morphological similarities among broadly sampled groups. These patterns of individual membership, population affinity and inter-regional associations have in many

cases reiterated observations made by prior tests of among-group relationships, which have relied upon morphoscopic assessments of the entire craniofacial complex or have applied supervised classification techniques to metric traits. Moreover, these observed cluster patterns largely correspond with the expectations of relatedness based upon recent molecular findings. What is of special note here is that this project has revealed both long-term evolutionary trends and short term population histories that reflect the complex processes of contact, exchange, and divergence. It has done so by implementing model-based clustering techniques, which have effectively teased-out the structured pattern of relationships that exist within the craniometric data under analysis. Most importantly, these clustering methods were able to detect this structure in the absence of prior information on biogeography and without requiring the implementation of user-guided procedures.

These model-based cluster analyses have produced solutions that highlight the diversity within African craniometric variation: the cluster patterns have distinguished, for example, the sub-Saharan populations from the widely sampled Bushmen or the North African Egyptians. The statistical strength with which these local clusters are isolated suggests that the diversity among the populations within Africa is high (enabling good separations among the African groups) and that the morphological distinctiveness of the entire region is also great relative to the other major continental groupings, e.g. Asia, Europe, and the Americas. The ease with which the clustering algorithms can detect morphological differences among the sub-Saharan African and the other non-African groups and the different degrees of statistical certainty associated with their clusterings further agrees with the genetic expectations for a monotonic decline in heterozygosity, relative to the distance from Africa. For example, South American populations have been found to carry only 64% of the neutral variation documented in African populations (Manica et al., 2007). In this regard, it is not only reasonable but highly intuitive that these

mixture analyses differentiated the Australo-Melanesian groups from the remaining populations (most often grouping them with the sub-Saharan Africans), but experienced increased difficulty in forming clusters that isolated the series localized within Europe or the majority of the American samples. For example, while the European populations generally cluster together with consistency, they are rarely assigned to their own unique cluster (i.e., in the case of Howells data alone, either a single cluster consisting of the Berg, Norse and Zalavar series or separate components for each population). They are instead lumped with "like" groups in composite clusters that cover a more broadly-based range of variation, including often the Egyptians and, to a lesser occurrence, Northern Asian and American populations. These more generalized patterns imply that this region is weakest in its degree of among-population differentiation, as increased homogenization makes for poor group separation. The Native American populations are also combined in clusters with other groups, most commonly, with Asian and Polynesian populations. While these cluster solutions suggest a similarly low level of regional differentiation, the Americas differ from Europe in that there is structure nested within this broader grouping, suggesting more heterogeneity among groups, within the region: Amerindians (e.g., the Arikara, Santa Cruz, Peruvian samples) are separated from the northern indigenous groups (Eskimo, Aleuts), which often are more closely clustered with the Western Eurasians (Buriats) and Northern Asian (North Japanese), and are more likely, instead, to be clustered with the Europeans and Polynesians. The partitions may reflecting their patterns of dispersal from Africa and waves of colonization of the Native American founders, as well as specialized cold adaptations that are characteristic of the circumpolar peoples (Sardi et al., 2005; Tamm et al., 2007; Fagundes et al., 2008; González-José et al., 2008; Perez et al., 2009; Betti et al., 2009; Hubbe, Hanihara & Harvati, 2009).

This evidence of increasing homogeneity outside of Africa does not imply, however, that sense cannot be made out of the clustering solutions that consider memberships among non-African populations; in fact, many patterns of allocation are repeated and the cluster relationships that they support have obvious evolutionary interpretations or represent significant, past demographic events that left an enduring biological mark of a given population's history. In most cases, regardless of the hierarchical level of analysis (e.g., continental or regional), the method of implementing the mixture analysis (MCLUST or EMMIX), or the intensity of the sampling (using either the Howells or combined Howells and Hanihara data sets), the patterns of association are consistent and clear. There is a strong tendency for the mixture models to form clusters that join, for example, the sub-Saharan Africans with the Australo-Melanesians, the Europeans with the North Africans, the Native Americans with the Asians, and the Polynesians with the Southern Asians. Links are also made between the African and Australo-Melanesians and the Andaman and South East Asians, the Northern Native Americans with the Western Eurasians and Northern East Asians, and Southern Native Americans with Polynesians and Asians. For some, these patterns, on initial inspection may suggest poor clustering solutions and, in turn, predict weak classification accuracy. This perspective certainly drove the comments made by Williams, Belcher and Armelagos (2005) on the results of their supervised classification of a series of Nubian crania using Howells 28 populations as reference groups: "We expected the Nubian population to cluster as a group and to be attributed to populations geographically near it" (p. 342). They took geographic proximity as the determining factor of morphological affinity, perhaps assuming that it either shaped the patterns of variation or erased morphological signals of population history. The poor classifications that they obtained (relative to their stated expectations) suggest, however, that this is not the case. It is argued here that the cluster findings produced by mixture analysis, using globally distributed samples demonstrate a remarkable level of resolution: the clustering solutions attest to the finite mixture

method's ability to tease out latent structure that reflects complex and ancestral population histories. These mixture-cluster findings support, therefore, the arguments levied against theories of cranial plasticity that have given significant weight to the effects of environmental change on craniofacial morphology; they reaffirm, in turn, the utility of craniometric data as a proxy for neutral genotypic relationships (Sparks & Jantz, 2002).

These patterns of craniometric variation also mirror, in many ways, the suggested patterns for global dispersal of anatomically modern humans and its point of origin, based on matched genetic data and evidence from the archaeological record (Jorde et al., 2008). This trajectory assumes (1) an earlier emergence of AMH in Africa, within the last 200,00 years, from a pre-existing population, (2) a population expansion within in Africa, (3) a later expansion out of Africa at an estimated time >60,000 years ago, and (4) a series of global colonization events, with populations dispersing, for example, through the Sinai to the Levant by 45,000-55,000 years ago and making their (first) entry into the Americas between 14,000 and 18,000 years ago (Klein, 1999; White et al., 2003; McDougall et al., 2005; Berg et al., 2005; Liu et al., 2006; González-José et al., 2008). This spatio-temporal or phylogenetic framework can be used to give meaning to several of the clustering patterns produced using Howells' (1996) and Hanihara's (2008) world-wide data sets. While the cluster solutions that create ties between the Africans and Australo-Melanesians are not introducing new ideas (c.f., Birdsell, 1972; Kirk, 1976), they do aid in further establishing the possibility of an evolutionary relationship among these groups: a hypothesis which has been already identified morphologically and attested in the fossil record but has received relatively limited genetic support. Many of the skeletal features that have been classified as "Australasian" are represented in late Pleistocene African fossils (Habgood 1989, 1991; Groves, 1989; Stringer, 1990, 1992; Brauer, 1992) and previous well-known studies of world-wide craniometric

diversity have identified similar links using multivariate statistical methods (e.g., Howells, 1989; Hanihara, 1996). Perhaps most importantly, one of the earliest modern skeletons discovered outside of African was found in Australia and has been dated to 42,000 years ago (Bowler et al., 2003), although environmentally based arguments have also been made that propose an even earlier date (>55kya) for the presence of modern humans in Australia (Miller et al., 1999). Molecular findings supporting this evidence have been, however, mixed. Nei and Roychoudhury (1993) proposed that these two groups shared an immediate ancestor but Cavalli-Sforza (1997) has subsequently reported findings that are in disagreement: using distances among major continental areas based on 120 classic polymorphisms, he found that the greatest calculated (24.7) distance was between Oceania and Africa. This implies a later and indirect colonization of Oceania and obviously problematizes the close and ancestral relationship that has been detected here between Africa and Australo-Melanesia. Only recently, however, Schillaci (2008) identified a pattern of phylogenetic relationships that revisits Nei and Roychoudhury's (1993) earlier discussion; he suggests that "the early modern humans from the Levant either contributed directly to the ancestry of an early lineage of Australasians, or that they share a recent common ancestor with them" (p. 814). Moreover, new attempts at developing more geographically explicit models of dispersal patterns from Africa and world-wide modern human settlement history have also lent their support. They have acknowledged the equal possibility of two routes for the out-of-Africa migration: one which passes through the Levant, and another which takes a southern path though the Horn of Africa, along the coast of the Indian Ocean to Southeast Asian and Australia (Lahr & Folley, 1994; Macaulay et al., 2005). While the former route represents the more conventional model, both options produce comparable geographic distances along landmasses (Liu et al., 2006). Similarly, competing models for the subsequent colonization of Melanesia have been recently addressed: these propose routes that pass through either the Malaysian Peninsula or the Indian subcontinent (Liu et al.,

2006). This connection made between these series of passages from Africa to Australia and Melanesia and South East Asia not only support their close evolutionary relationship but also helps to explain the morphological similarities detected among these groups. Hanihara (2000) found, for example, that "some subSaharan Africans share similar characteristics with Australians in terms of marked prognathism and flat frontal bones in the sagittal plane on the one hand, and with eastern Asians on the other hand, for flat nasal and zygomaxillary regions"(p. 105). When taken on a wider scale, these findings help to cement the emergence and dispersal models that adopt the out-of-Africa hypothesis over multi-regional evolution; furthermore, they can be taken to also have population-specific importance. These connections offer up, for example, a reasonable explanation for the peripheral clustering of the Andaman Islanders with the African and Australo-Melanesian groups. The patterns of cluster membership frequently places the local Andaman series (and Hanihara's like samples) as intermediates between the sub-Saharan African and Australo-Melanesian populations (most often adjacent to the Tolai) and Southern Asian groups (e.g., Ainu, Anyang, Atayal and Philippines), with whom they have been shown to have close genetic affinity (Thangaraj et al., 2003, 2005; Endicott et al., 2003, 2006; Reddy et al., 2007; Barik et al., 2008). It is interesting to also note at this point the position of the Polynesians in relation to these neighboring groups, both from the perspective of the mixture clustering solutions and genetic data. While the Western Pacific Island populations are geographically proximate to the Micronesians and Melanesians and, by extension, the Australians, their cluster patterns imply their greatest similarity is with East Asians and Native Americans. In fact, they were much more likely to be allocated to groups that contained the morphologically extreme Eskimo or the geographically distant Peruvians than to those clusters heavily populated by their more immediate southern neighbors. These findings are supported by recent genetic studies that have demonstrated that the Polynesians are "distinctive," they share little relationship with Melanesians, have a greater tendency to resemble

Micronesians, Taiwan Aboriginals and East Asians, and, so, most likely originated in Taiwan or East Asia (Friedlaender et al., 2008). As the latter group is taken to be responsible also for the settlement of the Americas, it is not surprising that previous craniometric distance studies have found the greatest association between Pacific and East Asian peoples (over, e.g., the Jomon or Arctic populations) and links were made here between the Pacific Island and Native American populations (c.f., Ishida & Dodo, 1993).

These evolutionary discussions need not be limited to population clusters that are close to Africa. Migration-based explanations can also aid in interpreting the clustering solutions for even the most distant of groups, like the circumpolar populations. In this case, the finite mixture analyses consistently grouped together the Northern Native Americans (e.g., the Eskimo, Aleut, North Canadian series), associated these with North Western Eurasians (like the North Japanese and Siberian Buriat), and created distinctly Eskimo and/or Buriat clusters. These patterns, although initially confusing for their multiple configurations, can be shown to reflect the complexity of the peopling of the New World and the exceptional selection pressure exerted by extremely cold climates on cranial shape. Perhaps still among the most influential publication in the debate over the settlement of the Americas is the 1986 work by Greeneberg and coworkers, in which they put forth an argument that considered multiple lines of evidence (i.e., genetic, dental, linguistic). Their model localized the ancestors of present-day Native Americans in Siberia and proposed three waves of migration occurring at three separate times, which yielded three different descendent groups: (1) the Paleoindians, now represented by contemporary Amerindians, who occupy most of North America, (2) the Na-Dene speakers, like the Navajo and Athabaskan tribes, who currently reside along the North Pacific coast, the interior of Alaska, and some areas in the American Southwest, and (3) the Eskimo and Aleuts, who inhabit the circumarctic regions.

While many recent craniometric explanations have drawn heavily upon this model, especially in discussions of the multiple migrations that it implies (c.f., Powell & Neves, 1999; Jantz & Owsley, 2001; Sardi et al., 2005; González-José et al., 2008), and genetic data has confirmed the proposed points of origin by tracing the ancestral lineages back to North Eastern Asian and Siberian sources, more recent molecular studies are showing a trend towards a single, complex scenario. This model involves a major migration, followed by reticulate gene flow between Beringia and Siberia (or Asia) and across the circumpolar region; it suggests, therefore, a large amount of shared (more recent) genetic heritage among the Eskimo, Aleuts, and Siberians (Santos et al., 1999; Tarazona-Santos & Santos 2002; Schurr, 2004; Zlojutro et al., 2006; Santos, Bonatto & Bertolini, 2007; Schroeder et al., 2007; Tamm et al., 2007; Wang et al., 2007). It is this back and forth exchange that may explain the significant differentiation seen in the extreme Northern populations, like the Aleut and Eskimo, from the other Native American groups (Howells, 1989; Hanihara, 1996, 2000). Ishida (1992) reported for craniofacial data that "While fundamentally having an extremely flat face, Siberian populations tend to be differentiated into two or three subgroups. On the other hand, other Mongoloid populations show greater variation in facial flatness. The less flat faces of the American Indians are almost equal to those of the Europeans" (p. 53). The greater East Asian morphogenetic ancestry that has been attributed to these circumpolar groups is also reiterated here. The clustering solutions produced findings that support this post-settlement gene-flow: it is represented by the frequent and closer association of Eskimo (rather than other indigenous groups) with Howells' (1989) Siberian (Buriat) and North Japanese series as well as the Aleuts and North Canadian samples made available by Hanihara (2008). Recent studies on the influence of climate on cranial morphology have paid special attention to the extreme cold adapted craniofacial complex, as an exception within the overall trends that otherwise suggest general selective neutrality (Roseman, 2004; Harvati & Weaver, 2006; Betti et al., 2009; Hubbe, Hanihara & Harvati, 2009). Of particular interest are

the results presented by Hubbe, Hanihara and Harvati (2009, p. 1720) who found that, while world-wide cranial morphology is phylogentically informative, North East Asian, North American, and Northern European populations do show a high susceptibility to selection related to cold temperatures. The differences observed among each of these regional groupings can be attributed to the fact that they "arrived at their morphology through distinctive adaptive pathways" (Hubbe, Hanihara and Harvati, 2009, p. 1720). When taken collectively with the genetic evidence for shared ancestry, these climatological findings can help to explain the two clustering trends observed here: some number of members from all of the circumpolar groups always clustered together and often subsets of the individual populations were also allocated into their own isolated cluster. While this pattern of isolation is especially true for the case for the Eskimo and Buriat series, the extent to which one cluster pattern dominated over the other (i.e. in terms of the number of individuals clustered alone or together) for all of the northern groups largely depended on the configuration of the groups tested (i.e., using the Howells or the combined Howells and Hanihara data sets) and the number of metric variables used (e.g., 12, 14 or all 57). Although the Paleoanthropological specimens obtained from Howells' (1995) TEST data set were included in select mixture analyses to serve primarily as a easily testable out-group when addressing this project's methodological questions, it is interesting to also consider these cluster results, and the Neanderthal allocations in particular, in the context of climate adaptation. In the past, the distinctive morphology observed among Neanderthal specimens has been interpreted as a product of cold environmental conditions (e.g., Coon, Garn & Birdsell, 1950; Sergi, 1958). In much of the current research, climatic explanations have largely been supplanted by discussions of genetic drift (Howell, 1952; Hublin, 1998) and the possible functional demands imposed upon the masticatory system (Rak, 1986; Trinkaus ,1987). In a recent study, however, Weaver, Roseman and Stringer (2007) could not reject climate as a potential driving force in the evolution of Neanderthal craniofacial morphology.

Hubbe, Hanihara and Harvati (2009) have proposed, in this same regard, that the morphological features indicative of adaptation to extremely cold climates that have been observed among Neanderthals (e.g., increased cranial breadth and nasal height, relative to the ancestral expression of these traits) may be consistent with the craniofacial patterns detected in modern human groups inhabiting cold regions. They noted in contrast, however, that, while "some aspects of Neanderthal morphology are consistent with an adaptation to extreme cold conditions," a clear relationship between cold adapted modern populations (who themselves were subject to diversifying selection) and Neanderthals is difficult to establish "given the constraints of their ancestral morphology" (2009, p. 1729). These findings may explain the unexpected cluster results produced when the Paleoanthropological samples were included in mixture analyses using Howells' (1989, 1995) global population and TEST data sets: that is, the highly divergent groups, like the Eskimo, maintained the cohesive clustering that has been repeatedly observed using the Howells or the Howells and Hanihara data alone, but the Paleoanthropological samples, who are arguably still more divergent in morphology than the modern human extremes, neither formed their own outlying group, as a unique "paleo" cluster, nor joined the groupings that contained, in high numbers, the populations to which these specimens may be expected to share the greatest amount of adaptive similarity. This lack of interpretable memberships may also suggest that other contributing, perhaps, statistical factors (e.g., poor treatment of extreme data outliers), were influencing the patterning of the "paleo" sample's allocation.

Of the "unexpected" cluster memberships produced by the model-based clustering methods adopted here, perhaps the most contentious is the consistent pairing of Howells' Egyptian series and Hanihara's North African group with the European samples instead of with the sub-Saharan samples. These findings

are potentially problematic given the geographic expectations for continental groupings, as prior discussions of the [mis]classification of the Nubian cases reported by Williams, Belcher and Armelagos (2005) have shown. They are also challenging to interpret in light of several tests of morphological similarity among Egyptians and other (African) populations (e.g., Keita, 1993), which have sought to answer questions of origins and ancestry or to quantify the effects of gene flow and migration as they relate to population history of the "Egyptians" and state formation in ancient Egypt (Keita, 2004; Irish 2006; Zakrzewski, 2007; Schillaci, Irish and Wood, 2009). Hypotheses suggesting, either, in situ development of local groups, descended from regional Neolithic peoples, or pivotal demographic events, like large scale migrations and even forcible incursions, have remained at the core of these debates over Egyptian biological relationships and African affinity. Keita's work in craniofacial relationships among African groups has held a place at the forefront of these discussions, as he has consistently reported findings of a collective African morphology: arguing that the Egyptian populations "are best viewed as part of an African descent group or lineage with tropical adaptations or relationships," identifying, specifically, the southern predynastic peoples as "Saharo-tropical variants" (1993, p.150-151). Similarly, Irish (2006) has provided evidence that corroborates theories of population continuity between the predynastic and dynastic periods, especially in early dynastic Upper Egypt. Schillaci, Irish and Wood (2009) and Zakrzewski (2007) have also independently concluded that "state formation was largely an indigenous process with some migration into the region state" (p. 242), but the identity of the possible migrants, as yet, remaining unknown. These arguments not only stand in opposition to the findings reported by Howells (1989) and to the observations made here using unsupervised clustering methods, but they are also in disagreement with the trends in genetic literature. Findings of genetic discontinuity in North Africa, and, in particular, the Nile Valley have been reported (Fadhlaoui-Zid et al., 2011), and many recent molecular studies, using globally sampled data, have

emphasized the biological similarity between North Africa, the Middle East, and Europe (Herráez et al., 2009). These latter findings imply, therefore, some complex past population relationship(s) that cannot be explained by modeling geographic distance alone and, so, they suggest that continued testing is needed.

The observed morphogenetic relationships between Europe and North Africa also serve as a good example of the implications of phenotypic or genetic cluster analyses to the ongoing conflict over the correct interpretation given to human diversity. For example, this project's consistent classifications of the Egyptians with the non-Africans, is of significant interest to the question of the utility of race in understanding and relaying information on human variation. As this clustering solution forms a more broadly characterized "Caucasian" group, it demonstrates how a race-based identification can better represents the make-up of a composite (here, North African + European) grouping than biogeographic ancestry. This problem of disconnect between presumed affinity based on geographic proximity and actual biological similarity is further highlighted by the clustering results and history of the Ainu, the indigenous inhabitants of Hokkaido. While the Ainu are geographically located in Japan, this population has been called "one of the most obvious outliers in an eastern Asian context" (Hanihara, Yoshida & Ishida, 2008, p. 284). The results of the mixture cluster analyses identify two trends in association for the Ainu, both of which have been proposed in previous morphological studies: one which suggests affinity with South East Asian populations (Turner, 1987, 1990) and another which suggests an association with Pacific island (Brace, Brace & Leonard, 1989, Brace & Hunt, 1990) and Native American populations (Jantz & Owsley, 2001). These findings imply, by the fact that the Ainu samples are split over these different groups, that perhaps the explanations given for the Ainu origins need not be competing. The Ainu population may reflect a complex history of admixture from these different sources, which has

contributed to their highly atypical morphology relative to the surrounding Japanese peoples (c.f., Hanihara, 1987, 1991, 1998, 2006; Hanihara, Ishida & Dodo, 1998).

In addition to these trends in evolutionary or historical neighbor-joining, distinct population-specific groupings were also formed, which took in subsets of the broader swaths of biogeographic variation. Clusters were produced that included, for example, only samples that represented some aspect of population-specific diversity found "nested" within the broader groupings of the Southeast Asians, Northern Eurasian and Circum Arctic populations, sub-Saharan Africans, Australo-Melanesians, Polynesians, Southern Native Americans, and Europeans. The strength of these relationships and the likelihood of producing clusters that have a direct correspondence with a local group or population increased substantially with the number of craniometric variables considered in the analysis. When all available variables were used, the cluster findings produced more closely approximated the results reported by Pontikos (2004) for his mixture analysis of Howells' (1989) data set: "separate clusters were identified in the best run, each of which corresponds strongly to either one of the original populations, or to a racial group. It is shown that cranial variation can be used to infer ethno-racial affiliation" (np). These results demonstrate how innovative uses of a model-based clustering approach can contribute to existing theories of origins and population history, by offering robust statistical evidence on inter-group relationships that may support one line of reasoning over another. Moreover, "by helping to frame the ways in which human genetic variation is conceptualized," a finite mixture approach can also be used to reveal new information or shed different light on old problems or conventional beliefs (Rosenberg et al., 2005, p. 0661).

It is precisely in this way that mixture models are being increasingly applied to world-wide data sets. The information that they provide has greatly expanded the contemporary understanding of the structure of human populations and it is, most notably, contributing to a revived dialogue on race and human variation. This is because these clustering approaches are enabling discussions that bring together the seemingly incompatible concepts of (1) clinical variation, (2) group-specific classification and (3) the resolution of latent population structure into statistically inferred clusters, which have been shown to correspond to proxies for world-wide biological diversity (e.g., populations, ancestries, races, and ethnicities). There is growing consensus in recent genetic literature that, even though the clinal nature of human diversity has been well-documented using genetic markers, this same molecular data can be assembled into clusters that represent five or six broad groups, that closely approximate racial divisions and correspond to continental geography (c.f., Risch et al., 2002; Rosenberg et al., 2002; Bamshad et al., 2003; Zhivotovsky et al., 2003; Jorde & Wooding, 2004; Tishkoff & Kidd, 2004). What is important to emphasize about the seminal genetic examples (e.g., Rosenberg et al., 2002; Bamshad et al., 2003) is that the cluster-based analyses that they present do not necessarily place their findings at odds with the long-standing approaches to human variation: these finite mixture cluster approaches can be taken to work in concert with evolutionary, clinal and geographic distance models. It is for this reason that recent molecular studies, for example, are able to present standard measures of F_{ST} alongside the results obtained from their genetic cluster analysis. This is quite remarkable since both were applied to very different goals, i.e., the clustering were methods were used specifically to ascertain population structure, to assign of individuals of unknown origin to populations, and to identify genetically distinctive populations (Rosenberg et al., 2001). It is argued here that the same conditions of "compatibility" can apply for phenotypic, and, specifically, craniometric, data given what is known about the (similar) pattern of distribution for metric traits.

Craniofacial, like genetic, differentiation among human populations at a world-wide scale is primarily dependent on geographic separation and, for more atypical morphologies, isolation. It is this phenomenon that has lead to the development and recognized utility of clinal and isolation by distance models (Relethford, 2004). It is not surprising then, that the present study has produced cluster results that show how craniofacial traits, represented by globally sampled data, are organized into clines. North-South and East-West divides become clear when both the Howells and combined Howells and Hanihara data sets are treated in terms of the three major continental groupings. Africa is distinguished from Europe, which is, in turn, distinguished from Asia. Similar partitions are also identifiable within these regions. For the Americas, the circumarctic populations are differentiated from the series originating in the United States and South American. For Asia, the South Eastern populations are separated from the Northern Asian and Western Eurasian groups. Despite these patterns, discrete or "discontinuous" population, racial and ethnic clusters can be defined even after geography has been taken into account. It is also shown here that the degree of heterogeneity in sampling does not produce significantly different results: a concern that was first argued by Serre and Pääbo (2004) for clustered genetic data and more recently implied by Relethford (2009) in the context of discussing the Howells data set and the success of craniometric discrimination methods. Rosenberg and colleagues (2005) have also provided convincing evidence attesting to the insignificant effects of sampling density, which they have taken as further support for the biological meaning of their genetically derived clusters.

Global craniometric, like genetic, variation may be best explained by a synthetic model, which assumes that the majority of the differences between groups can be accounted for by isolation by distance (IBD) and that only a small amount of these differences (the remaining variation) will be explained by breaks in the clines or gradations of character variation, which have arisen from barriers to trait dispersal (c.f.,

Handley et al., 2007). Taken in this way, the majority of the craniofacial variation observed is attributable to clines, so that clusters explain only a small fraction of the total craniometric variance. It is, however, the highly correlated information contained within this small, among-population component of variation that enables clustering and classification (Edwards, 2003; Jorde & Wooding, 2004) and has made metric identification methods successful (c.f., Ousley & Jantz, 2005). It is for this reason that classification accuracy increases with the inclusion of more traits; this is why considerable improvement in the clustering solutions (i.e., model fit, taken as the correspondence between a cluster and a true population) was observed here when the quantity of variables considered in the finite mixture analyses was greatly increased: moving from this project's standard subsets of 12 variables (for Howells' 28 population data set) or 14 variables (for the combined Howells and Hanihara data set) to all 57 variables, i.e., including the maximum number of available measurements. From methodological perspective, these findings also demonstrate how careful negotiations must be made between variable number and model parameterization when using mixture methods; the researcher must thoughtfully weigh the value of computational tractability and the risk of algorithmic failure, against the improvement in cluster separation and membership interpretability that may be gained by increasing the mixture model's complexity (McLachlan & Peel, 2000). It is nevertheless quite remarkable that statistically, biologically, and even anthropologically meaningful clusters were produced here using a small, well-chosen, but not statistically selected subset of variables. Furthermore, the inclusion of more predictors did not significantly alter the overall trends in the cluster patterns, but, instead, more clearly revealed the latent data structure already detected, by producing a solution with more sharply defined population-specific clusters, so that, in almost all cases, a cranial series was assigned wholly to a single (i.e., its "own") cluster.

Given this proven compatibility between clines and clusters, the demonstrated relationships that these statistically inferred (mixture) clusters have to race in genetic contexts, and this project's success in resolving *population*-based clusters out of mixtures of craniometric data using globally distributed samples, it is reasonable to conduct further, more focused, tests of population structure in order seek out racial patterning in craniofacial data. This project, first, evaluated races on a world-wide scale, to determine if broadly based findings, like those reported by the seminal genetic studies on race-informative clusters (c.f., Rosenberg et al., 2002; Bamshad et al., 2003), could be obtained using craniometrics. To this goal, the Howells (1989, 1996) data set was collapsed, from 28 populations, into the three major geographic classes (e.g., Africa, Asia and Europe), while the combined Howells (1898, 1996) and Hanihara (2008) dataset was evaluated in terms of broad geographic regions (like, Micronesia, North Africa, and the United States). When mixture analysis was applied to these two, now redefined, samples, major biogeographic divisions visibly emerged. These clusters largely reflected patterns corresponding to "continental races" or the racial and ethnic divisions adopted as standard in the cultural geographic and census literature (c.f., Bolnick, 2008; United Nations Statistics Division, 2011; U.S. Census Bureau, 2010, 2011). These findings suggest, therefore, that is it possible to detect (hidden and structured) biogeographic information with measureable degrees of statistical certainty and with reasonable clarity.

The ability to reveal this group structure on different levels of analysis (e.g., geographic populations, regions, continents) demonstrates the ease with which craniometric variation can be divided into hierarchical subsets and, in turn, attests to the nested organization or structure of this data. This "nesting" is a point of particular interest when classification by cluster analysis is the analytical goal, that is, for the purpose of ancestry estimation and biosocial race identification, and from the perspective of

[sub]population studies and forensic casework. It brings to the fore the potential concern over how best to define the "true" number of target groups or races, given the ease with which analyses can move between the biogeographic and biosocial levels that exist within this structured hierarchy. To wit, Relethford (2009) has written:

> Given enough markers, one can easily assign individuals to geographically widespread groupings such as Western Europe, East Asia, and sub-Saharan Africa. These analyses are simply another way of expressing the correlation between genetics and geography. There is no justification based on classification accuracy for preferring five, six, seven, or eight groups. From the perspective of racial classification, if we take these different groups as equivalent to races, then we see that the number of races, and the geographic cutoffs used to define them, are subjective decisions (p. 20).

Ousley, Jantz and Freid (2009) have also offered similar remarks: "Worldwide craniometric variation shows strong geographic patterning. However, if biological distinctiveness is an accepted criterion for biological races, a very large number of biological races can be discerned using craniometric data alone" (p. 74). Observations such as these can be taken critically, even as evidence in support of the wholesale abandonment of biological classification and including those socially beneficial projects that contribute positively to the needs of the biomedical and medico-legal communities. Or, they can be seen as temporary impediments that may well be surmounted in practice by good procedural choices (e.g., making the most appropriate use of statistics and skeletal reference/test samples) and with adequate consideration given the environment in which the analysis are being implemented. It is argued here that attending to "context" is critical for any analysis taking up ancestry and race, in order to produce the

most biologically meaningful cluster findings and to achieve the most practically relevant, and useful, classification results (c.f., Konigsberg, Algee-Hewitt & Steadman, 2009).

In order to evaluate how model-based clustering behaves when applied to subpopulations or within-region groups, and to assess the information value that these cluster findings can have in discussions of biosocial race, this project conducted further tests of population structure that applied finite mixture analysis to modern American samples, in the context of forensic identification casework conducted in North America. Craniometric data for five biosocial races (e.g., American Black, White, Hispanic, East Asian, and Native American) was obtained from the Forensic Anthropology Data Bank (Jantz 2008). The analyses of these populations were further supplemented by information obtained from the U.S. Census Bureau (2010, 2011) on the definitions of the major "American" races and their distributions within the current population-at-large. The results of the mixture analyses considering all five or some reduced configuration of these biosocial race groups suggest trends that are comparable to those reported for the world-wide tests. These American analyses found: (1) a close correspondence between the documented race and statistically inferred clusters, (2) distinctive and explainable patterns of overlap, and (3) evidence of population history and interaction, as revealed by the number of clusters produced, the differing degrees of cluster separation, the trends in cluster membership, and by the individual allocations. These findings attest to the presence of latent structure within the American population. They demonstrate, therefore, the ability of mixture modeling to detect this structure, with visible clarity and accompanied by robust statistical output, when only regional variation is considered. They also give credence to the utility of racial identifiers in discussions of craniofacial classification and, more specifically, in the practical identification of "unknown" skeletal remains, in context of medico-legal casework. Furthermore, they simultaneously highlight the challenges of making the most consistently

accurate identifications when composite groups are under analysis, for which a single biosocial identifier often oversimplifies the complex history of population origins, admixture events, and secular change that underlies their morphology. These potential problems are evident, for example, in a subset of the mixture analyses that considered American Black, White and Hispanic samples.

When admixture is considered, all three of these groups are known to have fluctuating levels of shared genetic history; in many of these cases, their expression of variation reflects different parental populations and often a mosaic of sources. Ancestry testing has revealed varying degrees of stratification especially among Black Americans and Hispanics. Although Nei and Livshits (1989) concluded that 20% of the American Black genetic make-up is "White" and Shriver and colleagues (1997) identified United States Hispanics as 60% European (or White), more recent studies have shown that these estimates are heavily influenced by the location of an individual's origin (c.f., Parra et al., 1998; Bryc et al., 2010). While, for example, Bryc et al., (2010) have reported for African Americans that "the median proportion of European ancestry is 18.5% (25th–75th percentiles: 11.6–27.7%)," they also noted that there is "very large variation among individuals." Considerable fluctuations in parental contributions as well as variation in parental sources have also been detected among Hispanics. Approximately 98% of the rural Mexican population of the State of Guerrero has a substantial Amerindian ancestry; at the same time, however, they are cohabiting with a genetically defined sub-population of Mestizos (Bonilla et al, 2005). Mexico City studies also show differing percentages of parental combinations, which have been taken to reflect the diversity of within-city populations and, in turn, the variability of the samples selected: Tiburcio et al. (1978) identified 27.6 % Amerindian, 70.8% European, and 1.4% African parental contributions; Lisker et al. (1988) reported 56.3% Indian, 33 % White, and 10.7% Black genes, specifically, among individuals originating from the state of Puebla; and Cerda-Flores et al., (2002) documented, for their sample of Mestizo populations, parental percentages

of 37-49% Amerindian, 50-60% Spanish-European, and 1-3% African. In addition to the complexity that variation due to admixture and population-overlaps may introduce, American Blacks and Whites are also know to exhibit different morphological trends that are closely tied to their year of birth. Jantz and colleagues have consistently demonstrated "that certain cranial variables exhibit striking change over the past 1.5 centuries in American blacks and whites" (Jantz, 2001, p. 784). These secular trends include modifications to cranial vault height, base length, total length and both facial and vault breadth (Jantz & Meadows Jantz, 2000; Wescott & Jantz, 2005). Of these key dimensions, Jantz (2001) has noted in particular that the distance from basion to bregma is the single measure "showing the most remarkable change" over time (p. 784). Using geometric morphometric techniques, Wescott and Jantz (2005) further refined the source of these effects, by demonstrating that "secular change in American crania is concentrated on the base and posterior aspect of the skull" (p. 234). These shifts in cranial architecture are said "to reflect a major restructuring of the cranial vault, yielding a high, narrower vault in modern people" (Jantz 2004, p. 11). The possible effects of admixture and secular change on the clustering solutions produced here were expressed, most obviously, in two ways: the FDB Hispanic cases were frequently classified as White and the American Black sample was often partitioned into multiple clusters, with some of these having a moderate relationship to time (birth year/cohort). The presence of these confounding factors in modern American data sets and the influence that they may exert over the quality or interpretability of any given mixture clustering solution reiterate the importance of considering the context of the analysis, so that, for example, when standard corrections are being introduced for the well-known issues of size and sex, efforts can be also made to improve model fitting by equally accounting for the potential problems in classification associated with recent changes in group morphology, like, in this case, the effects of birth year and admixture.

The finite mixture model-based clustering methods implemented throughout the course of this project were adopted for two important methodological reasons: (1) finite mixture analysis allows for the detection of cryptic population structure and the allocations of "unknown" individuals into statistically inferred clusters without requiring a reference or comparative data sets of craniometrics from "like" populations, and without the consideration of *a priori* groups identifiers (i.e., without "knowing," for each individual, their true memberships, in categories like race, ancestry, ethnicity, and sex), and (2) this model-based (parametric) clustering approach has the statistical benefits of assuming a well-defined mathematical model, investigating this model using well-established statistical techniques, and supporting findings this model produces by the use of information criteria or tests of significance (Marriott 1974). It is argued here that these mixture-based methods are particularly advantageous for the questions of craniometric clustering and classification because they are free of the concerns for "forced allocations," which has emerged as the predominant complaint given for the use of supervised classification techniques, like discriminant function analysis, in skeletal biology and, especially, forensic anthropology (e.g., Brown & Armelagos, 2001; Armelagos & Van Gerven, 2003; Williams, Belcher & Armelagos, 2005; Albanese & Saunders, 2006; Elliott & Collard, 2009). Moreover, these unsupervised methods can be implemented with reasonable ease; for example, the MCLUST (Fraley & Raftery, 2011) or EMMIX (McLachlan, 2011; McLachlan, Wang & Peel, 2011) packages are designed for use in the free and open-source software environment, R (R Development Core Team, 2010), and they each allow for different methods to be used in model identification. While it has been shown here that the Bayesian Information Criterion selection method, as it is integrated within MCLUST, returned the best models in terms of the interpretability of the cluster solutions from an anthropological perspective, the methods favored by EMMIX produced equally strong statistical results.

In order to corroborate these claims of tractability and utility, as well as the conclusions based upon findings returned by the unsupervised finite mixture clustering methods, the robust methods of the newly developed force-directed techniques as well as the mixture based (partial and fully) supervised classification procedures were employed in validation tests (Aurenhammer, 1991; McLachlan, 1992; Di Battista et al., 1999; McLachlan & Peel, 2000; Kauffmann & Wagner, 2001; Fraley & Raftery, 2006). These *post hoc* methods were also implemented as means to further explore the questions of population [sub]structuring and the (biosocial) reality of race using both approaches and data types that are familiar to biological anthropologists (discriminant function analysis and distance measures) but are being applied in previously untested ways. The use of these methods offered, therefore, the opportunity to investigate this project's test data sets further under objective validation criteria and to introduce additional, potentially more powerful, techniques for user-guided analyses. The results of the mixture discriminant functions tests were compared against the error rates obtained using traditional linear discriminant function analysis and, while classification success was similar in both cases, the mixture approach, in allowing for maximum flexibility in covariance structure, produced models with slightly better fit. These findings suggest that these alternative mixture-based -techniques should be investigated further as they have the potential to improve discriminatory power for future craniometric analyses. The obvious correspondence among the biological distances calculated between groups, the statistically inferred clusters, and the true identification information (e.g., population, race, sex) revealed in the Voronoi diagrams, not only attests to the promise of these force-directed graphing techniques for visualizing complex multi-level relationships, but also clearly substantiates the statistical validity, biological reality, and anthropological meaning of this project’s cluster solutions.

In this light, it is shown here that (1) there is detectable population structure, (2) while nested and hierarchical, this internal patterning can be nevertheless can be resolved into valid craniometric clusters, and that (3) the clusters have observable and testable relationships to the contemporary concepts of populations, biogeographic ancestries and biosocial races. This project offers external justification for the quantification of human craniometric variation using statistical techniques, the estimation of ancestry in forensic contexts and the reporting of race in the aid of the medico-legal community. Moreover, it suggests the possibility of new approaches that may give practical solutions for the, often problematic, analysis of human biological diversity and it invites continued discussions of race in anthropological theory and methods. First, understanding race and human variation and, second, coming to terms with the complex and often conflicted relationship of personal identity and Science are difficult and, justifiably, daunting tasks, especially in the present age of FORDISC, Daubert, and the Human Genome Project. Nevertheless, if anthropology intends to maintain its role, as Marks (1995) has written, as "the science that deals with human biological and cultural variation" (, p. 143), it must reclaim its position of authority. To remain relevant and encourage growth within all areas of the field of anthropology, practitioners must reengage the thoughtful scientific discussion of human biological diversity and revive discourse on the meaning and utility of biosocial race. "A fully functional multiracial society cannot be achieved without a sense of history and open, honest dialogue" (West, 1999).

List of Works Cited

2010 Census National Summary File of Redistricting Data - Technical Documentation. (2010). Retrieved from http://www.census.gov/prod/cen2010/doc/nsfrd.pdf.

Ackermann, R. R., & Cheverud, J. M. (2004). Detecting genetic drift versus selection in human evolution. Proceedings of the National Academy of Science, 52(17), 17,946-917,951.

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2004). Geometric morphometrics: Ten years of progress following the 'revolution'. Italian Journal of Zoology, 71, 5-16.

Aiello, L. C. . (1992). Allometry and the analysis of size and shape in human evolution. Journal of Human Evolution, 22(2), 127-147.

Aitkin, M., & Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. Journal of the Royal Statistical Society (Series B), 47, 67-75.

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. Journal of the Royal Statistical Society (Series A)(144), 419-461.

Akaike, H. (1973). Information Theory as an Extension of the Maximum Likelihood Principle. Paper presented at the Second International Symposium on Information Theory, Budapest.

Akaike, H. (1974). A new look at statistical model identification. IEEE Transactions on Automatic Control, AC - 19(6), 716-723.

Akaike, H. (1981). Likelihood of a model and information criteria. Journal of Econometrics, 16, 3-14.

Akaike, H. (1983). Information measures and model selection. Paper presented at the Proceedings of the International Statistical Institute 44th Session, The Hague.

Albanese, J., & Saunders, S. (2006). Is it Possible to Escape Racial Typology in Forensic Identification. In A. Schmitt, E. Cunha & J. Pinheiro (Eds.), Forensic Anthropology and Medicine. Complementary Sciences from Recovery to Cause of Death (pp. 281-316 (ch.212)). Totowa, NJ: Humana Press.

Albert, V., Jonsson, B., & Bernatchez, L. (2006). Natural hybrids in Atlantic eels (Anguilla anguilla, A. rostrata): evidence for successful reproduction and fluctuating abundance in space and time. Molecular Ecology, 15, 1903-1916.

Albrecht, G. H. (1992). Assessing the affinities of fossils using canonical variates and generalized distances. Human Evolution, 7, 49-69.

Alexander, R. (1979). Darwinism and Human Affairs. Seattle: University of Washington Press.

Algee-Hewitt, B., & Carson, E. A. (2009). The quantitative genetics of frontal curvature: evolutionary implications. Paper presented at the Annual Meeting of the American Association of Anthropological Genetics, Chicago, IL.

Algee-Hewitt, B., Wilson, R. J., & Meadows Jantz, L. (2007). The Donation Dilemma: Academic Ethics and Public Participation at the Anthropological Research Facility. Paper presented at the Annual Meeting of the American Academy of Forensic Sciences, San Antonio, TX.

Allocco, D. J., Song, Q., Gibbons, G. H., Ramoni, M. F., & Kohane, I. S. (2007). Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms. BMC Genomics, 8(68).

Amadon, D. (1949). The seventy-five percent rule for subspecies. The Condor, 51, 250-258.

Ancestry.com. (1997-2011). Retrieved August 20, 2011 www.ancestry.com

Anderberg, M. R. (1973). Cluster Analysis for Applications.

Anderberg, M. R. (1973). Cluster Analysis for Applications. New York: Academic Press.

Anderson, E. C., & Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. Genetics, 160, 1217-1229.

Anderson, M. R., & Moscou, S. (1998). Race and ethnicity in research on infant mortality. Family Medicine Journal, 30, 224-227.

Anderson, T. W. (1996). R.A. Fisher and multivariate analysis. Statistical Science, 11(1), 20-34.

Andreasen, R. O. (1998). A new perspective on the race debate. British Journal for the Philosophy of Science, 49, 199-225.

Andreasen, R. O. (2004). The cladistic race concept: a defense. Biology and Philosophy, 19, 425-442.

Andreasen, R. O. (2005). The meaning of 'race': folk conceptions and the new biology of race. Journal of Philosophy, 102, 94-106.

Andreasen, R. O. (2008). The concept of race in medicine. In M. Ruse (Ed.), The Oxford handbook of philosophy of biology (pp. 478-504). Oxford: Oxford University Press.

Angel, J. (1976). Colonial to modern skeletal change in the United States. American Journal of Physical Anthropology, 45, 723-736.

Angel, J. (1982). A new measure of growth efficiency-skull base height. American Journal of Physical Anthropology, 58, 723-736.

Annual Estimates of the Resident Population by Sex, Race, and Hispanic Origin for the United States: April 1, 2000 to July 1, 2009 (2009). (NC-EST2009-03). Retrieved from http://www.census.gov/popest/national/asrh/NC-EST2009-srh.html

Anthropologists, American Association of Physical. (1996). AAPA Statement on Biological Aspects of Race. American Journal of Physical Anthropology, 101, 569-570.

Anthropologists, American Association of Physical. (1999). Physical Anthropology statement on the biological aspects of race. American Anthropology, 100, 714-715.

Anthropology, American Board of Forensic. (2010). ABFA Home Page Retrieved June 10, 2010, from www.theabfa.org

Appiah, K. A. (1990). Racisms. In D. T. Goldberg (Ed.), Anatomy of racism. (pp. 3-17). Minneapolis: University of Minnesota Press.

Appiah, K. A. (1996). Race, culture, identity: misunderstood connections. In G. B. Peterson (Ed.), The Tanner lectures on human values (Vol. 17, pp. 51-136). Salt Lake City: The University of Utah Press.

Appiah, K. A. (1999). Why there are no races. In L. Harris (Ed.), Racism (pp. 267-280). Amherst: Humanity Press.

Arabie, P., Hubert, L. J., & De Soete, G. (Eds.). (1996). Clustering and Classification. Singapore: World Scientific Publishing.

Arendt, H. (1963). Eichmann in Jerusalem: A Report on the Banality of Evil. New York, NY: Viking.

Armelagos, G. J., & Goodman, A. H. (1998). Race, Racism and Anthropology. In A. H. Goodman & T. L. Leatherman (Eds.), Building in New Biocultural System: Political-Economic Perspective on Human Biology (pp. 359-377). Ann Arbor: University of Michigan Press.

Armelagos, G. J., Carlson, D. S., & Van Gerven, D. P. (1982). The Theoretical Foundations and Development of Skeletal Biology. In F. Spencer (Ed.), A History of Physical Anthropology, 1930-1980 (pp. 305-328). New York, NY: Academic Press.

Armelagos, George J., & Van Gerven, Dennis. (2003). A century of skeletal biology and paleopathology: Contrasts, contradictions, and conflicts. American Anthropologist, 105(1), 53-64.

Armstrong, G. L., Conn, L. A., & Pinner, R. W. (1999). Trends in infectious disease mortality in the United States during the 20th century. Journal of the American Medical Association, 6, 61-66.

Aspinall, P. J. (1998). Describing the "white"ethnic group and its composition in medical research. Social Science and Medicine, 47, 1797-1808.

Assakul, K., & Proctor, C. H. (1967). Testing independence in two-way contingency tables with data subject to misclassification. Psychometrika, 32, 67-76.

Assembly, United Nations General. (1960). Declaration on the Granting of Independence to Colonial Countries and Peoples. (resolution 1514 (XV)). Retrieved from http://www.un.org/Depts/dpi/decolonization/declaration.htm.

Association, American Anthropological. (1994). American Anthropological Association Statement on "Race" and Intelligence Retrieved Jan 3, 2011, from http://www.aaanet.org/stmts/race.htm

Association, American Anthropological. (1997). American Anthropological Association Response to OMB Directive 15: Race and Ethnic Standards for Federal Statistics and Administrative Reporting Retrieved August 18, 2010, from http://www.aaanet.org/gvt/ombdraft.htm

Association, American Anthropological. (1999 (adopted 1998)). American Anthropological Association Statement on "Race". American Anthropologist, 100(3), 712-713 (also available online at http://www.aaanet.org/stmts/racepp.htm).

Association, American Anthropological. (1999). American Anthropological Association Statement on Race. American Anthropology, 100, 712-713.

Association, American Anthropological. (2007). Science: 1890s-1930s. Eugenics and Physical Anthropology. Retrieved September 23, 2010, from http://www.understandingrace.org/history/science/eugenics_physical.html

Association, American Sociological. (2003). The importance of collecting data and doing social scientific research on race Retrieved January 3, 2011, from http://www2.asanet.org/media/asa_race_statement.pdf

Atran, S. (1994). Core domains versus scientific theories: Evidence from systematics and ItzaMaya folk biology. In L. A. Hirschfeld & S. A. Gelman (Eds.), Mapping the Mind: Domain Specificity in Cognition and Culture (pp. 316-340). New York: Cambridge University Press.

Auber, D. (2007). Tulip: Data Visualization Software (Version 3.5.0): GNU lesser general public license. Retrieved from http://tulip.labri.fr/TulipDrupal/

Aurenhammer, F. (1991). Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure. ACM Computing Surveys, 345-405.

Bacon, P. Jr. (2007, November 30). Rock Solid Support For Obama, But Will the Black Vote Follow? , The Washington Post. Retrieved from http://voices.washingtonpost.com/44/2007/11/rock-solid-support-for-obama-b.html

Baek, J., McLachlan, G. J., & Flack, L. (2010). Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32, 1298-1309.

Baker, L. D. (1994). Location of Franz Boas within the African-American Struggle. Critique of Anthropology, 14(2), 199-217.

Baker, L. D. (1998). From Savage to Nero: Anthropology and the Construction of Race, 1896-1954. Berkeley: University of California Press.

Bamshad, M. (2005). Genetic influences on health: Does race matter? . Journal of the American Medical Association, 294, 237-346.

Bamshad, M. J., Wooding, S., Watkins, W. S., Ostler, C. T., Batzer, M. A., & Jorde, L. B. (2003). Human population genetic structure and inference of group membership. American Journal of Human Genetics, 72, 578-589.

Bamshad, M., & Wooding, S. P. (2003). Signature of natural selection in the human genome. National Review of Genetics, 4, 99-111.

Bamshad, M., Kivisild, T., Watkins, S., Dixon, M., Ricker, C., Rao, B. B., . . . et al. (2001). Genetic Evidence on the Origins of Indian Caste Populations. Genome Research, 11, 994-1004.

Bamshad, M., Wooding, S., Salisbury, B. A., & Stephens, J. C. (2004). Deconstructing the relationship between genetics and race. Nature Reviews Genetics, 5, 598-609.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics, 49, 803-821.

Banks, D., House, L., McMorris, F. R., Arabie, P., & Gaul, W. (Eds.). (2004). Classification, Clustering, and Data Mining Applications. Berlin: Springer.

Banton, M. (1998). Racial Theories (Vol. 2nd). Cambridge: Cambridge University Press.

Baraybar, J. P., & Gasior, M. (2005). Forensic Anthropology and the Most Probable Cause of Death in Cases of Violations Against International Humanitarian Law: An Example from Bosnia and Herzegovina. Journal of Forensic Science, 51, 103-108.

Barbará, D., Couto, J., & Li, Y. (2002). COOLCAT: an entropy-based algorithm for categorical clustering. Paper presented at the Proceedings of the eleventh international conference on Information and knowledge management. .

Barbujani, G. (2005). Human races: Classifying people vs. understanding diversity. Current Genomics 6, 215-226.

Barbujani, G., Magagni, A., Minch, C., & Cavalli-Sforza, L. L. (1997). An apportionment of human DNA diversity. Proceedings of the National Academy of Sciences of the United States of America, 94, 4516-4519.

Barik, S.S., Sahani, R., Prasad, B.V., Endicott, P., Metspalu, M., Sarkar, B.N., . . . Rao, V.R. (2008). Detailed mtDNA genotypes permit a reassessment of the settlement and population structure of the Andaman Islands. American Journal of Physical Anthropology, 136, 19-27.

Bartholomew, D. J. (1980). Factor analysis for categorical data (with discussion),. Journal of the Royal Statistical Society Series B, 42(3), 293-321.

Bartholomew, D. J. (1984). The foundations of factor analysis. Biometrika, 71(2), 221-232.

Basford, K. E., & McLachlan, G. J. (1985). Estimation of allocation rates in a cluster analysis context. Journal of the American Statistical Association, 80, 286-293.

Basford, K. E., & McLachlan, G. J. (1985). Likelihood estimation with normal mixture models. Journal of Applied Statistical Science, 34, 282-289.

Basford, K. E., & McLachlan, G. J. (1985). The mixture method of clustering applied to three-way data. Journal of Classification, 2, 109-125.

Bass, J. (2006-2010). New York: Harper Collins.

Bass, W. M. (1995). Human Osteology: A Laboratory and Field Manual (Vol. 2). Springfield, MO: Missouri Archaeological Society.

Bass, W. M. (2005). Human Osteology: A Laboratory and Field Manual (Vol. 5th). Columbia, MO: Missouri Archeological Society.

Bass, W. M., & Jefferson, J. (2003). Death's Acre. New York: Putnam.

Bass, W. M., & Jefferson, J. (2007). Beyond the Body Farm. New York: Harper Collins.

Basu, S., Davidson, I., & Wagstaff, K. (2009). Constrained Clustering: Algorithms, Applications and Theory. Boca Raton: CRC Press.

Baudry, J. P. , Raftery, A. E., Celeux, G. , Lo, K., & Gottardo, R. . (2010). Combining Mixture Components for Clustering. Journal of Computational and Graphical Statistics, 9(2), 332-353.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics, 41(1), 164-171.

Baum, L., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. Annals of Mathematical Statistics, 37, 1554-1563.

Bay, S. D., & Pazzani, M. J. (1999). Detecting Change in Categorical Data: Mining Contrast Sets. Paper presented at the fifth ACM SIGKDD international conference on knowledge discovery and data mining, New York.

Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. Journal of the Royal Statistical Society (Series B), 37(1), 129-145.

Beals, K. L., Smith, C. L., & Dodd, S. M. (1993). Climate and the evolution of brachycephalization. American Journal of Physical Anthropology, 62, 425-437.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). The New S Language. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Beerli, P., & Felsenstein, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics, 152, 763-773.

Behboodian, J. (1970). On a mixture of normal distributions. Biometrika 57, 215-217.

Belcher, R., Williams, F., & Armelagos, G. J. (2002). Misidentification of Meroitic Nubians using Fordisc 2.0. American Journal of Physical Anthropology, 42, S. 34.

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software, 32(6), 1-29.

Benn-Torres, J., Bonilla, C. , Robbins, C. M., Waterman, L., Moses, T. Y., et al. (2008). Admixture and population stratification in African Caribbean populations. American Journal of Human Genetics, 72, 90-98.

Benson, D. A, Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2008). Nucleic Acids Res. 2008 from D25-30. Epub 2007 Dec 11

Berg, K., Bonham, V., Boyer, J., Brody, L., Brooks, L., Colins, F., . . . Warshauer-Baker, E. (2005). The Use of Racial, Ethnic, and Ancestral Categories in Human Genetics Research. American Journal of Human Genetics, 77, 517-532.

Bergstrom, B. (2002). Study Challenges Longtime View That Genetics Don't Shape Skulls: Associated Press.

Berlin, B., Breedlove, D. E., & Ravern, P. H. (1973). General principles of classification and nomenclature in folk biology. American Anthropologist, 75, 214-242.

Berryman, H. E. (1984). The Averbuch Skeletal Series: A Study of Biological and Social Stress at the Averbuch Site. In W. E. Klippel & W. M. Bass (Eds.), Averbuch: A Late Mississippian Manifestation in the Nashville Basin (Vol. 1). Washington, D. C.: National Park Service, Department of Interior.

Betti, L., Balloux, F., Amos, W., Hanihara, T., & Manica, A. (2009). Distance from Africa, not climate, explains within-population phenotypic diversity in humans. Proceedings of the Royal Society of London Series B, 276, 809-814.

Betti, L., Balloux, F., Hanihara, T., & Manica, A. (2010). The relative role of drift and selection in shaping the human skull. American Journal of Physical Anthropology, 141, 76-82.

Bhattacharya, C. G. (1967). A simple method of resolution of a distribution into Gaussian components. Biometrics 23, 115-135.

Biernacki, C., & Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. Computing Science and Statistics, 29, 451-457.

Biernacki, C., Celeux, G. , & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 719-725.

Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Statistics & Data Analysis, 41(3-4), 561-575.

Binford, L. (1968). Archeological perspectives. In S. Binford & L. Binford (Eds.), New Perspectives in Archeology (pp. 5-32). Chicago: Aldine.

BIOSIS. (2010). Retrieved August 19, 2010, from Thompson-Reuters http://thomsonreuters.com/products_services/science/science_products/a-z/biosis

Birdsell, J. (1987). Some reflections on fifty years in biological anthropology. Annual Reviews of Anthropology, 16(1), 1-12.

Birdsell, J.B. (1972). Human Evolution: An Introduction to the New Physical Anthropology. Chicago: Rand McNally.

Birkby, W. H. (1996). An evaluation of race and sex identification from cranial measurements. American Journal of Physical Anthropology, 24(1), 21-28.

Blackith, R., & Reyment, R. A. (1971). Multivariate morphometrics. New York: Academic Press.

Blakey, M. (1987). Skull doctors: intrinsic social and political bias in the history of American physical anthropology; with special reference to the work of Ales Hrdlicka. Critical Anthropology, 7, 7-35.

Blakey, M. L. (1997). Skull doctors revisited: intrinsic social and political bias in the history of American physical anthropology with special reference to the work of Alec Hrdlicka. In L. T. Reynolds & L. Lieberman (Eds.), Race and Other Misadventures: Essays in Honor of Ashley Montagu in His Ninetieth Year (pp. 64-95). Dix Hills, NY: General Hall.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. Psychological Bulletin, 83, 377-388.

Block, N. J. (1995). How heritability misleads about race. Cognition, 56, 99-128.

Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97(1-2), 245-271.

Blumenbach, J. F. (1969). On the natural varieties of mankind (T. Bendyshe, Trans.). New York: Bermann.

Blumenbach, J. F. (2005). On the natural variety of mankind. Boston: Adamant Media Corporation.

Boas, F. (1891). Physical characteristics of the Indians of the North Pacific coast. American Anthropologist, 4, 25-32.

Boas, F. (1894). Human faculty as determined by race. Proceedings of the American Association for the Advancement of Science, 43, 301-327.

Boas, F. (1894). The half-blood Indian, an anthropometric study. Popular Science Monthly, 45, 761-769.

Boas, F. (1899). The Cephalic Index. American Anthropologist, 1(3), 448-461.

Boas, F. (1903). Heredity in Head Form. American Anthropologist, 5(3), 530-538.

Boas, F. (1910). Changes in Bodily Form of Descendants of Immigrants United States Immigration Commission, Senate Document No. 208, 61st Congress. Washington DC: Government Printing Office.

Boas, F. (1911). The Mind of Primitive Man. New York, NY: The Macmillan Company.

Boas, F. (1912). Changes in Bodily Form of Descendants of Immigrants. New York, NY: Columbia University Press.

Boas, F. (1912). Changes in the Bodily Form of Descendants of Immigrants. American Anthropologist, 14, 530-562.

Boas, F. (1918). Notes on the anthropology of Sweden. American Journal of Physical Anthropology, 1, 415-426.

Boas, F. (1928). Materials for the Study of Inheritance in Man Columbia University Contributions to Anthropology (Vol. VI). New York, NY: Columbia University Press.

Boas, F. (1932). Studies in Growth: I. Human Biology, 4(307-350).

Boas, F. (1933). Studies in Growth: II. Human Biology, 5, 429-444.

Boas, F. (1935). Studies in Growth: III. Human Biology, 7(303-318).

Boas, F. (1938). The Mind of Primitive Man (Vol. Revised edition). New York, NY: The Macmillan Company.

Boas, F. (1940). Changes in bodily form of descendants of immigrants. In F. Boas (Ed.), Language and Culture (pp. 60-75). New York: The Free Press.

Boas, F. (1940). Race, Language, and Culture. New York, NY: Macmillan.

Bock, H. H. (1994). Classification and clustering: problems for the future. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand & B. Burtschy (Eds.), New Approaches in Classification and Data Analysis (pp. 3-24). Berlin: Springer-Verlag.

Bock, H. H. (1994). Information and entropy in cluster analysis. In H. Bozdogan (Ed.), Proceedings of the First US-Japan Conference on the Frontiers of Statistical Modeling (pp. 115-147). Deventer: Kluwer.

Bodmer, W. F., & Cavalli-Sforza, L. L. (1976). Genetics, evolution and man. San Francisco: W. H. Freeman.

Böhning, D. (1999). Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and Others. Boca Raton: Chapman & Hall / CRC.

Boldsen, J. L. (2000). Human demographic evolution-Is it possible to forget about Darwin? Odense: Danish Center for Demographic Research.

Bolnick, D. A. (2008). Individual ancestry inference and the reification of race as a biological phenomenon In B Koenig, S. Lee & S Richardson (Eds.), Revisiting race in a genomic age (pp. 70-85). New Brunswick: Rutgers University Press.

Bolnick, D., Fullwiley, D., Duster, T., Cooper, R., Fujimura, J., Kahn, J., . . . TallBear, K. (2007). The Science and Business of Genetic Ancestry Testing. Science, 318, 399-400.

Bonilla, C., Gutiérrez, G., Parra, E. J., Kline, C., & Shriver, M. D. (2005). Admixture analysis of a rural population of the state of Guerrero, Mexico. American Journal of Physical Anthropology, 128, 861-869.

Bonilla, C., Parra, E. J., Pfaff, C. L., Dios, S., Marshall, J. A., Hamman, R. F., . . . Shriver, M. D. (2004). Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. Annals of Human Genetics, 68, 139-153.

Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions. Statistical Science, 1, 181-242.

Bookstein, F. L. (1997). Morphometric tools for landmark data: geometry and biology. Cambridge: Cambridge University Press.

Bookstein, F. L., Gunz, P., Mitteroecker, P., Prossinger, H., Schaefer, K., & Seidler, H. (2003). Cranial integration in Homo: singular warps analysis of the midsagittal plane in ontogeny and evolution. Journal of Human Evolution, 44, 167-187.

Bookstein, F. L., Schafer, K., Prossinger, H., Seidler, H., Fieder, M., Stringer, C. B., . . . Marcus, L. F. (1999). Comparing frontal cranial profiles in archaic and modern Homo by morphometric analysis. Anatomical Record, 257, 217-224.

Border Death Database. (2011). Retrieved March 24, 2011, from Arizona Daily Star

Borg, I., & Gorenen, P. (1997). Modern Multidimensional Scaling. Theory and Applications. New York: Springer.

Bouguessa, M., Wang, S., & Sun, H. (2006). An objective approach to cluster validation. Pattern Recognition Letters, 27, 1419-1430.

Bowcock, A. M., Kidd, J. R., & Mountain, J. (1991). Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. Proceedings of the National Academy of Sciences of the United States of America, 88, 839-843.

Bowcock, A. M., Ruiz-Linares, A., & Tomfohrde, J. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. Nature, 368, 455-457.

Bowler, J. M., Johnson, H., Olley, J. M., Prescott, J. R., Roberts, R. G., Shawcross, W., & Spooner, N. A. (2003). New ages of human occupation and climatic change at Lake Mungo, Australia. Nature, 421(837-840).

Box, George E. P., & Draper, Norman. (1987). Empirical Model-Building and Response Surfaces. New York: Wiley.

Boyd, R., & Silk, J. (2009). How humans evolved (5th ed.). New York: W. W. Norton.

Boyd, W. C. (1950). Genetics and Races of Man: Blackwell.

Boyles, R. A. (1983). On the convergence of the EM algorithm. Journal of the Royal Statistical Society Series B, 45(1), 47-50.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52(3), 345-370.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse-Fisher information matrix. In O. Opitz, B. Lausen & R. Klar (Eds.), Information and Classification, Concepts, Methods and Applications (pp. 40-54). Berlin: Springer-Verlag.

Bozdogan, H. (1994). Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Information Measure of Complexity. In H. Bozdogan (Ed.), Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach (Vol. 2, pp. 69-113). Boston: Kluwer Academic.

Brace, C. L. (1964). A Nonracial Approach towards the Understanding of Human Diversity. In A. Montagu (Ed.), The Concept of Race (pp. 103-152). New York, NY: Free Press.

Brace, C. L. (1964). On The Race Concept. Current Anthropology, 5(4), 313-318+319-320.

Brace, C. L. (1982). The roots of the race concept in American physical anthropology. In F. Spencer (Ed.), A History of American Physical Anthropology (pp. 11-29). New York: Academic Press.

Brace, C. L. (1995). Region does not mean "race"-- Reality versus convention in forensic anthropology. Journal of Forensic Sciences, 40, 171-175.

Brace, C. L. (2000). Does race exist? A antagonist's perspective. Nova Retrieved January 3, 2011, from http://www.pbs.org/wgbh/nova/first/brace.html

Brace, C. L. (2005). "Race" Is a Four-Letter Word. The Genesis of the Concept. New York: Oxford UP.

Brace, C. L., & Hunt, K. D. (1990). A nonracial craniofacial perspective on human variation: A(ustralia) to Z(uni). American Journal of Physical Anthropology, 82, 341-360.

Brace, C. L., Rosenberg, K. R., & D., Hunt. K. (1987). Gradual change in human tooth size in the late Pleistocene and post- Pleistocene. Evolution, 41, 705-720.

Brace, C.L., Brace, M.L., & Leonard, W.L. (1989). Reflections on the face of Japan: a multivariate craniofacial and odontometric perspective. American Journal of Physical Anthropology, 78, 93-113.

Brame, R., Mulvey, E. P., & Piquero, A. R. (2001). On the development of different kinds of criminal activity. Sociological Methods and Research, 29(3), 319-341.

Brant, L. J., Morrell, C. H., & Pearson, D. J. (1990). Mixture models for investigating complex distributions. American Journal of Physical Anthropology, 81, 198.

Brattain, M. (2007). Race, Racism, and Anti-Racism: UNESCO and the Politics of Presenting Race to the Postwar Public. American Historical Review, 112, 1386-1413.

Brauer, G. (1992). The origins of modern Asians: By regional evolution or by replacement? In T. Akazawa, K. Aoki & T. Kimura (Eds.), The Evolution and Dispersal of Modern Humans in Asia (pp. 401-413). Tokyo: Sokusen-Sha.

Brauer, G., Collard, M. , & Stringer, C. . (2004). On the reliability of recent tests of the Out of Africa hypothesis for modern human origins. Anatomical Record, A279, 701-707.

Braun, L. (2006). Reifying human differences: the debate on genetics, race, and health. International Journal of Health Services, 36, 557-573.

Braun, L., Fausto-Sterline, A., Fullwiley, D., Hammonds, E. M., Nelson, A. , Quivers, W., . . . Shields, A. E. (2007). Racial Categories in Medical Practice: How Useful Are They? PLoS Medicine, 4(9), 1423-1428.

Breckenridge, J. M. (2000). Validating cluster analysis: Consistent replication and symmetry. Multivariate Behavioral Research, 35, 261-285.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks / Cole.

Brenda, N. (1927). Stoessiger: A Study Of The Badarian Crania Recently Excavated By The British School Of Archaeology In Egypt Biometrika, 19, 110-150.

Brian, P. (2011). Alas, Poor Yorick ... or Is It Othello? Can you tell a person's race from his or her skull? Retrieved January 19, 2011, from http://www.slate.com/id/2279942/

Brodkin, K. (1998). How Jews Became White Folks and What That Says About Race in America. New Brunswick, NJ and London: Rutgers University Press.

Brooks, S. P., & Morgan, B. J. T. (1995). Optimization using simulated annealing. The Statistician, 44(2), 241-257.

Bross, I. (1954). Misclassification in 2x2 tables. Biometrics, 10, 479-486.

Brown v. Board of Education, No. 347 US (1954).

Brown, P. (1989). Coobool Creek (Vol. 13). Canberra: The Australian National University.

Brown, P. (1992). Post-Pleistocene change in Australian Aboriginal tooth size: dental reduction or relative expansion? In T Brown & S Molnar (Eds.), Human Craniofacial Variation in Pacific Populations (pp. 33-52). Adelaide: Anthropology and Genetics Laboratory, University of Adelaide.

Brown, P. (1992). Recent human evolution in East Asia and Australasia. Philosophical Transactions of the Royal Society of London, Series B, 337, 235-242.

Brown, R. A., & Armelagos, G. J. (2001). Apportionment of racial diversity: A review. Evolutionary Anthropology, 10, 34-40.

Brues, A. M. (1972). Models of race and cline. American Journal of Physical Anthropology, 37, 389-399.

Brues, A. M. (1990). The once and future diagnosis of race. In G. Gill & S. Rhine (Eds.), Skeletal Attribution of race: methods for forensic anthropology. Maxwell Museum of Anthropological Papers No. 4 (pp. 1-7). Albuquerque, NM: University of New Mexico.

Brues, A. M. (2009). The Objective View of Race. In C. C. Gordon (Ed.), Race, Ethnicity, and Applied Bioanthropology (pp. 74-78). Chichester: John Wiley.

Bruner, E., Manzi, G., & Arsuaga, J. L. (2003). Encephalization and allometric trajectories in the genus Homo: Evidence from the Neandertal and modern lineages. Proceedings of the National Academy of Sciences of the United States of America 100, 15335-15340.

Bruner, E., Saracino, B., Ricci, F., Tafuri, M., Passarello, P., & Manzi, G. (2004). Midsagittal cranial shape variation in the genus Homo by geometric morphometrics. Collegium Anthropologicum 28, 99-112.

Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., & Froment, A. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proceedings of the National Academy of Sciences of the United States of America, 107(2), 786-791.

Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., . . . Ostrer, H. (2010). In the Light of Evolution IV: The Human Condition Sackler Colloquium: Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proceedings of the National Academy of Science, 107, 8954-8961.

Buikstra, J. E., & Beck, L. A. (2006). Bioarchaeology: the contextual analysis of human remains. Boston: Academic Press.

Buikstra, J. E., Frankenberg, S. R., & Konigsberg, L. W. (1990). Skeletal biological distance studies in American Physical Anthropology: Recent trends. American Journal of Physical Anthropology, 82, 1-7.

Bulbeck, D. (2005). Australian Aboriginal craniometrics as construed through FORDISC Retrieved February 06, 2011, from http://arts.anu.edu.au/bullda/oz_craniometrics.html

Bulbeck, D., Rayner, D., Groves, C., & Raghavan, P. (2003). The contribution of South Asia to the Peopling of Australasia" and the relevance of Basel's Naturhistorisch Museum anthropological collection to the project aims. Bulletin der Schweizerische Gesellschaft fur Anthropologie 9, 49-70.

Burchard, E. G., Ziv, E., Coyle, N., Gomez, S. L., Tang, H., Karter, A. J., . . . Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. New England Journal of Medicine, 348(12), 1170-1175.

Buretić-Tomljanović, A., Ostojić, S., & Kapović, M . (2006). Secular change of craniofacial measures in Croatian younger adults. American Journal of Human Biology, 18, 668-675.

Buretić-Tomljanović, A., Ristić, S., Brajenović-Milić, B., Ostojić, S., Gombač, E. , & Kapović, M. (2004). Secular change in body height and cephalic index of Croatian medical students (University of Rijeka). American Journal of Physical Anthropology, 123, 91-96.

Burnham, K. B., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods & Research 33, 261-305.

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer.

Burns, K. R. (1999). Forensic anthropology training manual. Upper Saddle River: Prentice-Hall.

Butler, R. W. (1986). Predictive Likelihood inference with applications (with discussion). Journal of the Royal Statistical Society Series B, 48, 1-38.

Buyske, S. (2003). R Package mmlcr: Mixed-Mode Latent Class Regression (Version 1.3.2). Retrieved from http://www.stat.rutgers.edu/~buyske/software.html

Byers, S. (2002 [2005]). Introduction to forensic anthropology. Boston: Allyn and Bacon.

Cadez, I, Heckerman, D, Meek, C, Smyth, P, & Whyte, S. (2003). Model-based clustering and visualization of navigation patterns on a web site. Data Mining and Knowledge Discovery, 7(4), 399-424.

Calcagno, J. M. (1986). Dental reduction in post-Pleistocene Nubia. American Journal of Physical Anthropology, 70, 349-363.

Calcagno, J. M., & Gibson, K. R. (1988). Human dental reduction: natural selection or the probable mutation effect. American Journal of Physical Anthropology, 77, 505-517.

Cameron, N., Tobias, P. V., Fraser, W. J., & Nagdee, M. (1990). Search for secular trends in calvarial diameters, cranial base height, indices and capacity in South African Negro crania. American Journal of Human Biology, 2, 53-61.

Camic, C., & Xie, Y. (1994). The Statistical Turn in American Social Science: Columbia University, 1890 to 1915. American Sociological Review, 59, 773-805.

Campbell, A. R., & Armelagos, G. J. (2007). Assessment of FORDISC 3.0's accuracy in classifying individuals from WW Howell's populations and the forensic data bank (Abstract). American Journal of Physical Anthropology 132(Suppl. 44), P83-P84.

Campbell, J. G., Fraley, C., Murtagh, F., & Raftery, A. E. (1998). Linear flaw detection in woven textiles using model-based clustering. Pattern Recognition Letters, 18, 1539-1548.

Campbell, M. C., & Tishkoff, S. A. (2010). The evolution of human genetic and phenotypic variation in Africa. Current Biology, 20(R166-R173).

Cann, R.L., Stoneking, M. , & Wilson, A.C. . (1987). Mitochondrial DNA and human evolution. Nature, 329, 31-36.

Canty, A., & Ripley, B. (2009). boot: Bootstrap R (S-Plus) Functions (Version 1.2-41).

Cardon, L. R. , & Palmer, L. J. . (2003). Population stratification and spurious allelic association. The Lancet, 361, 598-604.

Carey, J. W., & Steegmann, A. T. (1981). Human nasal protrusion, latitude, and climate. American Journal of Physical Anthropology, 56(313-319).

Carlson, D. S., & Van Gervan, D. P. (1977). Masticatory function and post-Pleistocene evolution in Nubia. American Journal of Physical Anthropology, 46, 495-506.

Carson, E. A. (2006). Genetic And Environmental Components Of Human Microevolution In Pleistocene And Holocene Australia. Ph. D., The University of New Mexico, Albuquerque.

Carson, E. A. (2006). Maximum likelihood estimation of human craniometric heritability. American Journal of Physical Anthropology, 13, 169-180.

Cartmill, M. (1998). The Status of the Race Concept in Physical Anthropology. American Anthropologist, 100, 651-660.

Cartmill, M., & Brown, K. (2003). Surveying the race concept: a reply to Lieberman, Kirk, and Littlefield. American Anthropologist, 105, 114-115.

Caspari, R. (2003). From Types to Populations: A Century of Race, Physical Anthropology and the American Anthropological Association. American Anthropologist, 105(1), 63-74.

Caspari, R. (2009). 1918: Three Perspectives on race and human variation. American Journal of Physical Anthropology, 139, 5-15.

Cavalli-Sforza , L.L. (1997). Genes, Peoples, and Languages. Proceedings of the National Academy of Sciences of the United States of America 94(15), 7719–7724.

Cavalli-Sforza, L. L. (1966). Population Structure and Human Evolution. Proceedings of the Royal Society Series B-Biological Sciences, 164(995), 362-379.

Cavalli-Sforza, L. L. (2005). The Human Genome Diversity Project: past, present and future. Nature Reviews, 6, 333-340.

Cavalli-Sforza, L. L., & Edwards, A. W. F. (1967). Phylogenetic Analysis Models and Estimation Procedures. American Journal of Human Genetics, 19(3P1), 233-257.

Cavalli-Sforza, L. L., & Feldman, M. W. . (2003). The application of molecular genetic approaches to the study of human evolution. Nature Genetics, 33, 233-275.

Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1994). The History and Geography of Human Genes Princeton, NJ: Princeton University Press.

Celeux, G. (1987). Reconnaissance de Mélanges de Densités de Probabilité et Applications à la Validation des résultats en Classification. Ph.D., L'Université de Paris IX Dauphine, Paris.

Celeux, G., & Diebolt. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly, v2, 73-82.

Celeux, G., & Govaert, G. (1992). A Classification EM algorithm for clustering and two stochastic versions. Computational Statistics & Data Analysis, 14(3), 315-332.

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern Recognition, 28(5), 781-793.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification, 13, 195-212.

Census Bureau Estimates Nearly Half of Children Under Age 5 are Minorities. Estimates find nation's population growing older, more diverse (press release). (2009). (cb09-75). Retrieved from http://www.census.gov/newsroom/releases/archives/population/cb09-75.html

Census National Summary File of Redistricting Data—Technical Documentation. (2010). Retrieved from http://www.census.gov/prod/cen2010/doc/nsfrd.pdf

Cerda-Flores, R. M., Villalobos-Torres, M. C., Barrera-Saldaña, H. A., Cortés-Prieto, L. M., Barajas, L. O., Rivas, F., . . . Chakraborty, R. . (2002). Genetic admixture in three Mexican mestizo populations based on D1S80 and HLA-DQA1 Loci. American Journal of Human Biology, 14, 257-263.

Chakraborty, R. (1978). Single-Locus and Multilocus Analysis of Genetic Differentiation of the Races of Man: A Critique. The American Naturalist, 112(988), 1134-1138.

Chakraborty, R. (1984). Relationship between heterozygosity and genetic distance in the three major races of man. American Journal of Physical Anthropology, 65(3), 249-258.

Charlier, C. V. L. . (1906). Researchers into the theory of probability. Lunds Universities Arskrift, (Ny folid 2.1 No. 5). Hakon Ohlsson, Lund.

Charlier, C. V. L., & Wicksell, S. D. . (1924). On the dissection of frequency functions. Arkiv for Mathematik Astronomi och Fysik, (BD. 18, No. 6). Lund.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society Series A, 158, 419-466.

Cheeseman, P., & Stutz, J. . (1996). Bayesian classification (AutoClass): theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining (pp. 61-83). Menlo Park, California: The AAAI Press.

Chen, C. (2007). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Molecular Ecology Notes, 7, 747-756.

Chen, C., Forbes, F., & François, O. . (2006). fastruct: Model-based clustering made faster. Molecular Ecology Notes, 6, 980-983.

Cheverud, J. M. (1982). Phenotypic, genetic, and environmental morphological integration in the cranium. . Evolution, 36, 499-516.

Cheverud, J. M. (1988). A comparison of genetic and phenotypic correlations. Evolution 42, 958-968.

Chhikara, R. S., & McKeon, J. (1984). Linear discriminant analysis with misallocation in training samples. Journal of the American Statistical Association, 79, 899-906.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. Journal of Econometrics, 75(1), 79-97.

Chretien, S., & Hero, A. O. (1998). Generalized Proximal Point Algorithms and Bundle Implementations: Communication and Signal Processing Laboratory, University of Michigan.

Christensen, A. F. (1998). Odontometric microevolution in the Valley of Oaxaca, Mexico. Journal of Human Evolution, 34, 333-360.

Christensen, A. M. (2004). The impact of Daubert: implications for testimony and research in forensic anthropology (and the use of frontal sinuses in personal identification). Journal of Forensic Sciences 49, 427-430.

Churchill, W. (1950). EUROPE: Thoughts & Actions. TIME. Retrieved from http://www.time.com/time/magazine/article/0,9171,856671,00.html

Clarke, A. E., Shim, J., Shostak, S., & Nelson, A. (2009). Biomedicalising genetic health, diseases and identities. In G. Atkinson & M. D. Lock (Eds.), The Handbook of Genetics & Society: Mapping the New Genomic Era (pp. 21-41). London: Routledge.

Clinton, W. J. . (2000, June 26). Text of Remarks on the Completion of the First Survey of the Entire Human Genome Project. The White House, Office of the Press Secretary, June 26, 2000 Retrieved December 17, 2010, from http://clinton5.nara.gov/WH/New/html/genome-20000626.html

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg & M. E. Sobel (Eds.), Handbook of statistical modeling for the social and behavioral science (pp. 311-360). New York: Plenum.

Coalición de Derechos Humanos Website. (2004) Retrieved 24 March, 2001, from http://www.derechoshumanosaz.net/index.php?option=com_content&task=view&id=20&Itemid=34

Cobb, W. M. (1942). Physical anthropology of the American Negro. American Journal of Physical Anthropology, 29, 113-223.

Cockerham, C. C. (1969). The variance of gene frequencies. Evolution, 23, 72-84.

Cockerham, C. C., & Weir, B. S. (1986). Estimation of inbreeding parameters in stratified populations. Annals of Human Genetics, 50, 271-281.

Cohen, A. C. (1968). Estimation in mixtures of two normal distributions. Technometrics 9, 15-28.

Collard, M., & Wood, B. (2000). How reliable are human phylogenetic hypotheses? . Proceedings of the National Academy of Science, 97, 5003-5006.

Collett, M. (1933). A study of Twelfth and Thirteenth Dynasty skulls form Kerma, Nubia. Biometrical, 25, 254-284.

Collins, F. S. (2004). What we do and don't know about 'race', 'ethnicity', genetics and health at the dawn of the genome era. Nature Genetics, 36(Suppl. 11), S13-15.

Collins, F. S. , & McKusick, V. A. (2001). Implications of the Human Genome Project for medical science. Journal of the American Medical Association, 285, 540-544.

Collins, F. S., & Mansoura, M. K. (2001). The Human Genome Project: Revealing the Shared Inheritance of All Humankind. Cancer, 92, S221-S225.

Collins, F. S., Green, E. D., Guttmacher, A. E., & Guyer, M. S. . (2003). A vision for the future of genomics research. Nature, 422(6934), 835-847.

Collins-Schramm, H. E., Kittles, R. A. , Operario, D. J. , Weber, J. L. , Criswell, L. A., et al. (2002). Markers that discriminate between European and African ancestry show limited variation within Africa. Human Genetics, 111, 566-569.

Comas, R. (1961). 'Scientific' Racism Again? Current Anthropology, 2(4), 310-340.

Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings. (2011). Retrieved from http://millenniumindicators.un.org/unsd/methods/m49/m49regin.htm.

Comstock, R. D., Castillo, E. M., & Lindsay, S. P. (2004). Four-year review of the use of race and ethnicity in epidemiologic and public health research. American Journal of Epidemiology, 159, 611-619.

Condit, C. M. (2007). How geneticists can help reporters to get their story right. Nature Reviews Genetics, 8, 815-820.

Condit, C. M., Parrott, R. L., Harris, T. M., Lynch, J. A., & Dubriwny, T. . (2004). The Role of "Genetics"in Popular Understandings of Race in the United States. Public Understanding of Science 13, 249-272.

Consortium, International HapMap. (2005). A haplotype map of the human genome. Nature 437, 1299-1320.

Consortium, International Human Genome Sequencing. (2001). Initial sequencing and analysis of the human genome. Nature, 409(9822), 860-921.

Consortium, The Chimpanzee Sequencing and Analysis. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature, 473(1), 69-87.

Coon, C. (1965). The Living Races of Man. New York: Alfred Knopf.

Coon, C. S. (1962). The Origins of Races. New York, NY: Alfred A. Knopf.

Coon, C.S. , Garn, S.M., & Birdsell, J.B. (1950). Races: a study of the problems of race formation in man. Springfield: Charles C. Thomas.

Cooper, R. S. (1994). A case study in the use of race and ethnicity in public health surveillance. Public Health Report, 109, 46-52.

Cooper, R. S., Kaufman, J. S., & Ward, R. (2003). Race and genomics. New England Journal of Medicine, 348, 1166-1170.

Corander, J., Waldmann, P., Marttinen, P., & Sillanpää, M. J. (2004). BAPS 2: enhanced possibilities for the analysis of genetic population structure Bioinformatics, 20(15), 2363-2369.

Cormack, R. M. (1971). A review of classification. Journal of the Royal Statistical Society Series A, 134, 321-367.

Cornell, s., & Hartmann, D. (1998). Ethnicity and race: making identities in a changing world. Thousand Oaks, California: Pine Forge Press.

Cornwell, P. (1994). The Body Farm. New York: Penguin.

Corruccini, R. S. . (1987). Shape in morphometrics: Comparative analyses. American Journal of Physical Anthropology, 73, 289-303.

Cosmides, L. . (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. . Cognition 31, 187-276.

Cosmides, L., & Tooby, J. . (1994). Origins of domain specificity: The evolution of functional organization. In L. A. Hirschfeld & S. A. Gelman (Eds.), Mapping the Mind: Domain Specificity in Cognition and Culture (pp. 85-116). New York: Cambridge University Press.

Council, National Research. (1988). Mapping and sequencing the human genome. Washington, DC: National Academies Press.

Crow, J. F. (2002). Unequal by nature: a geneticist's perspective on human differences Dedalus Winter (pp. 81-88).

Currat, M. , & Excoffier, L. . (2004). Modern humans did not admix with Neanderthals during their range expansion into Europe. PLoS Biology, 2, 2264-2274.

Cutler, A., & Windham, M. P. (1994). Information-Based Validity Functionals for Mixture Analysis. Paper presented at the Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach.

Dacunha-Castelle, D., & Gassia, E. (1997). The estimation of the order of a mixture model. Bernoulli, 3, 279-299.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. Nature Genetics, 29, 229-232.

Darnell, R. (2000). Re-envisioning Boas and Boasian Anthropology. American Anthropologist, 102(4), 896-900.

Darroch, J N, & Mosimann, J E. (1985). Canonical and principal components of shape. Biometrika, 72, 241-252.

Dasgupta, A., & Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. Journal of the American Statistical Association(93), 294-302.

Daubert v. Merrell Dow Pharmaceuticals, No. 2786 (133 S. Ct. 1993).

Davies, B. (2002). Integral transforms and their applications (3rd ed.). New York: Springer.

Davison, A. C. (1986). Approximate predictive likelihood. Biometrika, 73, 323-332.

Davison, A. C., & Hinkley, D. V. (1997). Bootstrap Methods and Their Application. Cambridge: Cambridge University Press.

Dawson, K. J., & Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genetic Research, 78, 59-77.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. Biometrika, 56, 463-474.

Day, N. E. (1969). Estimating the components of a mixture of two normal distributions. Biometrika, 56, 463-474.

Dayton, C. M. , & MacReady, G. B. . (1988). Concomitant-variable latent-class models. Journal of the American Statistical Association, 83(401), 173-178.

de Craen, S., Commandeur, J. F., Frank, L. E., & Heiser, W. J. (2006). Effects of group size and lack of sphericity on the recovery of clusters in k-means cluster analysis. Multivariate Behavioral Research, 41, 127-145.

Dean, M., Stephens, C., Winkler, C., Lomb, D. A., Ramsburg, M., Boaze, R., . . . O'Brian, S. J. (1994). Polymorphic admixture typing in human ethnic populations. American Journal of Human Genetics, 55, 788-808.

Dean, N. (2009). The Variable Selection for Model-based Clustering (clustvarsel) package. Retrieved from http://cran.r-project.org/web/packages/clustvarsel/vignettes/varsel.vignette.pdf

Dean, N., & Raftery, A E. (2010). Latent class analysis variable selection Annals of the Institute of Statistical Mathematics, 62, 11-35.

DeGiorgio, M., Jakobsson, M., & Rosenberg, N. A. (2009). Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proceedings of the National Academy of Science, 106(38), 16057-16062.

DeGroot, Morris H. (1987). A Conversation with George Box. Statistical Science, 2(3), 239-258.

Deka, R., Jin, L., Shriver, M. D., Yu, L. M., DeCroo, S., Hundrieser, J., . . . Chakraborty, R. (1995). Population genetics of dinucleotide (dC-dA)n.(dG-dT)n polymorphisms in world populations. American Journal of Human Genetics, 56(2), 461-474.

Dempster, A. P., Laird, N. M, & Rubin, D B. (1997). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society Series B, 39, 1-38.

Dempster, A. P., Laird, N. M., & Rubin, D. B. . (1977). Maximum Likelihood for Incomplete Data via the EM Algorithm (with discussion). Journal of the Royal Statistical Society, Ser. B, 39(1), 1-38.

DeSarbo, W. S. , & Cron, W. L. . (1988). A maximum likelihood methodology for clusterwise linear regression. Journal of Classification, 5(2), 249-282.

DeSarbo, W. S., Wedel, M., Vriens, M., & Ramaswamy, V. (1992). Latent Class Metric Conjoint Analysis. Marketing Letters, 3(3), 273-288.

Deshpande, O., Batzoglou, S. , Feldman, M.W. , & Cavalli-Sforza, L.L. . (2009). A serial founder effect model for human settlement out of Africa. Proceedings of the Royal Society B, 276, 291-300.

Devor, E. J. (1987). Transmission of human cranial dimensions. Journal of Craniofacial Genetics and Developmental Biology, 7, 95-106.

Devor, E. J. . (1987). Transmission of human craniofacial dimensions. Journal of Craniofacial Genetics and Developmental Biology, 7, 95-106.

Di Battista, G., Eades, P, Tamassia, R., & Tollis, I. G. (1999). Graph Drawing: Algorithms for the Visualization of Graphs. Upper Saddle River, NJ: Prentice Hall.

Diamond, J. (1994). Race Without Color. Discover, November, 82-89.

Diamond, J. M. . (1966). Zoological classification system of a primitive people. Science, 151, 1102-1104.

Dias, J. G. . (2004). Finite Mixture Models. Review, Applications and Computer-intensive Methods. PhD, University of Groningen, Groningen.

Dias, J. G., & Willekens, F. J. (2005). Model-based clustering of sequential data with an application to contraceptive use dynamics. Mathematical Population Studies, 12(3), 135-157.

Diebolt, J., & Robert, C. P. (1994). Estimation of Finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society Series B, 56, 363-375.

Dillon, W. R., & Kumar, A. . (1994). Latent structure and other mixture models in marketing: an integrative survey and overview. In R. Bagozzi (Ed.), Advanced Methods of Marketing Research (pp. 295-351). Cambridge, Massachusetts: Basil Blackwell.

Distell, T. R. (2000). Human Genetic Variation. Genome Biology, 1(5), comment 2004.2001-2004.2002.

Dobzhansky, T. (1937). Genetics and the Origin of Species. New York, NY: Columbia University Press.

Dobzhansky, T. (1944). On Species and Races of Living and Fossil Man. American Journal of Physical Anthropology, 2, 251-265.

Dobzhansky, T. (1955). A Review of Some Fundamental Concepts and Problems of Population Genetics. Cold Spring Harbor Symposium on Quantitative Biology, 20, 1-15.

Dobzhansky, T. (1961). "A Bogus 'Science' of Race Prejudice. Journal of Heredity, 52, 189-190.

Dobzhansky, T. (1962). Mankind Evolving. New Haven, CT: Yale University Press.

Dobzhansky, T. (1970). Genetics of the Evolutionary Process. New York, NY: Columbia University Press.

Doetsch, G. (1928). Die elimination des dopplereffekts auf spektroskopische feinstrukturen und exakte bestimung der komponenten. Zeitschrift fur Physik 49, 705-730.

Donahue, A., & Fink, K. (Writers). (2002). Burden of Proof, CSI: Crime Scene Investigation. Las Vegas: CBS productions.

Douglas, M. (1986). How Institutions Think (Frank W. Abrams Lectures). Syracuse: Syracuse University Press.

Drevdahl, D., Taylor, J. Y., & Phillips, D. A. (2001). Race and ethnicity as variables in Nursing Research, 1952-2000. Journal of Nursing Research, 50, 305-313.

Drevhadl, D., Taylor, J. Y., & Phillips, D. A. (2001). Race and Ethnicity as variable in Nursing Research, 1952-2000. Nursing Research, 50, 305-313.

Dryden, I. L. , & Mardia, K. V. . (1998). Statistical shape analysis. New York: Wiley.

Dubriwny, T. N., Bates, B. R., & Bevan, J. L. . (2004). Lay understandings of race: Cultural and genetic definitions. Community Genetics, 7, 185-195.

D'Unger, A V, Land, K C, L, McCall P, & S, Nagin D. (1998). How many latent classes of delinquent/criminal careers? Results from mixed Poisson regression analyses of the London, Philadelphia, and Racine cohorts studies. American Journal of Sociology, 103, 1593-1630.

Duster, T. (2003). Buried Alive: The Concept of Race in Science. In Heath, D., Lindee, S. M. and Goodman, A. H. (Ed.), Genetic Nature / Culture: Anthropology and Science Beyond the Two-Culture Divide (pp. 258-277). Berkeley: University of California Press.

Duster, T. (2006). Lessons from History: Why Race and Ethnicity Have Played A Major Role in Biomedical Research. Journal of Law, Medicine & Ethics, Fall, 1-11.

Ebersberger, I., Metzler, D., Schwarz, C., & S., Pääbo. (2002). Genomewide comparison of DNA sequences between humans and chimpanzees. American Journal of Human Genetics, 70, 1490-1497.

Edelbrock, C. (1979). Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. Multivariate Behavioral Research, 14, 367-384.

Edgar, H. J. H. (2009). Biohistorical Approaches to "Race" in the United States: Biological Distances Among African Americans, European Americans, and Their Ancestors. American Journal of Physical Anthropology, 139, 58-67.

Edgar, H. J., & Hunley, K. L. . (2009). Race reconciled?: How biological anthropologists view human variation. American Journal of Physical Anthropology, 139, 1-4.

Edwards, A. W. F. (2003). Human genetic diversity: Lewontin's fallacy. BioEssays, 25, 798-801.

Edwards, A. W. F., & Cavalli-Sforza, L. L. (1965). A Method for Cluster Analysis. Biometrics, 21(2), 362-375.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. The Annals of Statistics, 7(1), 1-26.

Efron, B. (1979). Computers and the theory of statistics: Thinking the unthinkable. SIAM Review, 21(4), 460-480.

Efron, B., & Tibshirani, R. J. (1993). An Introduction to the Bootstrap. New York: Chapman & Hall.

Efron, B., & Tibshirani, R. J. . (1997). Improvements on cross-validation: The .632+ bootstrap method. Journal of the American Statistical Association, 92, 548-560.

Ehrich, D. (2006). Aflpdat: a collection of r functions for convenient handling of AFLPdata. Molecular Ecology Notes 6, 603-604.

Ehrlich, P., & Holm, R. (1964). A biological view of races. In A. Montagu (Ed.), The concept of race (pp. 153-179). New York, NY: Free Press.

Einstein, A. (1922). Address to the French Philosophical Society at the Sorbonne (6 April 1922). Einstein Archive 36-378, 79-535.

Eldesbrunner, H. (1987). Algorithm in combinational geometry New York: Springer.

Elliott, C., & Brodwin, P. . (2002). Identity and genetic ancestry tracing. British Medical Journal, 325, 1469-1471.

Elliott, M. (2008). Fordisc And The Determination Of Ancestry From Craniometric Data. MA, Simon Frasier University, Burnaby, BC.

Elliott, M., & Collard, M. (2009). FORDISC and the determination of ancestry from cranial measurements. Biology Letters, 5, 849-852.

Endicott, P., Gilbert, M.T., Stringer, C., Lalueza-Fox, C., Willerslev, E., Hansen, A.J., & Cooper, A. . (2003). The genetic origins of the Andaman islanders. American Journal of Human Genetics, 72, 178-184.

Endicott, P., Metspalu, M. , Stringer, C., Macaulay, V., Cooper, A., & Sanchez, J.J. . Multiplexed SNP typing of ancient DNA clarifies the origin of Andaman mtDNA haplogroups amongst south Asian tribal populations. PLoS ONE 1, 81.

Ereshefsky, M. (2000). The Poverty of the Linnaean Hierarchy : A Philosophical Study of Biological Taxonomy. Cambridge: Cambridge University Press.

Esobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association, 90, 577-588.

Everitt, B. S. (1980). Cluster Analysis (2nd ed.). Portsmouth, NH: Heinemann.

Everitt, B. S. (1987). Introduction to Optimization Methods and their Application in Statistics. London: Chapman & Hall.

Everitt, B. S. (1993). Cluster Analysis (3rd ed.). New York: John Wiley and Sons.

Everitt, B. S., & Hand, D. J. (1981). Finite Mixture Distributions. London: Chapman & Hall.

Everitt, B. S., Landau, S., & Leese, M. (2001). Cluster Analysis (4th ed.). London: Arnold.

Eves, H. (1988). Return to Mathematical Circles. Boston: Prindle, Weber, and Schmidt.

Evvano, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology, 14, 2611-2620.

Excoffier, L. (2001). Analysis of population subdivision. In D. J. Balding, M. Bishop & C. Cannings (Eds.), Handbook of statistical genetics (pp. 271-307). Chichester: John Wiley and Sons.

Excoffier, L., & Hamilton, G. (2003). Comment on genetic structure of human populations. Science, 300, 1877.

Excoffier, L., Smouse, P. F., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction. Genetics, 131, 479-491.

Explore the 2010 Census Form. (2010), 2010, from http://2010census.gov/2010census/how/interactive-form.php

Fagundes, N.J.R., Kanitz, R. , Eckert, R., Valls, A.C.S., Bogo, M.R., Salzano, F.M., . . . Bonatto, S.L. . (2008). Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. American Journal of Human Genetics, 82, 583-592.

Fairgrieve, S. (199). Forensic osteological analysis: A book of case studies. Springfield: Charles C. Thomas.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics, 164, 1567-1587.

Falush, D., Stephens, M., & Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles Molecular Ecology Notes, 7, 574-578.

Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., Kidd, M., . . . Suerbaum, S. (2003). Traces of human migrations in Helicobacter pylori populations. Science, 299, 1582-1585.

Family Search. (2008) Retrieved September 2, 2010

FamilyLink. (2010) Retrieved September 2, 2010, from www.familylink.com http://corporate.familylink.com/about/index.htm

Faraway, J. (2002). Practical Regression and Anova using R. CRAN: Contributed Documentation Retrieved January 1, 2011, from http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Fawcett, C. D. (1901). A second study of the variation and correlation of the skull with special reference to the Negada crania. Biometrika, 24, 45-54.

FDB Forensic Data Forms. (2011) Retrieved January 20, 2011, from http://fac.utk.edu/databank.html

Feldesman, M. R., & Fountain, R. L. (1996). "Race" specificity and the femur/stature ratio. American Journal of Physical Anthropology, 100, 207-224.

Feldman, M. W., Lewontin, R. C., & King, M. C. (2003). Race: A genetic melting pot. Nature, 424(374).

Felman, M., & Lewontin, R. C. (2008). Race, Ancestry, and Medicine In B Koenig, S. Lee & S Richardson (Eds.), Revisiting Race in a Genomic Age (pp. 89-101). New Brunswick: Rutgers University Press.

Feng, Z. D., & McCulloch, C. E. (1994). On the likelihood ratio test statistic for the number of components in a normal mixture with unequal variances. Biometrics, 50, 1158-1162.

Feng, Z. D., & McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. Journal of the Royal Statistical Society Series B (Methodological), 58(609), 617.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In M. H. Rizvi, J. S. Rustagi & D. Siegmund (Eds.), Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday (pp. 287-302). New York: Academic Press.

Fessler, J. A., & Hero, A. O. (1994). Space-altering generalized expectation-maximization algorithm. IEEE Transaction of Signal Processing, 42, 2664-2677.

Fielding, A. H. (2007). Cluster and Classification Techniques for the Biosciences. New York: Cambridge University Press.

Fields, B. (1990). Slavery, Race, and Ideology in the United States of America. New Left Review, 181, 95-118.

Finch, W. H., & Schneider, M. K. (2006). Misclassification rates for four methods of group classification: Impact of predictor distribution, covariance inequality, effect size, sample size, and group size ratio. Educational and Psychological Measurement, 66, 240-257.

Fischer, A., Wiebe, V., S., Pääbo, & M., Przeworski. (2004). Evidence for a complex demographic history of chimpanzees. Molecular Biology and Evolution, 21, 799-808.

Fisher, R. A. (1930). The Genetical Theory of Natural Selection: Clarendon Press.

Fisher, R. A. (1936). The coefficient of racial likeness and the future of craniometry. Journal of the Royal Anthropological Institute, 66, 57-63.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.

Fisher, R. A. (1956). Statistical Methods and Scientific Inference. Edinburgh: Oliver and Boyd.

Flanagan, N., Healy, E., Ray, A., Philips, S., Todd, C., Jackson, I. J., . . . Rees, J. L. (2000). Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation. Human Molecular Genetics, 9, 2531-2537.

Fleagle, J. G. , Gilbert, C. C., & Baden, A. L. (2010). Primate cranial diversity. American Journal of Physical Anthropology, 142, 565-578.

Fleagle, J. G., Kay, R. F., & Simons, E. L. (1980). Sexual dimorphism in early anthropoids. Nature, 287, 328-330.

Flegal, K. L., Carroll, M. D., Odgen, C. L., & Johnson, C. (2002). Prevalence and trends in obesity among U.S. adults, 1999-2000. JAMA, 288(14), 1723-1727.

Flegal, K. M., Carroll, M. D., Kuczmarski, R. J., & Johnson, C. L. (1998). Overweight and obesity in the United States: Prevalence and trends, 1960-1994. International Journal of Obesity, 22, 39-47.

Fleiss, J. L., & Zubin, J. (1969). On the method and theory of clustering. Multivariate Behavioral Research, 4, 235-250.

FORDISC 3.0 Personal Computer Forensic Discriminant Functions. (2011) Retrieved Jan 20, 2011

Foreman, L., Smith, A., Evett, I., & 1997. (1997). Journal of the Royal Statistical Society Series A. Journal of the Royal Statistical Society Series A, 160, 429-469.

Forensic Anthropology Center, The William M. Bass Donated Skeletal Collection. (2011) Retrieved August 21, 2011, from http://fac.utk.edu/collection.html

Forensic Anthropology Data Bank Website. (2011) Retrieved January 30, 2011, from http://web.utk.edu/~fac/databank.html

Forensic Anthropology Data Bank. (2005). Retrieved September 25, 2010, from Department of Anthropology, University of Tennessee http://web.utk.edu/~fac/databank.shtml

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. Journal of the American Statistical Association, 87(418), 476-486.

Formicola, V. (1987). Neolithic transition and dental changes: the case of an Italian site. Journal of Human Evolution, 16(231-240).

Foster, J. W., & Sharp, R. R. (2002). Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. Genome Research, 12, 844-850.

Foster, M. W., & Sharp, R. R. (2003). Race, ethnicity, and genomics: Social classifications as proxies of biological heterogeneity. Genome Research, 12, 844-850.

Foucault, M. (1966). Les Mots et les Choses Paris: Gallimard.

Foucault, M. (1969). L'Archéologie du savoir Paris: Gallimard.

Foucault, M. (1971). L'Ordre du discours Paris: Gallimard.

Fowlkes, E. B. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. Journal of the American Statistical Association, 47, 561-575.

Fraley, C, & Raftery, A E. (1999). mclust: Software for Model-based Cluster Analysis. Journal of Classification,, 16, 297-306.

Fraley, C, & Raftery, A E. (2000). Model-Based Clustering, Discriminant Analysis, and Density Estimation Technical Report (pp. 1-43). Seattle: Department of Statistics, University of Washington.

Fraley, C, & Raftery, A E. (2002). MCLUST: Software for Model-Based Clustering, Discriminant Analysis and Density Estimation. Seattle, WA: University Washington.

Fraley, C, & Raftery, A. E. (2006). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering Technical Report no. 504. Seattle: Department of Statistics, University of Washington.

Fraley, C., & Raftery, A E. (1998). How Many Clusters? Which Clustering Method? – Answers via Model-based Cluster Analysis. Computer Journal, 41, 578-588.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis. Computer Journal, 41, 578-588.

Fraley, C., & Raftery, A. E. (2002). Model-based Clustering, Discriminant Analysis and Density Estimation. Journal of the American Statistical Association, 97, 611-631.

Fraley, C., & Raftery, A. E. (2003). Enhanced Software for Model-based Clustering, Density Estimation, and Discriminant Analysis: mclust. Journal of Classification, 20, 263-286.

Fraley, C., & Raftery, A. E. (2006). MCLUST version 3: an R package for normal mixture modeling and model-based clustering. Seattle, WA: Department of Statistics, University of Washington.

Fraley, C., & Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. Journal of Classification 24, 155-181.

Fraley, C., & Raftery, A. E. (2007). Model-Based Methods of Classification: Using mclust in Chemometrics. Journal of Statistical Software, 18(6).

Fraley, C., & Raftery, A. E. (2010). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering (Department of Statistics, Trans.). Seattle: University of Washington, Department of Statistics.

Fraley, C., & Raftery, A. E. (2010). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering (Vol. Technical Report No. 504). Seattle: Department of Statistics, University of Washington.

Fraley, C., & Raftery, A. E. (2011, May 29, 2011). Package 'mclust'. R documentation Retrieved June 01, 2011, from http://cran.r-project.org/web/packages/mclust/mclust.pdf

France, D. L. (2001). Lab Manual and Workbook for Physical Anthropology (Vol. 4th). Belmont, CA: Wadsworth Thomson Learning.

France, D. L., & Horn, A. D. (1988). Lab Manual and Workbook for Physical Anthropology. St. Paul, MN: West.

Franciscus, R. G., & Long, J. C. (1991). Variation in human nasal height and breadth. American Journal of Physical Anthropology, 85, 419-427.

Franco, P. P., & Shamos, M. I. (1985). Computational Geometry - An Introduction. New York: Springer-Verlag.

Frank, R. (2007). What to make of it? The (Re) emergence of a biological conceptualization of race in health disparities research. Social Science & Medicine, 64, 1977-1983.

Frayer, D. W. (1977). Metric dental changes in the European Upper Paleolithic and Mesolithic. American Journal of Physical Anthropology, 46, 109-120.

Fredrickson, G. M. (2002). Racism: A Short History. Princeton, New Jersey: Princeton University Press.

Freid, D., Spradley, MK, Jantz, RL, & SD, Ousley. (2005). The truth is out there: How NOT to use FORDISC. American Journal of Physical Anthropology, 128:S41, 103.

Frick, A., Ludwig, A., & Mehldau, H. (1995). A fast, adaptive layout algorithm for undirected graphs. In R Tamassia & I Tollis (Eds.), Graph Drawing '94: Lecture Notes in Computer Science (pp. 388-403). New York: Springer Verlag.

Friedlaender, J.S., Friedlaender, F.R., Reed, F.A., Kidd, K.K., Kidd, J.R. , et al. (2008). The Genetic Structure of Pacific Islanders. PLoS Genetics, 4(1), 0173-0190.

Friedman, H. P., & Rubin, J. (1967). On some invariant criteria for grouping data. Journal of the American Statistical Association, 62, 1159-1178.

Fruchterman, T. M. J., & Reigold, E. M. (1991). Graph Drawing by Force-Directed Placement. Software—Practice and Experience, 21(11), 1129-1164.

Frudakis, T., Kondragunta, V., Thomas, M. G., Gaskin, Z., Ginjupalli, S., Gunturi, S., . . . Nachimuthu, P. K. (2003). A classifier for the SNP-based inference of ancestry. Journal of Forensic Science. 48, 4(771-782).

Fryer, J. G., & Robertson, C. A. (1972). A comparison of some methods for estimating mixed normal distributions. Biometrika, 59(3), 639-648.

Fuentes, A. (2007). Core concepts in biological anthropology. Boston: McGraw Hill.

Fuentes, A. (2010). The new biological anthropology: Bringing Washburn's new physical anthropology into 2010 and beyond- The 2008 AAPA luncheon lecture. American Journal of Physical Anthropology, 143, 2-12.

Fujimura, J. H., & Rajagopalan, R. (2011). Different differences: The use of ‘genetic ancestry’ versus race in biomedical human genetic research Social Studies of Science (41), 1.

Fujimura, J. H., Duster, T, & Rajagopalan, R. (2008). Race, Genetics and Disease: Questions of Evidence, Matters of Consequence. Social Studies of Science, 35(5), 643-656.

Fujimura, J. H., Rajagopalan, R., Ossorio, P., & Doksum, K. (2010). Race and Ancestry: Operationalizing Populations in Human Genetic Variation Studies. In I. Whitmarsh & D. S. Jones (Eds.), What's the Use of Race? Modern Governance and the Biology of Difference. (pp. 169-183). Cambridge, MA: MIT Press.

Fullwiley, D. (2008). The Molecularization of Race and Institutions of Difference: Pharmacy and Public Science after the Genome. In B Koenig, S Soo-Jin Lee & S Richardson (Eds.), In Revisiting Race in a Genomic Age. (Studies in Medical Anthropology) (pp. 149-171). New Brunswick, NJ: Rutgers University Press.

Futuyma, D. (1998). Evolutionary Biology. Sunderland, MA: Sinauer.

Gage, T. (2000). Variability of gestational age distributions by sex and ethnicity: An analysis using mixture models. American Journal of Human Biology, 12(2), 181-191.

Gage, T. (2003). Classification of births by birth weight and gestational age: An application of multivariate mixture models. Annals of Human Biology, 30(5), 589-604.

Galton, F. (1869). Hereditary genius: An Inquiry into its Laws and Consequences. London: Macmillan.

Ganesalingam, S., & McLachlan, G. J. (1980). Error rate estimation on the basis of posterior probabilities. Pattern Recognition, 11, 405-413.

Garn, S. (1962). Human Races (Vol. Revised). Springfield, IL: Charles Thomas.

Garrett, H. E. (1961). The Equalitarian Dogma. Perspectives in Biology and Medicine, 4, 480-484.

Garrigan, D., & Hammer, M. F. (2006). Reconstructing human origins in the genomic era. Nature Reviews Genetics, 7, 669-680.

Gasman, D. (2004). The Scientific Origins of National Socialism. New Brunswick, NJ: Transaction Press.

Gasner, E., Hu, Y, Kobourov, S., & Volinksy, C. (2009). Putting Recommendations on the Map - Visualizing Clusters and Relations. Paper presented at the Proceedings of the third ACM conference on Recommender systems.

Geary, DC. (1998). Functional organization of the human mind: Implications for behavioral genetic research. Human Biology, 70(185-198).

Gelman, A., & Rubin, D. B. (1995). Avoiding Model Selection in Bayesian Social Research. Sociological Methodology 25, 165-173.

Geman, S., & Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. Annals of Statistics 10, 401-414.

GenBank. (2011) Retrieved January 7, 2011, from http://www.ncbi.nlm.nih.gov/genbank/

Genetics, The American Society of Human. (2008). Ancestry Testing Statement Retrieved March 5, 2011, from http://www.ashg.org/pdf/ASHGAncestryTestingStatement_FINAL.pdf

George, W. C. (1962). Biology and the Race Problem. Washington, DC: National Putnam Letters Committee.

Ghosh, J. H., & Sen, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In L. M. Le Cam & R. A. Olshen (Eds.), Proceedings of the Berkeley Conference in Honour of Jerzy Neymnan and Jack Kiefe (2 ed. ed., pp. 789-806). Monterey, C.A.: Wadsworth.

Giles, E. (1964). Sex determination by discriminant function analysis of the mandible. American Journal of Physical Anthropology, 22(2), 127-135.

Giles, E. (1966). Statistical techniques for sex and race determination: some comments in defense. American Journal of Physical Anthropology, 25, 85-86.

Giles, E. (1997). Hooton, E(arnest) A(lbert) (1887-1954). In F. Spencer (Ed.), History of physical anthropology: an encyclopedia (Vol. 1, pp. 499-500). New York, NY: Garland Publishing.

Giles, E. (1998). Earnest Albert Hooton and "Plain Statements" About Race In a Symposium "150 Years of Racial Thinking in American Science. Paper presented at the 150th Anniversary Meeting, American Association for the Advancement of Science.

Giles, E., & Elliot, O. (1962). Race identification from cranial measurements. Journal of Forensic Sciences, 7, 147-157.

Giles, E., & Elliot, O. (1963). Sex Determination by Discriminant Function Analysis of Crania. American Journal of Physical Anthropology, 21, 53-68.

Giles, E., & Elliott, O. (1962). Race Identification from Cranial Measurements. Journal of Forensic Sciences, 7, 147-157.

Gill, G. W. (1998). Craniofacial criteria in the skeletal attribution of race. In K. J. Reichs (Ed.), Advances in the identification of human remains (pp. 293-317). Springfield, IL: Charles C. Thomas.

Gill, G. W. (2000). Does race exist? A proponent's perspective. NOVA Online Retrieved September 3, 2010, from http://www.pbs.org/wgbh/nova/first/gill.html

Gillham, N. W. (2001). A Life of Sir Francis Galton: From American Exploration to the Birth of Eugenics: Oxford University Press.

Gissis, S. B. (2008). When is 'race' a race? 1946–2003. Studies in the History and Philosophy of Biological and Biomedical Sciences, 39, 437-450.

Giudici, P. (2003). Applied Data mining, statistical methods for business and industry. London: Wiley.

Glasgow, J. (2009). A Theory of Race. New York: Routledge.

Godfrey, L. R., Lyon, S. K., & Sutherland, M. R. (1993). Sexual dimorphism in large bodied primates: The case of the subfossil lemurs. American Journal of Physical Anthropology, 90, 315-334.

Goffinet, B., Loisel, P., & Laurent, B. (1992). Testing in normal mixture models when the proportions are known. Biometrika, 79, 842-846.

Goldstein, D. B., & Weale, M. E. (2001). Population genomics: Linkage disequilibrium holds the key. Current Biology, 11, R567-R579.

Goncalves, V. F., Prosdocimi, F., Santos, L. S., Ortega, J. M., & Pena, S. D. (2007). Sex-biased gene flow in African Americans but not in American Caucasians. Genetics and Molecular Research, 6, 156-161.

Gonzalez, E., Bamshad, M., Sato, N., Mummidi, S., Dhanda, R., Catano, G., . . . al., et. (1999). Race-specific HIV-1 disease modifying effects of CCR5 haplotypes. Proceedings of the National Academy of Science, 96, 12004-12009.

González-José, R., Bortolini, M. C., Santos, F. R. , & Bonatto, S. L. . (2008). The peopling of America: Craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. American Journal of Physical Anthropology, 137, 175-187.

Goodman, A. H. (1997). Bred in the Bone? . The Sciences, March/April, 20-25.

Goodman, A. H. (2000). Why genes don't count (for racial differences in health). American Journal of Public Health, 90, 1699-1702.

Goodman, A. H., & Armelagos, G. J. (1997). The Resurrection of Race: The Concept of Race in Physical Anthropology in the 1990s. In L. T. Reynolds & L. Lieberman (Eds.), Race and Other Misadventures: Essays in Honor of Ashley Montagu in His Ninetieth Year (pp. 174-186). Dix Hills, New York: General Hall.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 61(2), 215-231.

Gordon, A. D. (1980). Classification. London: Chapman and Hall.

Gordon, C. C. (Ed.). (2009). Ethnicity, and Applied Bioanthropology (Vol. NAPA Bulletin 13). Chichester: John Wiley and Sons.

Gordon, C. C., & Greiner, T. M. (2009). Biocultural Influences in the Analysis of Secular Trends. In C. C. Gordon (Ed.), Race, Ethnicity, and Applied Bioanthropology (Vol. NAPA Bulletin 13., pp. 22-33). Chichester: John Wiley and Sons.

Gosset, T. F. (1963). Race: The History of an Idea in America. Dallas: Southern Methodist Press.

Gossett, T. F. (1997). Race: The History of an Idea in America (Vol. New Edition). New York, NY: Oxford University Press.

Gould, S. J. (1996). The Mismeasure of Man (Vol. Revised and Expanded). New York, NY: W.W. Norton & Company.

Gower, J. C. (1972). Measures of taxonomic distance and their analysis. In J. S. Weiner & J. Huizinga (Eds.), The Assessment of Population Affinities in Man (pp. 1-14). Oxford: Clarendon.

Grant, M. (1916). The Passing of the Great Race. New York, NY: Charles Scribner's Sons.

Grant, M. (1933). The Conquest of a Continent. New York, NY: Charles Scribner's Sons.

Granzow, H., Klupp, B. G., Fuchs, W., Veits, J., Osterrieder, N., & Mettenleiter, T. C. (2001). Egress of alpha herpes viruses: comparative ultrastructural study. Journal of Virology, 75, 3675-3684.

Graves, J. J., Jr. (2006). What We Know and What We Don't Know: Human Genetic Variation and the Social Construction of Race; Is Race Real? Retrieved July 2, 2010, from http://raceandgenomics.ssrc.org/Graves/

Graves, J. L., Jr. (2001). The emperor's new clothes: biological theories of race at the millennium. New Brunswick, NJ: Rutgers University Press.

Gravlee, C. C. (2009). How Race Becomes Biology: Embodiment of Social Inequality. American Journal of Physical Anthropology, 139, 47-57.

Gravlee, C. C., & Sweet, E. (2008). Race, ethnicity, and racism in medical anthropology, 1977-2002. Medical Anthropology, 22, 27-51.

Gravlee, C. C., Bernard, H. R., & Leonard, W. R. (2003). Heredity, Environment, and Cranial Form: A Reanalysis of Boas's Immigrant Data. American Anthropologist, 105(1), 123-136.

Gravlee, C. C., Bernard, R., & Leonard, W. R. (2003). Boas's Changes in Bodily Form: The Immigrant Study, Cranial Plasticity, and Boas's Physical Anthropology. American Anthropologist, 105(2), 326-332.

Grayson, D. A. (1987). Statistical diagnosis and the influence of diagnostic error. Biometrics, 43, 975-984.

Green, J. C. (1959). The Death of Adam. New York, NY: Mentor Press.

Green, R. E., Krause, J., Briggs, A.W., Marcic, T. , Stensel, U., Kircher, M., . . . Pääbo, S. (2010). A draft sequence of the Neandertal genome. Science, 328, 710-722.

Green, R. E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., . . . Pääbo, S. (2006). Analysis of one million base pairs of Neanderthal DNA. Nature 444, 330-336.

Green, R. E., Malaspinas, A.-S. , Krause, J., Briggs, A., Johnson, P., Uhler, C., . . . Stenzel, U. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell, 134, 416-426.

Greenberg, J.H., Turner, C.G., II, & Zegura, S.L. (1986). The settlement of the Americas: a comparison of the linguistic, dental and genetic evidence. Current Anthropology, 27, 477-495

Gregory, Susanna. A Summer of Discontent.

Grene, M., & Depew, D. J. (2004). The philosophy of biology: an episodic history. Cambridge: Cambridge University Press.

Grivas, C. R., & Komar, D. A. (2008). Kumho, Daubert, and the nature of scientific inquiry: implications for forensic anthropology. Journal of Forensic Sciences 53, 771-776.

Groenen, P. J. F., & Jajuga, K. (2001). Fuzzy clustering with squared Minkowski distances. Fuzzy Sets and Systems,, 120, 227-237.

Groenen, P. J. F., Kaymak, U., & van Rosmalen, J. (2006). Fuzzy clustering with Minkowski distance functions (pp. 1-20): Econometric Institute Report EI 2006-24

Groh, M., & Ruiz, C. (1994). Determination of race and sex from the crania of a skeletal collection in Dade County, Florida. M. A., Florida Atlantic University, Boca Raton, FL.

Groh, M., & Ruiz, C. (1994). Determination of race and sex from the crania of a skeletal collection in Dade County, Florida. Masters Abstracts International, 33(2), 0368.

Groves, C.P. (1989). Regional approach to the problem of the origin of modern humans in Australasia. In P. Mellars & C. Stringer (Eds.), The Human Revolution: Behavioural and Biological Perspective in the Origins of Modern Humans (pp. 274-285). Edinburgh: Edinburgh University Press.

Guglielmino-Matessi, C. R., Gluckman, P., & Cavalli-Sforza, L. L. (1979). Climate and the evolution of skull metrics in man. American Journal of Physical Anthropology, 50, 549-564.

Guojun, G., Chaoqun, M., & Jianhong, W. (2007). Data Clustering: Theory, Algorithms, and Applications. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.

Habgood, P.J. (1992). The origin of anatomically modern humans in East Asia Continuity or Replacement: Controversies in Homo Sapiens Evolution (pp. 273-288). Rotterdam: Balkema.

Habgood, P.J. . (1989). An investigation into the usefulness of a cladistic approach to the study of the origin of anatomically modern humans. Human Evolution, 4, 241-252.

Haga, S. B., & Venter, J. C. (2003). Genetics: FDA races in wrong direction. Science, 301, 466.

Hagenaars, J. A., & McCutcheon, A. L. (2002). Applied latent class analysis. Cambridge: Cambridge University Press.

Haglund, W. D. (1997). Postmortem Involvement with Human Remains. In W. Haglund & M. Sorg (Eds.), Forensic Taphonomy, The Postmortem Fate of Human Remains (pp. 367-382). Boca Raton: CRC Press.

Hahn, R. A. (1992). The State of Federal Health Statistics on Racial and Ethnic Groups Journal of the American Medical Association, 276(2), 268-271.

Hahn, R. A., Mulinare, J., & Teutsch, S. M. (1992). Inconsistent Coding of Race and Ethnicity in Infants-Reply Journal of the American Medical Association, 267(23), 259-263.

Hald, A. (1948). Statistiske Metoder med Eksempler pá Anvendelser indenfor Teknikken. Copenhagen: Det Private Ingeniorfond.

Hald, A. (1952). Statistical Theory with Engineering Applications. New York: Wiley.

Haldane, J. B. S. (1932). The Causes of Evolution. Longman: Green and Co.

Hall, D. R., Smith, M. K., & Marciszewska, B. (Eds.). (20006). Tourism in the new Europe: the challenges and opportunities of EU enlargement. Cambridge, MA: CAB International.

Hallahan, D. P., Keller, C. E., Martinez, E. A., Gelman, J. A., & Fan, X. (2007). How variable are interstate prevalence rates of learning disabilities and other special education categories? A longitudinal comparison. Council for Exceptional Children, 73, 136-146.

Hammer, M. F., Karafet, T., Redd, A. J., Jarjanazi, H., Santachiara-Benerecetti, S., Soodyall, H., & Zegura, S. L. (2001). Hierarchical patterns of global human Y-chromosome diversity. Molecular Biology and Evolution, 18, 1189-1203.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. Bioinformatics, 21, 3201-3212.

Handley, L. J., Manica, A., Goudet, J., & Balloux, F. (2007). Going the distance: human population genetics in a clinal world. Trends in Genetics, 22(432-439).

Hanihara, K. (1987). Estimation of the number of early migrants to Japan: a simulative study. Journal of the Anthropological Society of Nippon 95, 391–403.

Hanihara, K. (1998). Reanalysis of local variations in the Ainu crania. Anthropological Science, 106(Suppl), 1-15.

Hanihara, K. . (1991). Dual structure model for the population history of the Japanese. Japan Review, 2, 1-33.

Hanihara, K., Kimura, K., & Minamidate, T. (1964). The sexing of Japanese skeleton by means of discriminant function. Nihon Hôigaku Zassi [Japanese Journal of Forensic Medicine] 18, 107-114.

Hanihara, T. (1993). Craniofacial features of Southeast Asians and Jomonese: a reconsideration of their microevolution since the late Pleistocene. Anthropological Science, 101, 25-46.

Hanihara, T. (1993). Population prehistory of East Asia and the Pacific as viewed from craniofacial morphology: the basic populations in East Asia, VII. American Journal of Physical Anthropology, 91, 173-187.

Hanihara, T. (1996). Comparison of craniofacial features of major human groups. American Journal of Physical Anthropology, 99, 389-412.

Hanihara, T. (1997). Craniofacial affinities of Mariana Islanders and Circum-Pacific peoples. American Journal of Physical Anthropology, 104, 411-425.

Hanihara, T. (2000). Frontal and facial flatness of major human populations. American Journal of Physical Anthropology, 111, 105-134.

Hanihara, T. (2006). Interpretation of craniofacial variation and diversification of East and Southeast Asians. In M. Oxenham & N. Tayles (Eds.), Bioarchaeology of Southeast Asia (pp. 91-111). Cambridge: Cambridge University Press.

Hanihara, T. (2008). Hanihara Craniometric Dataset.

Hanihara, T. (2008). Morphological Variation of Major Human Populations Based on Nonmetric Dental traits. American Journal of Physical Anthropology, 136, 169-182.

Hanihara, T. (2009). Geographic structure of craniometric variation and the estimates of possible dispersal routes of major human populations. In C L Mahoney & D A Springer (Eds.), Genetic Diversity (pp. 247-269). New York: Nova Science Publisher.

Hanihara, T. (2010). Estimate of divergence times for modern human populations based on craniometric variation. In K. Weiss (Ed.), Trends in Physical Anthropology - Focus on Civilizations and Cultures (pp. 153-174). New York: Nova Science Publisher, Inc.

Hanihara, T., & Ishida, H. (2001). Os incae: variation in frequency in major human population groups. Journal of Anatomy, 198, 137-152.

Hanihara, T., & Ishida, H. (2009). Regional differences in craniofacial diversity and the population history of Jomon Japan. American Journal of Physical Anthropology, 139, 311-322.

Hanihara, T., Ishida, H., & Dodo, Y. (1998). Place of the Hokkaido Ainu (northern Japan) among circumpolar and other peoples of the world: a comparison of the frequency variations of discrete cranial traits. International Journal of Circumpolar Health 57, 257-275.

Hanihara, T., Ishida, H., & Dodo, Y. (2003). Characterization of biological diversity through analysis of discrete cranial traits. American Journal of Physical Anthropology, 121, 241-251.

Hanihara, T., Yoshida, K., & Ishida, H. (2008). Craniometric variation of the Ainu: an assessment of differential gene flow from Northeast Asia into northern Japan, Hokkaido. American Journal of Physical Anthropology 137, 283-293.

Hansen, P., & Jaumard, B. (1997). Cluster analysis and mathematical programming. Math. Programming., 79(1-3 Ser. B), 191-215.

Hanson, H. (Writer). (2006-2011). Bones: Los Angeles: Far Field Productions and Josephson Entertainment in association with 20th Century Fox Television.

Harding, J. P. (1948). The use of probability paper for the graphical analysis of polymodal frequency distributions. Journal of Marine Biological Association, 28, 141-153.

Harel, D., & Koren, Y. (2004). Graph Drawing by High-Dimensional Embedding. Journal of Graph Algorithms and Applications, 8(2), 195-214.

Harell, F. E., & Lee, K. I. (1985). A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. In P. K. Sen (Ed.), Biostatistics: Statistics in biomedical, public health and environmental sciences (pp. 333-343). Amsterdam: Elsevier Science.

Harigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society Series C, 28(1), 100-108.

Harley, H. (1958). Maximum likelihood estimation from incomplete data. Biometrics, 14, 174-194.

Harmon, A. (2008, March 04). Gene Map Becomes a Luxury Item, The New York Times.

Harpending, H. C. (2000). Review of taboo: why black athletes dominate sports and why we're afraid to talk about it. Population and Environment, 22, 83-89.

Harpending, H., & Rogers, A.R. . (2000). Genetic perspectives on human origins and differentiation. Annual Review of Genomics and Human Genetics, 1(361-385).

Harris, H. (1969). Enzyme and Protein Polymorphism in Human Populations. British Medical Bulletin, 25, 5-13.

Harris, H., & Hopkinson, D. A. (1972). Average heterozygosity per locus in man: an estimate based on the incidence of enzyme polymorphisms. Annals of Human Genetics, 36, 9-20.

Harris, M. (1968). The rise of anthropological theory. New York, NY: Crowell and Company.

Harrison, S. (2006). Skull trophies of the Pacific war: transgressive objects of remembrance. Journal of the Royal Anthropological Institute, 12, 817-836.

Hart, D., & Ashmore, P. (2006). Changing students understanding of race. Anthropology News, 47(7), 10-11.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In L. M. Le Cam & R. A. Olshen (Eds.), In Proceedings of the Berkeley Conference in Honor of Jerzy Neymnani and Jack Kiefer (2nd ed., pp. 807-810). Monterey, C. A.: Wadsworth.

Hartigan, J. A. (1985). Statistical theory in clustering. Journal of Classification, 2(63), 76.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society, Series C (Applied Statistics), 28(1), 100-108.

Hartl, D. L. (2000). A Primer of Population Genetics (Vol. 3rd). Sunderland, MA: Sinauer Associates, Inc.

Hartl, D. L., & Clark, A. G. (1997). Principles of Population Genetics (Vol. 3). Sunderland, MA: Sinauer Associates.

Harvati, K., & Weaver, T. D. (2006). Human cranial anatomy and the differential preservation of population history and climate signatures. Anatomical Record, 228A, 1225-1233.

Harvati, K., & Weaver, T. D. (2006). Reliability of cranial morphology in reconstructing Neanderthal phylogeny. In K. Harvati & T. L. Harrison (Eds.), Neanderthals revisited: new approaches and perspectives (pp. 239-254). Berlin: Springer.

Haslanger, S. (2005). You mixed? Racial identity without racial biology. In S. Haslanger & C. Witt (Eds.), Adoption matters: philosophical and feminist essays (pp. 265-290). Ithaca: Cornell University Press.

Hasselband, V. (1966). Estimation of parameters of for a mixture of normal distributions. Technometrics, 8, 431-444.

Hasselband, V. (1969). Estimation of finite mixture s of distributions from the exponential family. Journal of the American Statistical Association, 64, 1459-1471.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1), 97-109.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. The Annals of Statistics, 13(2), 795-800.

Haughton, D. (1997). Packages for estimating finite mixtures: a review. Applied Statistics, 51, 194-205.

Hawkins, D. M. (1981). A New Test for Multivariate Normality and Homoscedasticity. Technometrics, 23, 105-110.

Hawkins, D. M., Muller, M. W., & ten Krooden, J. A. (1982). Cluster Analysis. In Hawkins. D. M. (Ed.), Topics in Applied Multivariate Analysis (pp. 303-356). Cambridge: Cambridge University Press.

Healy, M. (2008, March, 24, 2008). Genetic-Testing Consumers Have Tools But Little Guidance, Los Angeles Times.

Heckert, N. A. (2001). Dataplot website at the National Institute of Standards and Technology Retrieved July 20, 2010, from http://www.itl.nist.gov/div898/software/dataplot.html/,

Heckhausen, J., & Schulz, R. (1995). A life-span theory of control. Psychological Review, 102, 284-304.

Hedrick, P. W. (1971). A New Approach to Measuring Genetic Similarity. Evolution, 25(2), 276-280.

Hefner, J. T. (2007). The Statistical Determination of Ancestry Using Cranial Nonmetric Traits. Ph. D., University of Florida, Gainesville.

Hefner, J. T. (2009). Nonmetric Cranial Traits: New Approaches for the Determination of Ancestry. Journal of Forensic Sciences, 54(5), 985-995.

Hefner, J. T., Emanovsky, P., Byrd, J., & Ousley, S. D. (2007, February 19-24). The Value of Education and Experience in the Determination of Ancestry. Paper presented at the Proceedings of the American Academy of Forensic Science, San Antonio, TX.

Hefner, J. T., Ousley, S. D., & Warren, M. (2004). An historical perspective on nonmetric skeletal variation: Hooton and the Harvard list. Paper presented at the 56th Annual Meeting of the American Academy of Forensic Sciences, Dallas, TX.

Hefner, Joseph T. (2009). Cranial Nonmetric Variation and Estimating Ancestry. Journal of Forensic Sciences, 54(5), 985-995.

Henery, R. J. (1994). Classification. In D. Michie, D. J. Spiegelhalter & C. C. Taylor (Eds.), Machine Learning, Neural and Statistical Classification (pp. 6-16). New York: Ellis Horwood.

Hennessy, R. J., & Stringer, C. B. (2002). Geometric morphometric study of the regional variation of modern human craniofacial form. American Journal of Physical Anthropology, 117, 37-48.

Hennig, C. (2000). Identifiability of Models for Clusterwise Linear Regression. Journal of Classification, 17, 273-296.

Henschen, F. (1966). The human skull : a cultural history (S. Thomas, Trans.). London: Thomas and Hudson.

Herrera, K.J., Somarelli, J.A., Lowery, R.K. , & Herrera, R.J. . (2009). To what extent did Neanderthals and modern humans interact? . Biology Reviews, 84, 245-257.

Herrmann, N. P. (2002). Biological affinities of archaic period populations from west-central Kentucky and Tennessee. Ph.D., University Tennessee, Knoxville.

Herskovits, M. J. (1943). 1943 Franz Boas as Physical Anthropologist. In A. E. Kroeber (Ed.), American Anthropological Association Memoirs (Vol. 61, pp. 39-51). Washington DC: American Anthropological Association Memoir.

Hickman, C. B. (1997). The Devil and the One Drop Rule: Racial Categories, African Americans, and the U.S. Census. Michigan Law Review, 95(5), 1161-1265.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., . . . Cox, D. R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science, 307, 1072-1079.

Hinkes, M. J. (1993). Realities of Racial Determination in a Forensic Setting In C. C. Gordon (Ed.), Race, Ethnicity, and Forensic Anthropology (pp. 48-53). New York: Wiley.

Hirschfeld, L. (1996). Race in the Making: Cognition, Culture, and the Child's Construction of Human Kinds. Cambridge, Mass.: MIT. Press.

Hlusko, L. (2004). Integrating the genotype and phenotype in hominid paleontology. Proceedings of the National Academy of Science, 101, 2653-2657.

Ho, A. K., Sidanius, J., Levin, D. T., & Banaji, M. (2011). Evidence for hypodescent and racial hierarchy in the categorization and perception of biracial individuals. Journal of Personality and Social Psychology, 100, 492-506.

Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G., & McKeigue, P. M. (2004). Design and analysis of admixture mapping studies. American Journal of Human Genetics, 74, 965-978.

Holden, C. (2002). Going Head-to-Head over Boas's Data. Science, 298(5595), 942-945.

Holden, J., & Kelley, K. (2010). Effects of Initially Misclassified Data on the Effectiveness of Discriminant Function Analysis and Finite Mixture Modeling. Educational and Psychological Measurement 70(1), 36-55.

Holloway, R. L. (2002). Head to Head With Boas: Did He Err on the Plasticity of Head Form? Proceedings of the National Academy of Sciences, 99(23), 14622-14623.

Hooton, E. A. (1926). Progress in the Study of Race Mixtures with Special Reference to Work Carried on at Harvard University. Proceedings of the American Philosophical Society, 65(4), 312-325.

Hooton, E. A. (1930). The Indians of Pecos Pueblo: A Study of Their Skeletal Remains Papers of the Southwestern Expedition No. 4. New Haven, Conn.: Yale University Press.

Hooton, E. A. (1931). Up from the ape (2nd ed.). New York, NY: MacMillan.

Hooton, E. A. (1939). Crime and the Man. Cambridge, MA: Harvard University Press.

Hooton, E. A. (1939). The American Criminal: An Anthropological Study. Cambridge, Mass.: Harvard University Press.

Hooton, E. A. (1939). Twilight of Man. New York: C. P. Putnam and Sons.

Hooton, E. A. (1940). Why men behave like apes and vice versa. Princeton, NJ: Princeton University Press.

Hooton, E. A., & Dupertuis, C. W. (1951). Age changes and selective survival in Irish males. Ann Arbor, MI: American Association of Physical Anthropologists.

Hooton, E. A., Dawson, H., & Dupertuis, C. W. (1955). The physical anthropology of Ireland. Cambridge, Mass.: The Museum.

Horowitz, S., & Osborne, R. (1971). The genetic aspects of cranio-facial growth. In R. E. Moyers & W. M. Krogman (Eds.), Cranio-Facial growth in man (pp. 183-192). New York, NY: Pergamon.

Horowitz, S., & Osborne, R. H. (1971). The genetic aspects of cranio-facial growth. In R. E. Moyers & W. M. Krogman (Eds.), Craniofacial Growth in Man (pp. 183-192). Oxford: Pergamon.

Howell, F.C. (1952). Pleistocene glacial ecology and the evolution of "classic" Neandertal man. Southwest Journal of Anthropology 8, 377-410.

Howells, M. (1999). The Changing Face of Genealogy. Ancestry Magazine, 17.

Howells, W. W. (1937). The designation of the principal anthropometric landmarks on the head and skull. American Journal of Physical Anthropology, 22, 477-494.

Howells, W. W. (1966). The Jomon population of Japan: a study by discriminant analysis of Japanese and Ainu crania (Vol. 57). Cambridge, MA: Harvard University.

Howells, W. W. (1969). The use of multivariate techniques in the study of skeletal populations. American Journal of Physical Anthropology, 31, 311-314.

Howells, W. W. (1971). Application of Multivariate Analysis in Cranio-Facial Growth. In R. E. Moyers & W. M. Krogman (Eds.), Cranio-facial growth in man (pp. 209-218). New York, NY: Pergamon.

Howells, W. W. (1972). Analysis of patterns of variation in crania of recent man. In R. Tuttle (Ed.), The functional and evolutionary biology of primates (pp. 123-151). Chicago, IL: Aldine-Atherton Inc. .

Howells, W. W. (1973). Cranial Variation in man Papers of the Peabody Museum of Archeology and Ethnology (Vol. 67, pp. 1-259). Cambridge, MA: Harvard University.

Howells, W. W. (1973). Cranial variation in man: a study by multivariate analysis of pattern of difference among recent human populations Papers of the Peabody Museum (Vol. 67, pp. 1-259). Cambridge, Mass.: Peabody Museum, Harvard University.

Howells, W. W. (1973). Cranial Variation in Man: A Study by Multivariate Analysis of Patterns of Difference Among Recent Human Populations. Cambridge, Mass.: Peabody Museum of Archeology and Ethnology.

Howells, W. W. (1984). Introduction To Multivariate Statistical Methods. In G. N. Van Vark & W. W. Howells (Eds.), Multivariate Statistical Methods In Physical Anthropology (pp. 1-12). Dordrecht: D. Reidel Publishing Company.

Howells, W. W. (1989). Skull Shape and the Map: Craniometric Analyses in the Dispersion of Modern Homo. Cambridge, Mass.: Harvard University Press.

Howells, W. W. (1989). Skull shapes and the map. Craniometric analyses in the dispersion of modern Homo Papers of the Peabody Museum of Archeology and Ethnology (Vol. 79, pp. 1-189). Cambridge, MA: Harvard University.

Howells, W. W. (1995). Who's Who in Skulls. Ethnic Identification of Crania from Measurements. Papers of the Peabody Museum of Archeology and Ethnology (Vol. 82). Cambridge, Mass.: Peabody Museum.

Howells, W. W. (1996). Howells' Craniometric Data on the Internet. American Journal of Physical Anthropology, 101, 441-442.

Howells, W. W. (2010). Howells Craniometric Data. In http://konig.la.utk.edu/howells.htm (Ed.).

Hrdlicka, A. (1907). The Skeletal Remains Suggestion or Attributed to Early Man in North America Smithsonian Institution, Bureau of American Ethnology, Bulletin 33. Washington DC: Government Printing Office.

Hrdlicka, A. (1918). Physical Anthropology: its scope and aims; its history and present status in America B. History. American Journal of Physical Anthropology, 1, 133-182.

Hrdlicka, A. (1918). Physical Anthropology: its scope and aims; its history and present status in America. A. Physical anthropology, its scope and aims. American Journal of Physical Anthropology, 1, 3-23.

Hrdlicka, A. (1918). Physical Anthropology: its scope and aims; its history and present status in America. C. Recent history and present status of the science in North America. American Journal of Physical Anthropology, 1, 267-304; 377-414.

Hrdlicka, A. (1920). Shovel-Shaped Teeth. American Journal of Physical Anthropology, 3, 187-196.

Hrdlicka, A. (1927). Anthropology of the American negro. Historical notes. American Journal of Physical Anthropology, 10, 205-233.

Hu, Y. F. (2006). Efficient, High-Quality Force-Directed Graph Drawing. The Mathematical Journal 10(1), 37-71.

Hubbard, R., & Wald, E. (1993). Exploding the gene myth. Boston: Beacon Press.

Hubbe, M., & Neves, W. A. (2007). On the misclassification of human crania. Current Anthropology, 48, 285-288.

Hubbe, M., Hanihara, T., & Harvati, K. (2009). Climate signatures in the morphological differentiation of worldwide modern human populations. The Anatomical Record 292, 1720-1733.

Huberty, C. J. (1994). Applied discriminant analysis. New York: John Wiley and Sons.

Hubisz, M. J., Falush, D., Stephens, M., & Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information Molecular Ecology Resources 9, 1322-1332.

Hublin, J.-J. . (1998). Climatic changes, paleogeography, and the evolution of the Neandertals. In T. Akazawa, K. Aoki & O. Bar-Yosef (Eds.), Neandertals and modern humans in Western Asia (pp. 295-310). New York: Plenum Press.

Huelsenbeck, J. P., & Andolfatto, P. (2007). Inference of population structure under a Dirichlet process model. Genetics, 175, 1787-1802.

Huges, C., Juarez, C., Zephro, L., Fowler, G., & Chacon, S. (2010). Past or present? Differentiating California prehistoric native American remains from forensic cases: An empirical approach. International Journal of Osteoarchaeology (online preview), DOI: 10.1002/oa.1192.

Hull, D. (1998). Species, subspecies, and races. Sociological Research, 65, 351-367.

Humanos, Coalición de Derechos. (2010, September). List of Recovered Remains for Arizona-Mexico Border October 2009 - September 2010 Retrieved March 23, 2011, from http://www.derechoshumanosaz.net/images/pdfs/list%20of%20recovered%20remains%202009-2010%20%28web%20page%29.pdf

Humes, K. R., Jones, N. A., & Ramirez, R. R. (2011). Overview of Race and Hispanic Origin: 2010. Retrieved from http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf.

Hunley, K. L., Healy, M. E., & Long, J. C. (2009). The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: Implications for biological race. American Journal of Physical Anthropology, 139(35-46).

Hunt, E. E., Jr. (1959). Anthropometry, genetics and racial history. American Anthropologist, 61, 64-87.

Hunt, E. E., Jr. (1981). The old physical anthropology. American Journal of Physical Anthropology, 56, 339-346.

Hunt, L., & Jorgensen, J. (1999). Australian & New Zealand Journal of Statistics. Mixture model clustering using the Multimix program, 41, 153-171.

Hunter, D. R., Wang, S. , & Hettmansperger, T. P. . (2007). Inference for Mixtures of Symmetric Distributions The Annals of Statistics 35(1), 224-251.

Huxley, J. (1938). Clines: An auxiliary taxonomic principle. Nature, 142(47).

Huxley, J. (1942). Evolution, the Modern Synthesis. London: Allen and Unwin.

Huxley, J. S. (1940). The New Systematics: Oxford University Press.

Huxley, J. S. (1942). Evolution: the modern synthesis: The MIT Press.

Irish, J.D. (2006). Who were the ancient Egyptians? Dental affinities among Neolithic through post-dynastic peoples. American Journal of Physical Anthropology 129, 529-543.

Iscan, Y. M. (1988). Rise of Forensic Anthropology. Yearbook of Physical Anthropology, 31, 203-230.

Ishida, H. (1992). Flatness of facial skeletons in Siberian and other circum-Pacific populations. Zeitschrift fur Morphologie und Anthropologie, 79(1), 53-67.

Ishida, H., & Dodo, Y. (1993). Nonmetric cranial variation and the populational affinities of the Pacific peoples. American Journal of Physical Anthropology, 90, 49-57.

Ishida, H., Hanihara, T., Kondo, O., & Fukumine, T. (2009). Craniometric divergence history of the Japanese populations. Anthropological Science, 117, 147-156.

Jablonski, N. G., & Chaplin, G. (2000). The evolution of human skin coloration. Journal of Human Evolution, 39, 57-106.

Jackson, J. P. (2001). "In ways unacademical": the reception of Carleton S. Coon's The Origin of Races. Journal of the History of Biology, 34, 247-285.

Jacobs, K. (1994). Human dento-gnathic metric variation in Mesolithic/Neolithic Ukraine-possible evidence of demic diffusion in the Dnieper Rapids region. American Journal of Physical Anthropology, 95, 1-16.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Englewood Cliff, NJ: Prentice Hall.

Jajuga, K., Sokolowski, A., & Bock, H. H. (Eds.). (2002). Classification, Clustering and Data Analysis. Recent Advances and Applications. Berlin: Springer.

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics, 23, 1801-1806.

Jakobsson, M., Scholz, S.W., Scheet, P, Gibbs, J.R., VanLiere, J.M., Fung, H.C., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature, 451, 998-1003.

James, M. (2008). "Race". The Stanford Encyclopedia of Philosophy from http://plato.stanford.edu/archives/fall2008/entries/race/

Jamshidian, M., & Jennrich, R. I. (1993). Conjugate gradient acceleration of the EM algorithm. Journal of the American Statistical Association, 88, 221-228.

Jamshidian, M., & Jennrich, R. I. (1997). Acceleration of the EM algorithm using quasi-Newton methods. Journal of the Royal Statistical Society Series B, 59(3), 596-587.

Jansen, R. C. (1993). Maximum-likelihood in a generalized linear finite mixture model by using the EM algorithm. Biometrics, 49(1), 227-231.

Jantz, R. L. (2001). Cranial change in Americans: 1850–1975. Journal of Forensic Sciences, 46(784-787).

Jantz, R. L. (2003). The anthropometric legacy of Franz Boas. Economics and Human Biology, 1(2), 277-284.

Jantz, R. L. (2004). The Meaning And Consequences Of Morphological Variation. Paper presented at the American Anthropological Association (AAA) 2003 Annual Meeting Chicago, IL. http://understandingrace.com/resources/pdf/myth_reality/jantz.pdf

Jantz, R. L. (2008). Forensic Anthropology Data Bank (FDB): Forensic Anthropology Center.

Jantz, R. L., & Meadows Jantz, L. (2000). Secular change in craniofacial morphology. American Journal of Human Biology, 12, 327-338.

Jantz, R. L., & Moore-Jansen, P. H. (1998). A database for forensic anthropology: Structure, content and analysis Report of Investigations No. 47. Knoxville, TN: Department of Anthropology, University of Tennessee.

Jantz, R. L., & Ousley, S. D. (1992). FORDISC 1: Computerized Forensic Discriminant Functions (Version 1.0). Knoxville, TN: The University of Tennessee.

Jantz, R. L., & Ousley, S. D. (2008). The Uses and Abuses of Statistics and FORDISC in Forensic Anthropology. Paper presented at the 60th Annual Meeting of the American Academy of Forensic Sciences, Washington DC.

Jantz, R. L., & Owsley, D. W. (2001). Variation among early North American Crania. American Journal of Physical Anthropology, 114, 146-155.

Jantz, R. L., & Owsley, D. W. (2007). FORDISC 3.0: Theory, Methods and Applications. Paper presented at the American Academy of Forensics Sciences Annual Conference, San Antonio, TX.

Jantz, R. L., & Spencer, F. (1997). Boas, Franz (1858-1942). In F. Spencer (Ed.), History of physical anthropology: an encyclopedia (Vol. 1, pp. 186-190). New York, NY: Garland Publishing.

Jantz, R. L., Kimmerle, E. H., & Baraybar, J. P. (2008). Sexing and Stature Estimation Criteria for Balkan Populations. Journal of Forensic Sciences, 53(3), 601-605.

Jantz, RL. (2008). Forensic Anthropology Data Bank (FDB) Craniometric Dataset.

Jeffreys, H. (1932). An alternative to the rejection of observations. Proceedings of the Royal Society of London Series A, 137, 78-87.

Jeffreys, H. (1961). Theory of probability (Vol. 3rd): Oxford University Press.

Jiggins, C. D., Naisbit, R. E., Coe, R. L., & Mallet, J. (2001). Reproductive isolation caused by colour pattern mimicry. Nature 411, 302-305.

Jobling, M., Hurles, M., & Tyler-Smith, C. (2004). Human evolutionary genetics: Origins, peoples & disease. New York: Garland.

Johnson, N. L., & Kotz, S. (1970). Continuous univarate distributions. New York: Houghton Mifflin.

Johnson, R. A., & Wichern, D. W. (1992). Applied multivariate statistical analysis. Englewood Cliffs, NJ: Prentice-Hall.

Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). New York: Springer.

Jordan, M. I., & Xu, L. (1995). Convergence results for the EM approach to mixtures-of-experts architectures. Neural Networks, 8(9), 1409-1431.

Jorde, L. B. (200). Linkage disequilibrium and the search for complex disease genes. Genome Research, 10, 1435-1444.

Jorde, L. B., & Wooding, S. P. (2004). Genetic variation, classification and 'race'. Nature Genetics, 36, S28-S33.

Jorde, L. B., Bamshad, M., & Rogers, A. R. (1998). Using mitochondrial and nuclear DNA markers to reconstruct human evolution. BioEssays, 20, 126-136.

Jorde, L. B., Watkins, W. S., Bamshad, M. J., Dixon, M. E., Ricker, C. E., Seielstad, M. T., & Batzer, M. A. (2000). The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data. American Journal of Human Genetics, 66, 979-988.

Jorde, L. B., Watkins, W. S., Kere, J., Nyman, D., & Eriksson, A. W. (2000). Gene mapping in isolated populations: new roles for old friends? Human Heredity, 50, 57-65.

Josephson, S. C., Juell, K. E., & Rogers, A. R. (1996). Estimating Sexual Dimorphism by Method-of-Moments. American Journal of Physical Anthropology, 100, 191-206.

Jungers, W. L., Falsetti, A. B., & Wall, C. E. (1995). Shape, relative size, and size-adjustments in morphometrics. American Journal of Physical Anthropology, 38(137-161).

Jurmain, R., Kilgore, L., & Trevathan, W. (2005). Introduction to physical anthropology. Toronto: Thomson Wadsworth.

Kaburgai, M., Ishida, H., Goto, M., & Hanihara, T. (2010). Comparative studies of the Ainu, their ancestors and neighbors: assessment based on metric and nonmetric dental data. Anthropological Science, 118, 95-106.

Kaessmann, H., Wiebe, V., Weiss, G., & S., Pääbo. (2001). Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nature Genetics, 27, 155-156.

Kahn, J. (2005). Misreading race and genomics after BiDil. Nature Reviews Genetics, 37(655-656).

Kahn, J. (2010). What's the use of race in presenting forensic DNA evidence in court? . In I. Whitmarsh & D. S. Jones (Eds.), What's the Use of Race? Modern Governance and the Biology of Difference (pp. 27-48). Cambridge, MA: MIT Press.

Kallenberg, O. (1997). Foundations of Modern Probability. New York: Springer-Verlag.

Kamada, T., & Kawai, S. (1989). An Algorithm for Drawing General Undirected Graphs. Information Processing Letters, 31(1), 7-15.

Kamakura, W. A., & Russell, G. J. (1989). A probabilistic choice model for market segmentation and elasticity structure. Journal of Marketing Research, 25(4), 379-390.

Kass, R. E., & Raftery, A. E. (1995). Bayesian factors. Journal of the American Statistical Association 90(430), 773-795.

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association 90, 928-934.

Katz, B. M., & McSweeney, M. (1979). Misclassification errors and categorical data analysis. Journal of Experimental Education, 47, 331-338.

Kaufman, L., & Rousseeuw, P. J. (2005). Finding Groups in Data: an Introduction to Cluster Analysis. New York: J. Wiley.

Kaufmann, M, & Wagner, D. (Eds.). (2001). Drawing graphs: methods and models (Vol. 2025). New York: Springer.

Kawakubo, Y., Hanihara, T., Shigematsu, M., & Dodo, Y. (2009). Interpretation of the craniometric variation of northeastern Japan, the Tohoku region. Anthropological Science, 117, 57-65.

Kay, R. F. (1982). Sexual dimorphism in Ramapithecinae. Proceedings of the National Academy of Science, 79, 209-212.

Kayser, M., & de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. Nature Reviews Genetics, 12(3), 179-192.

Keita, S, & RA, Kittles. (1997). The persistence of racial thinking and the myth of racial divergence. American Anthropology, 99, 534-544.

Keita, S. (2007). On Merotic Nubian Crania: Fordisc 2.0 and Human Biological History. Current Anthropology, 48, 425-427.

Keita, S. O. Y., Kittles, R. A., Royal, C. D. M., Bonney, G. E., Furbert-Harris, P., Dunston, G. M., & Rotimi, C. N. (2004). Conceptualizing Human Variation. Nature Genetics, 36, S17-S20.

Keita, S. O., & Kittles, R. A. (1997). The persistence of racial thinking and the myth of racial divergence. American Anthropologist, 99, 534-544.

Keita, S., & Boyce, A. J. (2001). "Race": Confusion about zoological and social taxonomies, and their places in science. American Journal of Human Biology, 12, 569-575.

Keita, S.O.Y. (1993). Studies and Comments on Ancient Egyptian Biological Relationships. History in Africa, 20, 129-154.

Keita, S.O.Y. (2004). Exploring northeast African metric craniofacial variation at the individual level: a comparative study using principal components analysis. American Journal of Human Biology, 16, 679-689.

Kemkes-Grottenthaler, A. (2002). Aging through the ages: historical perspectives on age indicator methods. In R. D. Hoppa & J. W. Vaupel (Eds.), Paleodemography: Age Distributions from Skeletal Samples (Vol. 48-72). Cambridge: Cambridge University Press.

Kenji, Y., Takeuchi, J., Williams, G., & Milne, P. (2004). On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. Data Mining and Knowledge Discovery, 8(3), 275-300.

Kennedy, K. A. R. (1976). Human Variation in space and time. Dubuque, IA: Wm. C. Brown.

Kennedy, K. A. R. (1984). Biological adaptations and affinities of Mesolithic South Asians. In J. R. Lukacs (Ed.), The People of South Asia (pp. 29-58). New York: Plenum Press.

Kennedy, K. A. R. (1995). "But Professor, Why Teach Race Identification If Races Don't Exist?". Journal of Forensic Sciences, 40(5), 797-800.

Keogh, B. K. (2005). Revisiting classification and identification. Learning Disability Quarterly, 28, 100-102.

Keribin, C. (1998). Consistent estimate of the order of mixture models. Comptes Rendus de l'Académie des Sciences, Série I-Mathématiques, 326, 243-248.

Keribin, C. (2000). Consistent estimation of the order of mixture models. Sankhya, 62, 49-66.

Kerley, E. R. (1978). Recent Developments in Forensic Anthropology. Yearbook of Physical Anthropology, 20, 160-173.

Keverberg, Baron d. (182). Notes [appended to Quetelet 1827]. Nouveaux memoires de l'Académie royale des sciences et belles-lettres de Bruxelles, 175-192.

Kevles, D. J. (1985). In the name of eugenics: genetics and the uses of human heredity. Berkeley: The University of California Press.

Kidd, K. K., Pakstis, A. J., Speed, W. C., & Kidd, J. R. (2004). Understanding Human DNA Sequence Variation. Journal of Heredity, 95(5), 406-420.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum-likelihood estimator in the presence of infinitely many incidental parameters. Annals of Mathematical Statisticsl, 27, 888-906.

Kimmerle, E. H., Jantz, R. L., Konigsberg, L. W., & Baraybar, J. P. (2008). Skeletal estimation and identification in American and East European populations. Journal of Forensic Sciences, 53, 524-532.

King, R. C., Stansfield, W. D., & Mulligan, P. K. (2006). A Dictionary of Genetics (Vol. 7th). New York: Oxford University Press.

Kirk, R.L. (1976). Serum protein and enzyme markers as indicators of population affinities in Australia and the Western Pacific. In R.L. Kirk & A.G. Thorne (Eds.), The Origin of the Australians (pp. 329-346). Canberra: Australian Institute of Aboriginal Studies.

Kitcher, P. (2007). Does 'race' have a future? . Philosophy and Public Affairs, 35(4), 293-317.

Kittles, R. A., & Weiss, K. M. (2003). Race, ancestry, and genes: implications for defining disease risk. Annual Review of Genomics and Human Genetics, 4, 33-67.

Kittles, R. A., Santos, E. R., Oji-Njideka, N. S., & Bonilla, C. (2007). Race, Skin Color and Genetic Ancestry: Implications for Biomedical Research on Health Disparities. California Journal of Health Promotion, 5, 9-23.

Klein, J., & Takahata, N. (2002). Where do we come from? The molecular evidence for human descent. Berlin: Springer.

Klein, R. (1989). The Human Career: Human Biological and Cultural Origins. Chicago: University of Chicago Press.

Klein, R.G. . (1999). The human career: human biological and cultural origins (2 ed.). Chicago: University of Chicago Press.

Klepinger, L. (2006). Fundamentals of Forensic Anthropology. New York: Wiley-Liss.

Kobyliansky, E., & Livshits, G. (1985). A morphological approach to the problem of the biological similarity of Jewish and non-Jewish populations. Annals of Human Biology, 12, 203-212.

Koehler, A. B., & Murphree, E. S. (1988). A comparison of the Akaike and Schwarz criteria for selecting model order. Applied Statistics, 37, 187-195.

Koenig, B, Lee, S, & Richardson, S (Eds.). (2008). Revisiting Race in a Genomic Age. New Brunswick, NJ: Rutgers University Press.

Koenig, B. A., Soo-Jin Lee, S., & Richardson, S. S. (2008). Introduction: Race and Genetics in a Genomic Age. In B. A. Koenig, S. Soo-Jin Lee & S. Richardson (Eds.), Revisiting Race in a Genomic Age (pp. 1-17). New Brunswick, NJ: Rutgers University Press.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Paper presented at the International Joint Conference on Artificial Intelligence, Montreal, QC.

Kohavi, R., & John, G. (1997). Wrappers for feature selection. Artificial Intelligence, 97(1-2), 273-324.

Komar, D. A., & Grivas, C. R. (2008). Manufactured populations: What do contemporary reference skeletal collections represent? A comparative study using the Maxwell Museum documented collection. American Journal of Physical Anthropology, 137, 224-233.

Komlos, J. (1994). Stature, living standards, and economic development: essays in anthropometric history. Chicago: University of Chicago Press.

Konigsberg, L. W., & Ousley, S. D. (1995). Multivariate quantitative genetics of anthropometric traits from the Boas data. Human Biology, 67, 481-498.

Konigsberg, L. W., Algee-Hewitt, B., & Steadman, D. W. (2009). Estimation and Evidence in Forensic Anthropology: Sex and Race. American Journal of Physical Anthropology, 139(1), 77–90.

Konishi, S., & Kitagawa, G. (1996). Generalized information criteria in model selection. Biometrika, 83(4), 875-890.

Konishi, S., & Kitagawa, G. (2008). Information Criteria and Statistical Modeling. New York: Springer.

Koren, Y., Carmel, L., & Harel, D. (2003). Drawing Huge Graphs by Algebraic Multigrid Optimization. Multiscale Modeling and Simulation, 1(4), 645-673.

Kosiba, S. (2000). Assessing the Efficacy and Pragmatism of "Race" Designation in Human Skeletal Identification: A Test of Fordisc 2.0 Program (Abstract). American Journal of Physical Anthropology Supplement, 30, 200.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data Preprocessing for Supervised Leaning. International Journal of Computer Science, 1(2), 111-117.

Kouchi, M. (2000). Brachycephalization in Japan has ceased. American Journal of Physical Anthropology, 112, 337-347.

Kouchi, M., & Kozumi, K. (1992). Secular change: confounding factor or causative factor of inter-population differences in modern Japanese crania? Journal of the Anthropological Society of Nippon, 100, 405-416.

Kramer, A., & Konigsberg, L. W. (1999). Recognizing species diversity among large-bodied hominoids: a simulation test using missing data finite mixture analysis. Journal of Human Evolution, 36, 409-421.

Kramer, A., Algee-Hewitt, B., A., Dautartas., Driscoll, C., Eleazer, K., Godde, K., & Hulsey, B. (2009). Finite questions and fuzzy answers: exploring Neandertal/human admixture through baboon analogues. Paper presented at the Annual Meeting of the Paleoanthropological Society, Chicago, IL.

Kramer, A., Donnelly, S. M., Kidder, J. H., Ousley, S. D., & Olah, S. M. (1995). Craniometric Variation in Large-Bodied Hominoids - Testing the Single-Species Hypothesis for Homo-Habilis. Journal of Human Evolution, 29, 443-462.

Kroeber, A. I. (1923). Anthropology. New York, NY: Harcourt, Brace.

Krogman, W. (1976). Fifty years of physical anthropology: the men, the materials, the concepts, and the methods. Annual Reviews of Anthropology, 5, 1-14.

Krogman, W. M. (1943). What we do not know about race. Scientific Monthly, 57, 97-104.

Kruskal, J. B., & Wish, M. (1978). Multidimensional Scaling. London: Sage Press.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. Sociological methods and research, 33(2), 188-229.

Kuhn, T. (1962). The structure of scientific revolutions. Chicago: University of Chicago Press.

Kuiper, F. K., & Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. Biometrics, 31, 777-783.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 22(1), 79-86.

Kuznar, L. (1997). Reclaiming a Scientific Anthropology. Walnut Creek, CA: AltaMira Press.

Lachenbruch, P. A. (1966). Discriminant analysis when the initial samples are misclassified. Technometrics, 8, 657-662.

Lachenbruch, P. A. (1974). Discriminant analysis when the initial samples are misclassified II: Non-random misclassification models. Technometrics, 16, 419-424.

Lachenbruch, P. A. (1979). Note on initial misclassification effects on the quadratic discriminant function. Technometrics, 21, 129-132.

Lahr, M. M. (1996). The evolution of modern human diversity: a study of cranial variation. Cambridge: Cambridge University Press.

Lahr, M. M., & Wright, R. V. S. (1996). The question of robusticity and the relationship between cranial size and shape in Homo sapiens. Journal of Human Evolution, 31(2), 157-191.

Lahr, M.M. (1994). The multiregional model of modern human origins: A reassessment of its morphological basis. Journal of Human Evolution, 26, 23-56.

Lahr, M.M., & Foley, R. (1994). Multiple dispersals and modern human origins. Evolutionary Anthropology, 3, 48-60.

Laitine, E. K. (2007). Classification accuracy and correlation: LDA in failure prediction European Journal of Operational Research 183(1), 210-225.

Langley-Shirley, N., & Jantz, R. L. (2010). A Bayesian approach to age estimation in modern Americans from the clavicle. Journal of Forensic Sciences, 55, 571-583.

Lao, O., van Duijn, K., Kersbergen, P. , de Knijff, P., & Kayser, M. (2006). Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. American Journal of Human Genetics, 78, 680-690.

Larose, DT. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. London: Wiley.

Larsen, C. S. (1997). Bioarchaeology. Cambridge: Cambridge University Press.

Larsen, C. S. (1997). Bioarchaeology: Interpreting Behavior for the Human Skeleton. Cambridge: Cambridge University Press.

Larsen, C. S. (2002). Bioarchaeology: the lives and lifestyles of past people. Journal of Archaeological Research, 10(2), 119-165.

Lasker, G. W. (1970). Physical Anthropology: The Search for General Processes and Principles. American Anthropologist, 72(1), 1-8.

Latter, B. D. H. (1980). Genetic differences within and between populations of the major human subgroups. The American Naturalist, 116, 220-237.

LaViest, T. A. (1996). Why we should continue to study race…but do a better job: an essay on race, racism, and health. Ethnicity and Disease, 6, 21-29.

LaViest, T. A. (Ed.). (2002). Race, ethnicity, and health. San Francisco: Jossey-Bass.

Lazarsfeld, P. F. (1950). The interpretation and computation of some latent structures. In S. A. Stouffer (Ed.), Measurement and prediction, the American soldier: studies in social psychology in World War II Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In S. A. Stouffer (Ed.), Measurement and prediction, the American soldier: studies in social psychology in World War II (Vol. 4, pp. 362-412). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). Latent Structure Analysis. New York: Houghton Mifflin.

Le Blanc, S. A., & Black, B. (1974). A long term trend in tooth size in the eastern Mediterranean. American Journal of Physical Anthropology, 41, 417-422.

Leach, S., Lewis, M., Chenery, C., Müldner, G, & Eckardt, H. (2009). Migration and diversity in Roman Britain: A multidisciplinary approach to the identification of immigrants in Roman York, England. American Journal of Physical Anthropology, 140, 546-561.

Leathers, A., Edwards, J., & Armelagos, G. J. (2002). Assessment of classification of crania using Fordisc 2.0: Nubian X-Group Test. American Journal of Physical Anthropology, 99-100(Suppl. 34).

Lecis, R., Pierpaoli, M., & Biro, Z. S. . (2006). Bayesian analyses of admixture in wild and domestic cats (Felis silvestris) using linked microsatellite loci. Molecular Ecology, 15, 119-131.

Lee, A. (1901). Data for the Problem of Evolution in Man. VI. A First Study of the Correlation of the Human Skull. Philosophical Transactions of the Royal Society of London A, 196, 225-264.

Lee, D., & Preparata, F. P. (1984). Computational Geometry: A Survey. IEEE Transactions on Computers, 33(12), 1072-1101.

Lee, S. S., Mountain, J. L., & Koenig, B. A. . (2001). The meanings of "race"in the new genomics: implications for health disparities research. Yale Journal of Heath Policy, Law and Ethics, 1, 33-71.

Lei, P. W., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. Journal of Experimental Education, 72, 25-49.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. Journal of Statistical Software, 11(8), 1-18.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. Annals of Statistics, 20, 1350-1360.

Lett, J. W. (1997). Science, Reason, and Anthropology: The Principles of Rational Inquiry. New York, NY: Rowman & Littlefield.

Leutenegger, W., & Shell, B. (1987). Variability and sexual dimorphism in canine size of Australopithecus and extant hominoids. Journal of Human Evolution, 16, 359-367.

Lévi–Strauss, C. (1952). Race et histoire. Paris: Gonthier.

Lewin, R. (1999). Human Evolution: An Illustrated Introduction. Oxford: Blackwell Science.

Lewis, H. S. (2001). Boas, Darwin, Science, and Anthropology. Current Anthropology, 42(3), 381-406.

Lewontin, R. C. (1967). An estimate of average heterozygosity in man. American Journal of Human Genetics, 19, 681-685.

Lewontin, R. C. (1972). The apportionment of human diversity. Evolutionary Biology, 6, 381-398.

Lewontin, R. C. (1972). The apportionment of human diversity. In T. Dobzhansky, M. K. Hecht & W. C. Steere (Eds.), Evolutionary Biology 6 (pp. 381-398). New York: Appleton Century Crofts.

Lewontin, R. C. (1974). The Genetic Basis of Evolutionary Change. New York: Columbia University Press.

Lewontin, R. C. (1977). Genetic Diversity. Pediatrics, 59(3), 432.

Lewontin, R. C. (1978). Single- and Multiple-Locus Measures of Genetic Distance between Groups The American Naturalist, 112(988), 1138-1139.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., . . . Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science, 319, 1100-1104.

Li, W. H., & Sadler, L. A. (1991). Low nucleotide diversity in man. Genetics, 129, 513-523.

Lieberman, D. E., McBratney, B. M., & Krovitz, G. (2002). The evolution and development of cranial form in Homo sapiens. Proceedings of the National Academy of Sciences of the United States of America, 99(3), 1134-1139.

Lieberman, D. E., Pilbeam, D. R., & Wood, B. A. (1988). A probabilistic approach to the problem of sexual dimorphism in Homo habilis: A comparison of KNM-ER 1470 and KNM-ER 1813. Journal of Human Evolution, 17(5), 503-511.

Lieberman, L. (2001). How "Caucasoids" Got Such Big Crania and Why They Shrank. Current Anthropology, 42(1), 69-95.

Lieberman, L., & Jackson, F. L. C. (1995). Race and three models of human origin. American Anthropologist, 97, 231-242.

Lieberman, L., & Kirk, R. C. (2004). What Should We Teach about the Concept of Race? Anthropology & Education Quarterly, 35(1), 137-145.

Lieberman, L., & Reynolds, L. T. (1996). Race: The deconstruction of a scientific concept. In L. T. Reynolds & L. Lieberman (Eds.), Rae and other misadventures: Essays in honor of Ashley Montagu in his ninetieth year (pp. 142-173). Dix Hills, NY: General Hall.

Lieberman, L., Corcoran, M., Kirk, R., & Watterson-O'Neil, C. (2005). Are Physical Anthropology Textbooks Color-Blind? Transforming Anthropology, 13, 92-102.

Lieberman, L., Kirk, R. C., & Littlefield, A. (2003). Perishing paradigm: race-1931-99. American Anthropologist, 105, 110-113.

Lieberman, L., Stevenson, B. W., & Reynolds, L. T. (1989). Race and Anthropology: A Core Concept without Consensus. Anthropology & Education Quarterly, 20, 67-73.

Lin, T. C., & Pourhmadi, M. (1998). Nonparametric and non-linear models and data mining in time series: a case study on the Canadian lynx data. Applied Statistics, 47, 187-201.

Lindee, S. M., Goodman, A. H., & Heath, D. . (2003). Anthropology in an Age of Genetics. In S. M. Lindee, A. H. Goodman & D Heath (Eds.), Genetic Nature / Culture (pp. 1-19). Berkeley, CA: University of California Press.

Lindsay, B. G. (1983). The geometry of mixture likelihoods, Part II: The exponential family. The Annals of Statistics, 11(3), 783-792.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. The Annals of Statistics, 11(1), 86-94.

Lindsay, B. G. (1995). Mixture models: Theory, geometry, and applications (Vol. 5). Hayward, CA: Institute for Mathematical Statistics.

Lindstrom, M., & Bates, D. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. Journal of the American Statistical Association, 83(1014), 1022.

Linz, B., Balloux, F., Moodley, Y., et al. (2007). An African origin for the intimate association between humans and Helicobacter pylori. Nature, 445, 915-918.

Lisker, R., Pérez-Briceño, R., Granados, J., & Babinsky, V. (1988). Gene frequencies and admixture estimates in the state of Puebla, Mexico. American Journal of Physical Anthropology, 76, 331-335.

Little, B. B., Buschang, P. H., Reyes, M. E. P., Tan, S. K., & Malina, R. M. (2006). Craniofacial dimensions in children in rural Oaxaca, southern Mexico: secular change, 1968–2000. American Journal of Physical Anthropology, 131, 127-136.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83, 1198-1202.

Little, R. J. A., & Rubin, D B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika, 81, 633-648.

Little, R. J. A., & Rubin, D B. (2002). Statistical Analysis with Missing Data (2nd ed.). New York: Wiley.

Little, R. J. A., & Schluchter, M. D. (1985). Maximum likelihood estimation from mixed continuous and categorical data with missing values. Biometrika, 72, 497-512.

Littlefield, A., Lieberman, L., Reynolds, L. T., Azevedo, E. S., Beals, K. L., Brace, C. L., . . . Wolfe, L. D. (1982). Redefining Race: The Potential Demise of a Concept in Physical Anthropology [and Comments and Reply]. Current Anthropology, 23(6), 641-655.

Liu, H., Prugnolle, F., Manica, A., & et al. (2006). A geographically explicit genetic model of worldwide human settlement history. American Journal of Human Genetics, 79, 203-237.

Liu, R. J. A., & Rubin, D B. (1987). Statistical analysis with missing data. New York: Wiley.

Livingstone, F. B. (1958). Anthropological implications of sickle cell distribution in West Africa. American Anthropologist, 30, 533-562.

Livingstone, F. B. (1962). On the non-existence of human races. Current Anthropology, 3, 279-281.

illegal Immigration: Border-Crossing Deaths Have Doubled Since 1995; Border Patrol's Efforts to Prevent Deaths Have Not Been Fully Evaluated. (2006). GAO-06-770 Retrieved from http://www.gao.gov/new.items/d06770.pdf.

Lo, Y. (2005). Likelihood Ratio Tests of the Number of Components in a Normal Mixture with Unequal Variances. statistics & Probability Letters, 71, 225-235.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the Number of Components in a Normal Mixture. Biometrika, 88, 767-778.

Lockwood, C. A., Richmond, B. G., Jungers, W. L., & Kimbel, W. H. (1996). Randomization procedures and sexual dimorphism in Australopithecus afarensis. Journal of Human Evolution, 31(6), 537-548.

Logan, M. H., & Jantz, R. L. (2010). Why does head form change in children of immigrants? A reappraisal. American Journal of Human Biology, 22, 702-707.

Lombroso, C. (1895). L'Homme Criminel. Felix, Alcan.

Lombroso, C. (1897). L'uomo delinquente (Vol. 3). Horpli, Milan.

Londin, E. R., Keller, M. A., Maista, C., Smith, G., Mamounas, L. A., Zhang, R., . . . Corriveau, R. A. (2010). CoAIMs: A Cost-Effective Panel of Ancestry Informative Markers for Determining Continental Origins. PloS ONE, 5(10), e13443.

Long, J. C., & Kittles, R. A. (2003). Human Genetic Diversity and the Nonexistence of Biological Races. Human Biology, 75(4), 449-471.

Long, J. C., Li, J., & Healy, M. E. (2009). Human DNA sequences: more variation and less race. American Journal of Physical Anthropology, 139, 23-34.

López Herráez, D. , Bauchet, M. , Tang, K. , Theunert, C. , Pugach, I., et al. (2009). Genetic Variation and Recent Positive Selection in Worldwide Human Populations: Evidence from Nearly 1 Million SNPs. PLoS ONE 4(11), 1-16.

Lovvorn, M. B., Gill, G. W., Carlson, G. F., Bozell, J. R., & Steinacher, T. L. (199). Microevolution and the Skeletal Traits of a Middle Archaic Burial: Metric and Multivariate Comparison to Paleoindians and Modern Amerindians. American Antiquity, 64(3), 527-545.

Lubke, G., & Munthen, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. Structural Equation Modeling, 14, 26-47.

Lundy, B., Fausto-Sterling, A., Fullwiley, D., Hammonds, E. M., Nelson, A., et al. (2007). Racial Categories in Medical Practice: How Useful Are They? PLoS Medicine, 4(9), 1423-1428.

Lynch, J., & Condit, C. (2006). Genes and race in the news: A test of competing theories of news coverage. American Journal of Health Behaviour, 30(2), 125-135.

Lynch, M., & Walsh, B. (1998). Genetics and Analysis of Quantitative Traits. Sunderland, MA: Sinauer Associates.

Macaulay, V. , Hill, C. , Achilli, A. , Rengo, C. , Clarke, D. , Meehan, W., . . . Richards, M. . (2005). Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science, 308, 1034-1036.

MacDonald, K. B. (1988). Social and Personality Development: An Evolutionary Synthesis. New York: Plenum Press.

MacDonell, W. R. (1901). On Criminal Anthropometry and the Identification of Criminals. Biometrika, 1, 177-227.

MacEachern, S. (2000). Genes, Tribes, and African History. . Current Anthropology, 41, 357-384.

Machery, E., & Faucher, L. (2005). Social construction and the concept of race. Philosophy of Science, 72, 1208-1219.

MacNally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: the distinction between - and reconciliation of - 'predictive' and 'explanatory' models. Biodiversity and Conservation, 9, 655-671.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Paper presented at the Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.

Madow, W. G., Nisselson, H., & Olkin, I. (Eds.). (1983). Incomplete data in sample surveys, Vol 1: report and case studies. New York: Academic Press.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India 2(1), 49-55.

Maish, A., & Fukuzawa, S. (1998). Craniometric Trends of Ontario Iroquois in Comparison to FORDISC 2.0 Reference Populations. In Canadian Association of Physical Anthropology (Ed.), https://tspace.library.utoronto.ca/bitstream/1807/4382/1/CAPA Newsletter No 1 1998.pdf (Vol. 1, pp. 8).

Majumder, P. P. (1998). People of India: Biological diversity and affinities. Evolution Anthropology, 6, 100-110.

Majumder, P. P. (2001). Ethnic populations of India as seen from an evolutionary perspective. Journal of Biosciences, 26, 533-545.

Malécot, G. (1969). The Mathematics of Heredity (D. M. Yermanos, Trans.). San Francisco: Freeman.

Mallon, R. (2006). 'Race': normative, not metaphysical or semantic. Ethics, 116, 525-551.

Mangold, W. L., Nawrocki, S. P., & Scherbauer, J. (1993). The Shaffer Site (12 GR 109): Additional information on an Albee Phase Site in the White River Valley: Indiana University.

Manica, A., Amos, W., Balloux, F., & Hanihara, T. (2007). The effect of ancient population bottlenecks on human phenotypic variation. Nature, 448, 346-348.

Manica, A., Prugnolle, F., & Balloux, F. . (2005). Geography is a better determinant of human genetic differentiation than ethnicity. Human Genetics, 118(3-4), 366-371.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). Multivariate Analysis. New York: Academic Press.

Mardia. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhya B, 36, 115-128.

Marks, J. (1994). Blood will tell (won't it?): a century of molecular discourse in anthropological systematics. American Journal of Physical Anthropology, 94, 59-79.

Marks, J. (1995). Human biodiversity: genes, race, and history. New York, NY: Aldine de Gruyter.

Marks, J. (2002). What It Means To Be 98% Chimpanzee: Apes, People, and Their Genes. Berkeley: University of California Press.

Marks, J. (2008). Race: past, present, and future/ Jonathan Marks. In B. A. Koenig, S. Soo-Jin Lee & S. S. Richardson (Eds.), Revisiting Race in a genomic age (pp. 21-38). New Brunswick, NJ: Rutgers University Press.

Marks, J. . (1995). Dealing the Race Deck. Nature, 378, 143-144.

Marks, M. K. (1995). William M. Bass and the development of forensic anthropology in Tennessee. Journal of Forensic Sciences, 40, 741-741.

Marques-Bonet, T., Ryder, O. A., & Eichler, E. E. (2009). Sequencing primate genomes: what have we learned? Annual Review of Genomics and Human Genetics, 10, 355-386.

Marriott, F. H. C. (1974). The interpretation of multiple observations. London: Academic Press.

Martin, R. (1956). Lehrbuch der Anthropologie (3rd ed. Vol. 3). Suttgart: Gustav Fischer Verlag.

Martínez-Abadías, N., Esparza, M., & Sjøvold, T. (et al.). Heritability of human cranial dimensions: Comparing the evolvability of different cranial regions. Journal of Anatomy, 214, 19-35.

Martínez-Abadías, N., Esparza, M., Sjøvold, T., et al. (2009). Heritability of human cranial dimensions: Comparing the evolvability of different cranial regions. Journal of Anatomy, 214, 19-35.

Martínez-Abadías, N., González-José, R., González-Martín, A., et al. (2006). Phenotypic evolution of human craniofacial morphology after admixture: A geometric morphometric approach. American Journal of Physical Anthropology, 129, 387-398.

Masset, C. (1989). Age estimation on the basis of cranial sutures. In M. Y. Iscan (Ed.), Age Markers in the Human Skeleton (pp. 71-103). Springfield: Charles C. Thomas.

MATLAB V.7.10. (2010). Documentation for MathWorks Products, R2010a Retrieved August 5, 2010, from http://www.mathworks.com/access/helpdesk/help/helpdesk.html

Mayr, E. (1942). Systematics and the Origin of Species. New York, NY: Columbia University Press.

Mayr, E. (1950). Taxonomic Categories in Fossil Hominids. Cold Spring Harbor Symposium on Quantitative Biology, 15, 108-118.

Mayr, E. (1969). Principles of Systematic Zoology. New York, NY: McGraw-Hill.

Mayr, E. (1970). Populations, Species, and Evolution. Cambridge, MA: Belknap Press of Harvard University Press.

Mayr, E. (2002). The Biology of Race and the Concept of Equality. Daedalus, 131(1), 89-94.

McCann-Mortimer, P., Augoustinos, M., & Lecouteur, A. (2004). Race' and the Human Genome Project: Constructions of Scientific Legitimacy. Discourse & Society, 15(4), 409-432.

McCaullah, P., & Nelder, J. A. (1989). Generalized Linear Models (2nd ed.). Boca Raton:: Chapman & Hall.

McCoy, M. W., Bolker, B. M., Osenberg, C. W., Miner, B. G., & Vonesh, J. R. (2006). Size correction: comparing morphological traits among populations and environments. Oecologica, 148, 547-554.

McCready, J., & de Segonzac, J. (Writers). (2001). Wrath: Law & Order: Special Victims Unit NBC.

McCutcheon, A. L. (1987). Latent class analysis. Newbury Park, CA: Sage.

McDougall, I., Brown, F.H. , & Fleagle, J.G. . (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature, 433, 733-736.

McHenry, H. (1992). Body size and proportions in early hominids. American Journal of Physical Anthropology, 85(2), 407-431.

McHenry, H. (1998). Body proportions in Australopithecus afarensis and A. africanus and the origin of the genus Homo. Journal of Human Evolution, 35, 1-22.

McKeown, A. (1996). Craniometric sex estimation using a finite mixture missing data model. M. A., University of Tennessee, Knoxville, TN.

McKeown, A. (2000). Investigating variation among Arikara crania using geometric morphometry. Ph.D., University of Tennessee, Knoxville, TN.

McKusick, V. (2007). Mendelian Inheritance in Man and Its Online Version, OMIM. American Journal of Human Genetics, 80(4), 588-604.

McLachlan, G. J. (1972). Asymptotic results for discriminant analysis when initial samples are misclassified. Technometrics, 14, 415-422.

McLachlan, G. J. (1986). Assessing the performance of an allocation rule. In S. C. Choi (Ed.), Statistical Methods of Discrimination and Classification: Advances in Theory and Applications (pp. 261-272). New York: Pergamon Press.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Applied Statistics(36), 318-324.

McLachlan, G. J. (1992). Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley.

McLachlan, G. J. (2011). Package 'emmix': The EM Algorithm and Mixture Models. (Version 1.0.18). Retrieved from http://www.maths.uq.edu.au/~gjm/

McLachlan, G. J., & Baek, J. (2010). Clustering of high-dimensional data via finite mixture models. In A. Fink, Lausen B., W. Seidel & A. Ultsch (Eds.), Advances in Data Analysis, Data Handling and Business Intelligence (pp. 33-444). Berlin: Springer-Verlag.

McLachlan, G. J., & Basford, K. E. (1988). Mixture Models: Inference and Applications to Clustering. New York: Marcel Dekker.

McLachlan, G. J., & Kahn, N. (2004). On a resampling approach for tests on the number of clusters with mixture model-based clustering of tissue samples. Journal of Multivariate Analysis, 90, 90-105.

McLachlan, G. J., & Krishnan, T. (1997). The EM Algorithm and Extensions. New York: John Wiley & Sons.

McLachlan, G. J., & Peel, D. (1997). On a resampling approach to choosing the number of components in normal mixture models. In L. Billard & N. I. Fisher (Eds.), Computing Science and Statistics Vol. 28 (pp. 260-266). Fairfax Station, VA: Interface Foundation of North America.

McLachlan, G. J., & Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In A. Amin, D. Dori, P. Pudil & H. Freeman (Eds.), Lecture Notes in Computer Science Vol. 1451 (pp. 658-666). Berlin: Springer-Verlag.

McLachlan, G. J., & Peel, D. (2000). Finite Mixture Models. New York: John Wiley & Sons.

McLachlan, G. J., Do, K. A., & Ambroise, C. (2004). Analyzing Microarray Gene Expression Data. Hoboken, NJ: Wiley.

McLachlan, G. J., Ng, S. K., & Peel, D. (2003). On clustering by mixture models. In O. Opitz & M. Schwaiger (Eds.), Studies in Classification, Data Analysis, and Knowledge Organization: Exploratory Data Analysis in Empirical Research (pp. 141-148). Berlin: Springer-Verlag.

McLachlan, G. J., Peel, D, Adams, P., & Basford, K. E. (1995). An algorithm for assessing by resampling the P-value of the likelihood ratio test on the number of components in normal mixture models Research Report No. 31. Brisbane: Centre for Statistics, The University of Queensland.

McLachlan, G. J., Peel, D, Basford, K. E., & Adams, B. J. (2010). The Emmix Software For The Fitting Of Mixtures Of Normal And T-Components.: Technical Paper.

McLachlan, G. J., Peel, D, Basford, K. E., & Adams, D. C. (2010). THE EMMIX SOFTWARE FOR THE FITTING OF MIXTURES OF NORMAL AND t-COMPONENTS. Technical paper Retrieved July 17, 2010, from http://www.maths.uq.edu.au/~gjm/mix_soft/EMMIX_Demo/emmix.html

McLachlan, GJ, Wang, Kui, Ng, Angus, & Peel, D (2011). [Package 'emmix': the EM Algorithm and Mixture Models. Documentation for emmix version 1.0.18].

Mead, M., Dobzhansky, T., Tobach, E., & Light, R. E. (Eds.). (1968). Science and the concept of race. New York: Columbia University Press.

Meadows Jantz, L, & Jantz, R. L. (1999). Secular change in long bone length and proportion in the United States, 1800–1970. American Journal of Physical Anthropology, 110, 57-67.

Meadows Jantz, L, & Jantz, R. L. (2008). The Anthropology Research Facility: The Outdoor Laboratory of the Forensic Anthropology Center, University of Tennessee. In M. W. Warren, H. A. Walsh-Haney & L. E. Freas (Eds.), The Forensic Anthropology Laboratory (pp. 7-22). New York: CRC press.

Meadows Jantz, L. (2008). William M. Bass Donated Skeletal Collection Stature Dataset.

Meadows, L, & Jantz, R. L. (1995). Allometric secular change in the long bones from the 1800's to the present. Journal of Forensic Sciences, 40, 762-767.

Medawar, P. B. (1969). Induction and Intuition in Scientific Thought. Jayne Lectures for 1968. Philadelphia, PA: American Philosophical Society.

Meeting on the Biological Aspects of Race. (1964). Paper presented at the International Union of Anthropological and Ethnological Science, Moscow.

Mellars, P. (2006). Going East: New genetic and archaeological perspectives on the modern human colonization of Eurasia. Science, 313, 796-800.

Mellars, P. (2006). Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. . Proceedings of the National Academy of Sciences of the United States of America, 103, 9381–9386.

Meng, X. L, & Rubin, D B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika, 80(267-278).

Meng, X. L., & van Dyk, D. (1997). The EM algorithm - an old folk-song sung to a fast new tune (with discussion). Journal of the Royal Statistical Society B, 59(3), 511-567.

Mercer, A. (1990). Disease, mortality and population in transition: epidemiological-demographic change in England since the eighteenth century as part of a global phenomenon. Leicester: Leicester University Press.

Merilä, J., & Crnokrak, P. (2001). Comparison of genetic differentiation at marker loci and quantitative traits. Journal of Evolutionary Biology, 14, 892-903.

Merriwether, D. A., Clark, A. G., Ballinger, S. W., et al. (1991). The structure of human mitochondrial DNA variation. Journal of Molecular Evolution, 33, 543-555.

Michalakis, Y., & Excoffier, L. (1996). A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics, 142, 1061-1064.

Mielke, J. H., Konigsberg, L. W., & Relethford, J. H. (2006). Human Biological Variation. Oxford: Oxford University Press.

Miller, G.H., Magee, J.W. , Johnson, B.J. , Fogel, M.L. , Spooner, N.A., McCulloch, M.T., & Ayliffe, L.K. . (1999). Pleistocene extinction of Genyornis newtoni: human impact on Australian megafauna. Science, 283, 205-208.

Miller, J. M. A. (2000). Craniofacial variation in Homo habilis: An analysis of the evidence for multiple species. American Journal of Physical Anthropology, 112(1), 103-128.

Milligan, G. W., Soon, S. C., & Sokol, L. M. (1983). The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 5, 40-47.

Mitchell, L. (1999). The History of Everyone and Everything. Stanford Magazine. Retrieved from http://www.stanfordalumni.org/news/magazine/1999/mayjun/articles/cavalli_sforza.html

Mitchell, T. (1997). Machine Learning. New York: McGraw-Hill.

Mitteroecker, P., Gunz, P., Berhard, M., Schaefer, K., & Bookstein, F. L. (2004). Comparison of cranial ontogenetic trajectories among great apes and humans. Journal of Human Evolution, 32, 243-247.

Mitton, Jeffry B. (1977). Genetic Differentiation of Races of Mas As Judged by Single-Locus and Multilocus Analyses. The American Naturalist, 111(978), 203-212.

Mitton, Jeffry B. (1978). Measurement of Differentiation: Reply to Lewontin, Powell, and Taylor. The American Naturalist, 112(988), 1142-1144.

McKeigue, P. M. (2005). Prospects for admixture mapping of complex traits. American Journal of Human Genetics, 76, 1-7.

Molnar, S. (1975). Races, Types and Ethnic Groups. Englewood, NJ: Prentice Hall.

Montagu, A. (1941). The Concept of Race in the Light of Genetics. Journal of Heredity, 32, 243-247.

Montagu, A. (1942). Man's Most Dangerous Myth: The Fallacy of Race. New York, NY: Columbia University Press.

Montagu, A. (1951). Statement on race: an extended discussion in plain language of the UNESCO Statement by experts on race problems. New York, NY: Henry Schuman.

Montagu, A. (1964). The concept of race. New York, NY: The Free Press.

Montagu, A. (1966). A Handbook of Anthropometry. Springfield: Charles C. Thomas.

Montagu, A. (1997). Man's Most Dangerous Myth: The Fallacy of Race (Vol. 6th). Walnut Creek, CA: AltaMira Press.

Montana, G., & Hoggart, C. (2007). Statistical software for gene mapping by admixture linkage disequilibrium. Brief. BioInform, 8, 393-395.

Moore, M., Algee-Hewitt, B., & Wilson, R. J. (2009). The Effects of Ethanol Abuse on Bone Mineral Density in the Proximal Femur. Paper presented at the Annual Meeting of the American Academy of Forensic Sciences, Boulder, CO.

Moore-Jansen, P. H., Ousley, S. D., & Jantz, R. L. (1994). Data collection procedures for forensic skeletal material Report investigation no. 48. Knoxville, TN: Department of Anthropology, University Tennessee.

Moore-Jansen, P. M., Ousley, S. D., & Jantz, R. L. (1994). Data Collection Procedures for Forensic Skeletal Material. (Vol. Report of Investigations 48). Knoxville, TN: Department of Anthropology, The University of Tennessee.

Morant, G. M. (1935). A Study of Predynastic Egyptian Skulls from Badari Based on Measurements Taken by Miss. B.N. Stroessiger and Prof. D.E. Derry. Biometrika, 27(3-4), 293-309.

Morton, S. G. (1844). Crania Aegyptica. Philadelphia: Penington.

Morton, S. G. (1849). Observations on the size of the brain in various races and families of man. Proceedings of the Academy of Natural Sciences, Philadelphia, 4, 221-224.

Mosimann, J. N., & James, F. C. (1979). New statistical methods for allometry with application to Florida red-winged blackbirds. Evolution, 331, 444-459.

Mountain, J. L., & Cavalli-Sforza, L. L. (1997). Multilocus genotypes, a tree of individuals, and human evolutionary history. American Journal of Human Genetics, 61, 705-718.

Mountain, J. L., & Risch, N. (2004). Assessing genetic contributions to phenotypic differences among ‘racial ‘and ‘ethnic ‘groups. Nature Reviews Genetics, 36, S48-S53.

Mountain, J. L., & Risch, N. Assessing genetic contributions to phenotypic differences among 'racial' and 'ethnic' groups. National Review of Genetics, 36(11 - supplement), 2004.

Mousseau, T. A., & Roff, D. A. (1987). Natural selection and heritability of fitness. Heredity, 59, 181-197.

Mueller, W. S. (2005). Race and Ethnicity: Human Diversity and the UNESCO Statement on Race (1950-51). Paper presented at the Sixty Years of UNESCO's history: Proceedings of the International Symposium in Paris, Paris.

Mukhopadhyay, C. C., & Moses, Y. (1997). Reestablishing race in anthropological discourse. American Anthropologist, 99(3), 517-533.

Muller, M. W., Ten Krooden, J. A., & Hawkins, D. M. (1982). Cluster analysis. In D. M. Hawkins (Ed.), Topics in Applied Multivariate Analysis (pp. 303-353). Cambridge: Cambridge University Press.

Murtagh, F., & Raftery, A. E. (1984). Fitting straight lines to point patterns. Pattern Recognition, 17, 479-483.

Naar, N. A., Hilgenberg, D., & Armelagos, G. J. . (2006). FORDISC. 2.0 the ultimate test: what is the truth? . American Journal of Physical Anthropology, S42, 136.

Nakahashi, T. (1987). Human skeletal remains of the Edo period excavated from the Tenpukuji site, Fukuoka. Journal of the Anthropological Society of Nippon, 95, 89-106.

Natalie, R. S. (1009). Age And Sex Estimation From The Human Clavicle: An Investigation Of Traditional And Novel Methods. Ph. D., University of Tennessee, Knoxville, TN.

National Center for Biotechnology Information (NCI) Human Genome Resources. (2010) Retrieved December 16, 2010, from http://www.ncbi.nlm.nih.gov/genome/guide/human/

National Human Genome Research Institute's (NHGRI) Ethical, Legal and Social Implications (ELSI) Research Program. (2010) Retrieved July 1, 2011, from http://www.genome.gov/10001618

Nawrocki, S., Pless, J., Hawley, D., & Wagner, S. (1997). Fluvial transport of human crania. In W. Haglund & M. Sorg (Eds.), Forensic Taphonomy (pp. 529-552). New York: CRC Press.

Nebert, D. W., & Menon, A. G. (2001). Pharmacogenomics, ethnicity, and susceptibility genes. Pharmacogenomics 1, 19-22.

Nei, M. (1981). Gene Differences Between Human Races The American Naturalist, 117(1), 88-89.

Nei, M. (1987). Molecular Evolutionary Genetics. New York, NY: Columbia University Press.

Nei, M., & Livshits, G. (1989). Genetic-Relationships of Europeans, Asians, and Africans and the Origin of Modern Homo-Sapiens. Human Heredity, 39, 276-281.

Nei, M., & Roychoudhury, A. K. (1972). Gene Differences Between Caucasian, Negro, and Japanese Populations. Science, 177, 434-436.

Nei, M., & Roychoudhury, A. K. (1974). Genetic Variation Within and Between the Three Major Races of Man, Caucasoids, Negroids and Mongoloids. American Journal of Human Genetics, 26, 421-443.

Nei, M., & Roychoudhury, A. K. (1982). Evolutionary relationships and evolution of human races. Evolutionary Biology, 14, 1-59.

Nei, M., & Roychoudhury, A. K. (1982). Genetic-Relationship and Evolution of Human Races. Evolutionary Biology, 14, 1-59.

Nei, M., & Roychoudhury, A.K. . (1993). Evolutionary relationships of human populations on a global scale. Molecular Biology and Evolution, 10, 927-943.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society Series A, 135(3), 370-384.

Nelkin, D. (2001). Molecular metaphors: The gene in popular discourse. Nature Reviews Genetics, 2, 555-559.

Nelkin, D., & Lindee, S. M. (1995). The DNA mystique: The gene as a cultural icon. New York: W. H. Freeman.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. American Journal of Mathematics, 8, 343-366.

Newman, M. T. (1963). Geographic and microgeographic races. Current Anthropology, 4, 189-192.

Nexus program. (2011) Retrieved February 28, 2011, from http://www.cbp.gov/xp/cgov/travel/trusted_traveler/nexus_prog/

Nguyen, G. C., Torres, E. A., M., Reguerio., Bromfield, G., et al. (23006). Inflammatory Bowel Disease Characteristics Among African Americans, Hispanics, and Non-Hispanic Whites: Characterization of a Large North American Cohort. American Journal of Gastroenterology, 101, 1012-1023.

NIST/SEMATECH e-Handbook of Statistical Methods. (2010) Retrieved July 20, 2010, from http://www.itl.nist.gov/div898/handbook/

No scientific basis for race bias found by world panel of experts. (1950, July 17, 1950). New York Times.

Noak, A. (2009). Modularity clustering is force-directed layout. Physical Review 79, 1-8.

Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J. , . . . Rubin, E.M. (2006). Sequencing and analysis of Neanderthal genomic DNA. Science, 314, 1113-1118.

Nordborg, M., & Travare, S. (2002). Linkage disequilibrium: what history has to tell us. Trends in Genetics, 18, 83-90.

Occam, William of. (1951). Summa totius logicae (P. Boehner, Trans.). New York: Louvain.

Ogen, C. L., Carroll, M. D., Curtin, L. R., Macdowll, M. A., Tabak, C. J., & Flegal, F . F. (2006). Prevalence of overweight and obesity in the United States, 1999–2004. Journal of the American Medical Association, 295(13), 1547-1555.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). Journal of the Royal Statistical Society Series B, 57, 99-138.

O'Higgins, P. (2000). The study of morphological variation in the hominid fossil record: biology, landmarks and geometry Journal of Anatomy, 197.Pt.1, 103-120.

Oka, M. (1954). Ecological studies on the kidai by the statistical method. II. On the growth of the kidai (Taitus tumifrons). Bulletin of the Faculty of Fisheries Hokkaido, 2, 8-25.

Olden, K., & Wilson, S. (2000). Environmental health and genomics: Visions and implications. Nature Reviews Genetics, 1, 149-153.

OMB Directives: Racial and Ethnic Classifications Used in Census 2000 and Beyond. (1997). Retrieved from http://www.census.gov/population/www/socdemo/race/racefactcb.html

Online Mendelian Inheritance in Man, OMIM (2001) Retrieved July 1, 2011

Oppenheimer, G. M. (2001). Paradigm lost: race, ethnicity, and the search for a new population taxonomy. American Journal of Public Health, 91, 1049-1055.

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. In L. M. LeCan, J. Neyman & E. Scott (Eds.), Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1). Berkeley, CA: University of California Press.

Ossorio, P., & Duster, T. (2005). Race and genetics: Controversies in biomedical, behavioral, and forensic sciences. American Psychologist, 60(1), 115-128.

Ostrowski, M. F., David, J., Santoni, S., et al. (2006). Evidence for a large-scale population structure among accessions of Arabidopsis thaliana: possible causes and consequences for the distribution of linkage disequilibrium. Molecular Ecology, 15, 1507-1517.

Ousley, S. D., & Jantz, R. L. (1997). The forensic data bank: documenting skeletal trends in the United States. In K. J. Reichs (Ed.), Forensic Osteology (pp. 441-458). Springfield, IL: C. C. Thomas.

Ousley, S. D., & Jantz, R. L. (2002). Social races and human populations: why forensic anthropologists are good at identifying races. American Journal of Physical Anthropology Supplement, 34, 83-84.

Ousley, S. D., & Jantz, R. L. (2005). FORDISC 3.0: Personal Computer Forensic Discriminant Functions. Knoxville: University of Tennessee, Knoxville.

Ousley, S. D., Jantz, R. L., & Freid, D. (2009). Understanding race and human variation: Why forensic anthropologists are good at identifying race. American Journal of Physical Anthropology, 139(68-76).

Ousley, S., & Jantz, R. (1996). FORDISC 2.0. Knoxville, TN: Department of Anthropology, University of Tennessee.

Owens, K., & King, M. C. (1999). Genomic views of human history. Science, 15(286), 451.

Palmer, B. . (2011, January 5). Alas, Poor Yorick ... or Is It Othello? Can you tell a person's race from his or her skull? Slate Explainer Retrieved January 19, 2011, from http://www.slate.com/id/2279942/

Parra, E. J., Kittles, R. A., & Shriver, M. D. (2004). Implications of correlations between skin color and genetic ancestry for biomedical research. Nature Genetics, 36(11), Supplement.

Parra, E. J., Marcini, A., Akey, L., Martinson, J., Batzer, M. A., Cooper, R., . . . Shriver, M. D. (1998). Estimating African American admixture proportions by use of population-specific alleles. American Journal of Human Genetics, 63, 1839-1851.

Parra, F. C., Amado, R. C., Lambertucci, J. R., Rocha, J., Antunes, C. M., & Pena, S. D. (2003). Color and genomic ancestry in Brazilians. Proceedings of the National Academy of Sciences of the United States of America, 100, 177-182.

Passel, J. S., & Chohn, D. (2011). How Many Hispanics? Comparing New Census Counts with the Latest Census Estimates Pew Hispanic Center Report #139.

Pasternak, J. J. (2005). An Introduction to Human Molecular Genetics: Mechanisms of Inherited Diseases, Second Edition. New York: Wiley-Liss.

Pearson, J. D., Morrell, C. H., & Brant, L. J. (1992). Mixture models for investigating complex distributions. Journal of Quantitative Anthropology. Journal of Quantitative Anthropology, 3, 325-345.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London Series A, 185, 71-110.

Pearson, K. (1895). Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material. Philosophical Transactions of the Royal Society of London Series A, 186, 343-414.

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. Philosophical Transactions of the Royal Society A, 187, 253-318.

Pearson, K. (1926). On the coefficient of racial likeness. Biometrika, 18, 105-117.

Pearson, K., & Davin, A. G. (1924). On the biometric constants of the human skull. Biometrika, 16, 328-363.

Pearson, K., & Woo, T. L. (1935). Further investigations on the morphometric characters of the individual bones of the human skull. Biometrika, 27, 424-465.

Peel, D, & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. Statistics and Computing 10, 335-344.

Peel, D, & McLachlan, GJ. (1999). User's Guide to EMMIX Version 1.3 Retrieved 08/01/10, from http://www.maths.uq.edu.au/~gjm/mix_soft/EMMIX_Demo/emmix.html

Pella, J., & Masuda, M. (2006). The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. Canadian Journal of Fisheries and Aquatic Sciences, 63, 576-596.

Peng, R. , Jackobs, R. A., & Tanner, M. A. (1996). Bayesian inference in mixture-of- experts and hierarchical mixture-of experts models with an application to speech recognition. Journal of the American Statistical Association, 91(435), 953-960.

Perez, I. S., Bernal, V., & Gonzalez, P. N. (2007). Evolutionary relationship among prehistoric human populations: An evolution of relatedness patterns based on facial morphometric data using molecular data. Human Biology, 78(25-50).

Perez, S.I., Bernal, V. , Gonzalez, P.N., Sardi, M. , & Politis, G.G. . (2009). Discrepancy between Cranial and DNA Data of Early Americans: Implications for American Peopling. PLoS ONE 4(5), e5746.

Pierpont, C. R. (2004). The Measure of America: How a Rebel Anthropologist Waged War on Racism, The New Yorker, pp. 48-63.

Pietrusewsky, M. (2000). Metric Analysis of Skeletal Remains: Methods and Applications. In M. A. Katzenberg & S. Saunders (Eds.), Metric Analysis of Skeletal Remains: Methods and Applications (Vol. 375-416). New York: Wiley-Liss.

Pietrusewsky, M. (2008). Metric Analysis of Skeletal Remains: Methods and Applications. In M. A. Katzenberg & S. Saunders (Eds.), Metric Analysis of Skeletal Remains: Methods and Applications (2nd ed., Vol. 487-532). New York: Wiley-Liss.

Pirie, N. W. (1952). Concepts Out of Context: The Pied Pipers of Science. The British Journal for the Philosophy of Science, 2(8), 269-280.

Plagnol, V., & Wall, J.D. . (2006). Possible ancestral structure in human populations. PLoS Genetics, 2, 972-979.

Plavcan, J. M. (1994). Comparison of four simple methods for estimating sexual dimorphism in fossils. American Journal of Physical Anthropology, 94, 465-476.

Plessy v. Ferguson, No. 164 U.S. (1896).

Poisson, S. D. (1837). Recherches sur la probabilité des jugements en matière criminelle et en matière civile, precedes des règles generals du calcul des probabilities. Paris.

Pontikos, D. (2004). Model-based clustering of world craniometric variation, 2010, from http://dienekes.50webs.com/arp/articles/antrhopologica/clustering.html

Powell, J. B., & Taylor, C. E. (1978). Are human races "substantially" different genetically? The American Naturalist, 112, 1139-1142.

Powell, J.F., & Neves, W.A. (1999). Craniofacial morphology of the first Americans: pattern and process in the peopling of the New World. Yearbook of Physical Anthropology 42, 153–188.

Press, Associated. (2002). What Cranial Shape Tells Us. Wired Magazine Retrieved September 26, 2010, from www.wired.com

Preston, E. J. (1953). A graphical method for the analysis of statistical distributions into two normal distributions. Biometrika, 40, 460-464.

Prevention, Centers for Disease Control and. (1993). Use of race and ethnicity in public health surveillance. MMWR 1993; 42 (No. RR-10) Retrieved August 19, 2010, from http://www.cdc.gov/mmwr/preview/mmwrhtml/00021729.htm

Prevention, Centers for Disease Control and. (2000). Race and Ethnicity Code Set version 1.0 Retrieved September 18, 2010, from http://www.cdc.gov/nedss/DataModels

Prichard, J. K., & Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics, 65, 220-228.

Prichard, J. K., Wen, W., & Falush, D. (2009). Documentation for structure Software version 2.3 Retrieved April 29, 2011, from http://pritch.bsd.uchicago.edu/software/documentation.pdf

Priebe, C. E. (1994). Adaptive Mixtures. Journal of the American Statistical Association, 89, 796-806.

Pritchard, J. K., & Prezworski, M. (2001). Linkage disequilibrium in humans: models and data. American Journal of Human Genetics, 67, 1-14.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155, 945-959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., & Donnelly, P. (2000). Association mapping in structured populations. American Journal of Human Genetics, 67, 107-181.

Pritchard, JK, Stephens, M, & Donnelly, P. (2000). Inference of population structure using mutilocus genotype data. Genetics, 155(945-959).

Project, Human Genome. (2001). The human genome. Nature, 409, 812.

Provine, W. B. (1971). The Origins of Theoretical Population Genetics. Chicago: University of Chicago Press.

Prugnolle, F., Manica, A., & Balloux, F. (2005). Geography predicts diversity of human populations. Current Biology, 15, R157-R160.

Prugnolle, F., Manica, A., Charpentier, M., Guegan, J. F., Buernier, V., & Balloux, F. (2005). Pathogen-driven selection and worldwide HLA class l diversity. Current Biology, 15, 1022-1027.

Putnam, C. (1962). Race and Reason: A Yankee View. Washington, DC: Public Affairs Press.

Pyle, D. (1999). Data Preparation for Data Mining. Los Altos, CA: Morgan Kaufmann Publishers.

Quandt, R. E. (1972). A new approach to estimating switching regressions. Journal of the American Statistical Association, 67(338), 306-310.

Quandt, R. E., & Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions (with discussion). Journal of the American Statistical Association, 73(364), 730-752.

Quetelet, A. (1827). Recherches sur la population, les naissances, les décès, les prisons, les depots de mendicité, etc.. dans le royaume des Pays-Bas. Nouveaux memoires de l'Académie royale des sciences et belles-lettres de Bruxelles, 4, 175-192.

Quetelet, A. (1835). Sur l'homme et le développement de ses faculties, ou Essai de physique sociale. Paris.

Quetelet, A. (1870). Anthropometrie, ou mesure des différentes faculties de l'homme (Vol. 1 and 2). Brussels.

Quetelet, A. . (1846). Lettres à S.A.R. le Duc Regnant de Saxe-Cobourg et Gotha, sur la théorie des probabilities, appliqué aux sciences morales et politiques. Brussels.

Quinn, B. G., McLachlan, G. J., & Hjort, N. L. (1987). A Note on the Aitkin-Rubin Approach to Hypothesis Testing in Mixture Models. Journal of the Royal Statistical Society, 49(3), 311-314.

Quinn, N. R., & Breuer, M. A. (1979). A Force Directed Component Placement Procedure for Printed Circuit Boards. IEEE Transactions on Circuits and Systems - CAS, 16(6), 377-388.

Quintana-Murci, L., Semino, O., Bandelt, H. J., Pas-sarino, G., McEleavey, K., & Santachiara-Benercetti, A. S. (1999). Genetic evidence of an early exit of Homo sapiens from Africa through eastern Africa. National Review of Genetics, 23(4), 437-441.

R., González-José, Van der Molen, S., González-Pérez, E., & Hernández, M. (2004). Patterns of phenotypic covariation and correlation in modern humans as viewed from morphological integration. American Journal of Physical Anthropology, 123, 67-77.

Rabinow, P. (1996). Artificiality and Enlightenment: From Sociobiology to Biosociality. In P. Rabinow (Ed.), Essays on the Anthropology of Reason (pp. 91-111). Princeton: Princeton University Press.

Race, Ethnicity, & Genetics Working, Group. (2005). The Use of Racial, Ethnic, and Ancestral Categories in Human Genetics Research. American Journal of Human Genetics, 77(4), 519-532.

Rafter, N. (2004). Earnest A. Hooton and the Biological Tradition in American Criminology. Criminology, 42, 735-772.

Raftery, A. E. (1995). Bayesian model selection in social research with discussion. Sociological Methodology, 25, 111-163 (discussion 163-196).

Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. Journal of the American Statistical Association, 101, 168-178.

Rak, Y. (1986). The Neanderthal: a new look at an old face. Journal of Human Evolution, 15, 151-164.

Ramachandran, S., Despande, O., Roseman, C. C., et al. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Science, 102, 15942-15947.

Ramachandran, S., Tang, H., Gutenkunst, R. N., & Bustamante, C. D. (2010). Genetics and Genomics of Human Population Structure. In M. R. Speicher, S. E. Antonarakis & A. G. Motulsky (Eds.), Human Genetics: Problems and Approaches (4th ed., pp. 589-616). New York: Springer.

Ramierz, E., & de la Cruz, G. (2003). The Hispanic population in the United States: March 2002 (Vol. June Report No. P20-545). Washington, DC: U. S. Census Bureau.

Rannala, B., & Mountain, J. L. (1997). Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Science, 94, 9197-9201.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society Series B, 10, 159-203.

Rao, C. R. (1965). Linear statistical inference and its applications New York: John Wiley and Sons.

Raposo do Amaral, C. M., Krieger, H., Cabello, H., et al. (1989). Heritability of quantitative orbital traits. Human Biology, 61, 551-557.

Reddy, B.M., Langstieh, B.T., Kumar, V., Nagaraja, T., Reddy, A.N., Meka, A., . . . Singh, L. (2007). Austro-Asiatic tribes of northeast India provide hither to missing genetic link between south and southeast Asia. PLoS ONE, 2, e1141.

Redner, R. A. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. Annals of Statistics, 9, 225-228.

Redner, R. A., & Walker, H. E. (1984). Mixture densities, maximum likelihood and the EM algorithm. SIAM Review, 26(2), 195-239.

Rehfeldt, G. E., & Gallo, L. A. (2001). Introduction of ponderosa pine and Douglas-fir to Argentina - Using quantitative traits for retrospective identification and prospective selection of provenances. New Forests, 2, 35-44.

Reichs, K. J. (1995). A Professional Profile of Diplomats of the American Board of Forensic Anthropology: 1984-1992. Journal of Forensic Sciences, 40, 176-182.

Reichs, K. J. (1998). Forensic Anthropology: A Decade of Progress. In K. J. Reichs (Ed.), Forensic Osteology: Advances in the Identification of Human Remains (pp. 13-38). Springfield: Thomas, C.C.

Relethford, J. H. (1984). Morphological size and shape variation among local human populations Journal of Human Evolution, 13(2), 191-194.

Relethford, J. H. (1994). Craniometric variation among modern human populations. American Journal of Physical Anthropology, 95, 53-62.

Relethford, J. H. (1997). Hemispheric difference in human skin color. American Journal of Physical Anthropology, 104, 449-457.

Relethford, J. H. (2001). Global analysis of regional differences in craniometric diversity and population substructure. Human Biology, 73, 629-636.

Relethford, J. H. (2002). Apportionment of global human genetic diversity based on craniometrics and skin color. American Journal of Physical Anthropology, 118, 393-398.

Relethford, J. H. (2004). Boas and Beyond: Migration and Craniometric Variation. American Journal of Human Biology, 16(4), 379-386.

Relethford, J. H. (2004). Global patterns of isolation by distance based on genetic and morphological data. Human Biology, 76(499-513).

Relethford, J. H. (2007). The use of quantitative traits in anthropological genetic studies of population structure and history. In M. H. Crawford (Ed.), Anthropological genetics: theory, methods, and applications (pp. 187-209). Cambridge: Cambridge University press.

Relethford, J. H. (2009). Race and global patterns of phenotypic variation. American Journal of Physical Anthropology, 139, 16-22.

Relethford, J. H. (2010). Population-Specific Deviations of Global Human Craniometric Variation From a Neutral Model American Journal of Physical Anthropology, 142, 105-111.

Relethford, J. H., & Blangero, J. (1990). Detection of differential gene flow from patterns of quantitative variation. Human Biology, 62, 5-25.

Relethford, J. H., & Blangero, J. (1990). Detection of differential gene flow patterns of quantitative variation. Human Biology, 621, 5-25.

Relethford, J. H., & Harpending, H. C. (1994). Craniometric Variation, Genetic Theory, and Modern Human Origins. American Journal of Physical Anthropology, 95, 249-270.

Relethford, J. H., Crawford, M. H., & Blangero, J. (1997). Genetic drift and gene flow in post-Famine Ireland. Human Biology, 69, 443-465.

Relethford, J.H. (2008). Genetic evidence and the modern human origins debate. Heredity, 100, 555-563.

Render, R. A. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. Ann. Statist., 9, 225-228.

Renka, R. J. (1996). Algorithm 751: TRIPACK: a constrained two-dimensional {Delaunay} triangulation package. ACM Transactions on Mathematical Software, 22, 1-8.

Renka, R. J., & Gebhardt, A. (2009). Triangulation of irregularly spaced data. Documentation for the R package "tripack" version 1.3-4 Retrieved July 30, 2011, from http://cran.r-project.org/web/packages/tripack/tripack.pdf

Retzius, G., & Fürst, C. M. (1902). Anthropologia Suecica, Beiträge zur Anthropologie der Schweden Stockholm.

Rhine, S. (1990). Nonmetric Skull Racing. In G. Gill & S. Rhine (Eds.), Skeletal attribution of race: methods for forensic anthropology. Maxwell Museum of Anthropological Papers No. 4 (pp. 9-20). Albuquerque, NM: University of New Mexico.

Rhine, S. (1993). Skeletal Criteria for Racial Attribution. National Association for the Practice of Anthropology Bulletin 13(1), 54-67.

Richard Vandergeer, Second Lieutenant, USAF. (2000) Retrieved February 8, 2011, from http://www.arlingtoncemetery.net/rvandegeer.htm

Richardson, R. C. (2000). Review of M. Levin's why race matters. Ethics, 110, 847-848.

Richerson, P. J. (2010). Culture Is an ACTIVE Part of Biology: Symposium on the Question 'How Is Culture Biological? Politics and Culture, 1, www.politicsandculture.org

Richerson, P. J., & Boyd, R. (2005). Not By Genes Alone: How Culture Transformed Human Evolution. Chicago: University of Chicago Press.

Ridley, M. (2004). Evolution. Oxford: Blackwell.

Riley, J. C. (2001). Rising life expectancy: a global history. Cambridge: Cambridge University Press.

Riley, J. C. (2005). The timing and pace of health transitions around the world. Population and Development Review, 31, 741-748.

Ripley, B. D. (1996). Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press.

Risch, N. , Burchard, E. , Ziv, E., & Tang, H. (2002). Categorization of humans in biomedical research: genes, race and disease. Genome Biology, 3(7), comment 2007.

Risch, N., Burchard, E. G., Ziv, E. , & Tang, H. (2007). Categorization of Humans in Biomedical Research: Genes, Race and Disease. Genome Biology, 3(7 - comment), 1-12.

Risch, N., Burchard, E. G., Ziv, E., & Tang, H. (2002). Categorization of humans in biomedical research: genes, race and disease. Genome Biology, 3, 1-12.

Roach, M. . (2003). Stiff: The Curious Lives of Human Cadavers. New York: W. W. Norton.

Roberts, D. (2010). Race and the New Biocitizen. In D. S. Jones & I. Whitmarsh (Eds.), What's the Use of Race? Modern Governance and the Biology of Difference (pp. 259-276). Cambridge, MA: MIT Press.

Roberts, D. F., & Kahlon, D. P. S. (1976). Environmental correlations of skin colour. Annals of Human Biology, 3, 11-22.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. Journal of the American Statistical Association, 89, 487-495.

Roeder, K., & Wasserman, L. (1997). Practical Bayesian density estimation using mixture of normals. Journal of the American Statistical Association, 92, 894-902.

Roeder, K., Escobar, J., Kadane, B., & Balazs, I. (1998). Measuring heterogeneity in forensic databases using hierarchical Bayes models. Biometrika, 85, 269-287.

Rohlf, F. J., & Marcus, L. F. (1993). A revolution in morphometrics. Trends in Ecology and Evolution, 8(4), 129-132.

Romualdi, C., Balding, D., Nasidze, I. S., Risch, G., Robichaux, M., Sherry, S. T., . . . Barbujani, G. (2002). Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Research, 12, 602-612.

Rose, J. C., & Ungar, P. S. (1998). Gross dental wear and dental microwear in historical perspective. In K . W. Alt, F. W. Rosing & M. Teschler-Nicola (Eds.), Dental Anthropology: Fundamentals, Limits, Perspectives. (pp. 349-386). Suttgart: Gustav Fischer publishers.

Roseman, C. C. (2004). Detecting interregionally diversifying natural selection on modern human cranial form by using matched molecular and morphometric data. Proceedings of the National Academy of Sciences of the United States of America, 101(35), 12824-12829.

Roseman, C. C., & Weaver, T. D. (2004). Multivariate apportionment of global human craniometric diversity. American Journal of Physical Anthropology, 125, 257-263.

Roseman, C. C., & Weaver, T. D. (2007). Molecules versus morphology? Not for the human cranium. BioEssays, 29, 1185-1188.

Rosenberg, N. A. (2004). DISTRUCT: a program for the graphical display of population structure. Molecular Ecology Notes, 4, 137-138.

Rosenberg, N. A., Burke, T., Elo, K., Feldman, M. W., Friedlin, P., Groenen, M. A. M., . . . Weigend, S. (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics, 159, 699-713.

Rosenberg, N. A., Mahajan, S., Gonzalez-Quevedo, C., Blum, M. G. B., Nino-Rosales, L., & et al. (2006). Low levels of genetic divergence across geographically and linguistically diverse populations from India. PLoS Genet, 2(12), 215.

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., et al. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genetics, 1(6), e70.

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., & et al. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet, 1(6), 70.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. Science, 298, 2381-2385.

Rösing, F. W. (1992). Diachronic trends in Egypt. Acta Musei Nationalis Prague 46, 191-195.

Ross, A. H., & Konigsberg, L. W. (2002). New formulae for estimating stature in the Balkans. Journal of Forensic Sciences, 47, 165-167.

Ross, A. H., Slice, D. E., & Pachar, J. V. (2005, February 21-27). Forensic Identifications and the Complexity of Determining Biological Affinities of "Hispanic"Crania. Paper presented at the Proceedings of the 57th Annual Meeting of the American Academy of Forensic Sciences, New Orleans, LA.

Ross, A. H., Slice, D. E., Ubelaker, D. H., & Falsetti, A. B. (2004). Population affinities of 19th century Cuban crania: implications for identification criteria in south Florida Cuban Americans. Journal of Forensic Sciences, 49, 1-6.

Rothhammer, F., & Silva, C. (1990). Craniometric variation among South American prehistoric populations: Climatic, altitudinal, and geographic contributions. American Journal of Physical Anthropology, 82, 9-17.

Rotimi, C. N. (2004). Are medical and nonmedical uses of large-scale genomic markers conflating genetics and 'race'? Nature Reviews Genetics, 36(11 Suppl), S43-S47.

Roubens, M. (1978). Pattern Classification Problems and Fuzzy Sets. Fuzzy Sets and Systems(1), 239-253.

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, 290, 2323-2326.

Royall, R. (1997). Statistical evidence: A likelihood paradigm. London: Chapman & Hall.

Roychoudhury, A. K. (1978). Genetic distance between the American Indians and the three major races of man. Human Heredity, 28(5), 380-385.

Rubin, D B. (1996). Multiple Imputation after 18+ Years" (with discussion). Journal of the American Statistical Association, 91, 473-489.

Rubin, D. B. (1976). Inference and missing data. Biometrika(63), 581-592.

Rubin, D. B. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. Paper presented at the ASA Proceedings of the Section on Survey Research Methods.

Rubin, D. R. (1972). A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. Applied Statistics, 21, 136-141.

Ruff, C. . (2002). Variation In Human Body Size And Shape. Annual Review of Anthropology 31, 211-232.

Ruff, C. B. (1980). Age differences in craniofacial dimensions among adults from Indian Knoll, Kentucky. American Journal of Physical Anthropology, 53, 101-108.

Rui, X., & Wunsch, D. (2009). Clustering. Hoboken, NJ: Wiley.

Rushton, J. P. (1995). Race, Evolution, and Behavior: A Life History Perspective. New Brunswick, NJ: Transaction Publishers.

Ryman, N., Chakraborty, R., & Nei, M. (1983). Differences in the relative distribution of human genetic diversity between electrophoretic and red and white cell antigens. Human Heredity, 33, 93-102.

Saks, M. J., & Koehler, M. (2005). The coming paradigm shift in forensic identification science. Science, 309, 892-895.

Salkin, Allen. (2008, September 12). When in Doubt, Spit it Out, The New York Times.

Sankar, P. (2010). Forensic DNA Phenotyping: Reinforcing Race in Law Enforcement. In I. Whitmarsh & D. S. Jones (Eds.), What's the Use of Race? Modern Governance and the Biology of Difference (pp. 49-62). Cambridge, MA: MIT press.

Sankar, P., Cho, M. K., & Mountain, J. (2007). Race and Ethnicity in genetic research. American Journal of Medical Genetics Part A, 143A, 961-970.

Santos, F.R. , Bonatto, S.L., & Bortolini, M.C. (2007). Molecular evidence from contemporary indigenous populations to the peopling of America. In C. Santos & M. Lima (Eds.), Advances in molecular biology and evolution and its applications to biological anthropology (pp. 333-346). Kerela, India: Research Signpost.

Santos, F.R., Pandya, A., Tyler-Smith , C., Pena, S.D., Schanfield, M., Leonard, W.R., . . . Mitchell, R.J. (1999). The central Siberian origin for native American Y chromosomes. American Journal of Human Genetics, 64, 619-628.

Sapir, G. I. (2002). Legal aspects of forensic science. In R. Saferstein (Ed.), Forensic Science Handbook, (2nd ed., Vol. 1, pp. 2-24). Upper Saddle River, NJ: Prentice Hall.

Sardi, M. L., Rozzi, F. R., González-José, R. , & Pucciarelli, H. M. . (2005). South Amerindian craniofacial morphology: Diversity and implications for Amerindian evolution. American Journal of Physical Anthropology, 128, 747-756.

Sarich, V., & Miele, F. (2004). Race: the reality of human differences. Boulder, CO: Westview Press.

SAS/SAT(R) 9.2 User's Guide, Second Edition. (2011). SAS online documentation Retrieved June 1, 2011

Satta, Y., & Takahata, N. . (2002). Out of Africa with regional interbreeding? Modern human origins. BioEssays, 24, 871-875.

Sauer, N. J. (1992). Forensic anthropology and the concept of race: if races don't exist, why are forensic anthropologists so good at identifying them? Social Science and Medicine, 34, 107-111.

Schaefer, K., Mitteroecker, P., Gunz, P., Berhard, M., & Bookstein, F. L. (2004). Craniofacial sexual dimorphism patterns and allometry among extant hominids. Annals of Anatomy, 186, 471-478.

Schafer, J. L. (1999). NORM Version 2.02 for Windows 95/98/NT Retrieved June 01, 2011, from http://www.stat.psu.edu/~jls/misoftwa.html#top

Schafer, J. L. (1999`). Multiple imputation: A primer. Statistical Methods in Medical Research, 8(1), 3-15.

Schafer, J. L. (2000). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Schafer, J. L. (2010). mix: Estimation/multiple Imputation for Mixed Categorical and Continuous Data (Version 1.0-8): Maintained by Brian Ripley. Retrieved from http://cran.r-project.org/web/packages/mix/index.html

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research, 33(545-571).

Schillaci, M. (2008). Human cranial diversity and evidence for an ancient lineage of modern humans. Journal of Human Evolution, 54, 814-826.

Schillaci, M. A., Irish, J. D. , & Wood, C. C. . (2009). Further analysis of the population history of ancient Egyptians. American Journal of Physical Anthropology, 139, 235–243.

Schlattmann, P. (2009). Medical Application of Finite Mixture Models. Berlin: Springer-Verlag.

Schroeder, K. B., Schurr, T. G., Long, J. C., Rosenberg, N. A., Crawford, M. H. , Tarskaia, L. A., . . . Smith, D. G. (2007). A private allele ubiquitous in the Americas. Biology Letters, 3, 218-223.

Schurr, T. (2004). The peopling of the New World: perspectives from molecular anthropology. Annual Review of Anthropology, 33, 551-583.

Schwartz, R. S. (2001). Racial Profiling in medical research. New England Journal of Medicine, 344, 1392-1293.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Schwidetzky, I., & Rösing, F. W. (1990). Vergleichend-statistische Untersuchungen zur Anthropologie von Neolithikum und Bronzezeit. Homo, 40, 4-45.

Science and the concept of race. (1968). New York: Columbia University Press.

Scientific Working Group for Forensic Anthropology (SWGANTH). (2011) Retrieved February 15, 2011, from http://www.swganth.org/201.html

Scott, A. J., & Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. Biometrics(27), 387-397.

Seidemann, R. M, Stojanowski, C. M., & Rich, F. I. (2009). The identification of a human skull recovered from an eBay sale. Journal of Forensic Sciences. Journal of Forensic Sciences, 54(6), 1247-1253.

Seielstad, M. T., Minch, E., & Cavalli-Sforza, L. L. (1998). Genetic evidence for a higher female migration rate in humans. Nature Genetics, 20, 278-280.

Self, S. G., & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association, 82, 605-610.

Sergi, S. (1958). Die Neandertalischen palaeanthropen in Italien. In G. H. R. Koenigswald (Ed.), Hundert jahr Neanderthaler (pp. 38-51). Utrecht: Kemink en Zoon.

Serre, D. , & S., Pääbo. (2004). Evidence for gradients of human genetic diversity within and among continents. Genome Research, 14(9), 1679-1685.

Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M., Paunovic, M., Mennecier, P., . . . Pääbo, S. (2004). No evidence of Neanderthal mtDNA contribution to early modern humans. PLoS Biology 2, 0313-0317.

Sesardic, Neven. (2010). Race: A Social Destruction of a Biological Concept. Biology & Philosphy, 25, 143-162.

Shanklin, E. (1994). Anthropology and race. Belmont, CA: Wadsworth.

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. Urbana, IL: University of Illinois Press.

Shapiro, H. L. (1954). Earnest Albert Hooton, 1887-1954. American Anthropologist, 56, 1081-1084.

Sherwood, R. J., Duren, D. L, Demerath, E. W., et al. (2008). Quantitative genetics of modern human cranial variation. Journal of Human Evolution, 54, 909-914.

Shields, A. E., Fortun, M., Hammonds, E. M., A., King. P., Lerman, C., Rapp, R., & Sullivan, P. F. (2005). The use of race variables in genetic studies of complex traits and the goal of reducing health disparities: a transdisciplinary perspective. American Psychology, 60(1), 77-103.

Shriver, M. D., & Kittles, R. A. (2004). Genetic ancestry and the search for personalized genetic histories. Nature Reviews Genetics, 5, 611-618.

Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., . . . Jones, K. W. (2004). The genomic distribution of population substructure in four populations using 8525 autosomal SNPs. Human Genomics, 1, 274-286.

Shriver, M. D., Parra, E. J., Dios, S., Bonilla, C., Norton, H., Jovel, C., . . . Kittles, R. A. (2003). Skin pigmentation, biogeographical ancestry and admixture mapping. Human Genetics, 112, 387-399.

Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R., & Farrell, R. E. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. American Journal of Human Genetics 60, 957-964.

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. London: Chapman & Hall.

Simon, H. A. (1962). The architecture of complexity. Proceedings of the American Philosophical Society, 106(6), 467-482.

Sjøvold, T. (1984). A report on the heritability of some cranial measurements and non-metric traits. In G. N. Van Vark & W. W. Howells (Eds.), Multivariate statistics in physical anthropology (pp. 223-246). Dordrecht: D. Reidel Publishing Co.

Sledzik, P. S., & Ousley, S. D. (1991). Analysis of six Vietnamese trophy skulls. Journal of Forensic Sciences, 36, 520-530.

Slice, D. E. (Ed.). (2005). Modern Morphometrics in Physical Anthropology. Norwell, MA: Kluwer Academic Publishers.

Slice, D. E., & Ross, A. (2009). 3D-ID: geometric morphometric classification of crania for forensic scientists (Version 1.0). Retrieved from http://www.3d-id.org

Slice, D. E., & Ross, A. H. (2004). Population affinities of "Hispanic" crania: implications for forensic identification. Paper presented at the American Academy of Forensic Science, Dallas, TX.

Slicing Soup (Biotechnology editorial). (2002). Nature, 20, 637.

Smay, D. B., & Armelagos, G. J. (2000). Galileo Wept: A Critical Assessment of the Use of Race in Forensic Anthropology. Transforming Anthropology, 9(2), 19-29.

Smedley, A. (2007). Race in North America: Origin and Evolution of a Worldview (Vol. 3rd). Boulder: Westview Press.

Smedley, A. (2007). The history of the idea of race...and why it matters. Paper presented at the American Association of Anthropology conference "Race Human Variation and Disease: Consensus and Frontiers, Warrenton, VA. www.understandingrace.org/resources/pdf/disease/smedley.pdf

Smith, B. H., Garn, S. M., & Hunter, W. S. (1986). Secular trends in face size. Angle Orthodontist, 56, 196-204.

Smith, G. (2005). The Genomics Age: How DNA Technology Is Transforming the Way We Live and Who We Are. New York: AMACOM.

Smith, H. F. (2009). Which cranial regions reflect molecular distances reliably in humans? Evidence from three-dimensional morphology. American Journal of Human Biology, 21, 36-47.

Smyth, P. (2000). Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood. Statistics and Computing, 10, 63-72.

Snow, C. C. (1982). Forensic Anthropology Annual Review of Anthropology, 11, 97-131.

Snow, C. C., Hartman, S., Giles, E., & Young, F. A. (1979). Sex and Race Determination by Calipers and Computer: A Test of the Giles and Elliot Discriminant Functions in 52 Forensic Cases. Journal of Forensic Sciences, 24, 448-460.

Sofaer, J. A. (1973). A model relating developmental interaction and differential evolutionary reduction of tooth size. Evolution, 27, 427-434.

Sofaer, J. A., Bailit, H. L., & MacLean, C. J. (1971). A developmental basis for differential tooth reduction during hominid evolution. Evolution, 25, 506-517.

Sokal, R. R., & Sneath, P. H. A. (1963). Principles of Numerical Taxonomy. San Francisco: W. H. Freeman.

Soong, R. (1999). Racial Classifications in Latin America, 2010, from http://www.zonalatina.com/Zldata55.htm

Soromenho, G. (1994). Comparing approaches for testing the number of components in a finite mixture model. Computational Statistics, 9, 65-78.

Sparks, C. S. (2001). Reassessment of Cranial Plasticity in Man: A Modern Critique of Changes in Bodily Form of Descendants of Immigrants. M.A., Department of Anthropology, University of Tennessee, Knoxville, TN.

Sparks, C. S., & Jantz, R. L. (2002). A Reassessment of Human Cranial Plasticity: Boas Revisited. Proceedings of the National Academy of Sciences of the United States of America, 99(23), 14636-14639.

Sparks, C. S., & Jantz, R. L. (2003). Changing Times, Changing Faces: Franz Boas's Immigrant Study in Modern Perspective. American Anthropologist, 105(2), 333-337.

Spencer, F. (1981). Charter members of the American Association of Physical Anthropologists. American Journal of Physical Anthropology, 56(4), 531-535.

Spencer, F. (1981). The Rise of Academic Physical Anthropology in the United States, 1880-1980: A Historical Overview. American Journal of Physical Anthropology, 56, 353-364.

Spencer, F. (1997). A History of Physical Anthropology. An Encyclopedia. London: Garland Publishing.

Spitze, K. (1993). Population structure in Daphnia obtusa: quantitative genetic and allozymic variation. Genetics, 135, 367-374.

Spradley, M. K. (2006). Biological anthropological aspects of the African diaspora: geographic origins, secular trends, and plastic versus genetic influences utilizing craniometric data. Ph.D., University of Tennessee, Knoxville, TN.

Spradley, M. K., & Algee-Hewitt, B. (2007). Estimating Geographic Ancestry of Hispanic Crania using Geometric Morphometrics. Paper presented at the Annual Meeting of the American Academy of Forensic Sciences, San Antonio, TX.

Spradley, M. K., & Jantz, R. L. (2005). Biological variation of Hispanic (Spanish speaking) peoples. Paper presented at the 57th Annual Meeting of the American Academy of Forensic Sciences, New Orleans, LA.

Spradley, M. K., Jantz, R. L., Robinson, A., & Peccerelli, F. (2008). Demographic Change and Forensic Identification: Problems in Metric Identification of Hispanic Skeletons. Journal of Forensic Sciences, 53, 21-28.

Spradley, M. K., Ousley, S. D., & Jantz, R. L. (2008). Evaluating Cranial Morphometric Relationships using Discriminant Function Analysis. (Abstract) American Journal of Physical Anthropology, 135(S46), 199.

St. Hoyme, L. E., & Yasar Iscan, M. (1989). Determination of sex and race: Accuracy and assumptions Reconstruction of Life from the Skeleton (pp. 53-93). New York: Alan R. Liss, Inc.

Stanton, W. (1960). The leopard's spots: scientific attitudes toward race in America 1815-59. Chicago: University of Chicago Press.

State and Country QuickFacts. Data derived from Population Estimates, Census of Population and Housing, Small Area Income and Poverty Estimates, State and County Housing Unit Estimates, County Business Patterns, Nonemployer Statistics, Economic Census, Survey of Business Owners, Building Permits, Consolidated Federal Funds (2011). Retrieved from http://quickfacts.census.gov/qfd/states/00000.html.

State and County Quick Facts. Data derived from Population Estimates, Census of Population and Housing, Small Area Income and Poverty Estimates, State and County Housing Unit Estimates, County Business Patterns, Nonemployer Statistics, Economic Census, Survey of Business Owners, Building Permits, Consolidated Federal Funds Report. (2011). Retrieved from http://quickfacts.census.gov/qfd/states/00000.html.

Steadman, D. W. (2009). Personal Identification: Theory and Applications. The Case Study Approach. In D. W. Steadman (Ed.), Hard evidence: case studies in forensic anthropology. (2nd ed., pp. 1-7). Upper Saddle River, NJ: Prentice Hall.

Steadman, D. W. (Ed.). (2009). Hard evidence: Case studies in forensic anthropology (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Steadman, D. W., & Andersen, S. A. (2009). The Marty Miller Case: Introducing Forensic Anthropology. In D. W. Steadman (Ed.), Hard evidence: case studies in forensic anthropology (2nd ed., pp. 8-28). Upper Saddle River, NJ: Prentice Hall.

Steadman, D. W., & Konigsberg, L. W. (2009). Multiple points of similarity. In D. W. Steadman (Ed.), Hard evidence: case studies in forensic anthropology (2nd ed., pp. 68-79). Upper Saddle River, NJ: Prentice Hall.

Steadman, D. W., Adams, B. J., & Konigsberg, L. W. (2006). Statistical basis for positive identification in forensic anthropology. American Journal of Physical Anthropology, 131, 15-26.

Steel, D. G., & Powell, J. F. (1999). Peopling of the Americas: A Historical and Comparative Perspective. In R. Bonnichsen (Ed.), Who Were the First Americans (pp. 97-126). Corvallis: Center for the Study of the First Americans, Oregon State University.

Steele, D. G. , & Bramblett, C. A. (1988). The anatomy and biology of the human skeleton. College Station: Texas A&M University Press.

Steele, R. J., & Raftery, A. E. (2009). Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models Technical Report no. 559. Seattle, WA: Department of Statistics, University of Washington.

Steele, R. J., Wang, N., & Raftery, A. E. (2010). Inference from multiple imputation for missing data using mixtures of normals. Statistical Methodology, 7, 351-365.

Steele, R. J., Wang, N., & Raftery, A. E. (2010). Package 'MImix'. R documentation Retrieved June 1, 2011, from http://cran.r-project.org/web/packages/MImix/MImix.pdf

Stefflova, K. , Dulik, M. C., Barholtz-Sloan, J. S., Pai, A. A., & Walker, A. H. (2011). Dissecting the Within-Africa Ancestry of Populations of African Descent in the Americas. PLoS ONE, 6(1), e14495.

Stein, G. (1971). Everybody's Autobiography. New York: Cooper Square Publishers.

Stein, G. (1988). Biological Science and the Roots of Nazism. American Scientist, 76(1), 50-58.

Stephens, J. C. , Schneider, J. A., Tanguay, D. A., Choi, J., Archarya, T., Stanley, S. E., & Jiang, R. (2001). Haplotype variation and linkage disequilibrium in 313 human genes. Science, 293, 489-493.

Stewart, T. (1979). Essentials of Forensic Anthropology. Springfield, IL: Charles C. Thomas.

Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885-1920. Journal of the American Statistical Association, 68, 872-879.

Stigler, S. M. (1986). The history of Statistics: the measurement of uncertainty before 1900. Cambridge, MA: Harvard University Press.

Stigler, S. M. (1999). Statistics on the Table. Cambridge, Mass.: Harvard University Press.

Stocking, G. W., Jr. (1960). Franz Boas and the founding of the American Anthropological Association. American Anthropologist, 62, 1-17.

Stocking, G. W., Jr. (1968). Race, Culture, and evolution: essays on the history of anthropology. New York, NY: Free Press.

Stoessiger, B. N. (1927). A Study of the Badarian Crania Recently Excavated by the British School of Archeology in Egypt. Biometrika, 19, 110-150.

Stojanowski, C. M., & Duncan, W. N. (2008). Anthropological contributions to the cause of the Georgia Martyrs Occasional Papers of the Georgia Southern Museum, Number 3.: Georgia Southern Museum.

Stojanowski, C. M., & Duncan, W. N. (2009). Historiography and forensic analysis of the Fort King George "skull": Craniometric assessment using the specific population approach. American Journal of Physical Anthropology, 140, 275-289.

Stone, G., Atkinson, R., Rokas, A., Csoka, G., & Nieves-Aldrey, J. L. (2001). Differential success in northwards range expansion between ecotypes of the marble gallwasp Andricus kollari: a tale of two lifecycles. Molecular Ecology, 10, 761-778.

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion Journal of the Royal Statistical Society Series B, 38, 44-47.

Stoneking, M., Fontius, J. J., Clifford, S. L., Soodyall, H., Arcot, S. S., Saha, N., . . . Batzer, M. A. (1997). Alu insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. Genome Research, 7, 1061-1071.

Strait, D. S., Moniz, M. A., & Strait, P. T. (1996). Finite Mixture Coding: A New Approach To Coding Continuous Characters. Systematic Biology, 45(1), 67-78.

Strauss, A., & Hubbe, M. (2010). Craniometric Similarities Within and between Human Populations in Comparison with Neutral Genetic Data. Human Biology, 82(3), 315-330.

Strengthening Forensic Science in the United States: A Path Forward. (2009). National Research Council (NRC), Committee on Identifying the Needs of the Forensic Sciences Community: U. S. Department of Justice Retrieved from www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf

Stringer, C. B. (1990). The emergence of modern humans. Scientific American, 263, 98-104.

Stringer, C. B. (1992). Replacement, continuity and the origin of Homo sapiens. In G. Brauer & F. H. Smith (Eds.), Continuity or Replacement: Controversies in Homo sapiens Evolution (pp. 9-24). Rotterdam: A.A. Balkema.

Stringer, C. B. (2002). Modern human origins: progress and prospects. Philosophical Transactions of the Royal Society of London Series B, 357, 563-579.

Stringer, C. B., & Andrews, P. (1988). Genetic and fossil evidence for the origin of modern humans. Science, 239, 1263-1268.

Štrkalj, G. (2007). The status of the concept of race in contemporary biological anthropology: a review. Anthropologist, 9, 73-78.

Subcommittee on Forensic Science (SoFS) website. (2011) Retrieved February 15, 2011, from http://www.forensicscience.gov/

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. Scandinavian Journal of Statistics 1(2), 49-58.

Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. Communications in Statistics – Simulation and Computation 5(1), 55-64.

Sutcliffe, J. P. (1993). Concept, class, and category in the tradition of Aristotle. In I. van Mechelen, J. Jampton, R. S. Michalski & P. Theuns (Eds.), Categories and Concepts (pp. 35-65). London: Academic Press.

Suzuki, H. (1969). Microevolutional changes in the Japanese population from the prehistoric age to the present day. Journal of the Faculty of Sciences of the University of Tokyo, Sec V., 3(4), 279-308.

Swagatam, D., Ajith, A., & Konar, A. (2009). Metaheuristic Clustering. Heidelberg: Springer.

Symons, M. J. . (1981). Clustering criteria and multivariate normal mixtures. Biometrics, 37, 35-43.

Tabachnick, B. G., & Fidell, L. S. (2001). Using Multivariate Statistics, Fourth Edition. Needham Heights, MA: Allyn & Bacon.

Tamm, E., Kivisild, T., Reidla, M., Metspalu, M. , Smith, D. G., Mulligan, C. J., . . . Malhi, R. S. (2007). Beringian standstill and spread of Native American founders. PLoS ONE 5, 1-6.

Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. American Journal of Human Genetics, 79, 1-12.

Tang, H., Peng, J., Wang, P., & Risch, N. (2005). Estimation of individual admixture: Analytical and study design considerations. Genetic Epidemiology, 28, 289-301.

Tang, H., Quertermous, T., Rodrigues, B., Kardia, S. L., Zhu, X., Brown, A., . . . Risch, N. J. (2005). Genetic structure, self-identified race/ethnicity, and confounding in care-control association studies. American Journal of Human Genetics, 76, 268-275.

Tarazona-Santos, E., & Santos, F. R. (2002). The peopling of the Americas: a second major migration? American Journal of Human Genetics, 70, 1377-1380.

Tarter, M. E., & Lock, M. D. (1993). Model Free Curve Estimation. London: Chapman and Hall.

Tatsuoka, M. M. (1971). Multivariate analysis. New York: John Wiley and Sons.

Tattersall, I. (2009). Out of Africa: Modern human origins Proceedings of the National Academy of Sciences of the United States of America 106, 16018-16021.

Taylor, J. V., Roh, L., & Goldman, A. D. (1984). Metropolitan forensic anthropology team (MFAT) case studies in identification: 2. Identification of a Vietnamese trophy skull. Journal of Forensic Science, 29(1253-1259).

Team, R Development Core. (2010). R: a language and environment for statistical computing (Version 2.13.0). Vienna, Austria: R foundation for statistical computing. Retrieved from http://www.R-project.org

Templeton, A. (2002). Out of Africa again and again. Nature, 416, 45-51.

Templeton, A. R. (1998). Human Races: A genetic and evolutionary perspective. American Anthropologist, 100, 632-650.

Templeton, Alan R. (1999). Human Races: A Genetic and Evolutionary Perspective. American Anthropologist, 100(3), 632-650.

Tenebaum, J. B., de Silva, V. V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290, 2319-2323.

Thangaraj, K., Chaubey, G., Kivisild, T., Reddy, A. G., Singh, V. K., Rasalkar, A. A., & Singh, L. (2005). Reconstructing the origin of Andaman islanders. Science, 308, 996.

Thangaraj, K., Singh, L., Reddy, A. G., Rao, V. R., Sehgal, S. C., Underhill, P. A., . . . Hagelberg, E. (2003). Genetic affinities of the Andaman islanders, a vanishing human population. Current Biology, 13, 86-93.

The Genographic Project. (2011) Retrieved January 20, 2011, from https://genographic.nationalgeographic.com/genographic/lan/en/about.html

The International Hap Map Project. (2011) Retrieved July, 1, 2011, from http://hapmap.ncbi.nlm.nih.gov/

The Mankind Quarterly. (1961-) Retrieved September 13, 2010, from http://www.mankindquarterly.org/about.html

Thio, C. L., Thomas, D. L., Goedert, J. J., Vlahov, D., Nelson, K. E., Hilgartner, M. W., . . . Carrington, M. (2002). Racial differences in HLA class II associations with hepatitis C virus outcomes. Journal of Infectious Diseases, 184, 16-21.

Thomas, Clinton. (2011, January 5). Comment on "Alas, Poor Yorick ... or Is It Othello? Can you tell a person's race from his or her skull?" Slate Explainer Retrieved January 19, 2011, from http://www.slate.com/id/2279942/

Thomas, D. C., Haile, R. W., & Duggan, D. (2005). Recent developments in genomewide association scans: a workshop summary and review. American Journal of Human Genetics, 77, 337-345.

Thorisson, G. A., Smith, A. V., Kirishnan, L., & Stein, L. D. (2005). The International HapMap Project Web site. Genome Research, 15, 1591-1593.

Tibshirani, R., & Leisch, F. (2009). bootstrap: Functions for the Book "An Introduction to the Bootstrap." (Version 1.0-22).

Tishkoff, S. A., & Gonder, M. K. (2006). Human origins within and out of Africa. In M Crawford (Ed.), Anthropological Genetics: theory methods and applications. (pp. 337-379). Cambridge: Cambridge University Press.

Tishkoff, S. A., & Verrelli, B. C. (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. Annual Review of Genomics and Human Genetics 4, 293-340.

Tishkoff, S. A., & Williams, S. M. (2002). Genetic Analysis of African Populations: Human Evolution and Complex Disease. Nature Reviews Genetics, 3, 611-621.

Tishkoff, S. A., Floyd, A., Reed, F., Friedlaender, R., Ehret, C., Ranciaro, A., . . . Williams, S. M. (2009). The Genetic Structure and History of Africans and African Americans. Science, 324, 1035-1044.

Tishkoff, S. A., Pakstis, A. J., Stoneking, M., Kidd, J. R., Destro-Bisol, G., Sanjantila, A., . . . Clark, A. G. (2000). Short tandem-repeat polymorphism/Alu haplotype variation at the PLAT locus: Implications for modern human origins. American Journal of Human Genetics, 67(4), 901-925.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). Statistical Analysis of Finite Mixture Distributions. New York: John Wiley & Sons.

Tomanio, J., Bryan, C., Yarnall, K., Vessels, J., Ibinson, J., Snowberger, M., & Haub, C. (2011). The Face Of Seven Billion National Geographic Magazine, 219(3), supplement poster.

Tooby, J., & Cosmides, L. (1990). On the universality of human nature and the uniqueness of the individual: The role of genetics and adaptation. Journal of Personality and Social Psychology, 58, 17-67.

Tran, T. N., Wehrens, R., & Buydens, L. M. C. (2006). SMIXTURE: A Strategy of Mixture Models Clustering of Multivariate Images. Journal of Chemometrics, 19, 607-614.

Trial of the Major War Criminals before the International Military Tribunal Nuremberg, 14 November 1945 - 1 October 1946, (1945-1946).

Trinkaus, E. (2005). Early modern humans. Annual Review of Anthropology, 34, 207-230.

Trinkaus, E. . (1987). The Neanderthal face: evolutionary and functional perspectives on a recent hominid face. Journal of Human Evolution, 16, 429-443.

Tryon, R. C. (1939). Cluster analysis. Ann Arbor: Edwards Brothers.

Tukey, J. (1977). Exploratory Data Analysis. Boston: Addison-Wesley.

Turakulov, R., & Easteal, S. (2003). Number of SNPs loci needed to detect population structure. Human Heredity, 55, 37-45.

Turner, C. G., II. (1987). Late Pleistocene and Holocene population history of East Asia based on dental variation. American Journal of Physical Anthropology, 73, 305-321.

Turner, C. G., II. (1990). Major features of sundadonty and sinodonty, including suggestions about East Asian microevolution, population history, and late Pleistocene relationships with Australian Aboriginals. American Journal of Physical Anthropology, 82, 295-317.

Tyler, T. . (2006). Viewing CSI and the Threshold of Guilt: Managing Truth and Justice in Reality and Fiction. The Yale Law Journal, 115, 1050-1085.

Ubelaker, D. H. (1998). FORDISC 2.0: Personal Computer Forensic Discriminant Functions (Book Review). International Journal of Osteoarchaeology, 8, 128-133.

Ubelaker, D. H. (1998). FORDISC 2.0: Personal computer forensic discriminant functions. International Journal of Osteoarchaeology, 8, 128-133.

Ubelaker, D. H. (1999). Ales Hrdlicka's Role in the History of Forensic Anthropology. Journal of Forensic Sciences, 44(4), 724-730.

Ubelaker, D. H. (1999). The Legacy of Aleš Hrdlička in Smithsonian Forensic Anthropology. In M. Prokopec (Ed.), Aleš Hrdlička - 130th Anniversary of Birth (pp. 43-46). Humpolec: Czech Anthropological Society

Ubelaker, D. H. (2000). A History of Smithsonian–FBI Collaboration in Forensic Anthropology, Especially in Regard to Facial Imagery. Forensic Sciences Communications, 2(4), np.

Ubelaker, D. H. (2000). Aleš Hrdlička: Czech-American Physical Anthropologist Retrieved September 6, 2010, from http://www.svu2000.org/publications/UBERLAKER.PDF

Ubelaker, D. H. (2008). Issues in the Global Applications in Methodology of Forensic Anthropology. Journal of Forensic Sciences, 53, 606-617.

Ubelaker, D. H., Ross, A. H., & Graver, S. M. (2002). Application of Forensic Discriminant Functions to a Spanish Cranial Sample. Forensic Science Communications 4(3), np.

Underhill, P. A., Shen, P. D., Lin, A. A., et al. (2000). Y chromosome sequence variation and the history of human populations. Nature Genetics, 26, 358-361.

Usher, B. (2002.). Reference samples: the first step in linking biology and age in the human skeleton. In R. Hoppa & J. Vaupel (Eds.), Paleodemography: age distributions from skeletal samples (pp. 29-47). London: Cambridge University Press.

Van Ness, J. W., & Yang, J. J. (1998). Robust discriminant analysis: Training data breakdown point. Journal of Statistical Planning and Inference, 67, 67-83.

Van Vark, G. N. (1985). Multivariate analysis in physical anthropology. In P. R. Krishnaiah (Ed.), Multivariate Analysis VI (pp. 599-611). New York: Elsevier Science Publishing Co. .

Van Vark, G. N., & Howells, W. W. (Eds.). (1984). Multivariate Statistics in Physical Anthropology. Dordrecht: D. Reidel Publishing Co.

Van Vark, G. N., Bilsborough, A., & Henke, W. (1992). Affinities of European Upper Palaeolithic Homo sapiens and later Human evolution. Journal of Human Evolution, 23, 401-417.

Vaupel, J. W., & Yashin, A. I. (1985). The deviant dynamics of death in heterogeneous populations. In N. B. Tuma (Ed.), Sociological Methodology (pp. 176-185). London: Jossey-Bass.

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S (4th ed.). New York: Springer.

Venter, C. (2000). Text of Remarks on the Completion of the First Survey of the Entire Human Genome Project. Retrieved from http://clinton5.nara.gov/WH/New/html/genome-20000626.html.

Verano, J. W. (23003). Mummified trophy heads from Peru: diagnostic features and medicolegal significance. Journal of Forensic Science, 48, 525-530.

Vermunt, J. K. (2003). Applications of latent class analysis in social science research. Lecture Notes in Artificial Intelligence, 2711(22-36).

Vicente, J. G., Conway, J., Roberts, S. J., & Taylor, J. D. (2001). Identification and origin of Xanthomonas campestris pv. campestris races and related pathovars. Phytopathology, 91, 492-499.

Vidarsdottir, U. S., O'Higgins, P., & Stringer, C. B. (2002). A geometric morphometric study of regional differences in the ontogeny of the modern human facial skeleton. Journal of Anatomy, 201, 211-229.

Vogele, J. (1999). Urban mortality change in Britain and Germany 1870–1913. Liverpool: Liverpool University Press.

Von Cramon-Taubadel, N. (2009). Congruence of individual cranial bone morphology and neutral molecular affinity patterns in modern humans. American Journal of Physical Anthropology, 140, 205-215.

Von Cramon-Taubadel, N. (2009). Revisiting the homoiology hypothesis: the impact of phenotypic plasticity on the reconstruction of human population history from craniometric data. Journal of Human Evolution, 57, 179-190.

Von Cramon-Taubadel, N., & Lycett, S. J. (2008). Brief communication: human cranial variation fits iterative founder effect model with African origin. American Journal of Physical Anthropology, 136, 108-113.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307-333.

Wade, N. (2002). A New Look At Old Data May Discredit a Theory on Race, New York Times.

Wade, N. (2002, December 20). Gene Study Identifies Five Main Human Populations, Linking Them to Geography, New York Times, p. A37.

Wade, P. (2004). Human Nature and Race. Anthropological Theory 4, 157-172.

Waldron, R. (1994). Counting the Dead: the epidemiology of skeletal populations. New York: John Wiley & Sons.

Waldron, T. (1987). The relative survival of the human skeleton: implications for paleodemography. In A. Boddington, A. N. Garland & R. C. Janeway (Eds.), Death, decay, and reconstruction: approaches to archaeology and forensic science (pp. 55-64). Manchester: Manchester University Press.

Walker, P. L., Johnson, J., & Lambert, P. (1988). Age and Sex Biases in the Preservation of Human Skeletal Remains. American Journal of Physical Anthropology 76, 183-188.

Wall, J. D. . (2000). Detecting ancient admixture in humans using sequence polymorphism data. Genetics, 154, 1271-1279.

Wallace, C. S., & Downe, D. L. (1994). Intrinsic classification by MML - the Snob program. Paper presented at the 7th Australian Joint Conference on Artificial Intelligence.

Wand, Y., & Wang, R. Y. (1996). Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the ACM, 39(11), 86-95.

Wang, R. Y., Reddy, M. P., &. Kon, H. B. (1995). Toward Quality Data: An Attribute based Approach, Decision Support Systems(13), 349-372.

Wang, S., Lewis, C. M., Jr., Jakobsson, M., Ramachandran, S., Ray, N., & et al. (2007). Genetic variation and population structure in Native Americans. PLoS Genetics, 3(11), 185.

Wantanabe, S. (1960). Knowing And Guessing. New York Wiley.

Waser, N. (2000). Review: A Primer of Population Genetics. Heredity, 85(510).

Washburn, S. L. (1950). Constitutional Types and Physical Anthropology. American Anthropologist, 52, 437.

Washburn, S. L. (1951). The new physical anthropology Transactions of the New York Academy of Science (Vol. 13, pp. 298-304).

Washburn, S. L. (1951). The New Physical Anthropology. Yearbook of Physical Anthropology, 7, 124-130.

Washburn, S. L. (1963). The Study of Race. American Anthropologist, 65, 521-532.

Washburn, S. L. (1983). Evolution of a Teacher. Annual Review of Anthropology, 12, 1-24.

Wasserman, L. (2000). Bayesian model selection and model averaging. Journal of Mathematical Psychology, 44, 92-107.

Watanabe, S. (1985). Pattern Recognition - Human And Mechanical. New York: Wiley.

Watkins, W. S., Rogers, A. R., Ostler, C. T., Wooding, S., Bamshad, M. J., Brassington, A. M., . . . Prasad, B. V. (2003). Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome Research, 13(7), 1607-1618.

Way, T. W., Sahiner, B., Hadjiiski, L. M., & Chan, H. (2010). Effect of finite sample size on feature selection and classification: A simulation study Medical Physics, 37(2), 907-920.

Weakliem, D. (1998). A critique of the Bayesian information criterion. Sociological Methods and Research, 27, 359-397.

Weaver, T. D., & Roseman, C. C. (2008). New developments in the genetic evidence for modern human origins. Evolutionary Anthropology, 17, 69-80.

Weaver, T. D., & Roseman, C. C. . (2005). Ancient DNA, late Neandertal survival, and modern-human-Neandertal genetic admixture. Current Anthropology, 46, 677-683.

Weaver, T. D., Roseman, C. C., & Stringer, C. B. (2007). Were neandertal and modern human cranial differences produced by natural selection or genetic drift? Journal of Human Evolution, 53(2), 135-145.

Webb, A. (1999). Statistical pattern recognition. New York: Oxford University Press.

Weber, G. W., & Bookstein, F. L. (2001). Virtual Anthropology. A guide to a new interdisciplinary field (1st ed.). New York: Springer.

Wedel, M. (2001). GLIMMIX: Software for estimating mixtures and mixtures of generalized linear models. Journal of Classification, 18, 129-135.

Wedel, M., & DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. Journal of Classification, 12(1), 21-55.

Wedel, M., DeSarbo, W. S., Bult, J. R., & Ramaswamy, V. (1993). A latent class Poisson regression model for heterogeneous count data. Journal of Applied Econometrics, 8, 397-411.

Wei, G. C. G., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association, 85(411), 699-704.

Weihs, C., & Gaul, W. (Eds.). (2004). Classification - the Ubiquitous Challenge. Berlin: Springer.

Weir, B. S. (1996). Genetic Data Analysis. Sunderland, MA: Sinauer Associates, Inc.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. Evolution, 38, 1358-1370.

Weisensee, K. E., & Jantz, R. L. (2011). Secular changes in craniofacial morphology of the Portuguese using geometric morphometrics. American Journal of Physical Anthropology, 145(Online Preview), NA.

Weiss, k. M., & Clark, A. G. (2002). Linkage disequilibrium and the mapping of complex human traits. Trends in Genetics, 18, 19-24.

Weiss, K. M., & Long, J. C. (2009). Non-Darwinian estimation: My ancestors, my genes' ancestors. Genome Research, 19, 703-710.

Weiss, R. (2008, March 25). Genetic Testing Gets Personal, The Washington Post.

Weldon, W. F. R. (1892). On certain correlated variations in Carcinus moenas. Proceedings of the Royal society of London, 54, 318-329.

Weldon, WFR. (1892). Certain correlated variations in Crangon vulgaris. Proceedings of the Royal Society of London Series B, 51, 2-21.

Wescott, D. J., & Jantz, R. L. (2005). Assessing craniofacial secular change in American whites and blacks using geometric morphometry. In D. E. Slice (Ed.), Modern Morphometrics in Physical Anthropology, Volume V: Developments in Primatology: Progress and Prospects (pp. 231-246). New York: Kluwer Academic Press.

West, C., Trawalter, S., & Richeson, J.A. (2008). Let’s Talk About Race, Baby! When Whites’ and Blacks’ Interracial Contact Experiences Diverge Journal of Experimental Social Psychology, 44(4), 1214-1217.

West, M. (1993). Approximating posterior distributions by mixtures. Journal of the Royal Statistical Society Series B, 55, 409-422.

Wheat, A. D. (2009). Assessing Ancestry Through Nonmetric Traits of the Skull: A Test of Education and Experience. M.A., Texas State University, San Marcos.

Wherens, R., Buydens, L. M. C., Fraley, C, & Raftery, A. E. (2004). Model-based Clustering for Image Segmentation and Large Datasets via Sampling. Journal of Classification, 21, 231-253.

White, T. D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G. D., Suwa, G., & Howell, F. C. . (2003). Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature, 423, 742-747.

Whitlock, M. C. (1999). Neutral additive genetic variance in a metapopulation. Genetical Research, 74, 215-221.

Whitmarsh, I., & Jones, D. (2010). Governance and the Uses of Race. In I. Whitmarsh & D. Jones (Eds.), What's the Use of Race? Modern Governance and The Biology of Difference (pp. 1-23). Cambridge: MIT Press.

Whitmarsh, I., & Jones, D. (Eds.). (2010). What's the Use of Race? Modern Governance and The Biology of Difference. Cambridge: MIT Press.

Whitmarsh, I., & S., Jones D. (Eds.). (2010). What's the Use of Race? Modern Governance and the Biology of Difference. Cambridge, MA: MIT Press.

Whitmarsh, Ian, & Jones, David. (2010). 2010 Governance and the Uses of Race. In I. Whitmarsh & D. Jones (Eds.), What's the Use of Race? Modern Governance and The Biology of Difference (pp. 1-23). Cambridge: MIT Press.

Who Do You Think You Are? (2010), from http://www.nbc.com/who-do-you-think-you-are/about/index.shtml

Wienker, C. W. (2001). The Anthropological Perspective on Race: An Historical Overview. Perspectives: American Association of Behavioral and Social Sciences Journal, 4, 34.

Wienker, C. W. (2005). Using biological data to challenge the reality of race. Perspectives: Electronic Journal of the American Association of Behavioral and Social Sciences, 8. Retrieved from http://aabss.org/Perspectives2005/

Wiersema, J., Love, J. C., & Naul, L. G. (2009). The influence of the Daubert guidelines of anthropological methods of scientific identification in the medical examiner setting. In D. W. Steadman (Ed.), Hard Evidence: Case Studies in Forensic Anthropology, (pp. 80-90). Upper Saddle River: Prentice Hall.

Willekens, F. J. (1999). Modeling approaches to the indirect estimation of migration flows: From entropy to EM. Mathematical Population Studies, 7(3), 239-278.

Willey, G. R., & Philips, P. (1958). Method and Theory in American Archeology. Chicago: University of Chicago Press.

Williams, D. R. (1994). The concept of race in Health Sciences Research: 1966 to 1990. Health Services Research, 29, 261-274.

Williams, D. R. (1997). Race and health: basic questions, emerging directions. Annals of Epidemiology, 7, 322-333.

Williams, F. L., & Armelagos, G. J. (2007). Reply to M Hubbe and WA Neves "On the misclassification of human crania." Current Anthropology, 48, 285-288.

Williams, F. L., Belcher, R. L., & Armelagos, G. J. (2005). Forensic misclassification of ancient Nubian crania: Implications for assumptions about human variation. Current Anthropology, 46, 340-346.

Williams, F. L., Belcher, R. L., & Armelagos, G. J. (2007). Reply to SOY Keita . "On Meroitic Nubian Crania, Fordisc 2.0, and Human Biological History". Current Anthropology, 48, 425-427.

Williams, P. B., Erickson, P., & Niven, L. (2000). Retrieving History: The 18th Century Mortuary History of the Little Dutch Church, Halifax Retrieved February 2, 2011, from http://canadianarchaeology.com/caa/node/2150

Williams-Blangero, S., & Blangero, J. (1992). Quantitative genetic analysis of skin reflectance: a multivariate approach. Human Biology, 64, 35-49.

Wilmoth, J. R. (2000). Demography of longevity: past, present, and future trends. Experimental Gerontology, 35, 1111-1129.

Wilson, E. O., & Brown, W., Jr. (1953). The subspecies concept and its taxonomic application. Systematic Zoology, 22, 97-111.

Wilson, J. F., Weale, M. E., Smith, A. C., Gratrix, F., Fletcher, B., Thomas, M. G., . . . Goldstein, D. B. (2001). Population genetic structure of variable drug response. National Review of Genetics, 29(265-269).

Wilson, R. J., & Algee-Hewitt, B. (2008). Alcoholism Related Weight Loss in Bone. Paper presented at the International Association of Forensic Science's 18th Triennial Meeting, New Orleans, LA.

Wilson, R. J., & Algee-Hewitt, B. (2009). [Inter]Facing Age: a test of the ADBOU age estimation software in a forensic context Paper presented at the Annual Meeting of the American Academy of Physical Anthropologists, Chicago, Il.

Wilson, R. J., Algee-Hewitt, B., & Meadows Jantz, L. (2007). Demographic trends within the Forensic Anthropology Center's body donation program. Paper presented at the Annual Meeting of the American Academy of Physical Anthropologists, Philadelphia, PA.

Wilson, R. J., Herrmann, N. P., & Meadows Jantz, L. (2010). Evaluation of Stature Estimation from the Database for Forensic Anthropology. Journal of Forensic Sciences, 55(3), 684-689.

Witherspoon, D. J., Marchani, E. E., Watkins, W. S., Ostler, C. T., Wooding, S. P. , et al. (2006). Human population genetic structure and diversity inferred from polymorphic L1(LINE-1) and Alu insertions. Human Heredity, 62(30-46).

Witherspoon, D. J., Wooding, S., Rogers, A. R., Marchani, E. E., Watkins, W. S., Batzer, M. A. , & Jorde, L. B. (2007). Genetic similarities within and between human populations. Genetics, 176, 351-359.

Witherspoon, D. J., Wooding, S., Rogers, A. R., Marchani, E. E., Watkins, W. S., Batzer, M. A., & Jorde, L. B. (2007). Genetic similarities and between human populations. Genetics, 176, 351-359.

Witzig, R. (1996). The medicalization of race: scientific legitimization of a flawed social construct. Annals of Internal Medicine, 125, 675-679.

Wolfe, J. H. (1965). A computer program for the computation of maximum likelihood analysis of types. . San Diego: U.S. Naval Personnel Research Activity.

Wolfe, J. H. (1967). NORMIX: computational methods for estimating the parameters of multivariate normal mixtures of distributions. San Diego: U.S. Naval Personnel Research Activity.

Wolfe, J. H. (1970). NORMIX. Retrieved from http://alumnus.caltech.edu/~wolfe/normix.htm

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, 5, 329-350.

Wolfe, J. H. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal populations Technical Bulletin STB 72-2. . San Diego: U.S. Naval Personnel and Training Research Laboratory.

Wolpoff, M. H. (1989). Multiregional Evolution: the Fossil Alternative to Eden. In P. Mellars & C. B. Stringer (Eds.), The Human Revolution: Behavioural and Biological Perspectives on the Origins of Modern Humans (pp. 62-108). Edinburgh: Edinburgh University Press.

Wolpoff, M. H., & Caspari, R. (1997). Race and Human Evolution: A Fatal Attraction. New York, NY: Simon and Schuster.

Wolpoff, M. H., & Lee, S. H. (2006). Variation in the Habiline Crania - Must it be Taxonomic? . Human Evolution, 21, 71-84.

Wolpoff, M. H., Hawks, J. , & Caspari, R. . (2000). Multiregional, not multiple origins. American Journal of Physical Anthropology, 112, 129-136.

Wolpoff, M., Hawks, J., Frayer, D. W., & Hunley, K. . (2001). Modern human ancestry at the peripheries: a test of the replacement theory. Science, 291, 293-297.

Woodbury, G. (1932). North America: The Indians of Pecos Pueblo. E.A. Hooton. American Anthropologist, 32, 142-143.

Wright, R. L., Williamson, M. A., & Crass, D. C. (2003). Skeletal Analysis of a Nineteenth Century Grave from Sapelo Island. Early Georgia, 31(1), 91-98.

Wright, R. V. S. (1992). Correlation between Cranial Form and Geography in Homo Sapiens: CRANID - A Computer Program for Forensic and Other Applications. Archaeology in Oceania, 27(3), 105-112.

Wright, S. (1931). Evolution in Mendelian populations. Genetics, 16, 97-159.

Wright, S. (1943). Isolation by distance. Genetics, 28(114-138).

Wright, S. (1951). The genetical structure of populations. Annals of Eugenics, 15, 323-354.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution, 19, 395-420.

Wright, S. (1969). Evolution and the Genetics of Populations, Volume 4: The Theory of Gene Frequencies (Vol. 2). Chicago: University of Chicago Press.

Wright, S. (1978). Evolution and the Genetics of Populations, Volume 4: Variability within and among Natural Populations (Vol. 4). Chicago: University of Chicago Press.

Wright, S. (1978). Variability within and among natural populations Evolution and the Genetics of Populations, Volume 4 (Vol. 4, pp. 449-450). Chicago: University of Chicago Press.

Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. The Annals of Statistics, 11(95-103).

Xie, Yu. (1988). Franz Boas and Statistics. Annals of Scholarship, 5, 269-296.

Xing, J., Watkins, W. S., Witherspoon, D. J., Zhang, Y., Guthery, S. L., Thara, R., . . . Jorde, L. B. (2009). Fine-scales human genetic structure revealed by SNP microarrays 2009. Genome Research, 19, 815-825.

Yamaoka, Y., Mégraud, F, Otto, K., Reichard, U., Katzowitsch, E., Wang, X, . . . Suerbaum, S. (2003). Traces of human migrations in Helicobacter pylori populations. Science, 299, 1582-1585.

Young, F. W., Valero-Mora, P., & Friendly, M. (2006). Young, F. W. Valero-Mora, P. and Friendly M. (2006) Visual Statistics: Seeing your data with Dynamic Interactive Graphics. New York: Wiley.

Yu, N., Chen, F. C., Ota, S., Jorde, L. B., Pamilo, P., Patthy, L., . . . Li, W. H. (2002). Larger genetic differences within Africans than between Africans and Eurasians. Genetics, 161, 269-274.

Yu, N., Jensen-Seaman, M. I., Chmnick, L., Kidd, J. R., Deinard, A. S., Ryder, O. A., . . . Li, W. H. (2003). Low nucleotide diversity in chimpanzees and bonobos. Genetics, 164, 1511-1518.

Yunxin, T., & Dechang, P. (2009). Unifying Density-Based Clustering and Outlier Detection. Paper presented at the Proceedings of the 2009 Second International Workshop on Knowledge Discovery and Data Mining Washington, D.C.

Zack, N. (2002). Philosophy of science and race. New York, NY: Routledge.

Zakrzewski, S. R. . (2007). Population continuity or population change: formation of the ancient Egyptian state. American Journal of Physical Anthropology, 132, 501-509.

Zhivotovsky, L., Rosenberg, N., & Feldman, M.W. . (2003). Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. American Journal of Human Genetics, 72, 1171-1186.

Zhu, X., Zhang, S., Tang, H., & Cooper, R. (2006). A classical likelihood based approach for admixture mapping using EM algorithm. Human Genetics, 120, 431-445.

Ziliak, S. T., & McCloskey, D. N. (2008). The cult of statistical significance: how the standard error costs us jobs, justice, and lives. Ann Arbor (MI): The University of Michigan Press.

Zlojutro, M., Rubicz, R., Devor, E. J., Spitsyn, V. A., Makarov, S. V., Wilson, K. , & Crawford, M. H. (2006). Genetic structure of the Aleuts and Circumpolar populations based on mitochondrial DNA sequences: A synthesis. American Journal of Physical Anthropology, 129, 446-464.

Vita

Bridget Frances Beatrice Algee-Hewitt was born on a Friday, in Halifax, Nova Scotia. She and her younger sister, Hannah, spent their childhood in Moncton, New Brunswick, where their parents, Janet and Boyd Algee, are still practicing architects, artists, and gallery owners. After a non-traditional early education, Bridget attended Mount Allison University, where she received a Bachelors of Arts, First Class Honours with Distinction, in Classics. She was awarded a Masters of Arts, *magna cum laude*, from Bryn Mawr College in Classical and Near Eastern Archaeology, before pursuing her Doctor of Philosophy in Biological Anthropology at the University of Tennessee. She held the post of Assistant Professor in Biomedical Sciences at Grand Valley State University, whilst completing her doctoral degree. She will be returning to Knoxville to take up a new position as the Haslam Post-Doctoral Fellow in the Department of Anthropology at the University of Tennessee. She and her significant-other, Mark Algee-Hewitt, anticipate the day when they both hold faculty appointments in the same country. Bridget credits the completion of her dissertation, with all probability, to the fact that she comes from a robust line of bootstrappers.