

**The nature and significance of worldwide human dermatoglyphic variation in
finger ridge-counts**

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

**Kathleen Ann Hauther
December 2023**

Copyright © 2023 by Kathleen Ann Hauther
All rights reserved.

ACKNOWLEDGEMENTS

I would like to thank the individuals who donated their handprints as well as the 116+ researchers and assistants who collected the dermatoglyphic data. Without their efforts, this work would not be possible. I would like to acknowledge the late Mr. Heinz Brehme and Dr. Miyo Yokota for assessing the handprints for dermatoglyphic features like finger patterns and finger ridge-counts.

Many thanks to my co-chairs, Dr. Lee Meadows Jantz and Dr. Steven G Lautzenheiser, for their support and direction to complete the dissertation. Dr. Lee Meadows Jantz is a force in her own research and in her collaborative research through her ability to quickly identify and dismiss unproductive avenues of research and through her commitment to research with integrity. As a result, researchers and students actually use her research in coursework, academic research, and forensic applications, an enigmatic goal I have sought ever since entering the field of Anthropology. This initially drew me to want to work with Dr. Jantz to learn her secrets, and through many years working with her, I have identified an important piece to her success. Should Dr. Jantz agree with work with you, she truly supports her collaborators, students, and mentees by challenging them to be their best selves and gives them the freedom to pursue or not opportunities she sets in their path. This freedom has allowed me to make my own mistakes and also to grow in my own manner. I am very lucky to know Dr. Jantz and to have benefited from her mentorship. I am also grateful to have learned some of her secrets so I can incorporate them in my own work and be my best self.

Thank you to Dr. R Alex Bentley whose enthusiasm for the anthropological applications kept the entire committee excited throughout the project, and thank you for taking your time to

explain the capabilities in interpretation of the results and encouraging me to ‘go out on a limb’, step outside of my comfort zone, and engage with the possibilities in this research. Thank you to Dr. Richard L Jantz for openly sharing his worldwide dermatoglyphics data set and for allowing me to participate in its renaissance. Also, thank you for keeping me grounded in theory and practicality so the research could be useful to others. Thank you to Dr. Michael Jantz who jumped into the committee with both feet just as time seemed to speed up and who encouraged thinking big in applications for this research.

Thank you to Dr. Jim Fordyce who lowered the barrier to entry for me for statistics in biology by breaking open the black boxes and revealing manageable, reasonable parts. Without his certainty that I, too, could partake in this avenue of research, I likely would not have started the journey into the realm of statistics. In another vein of making the impossible possible, thank you to Ms. Pam Poe, Ms. Charlene Weaver, Ms. Pam Bailey, and Ms. Helen Spencer for expertly navigating the hidden channels of the university and opening doors for me I didn’t even know existed. My growth in this program was directly related to their help.

Many thanks to Dr. Pierre Legendre, Dr. Daniel Borcard, Dr. Stéphane Dray, and Dr. François Gillet for redirection of my efforts for spatial analysis and to Dr. Bruce MacLennan for instruction on support vector machines. Dr. Xianfeng Ma and Dr. Pankaj Kumar, of the High Performance and Scientific Computing group, supported my work on Infrastructure for Scientific Applications and Advanced Computing (ISAAC) Open Enclave and Open OnDemand RStudio Server at the University of Tennessee, Knoxville.

This work was supported by the University of Tennessee, Knoxville, as well as the Kneberg-Lewis Scholarship from the Department of Anthropology at the University of Tennessee.

ABSTRACT

Within humans, dermatoglyphic features of friction ridge skin found on fingers, palms, soles, and toes, exhibit a wide range of variation across populations and data types that is maintained over time but ostensibly serves no direct functional purpose. Unlike any other aspect of human biology, dermatoglyphic data can both discriminate between individuals and remain unchanged throughout life due to fixation during intrauterine development. However, no agreement exists on the nature and significance of the observed variation in dermatoglyphic features partly because traditional statistical analyses cannot accommodate the structure of dermatoglyphic data.

The goals of this dissertation were to identify the patterns of human dermatoglyphic variation in worldwide populations by comparing finger ridge-counts, directional asymmetry, fluctuating asymmetry, and summary response variables at different geographic and biological scales. The worldwide dermatoglyphics data set was introduced in entirety for the first time and contained more than 51,000 individuals from over 300 ethnic groups, and modeling strategies from data mining were applied like description, clustering, and prediction.

Univariate finger ridge-counts performed better than summary total ridge-count and absolute ridge-count response variables, likely due to modeling constraints. Summary indices of asymmetry and diversity performed better than univariate Directional Asymmetry variables for each digit or Fluctuating Asymmetry variables for each digit, likely due to the structure of the data variables. From the supervised methods, the support vector regression models performed better at predicting the variation in dermatoglyphic features compared to distribution-dependent generalized linear models. However, these modeling frameworks often did not perform better

than baseline models estimating the central tendency or performing linear regression. From the unsupervised methods, the k-means clustering assignments consistently captured the most variation in relevant dermatoglyphic features, but the remaining explanatory variables differentially captured variation depending on the specific dermatoglyphic feature. Principal component analysis loadings revealed developmentally linked drivers of variation in finger ridge-counts like finger pattern and size, digit 1 variation, radial-ulnar gradients, and 2D:4D contrasts. This dissertation on the nature and significance of worldwide human dermatoglyphic variation provides a framework for resurgence in anthropological dermatoglyphics to discover the evolutionary forces acting on and maintaining variation in discriminatory and persistent dermatoglyphic features.

TABLE OF CONTENTS

CHAPTER ONE INTRODUCTION, BACKGROUND, AND RESEARCH QUESTIONS	1
1.1 Introduction.....	1
1.1.1 Dermatoglyphic features	5
1.1.2 Development of discriminatory and persistent dermatoglyphic morphology.....	10
1.2 Previous human dermatoglyphics research.....	15
1.2.1 How to analyze dermatoglyphic features of finger patterns and finger ridge-counts..	16
1.2.2 Descriptive approaches of analyzing dermatoglyphic features.....	20
1.2.3 Comparison of researcher-defined groups to identify biomarkers or patterns in variation	24
1.2.4 Comparison of dermatoglyphic features to other biological data types.....	32
1.2.5 How to analyze dermatoglyphic features of asymmetry and diversity	36
1.3 Data mining.....	45
1.4 Research questions and dissertation organization.....	55
CHAPTER TWO METHODOLOGY	59
2.1 k-means clustering	59
2.2 Spatial analysis via distance-based Moran’s eigenvector maps (dbMEM)	64
2.3 Principal component analysis (PCA)	68
2.4 Baseline models	69
2.5 Generalized linear modeling (GLM)	72
2.6 Support vector regression (SVR).....	75
2.7 Model selection and evaluation of model performance.....	79
2.7.1 k-means clustering internal performance using the Dunn Index	80
2.7.2 Generalized linear modeling with LASSO regularization	80
2.7.3 Support vector grid search model evaluation via hyperparameter optimization	81
2.7.4 Regression performance metrics.....	83
CHAPTER THREE MATERIALS AND METHODS.....	87
3.1 Data set.....	87
3.1.1 Data set, original	90
3.1.2 Data set, final	92
3.1.2.1 Response variables, dermatoglyphic features	98
3.1.2.2 Explanatory variables.....	112
3.2 Ethics, privacy, and data management.....	116
3.3 Experimental procedures	119
3.3.1 Computing environment	119
3.3.2 k-means clustering	120
3.3.3 Spatial analysis via distance-based Moran’s eigenvector maps (dbMEM)	125
3.3.4 Principal component analysis (PCA)	131
3.3.5 Baseline models	132
3.3.6 Generalized linear modeling (GLM)	137
3.3.7 Support vector regression (SVR).....	142
CHAPTER FOUR RESULTS AND DISCUSSION	146
4.1 Results for finger ridge-counts and summaries of ridge-counts	146

4.1.1 k-means clustering	146
4.1.2 Spatial analysis via distance-based Moran's eigenvector maps (dbMEM)	149
4.1.3 Principal component analysis (PCA)	161
4.1.4 Baseline models for generalized linear modeling	186
4.1.5 Generalized linear modeling (GLM)	208
4.1.6 Baseline models for support vector regression	218
4.1.7 Support vector regression (SVR)	227
4.2 Results for directional asymmetry, fluctuating asymmetry, and indices of asymmetry and diversity	228
4.2.1 k-means clustering	229
4.2.2 Spatial analysis via distance-based Moran's eigenvector maps (dbMEM)	240
4.2.3 Principal component analysis (PCA)	249
4.2.4 Baseline models for generalized linear modeling	298
4.2.5 Generalized linear modeling (GLM)	315
4.2.6 Baseline models for support vector regression	336
4.2.7 Support vector regression (SVR)	349
4.3 Discussion for finger ridge-counts and summaries of ridge-counts	351
4.3.1 Response variables	352
4.3.2 Explanatory variables	355
4.3.3 Methods	362
4.4 Discussion for directional asymmetry, fluctuating asymmetry, and indices of asymmetry and diversity	364
4.4.1 Response variables	364
4.4.2 Explanatory variables	369
4.4.3 Methods	374
CHAPTER FIVE CONCLUSIONS AND RECOMMENDATIONS	376
5.1 Conclusions of research	376
5.2 Limitations of study	380
5.3 Recommendation for future research	381
REFERENCES	384
APPENDICES	402
Appendix A Supplementary information and table, Chapter One	402
Appendix B Supplementary information and tables, Chapter Three	409
Appendix C Supplementary figures, Chapter Four	446
Appendix C.1 Supplementary figures for finger ridge-counts and summaries of ridge-counts	446
Appendix C.1.1 Supplementary figures for Spatial analysis via distance-based Moran's eigenvector maps (dbMEM)	446
Appendix C.1.2 Supplementary figures for Baseline models for generalized linear modeling	463
Appendix C.1.3 Supplementary figures for Baseline models for support vector regression	510
Appendix C.1.4 Supplementary figures for Support vector regression (SVR)	515

Appendix C.2 Supplementary figures for directional asymmetry, fluctuating asymmetry, and indices of asymmetry and diversity.....	518
Appendix C.2.1 Supplementary figures for Spatial analysis via distance-based Moran's eigenvector maps (dbMEM)	518
Appendix C.2.2 Supplementary figures for Baseline models for generalized linear modeling	526
Appendix C.2.3 Supplementary figures for Baseline models for support vector regression	552
Appendix C.2.4 Supplementary figures for Support vector regression (SVR).....	579
VITA	598

LIST OF TABLES

Table 1. Strategy of dissertation divided into response variable type, explanatory variable type, and method.....	60
Table 2. New classes and descriptions for attributes from worldwide dermatoglyphics data set.	94
Table 3. Response variables and their variable type.....	99
Table 4. Performance metric evaluation of TRC, ARC, and finger ridge-counts	102
Table 5. Performance metric evaluation of Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry and diversity.....	107
Table 6. Explanatory variables and their variable type.....	113
Table 7. Different sample sizes in different spatial analyses.....	126
Table 8. Dunn Index values for optimal K values for finger ridge-counts	148
Table 9. Truncation distances for finger patterns and finger ridge-counts within Landmass.....	153
Table 10. Number of dbMEMs and the adjusted R^2 and the number of significant dbMEMs and the adjusted R^2 for finger patterns and finger ridge-counts within Landmass.....	155
Table 11. Variance explained by the significant canonical axes for finger patterns and finger ridge-counts.....	155
Table 12. Violation of standard regression assumptions of residuals for baseline linear models for GLM of TRC, ARC, and finger ridge-counts	191
Table 13. Goodness-of-fit for full and reduced baseline linear models of TRC, ARC, and finger ridge-counts, of same GLM data set	196
Table 14. Performance metrics of final GLM formulation as a result of LASSO regularization and lambda optimization as compared to baselines for TRC, ARC, and finger ridge-counts	199
Table 15. Goodness-of-fit for full and reduced GLM of TRC, ARC, and finger ridge-counts..	211
Table 16. Violation of standard regression assumptions of residuals for baseline linear models for SVR of TRC and ARC.....	220
Table 17. Goodness-of-fit for full and reduced baseline linear models of TRC, ARC, and finger ridge-counts, of same SVR data set	224
Table 18. Sample sizes and performance metrics of final optimized SVR models for TRC and ARC	226
Table 19. Dunn Index values for optimal K values for Directional Asymmetry	232
Table 20. Dunn Index values for optimal K values for Fluctuating Asymmetry.....	232
Table 21. Truncation distances for Directional Asymmetry and Fluctuating Asymmetry within Landmass	241
Table 22. Number of dbMEMs and the adjusted R^2 and the number of significant dbMEMs and the adjusted R^2 for Directional Asymmetry and Fluctuating Asymmetry within Landmass	243
Table 23. Variance explained by the significant canonical axes for Directional Asymmetry and Fluctuating Asymmetry	245
Table 24. Violation of standard regression assumptions of residuals for baseline linear models for GLM of Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry of diversity.....	303

Table 25. Goodness-of-fit for full and reduced baseline linear models of Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry and diversity, of same GLM data set	308
Table 26. Performance metrics of final GLM formulation as a result of LASSO regularization and lambda optimization as compared to baselines for Directional Asymmetry, Fluctuating Asymmetry, and indices of asymmetry and diversity	310
Table 27. Goodness-of-fit for full and reduced GLM of Directional Asymmetry variables for each digit	318
Table 28. Goodness-of-fit for full and reduced GLM of Fluctuating Asymmetry variables for each digit	318
Table 29. Goodness-of-fit for full and reduced GLM of Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt	319
Table 30. Violation of standard regression assumptions of residuals for baseline linear models for SVR of Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry of diversity	338
Table 31. Goodness-of-fit for full and reduced baseline linear models of Directional Asymmetry and Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity, of same SVR data set	343
Table 32. Sample sizes and performance metrics of final optimized SVR models for Directional Asymmetry and Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity	346
Table S1. Abbreviations used throughout this paper.	406
Table S2. Abbreviation, Sample Number, Ethnic Group, and Region present in overall data set	410
Table S3. Software and algorithms required to execute code for dermatoglyphic analyses	422
Table S4. Distribution of finger ridge-counts in overall data set compared to 10% subset	424
Table S5. Distribution of Directional Asymmetry variables separated by digit in overall data set compared to 10% subset	427
Table S6. Distribution of Fluctuating Asymmetry variables separated by digit in overall data set compared to 10% subset	428
Table S7. Different sample sizes in different PCA	430
Table S8. Summary of response variables TRC, ARC, and finger ridge-counts before any response variable-specific modification for GLM analysis and prediction	432
Table S9. Summary of response variables Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry of diversity before any response variable-specific modification for GLM analysis and prediction	436
Table S10. Summary of response variables TRC and ARC before any response variable-specific modification for SVR analysis and prediction	439
Table S11. Summary of response variables Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry of diversity before any response variable-specific modification for SVR analysis and prediction	440
Table S12. Change in SVR model performance through grid searching for TRC and ARC	517

Table S13. Change in SVR model performance through grid searching for Directional Asymmetry and Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity	593
--	-----

LIST OF FIGURES

Figure 1. Factors affecting congruence of dermatoglyphic distances with other biological distances.....	4
Figure 2. Finger pattern types including focal point locations (core and triradius) for finger ridge-count.....	7
Figure 3. Biochemistry lacking mesenchymal activity and morphology lacking extended downgrowth of dermal ridge formation in volar skin vs hair follicle formation in hairy skin	14
Figure 4. Data analysis process using data mining approach	47
Figure 5. Map of Ethnic Group locations, colored by Region.....	88
Figure 6. Process from handprint to data file.....	89
Figure 7. Boxplots of finger ridge-counts separated by digit and side	101
Figure 8. Boxplots of total ridge-count separated by Region	103
Figure 9. Boxplots of absolute ridge-count separated by Region	104
Figure 10. Boxplots of Directional Asymmetry variables separated by digit in Data	106
Figure 11. Boxplots of Fluctuating Asymmetry variables separated by digit in Data.....	109
Figure 12. Boxplots of Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt in Data.	110
Figure 13. Boxplot of distribution of finger ridge-counts in overall data set compared to 10% subset.....	121
Figure 14. Boxplot of distribution of Directional Asymmetry variables separated by digit in overall data set compared to 10% subset	122
Figure 15. Boxplot of distribution of Fluctuating Asymmetry variables separated by digit in overall data set compared to 10% subset	123
Figure 16. Finger ridge-counts optimal K for three approaches	147
Figure 17. Stacked bar chart for $K = 2$ from the average silhouette method for finger ridge-counts	150
Figure 18. Stacked bar chart for $K = 4$ for finger ridge-counts	151
Figure 19. Stacked bar chart for $K = 16$ from the gap statistic method for finger ridge-counts.	152
Figure 20. Variation partitioning of undetrended finger pattern and finger ridge-counts response variables into an environmental component, a linear trend, a broad scale, and a fine scale	160
Figure 21. PC 1-6 loadings from PCA of covariance matrix of finger ridge-counts.....	163
Figure 22. Principal component analysis of covariance matrix of finger ridge-counts	164
Figure 23. PC 1-6 loadings from PCA of covariance matrix of female finger ridge-counts.....	166
Figure 24. PC 1-6 loadings from PCA of covariance matrix of male finger ridge-counts	167
Figure 25. Principal component analysis of covariance matrix of female and male finger ridge-counts	168
Figure 26. PC 1-6 loadings from PCA of covariance matrix of Landmass Africa finger ridge-counts	170
Figure 27. Principal component analysis of covariance matrix of Landmass Africa finger ridge-counts	171

Figure 28. Principal component analysis of covariance matrix of female and male Landmass Africa finger ridge-counts.....	173
Figure 29. PC 1-6 loadings from PCA of covariance matrix of Landmass Americas finger ridge-counts	174
Figure 30. Principal component analysis of covariance matrix of Landmass Americas finger ridge-counts.....	176
Figure 31. Principal component analysis of covariance matrix of female and male Landmass Americas finger ridge-counts.....	177
Figure 32. PC 1-6 loadings from PCA of covariance matrix of Landmass Asia finger ridge-counts	179
Figure 33. Principal component analysis of covariance matrix of Landmass Asia finger ridge-counts	180
Figure 34. Principal component analysis of covariance matrix of female and male Landmass Asia finger ridge-counts	182
Figure 35. Principal component analysis of covariance matrix of female and male Landmass Asia finger ridge-counts with Regions Australia, Melanesia, and Oceania removed.....	183
Figure 36. PC 1-6 loadings from PCA of covariance matrix of Landmass Europe finger ridge-counts	185
Figure 37. Principal component analysis of covariance matrix of Landmass Europe finger ridge-counts	187
Figure 38. Principal component analysis of covariance matrix of female and male Landmass Europe finger ridge-counts	188
Figure 39. GLM coefficient value for each explanatory variable, colored by TRC, ARC, or finger ridge-counts response variable.....	214
Figure 40. GLM coefficient value for each explanatory variable, colored by TRC, ARC, or finger ridge-counts response variable, zoomed to range -0.1 to 0.1.....	216
Figure 41. GLM coefficient value for each explanatory variable, colored by TRC, ARC, or finger ridge-counts response variable, zoomed to range -0.002 to 0.002.....	217
Figure 42. Directional Asymmetry optimal K for three approaches.....	230
Figure 43. Fluctuating Asymmetry optimal K for three approaches	231
Figure 44. Stacked bar chart for $K = 2$ from the average silhouette method for Directional Asymmetry.....	234
Figure 45. Stacked bar chart for $K = 2$ from the average silhouette method for Fluctuating Asymmetry.....	235
Figure 46. Stacked bar chart for $K = 4$ for Directional Asymmetry	236
Figure 47. Stacked bar chart for $K = 4$ for Fluctuating Asymmetry	237
Figure 48. Stacked bar chart for $K = 10$ from the within-cluster sum of squares method for Directional Asymmetry.....	238
Figure 49. Stacked bar chart for $K = 8$ from the within-cluster sum of squares method for Fluctuating Asymmetry	239
Figure 50. Variation partitioning of undetrended Directional Asymmetry and Fluctuating Asymmetry response variables into and environmental component, a linear trend, a broad scale, and a fine scale.....	248
Figure 51. PC 1-5 loadings from PCA of covariance matrix of Directional Asymmetry	251

Figure 52. Principal component analysis of covariance matrix of Directional Asymmetry.....	253
Figure 53. PC 1-5 loadings from PCA of covariance matrix of Fluctuating Asymmetry	254
Figure 54. Principal component analysis of covariance matrix of Fluctuating Asymmetry	255
Figure 55. PC 1-5 loadings from PCA of covariance matrix of female Directional Asymmetry.....	257
Figure 56. PC 1-5 loadings from PCA of covariance matrix of male Directional Asymmetry..	258
Figure 57. Principal component analysis of covariance matrix of female and male Directional Asymmetry.....	259
Figure 58. PC 1-5 loadings from PCA of covariance matrix of female Fluctuating Asymmetry	261
Figure 59. PC 1-5 loadings from PCA of covariance matrix of male Fluctuating Asymmetry .	262
Figure 60. Principal component analysis of covariance matrix of female and male Fluctuating Asymmetry.....	263
Figure 61. PC 1-5 loadings from PCA of covariance matrix of Landmass Africa Directional Asymmetry.....	265
Figure 62. Principal component analysis of covariance matrix of Landmass Africa Directional Asymmetry.....	266
Figure 63. Principal component analysis of covariance matrix of female and male Landmass Africa Directional Asymmetry	268
Figure 64. PC 1-5 loadings from PCA of covariance matrix of Landmass Africa Fluctuating Asymmetry.....	269
Figure 65. Principal component analysis of covariance matrix of Landmass Africa Fluctuating Asymmetry.....	271
Figure 66. Principal component analysis of covariance matrix of female and male Landmass Africa Fluctuating Asymmetry	272
Figure 67. PC 1-5 loadings from PCA of covariance matrix of Landmass Americas Directional Asymmetry.....	274
Figure 68. Principal component analysis of covariance matrix of Landmass Americas Directional Asymmetry.....	275
Figure 69. Principal component analysis of covariance matrix of female and male Landmass Americas Directional Asymmetry	277
Figure 70. PC 1-5 loadings from PCA of covariance matrix of Landmass Americas Fluctuating Asymmetry.....	278
Figure 71. Principal component analysis of covariance matrix of Landmass Americas Fluctuating Asymmetry.....	280
Figure 72. Principal component analysis of covariance matrix of female and male Landmass Americas Fluctuating Asymmetry	281
Figure 73. PC 1-5 loadings from PCA of covariance matrix of Landmass Asia Directional Asymmetry.....	283
Figure 74. Principal component analysis of covariance matrix of Landmass Asia Directional Asymmetry.....	284
Figure 75. Principal component analysis of covariance matrix of female and male Landmass Asia Directional Asymmetry.....	286
Figure 76. PC 1-5 loadings from PCA of covariance matrix of Landmass Asia Fluctuating Asymmetry.....	287

Figure 77. Principal component analysis of covariance matrix of Landmass Asia Fluctuating Asymmetry.....	289
Figure 78. Principal component analysis of covariance matrix of female and male Landmass Asia Fluctuating Asymmetry	291
Figure 79. PC 1-5 loadings from PCA of covariance matrix of Landmass Europe Directional Asymmetry.....	292
Figure 80. Principal component analysis of covariance matrix of Landmass Europe Directional Asymmetry.....	294
Figure 81. Principal component analysis of covariance matrix of female and male Landmass Europe Directional Asymmetry	295
Figure 82. PC 1-5 loadings from PCA of covariance matrix of Landmass Europe Fluctuating Asymmetry.....	297
Figure 83. Principal component analysis of covariance matrix of Landmass Europe Fluctuating Asymmetry.....	299
Figure 84. Principal component analysis of covariance matrix of female and male Landmass Europe Fluctuating Asymmetry	300
Figure 85. GLM coefficient value for each explanatory variable, colored by Directional Asymmetry response variable.....	321
Figure 86. GLM coefficient value for each explanatory variable, colored by Directional Asymmetry response variable, zoomed to range -2 to 2.....	323
Figure 87. GLM coefficient value for each explanatory variable, colored by Directional Asymmetry response variable, zoomed to range -0.1 to 0.1.....	324
Figure 88. GLM coefficient value for each explanatory variable, colored by Fluctuating Asymmetry response variable.....	327
Figure 89. GLM coefficient value for each explanatory variable, colored by Fluctuating Asymmetry response variable, zoomed to range -0.1 to 0.1.....	328
Figure 90. GLM coefficient value for each explanatory variable, colored by Fluctuating Asymmetry response variable, zoomed to range -0.002 to 0.002.....	329
Figure 91. GLM coefficient value for each explanatory variable, colored by Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt response variable	332
Figure 92. GLM coefficient value for each explanatory variable, colored by Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt response variable, zoomed to range -0.4 to 0.4333	
Figure 93. GLM coefficient value for each explanatory variable, colored by Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt response variable, zoomed to range -0.04 to 0.04	335
Figure S1. dbMEM analysis of detrended Landmass Africa finger patterns with the two significant dbMEM variables.....	447
Figure S2. dbMEM analysis of detrended Landmass Africa finger ridge-counts with the six significant dbMEM variables.....	448
Figure S3. dbMEM analysis of detrended Landmass Americas finger patterns with the five significant dbMEM variables.....	449

Figure S4. dbMEM analysis of detrended Landmass Americas finger ridge-counts with the five significant dbMEM variables.....	450
Figure S5. dbMEM analysis of detrended Landmass Asia finger patterns with the three significant dbMEM variables.....	451
Figure S6. dbMEM analysis of detrended Landmass Asia finger ridge-counts with the two significant dbMEM variables.....	452
Figure S7. dbMEM analysis of detrended Landmass Europe finger patterns with the one significant dbMEM variable	453
Figure S8. dbMEM analysis of detrended Landmass Europe finger ridge-counts with the one significant dbMEM variable	454
Figure S9. Map of two significant dbMEM variables for Landmass Africa finger patterns	455
Figure S10. Map of five significant dbMEM variables for Landmass Americas finger patterns.....	456
Figure S11. Map of three significant dbMEM variables for Landmass Asia finger patterns	457
Figure S12. Map of one significant dbMEM variable for Landmass Europe finger patterns	458
Figure S13. Map of six significant dbMEM variables for Landmass Africa finger ridge-counts	459
Figure S14. Map of five significant dbMEM variables for Landmass Americas finger ridge-counts	460
Figure S15. Map of two significant dbMEM variables for Landmass Asia finger ridge-counts	461
Figure S16. Map of one significant dbMEM variable for Landmass Europe finger ridge-counts	462
Figure S17. Linear model assumptions check for baseline linear model of ARC response variable, from GLM data set.....	464
Figure S18. Histogram of residuals for baseline linear model of ARC response variable, from GLM data set.....	465
Figure S19. Linear model assumptions check for baseline linear model of Lrc1 response variable, from GLM data set.....	466
Figure S20. Histogram of residuals for baseline linear model of Lrc1 response variable, from GLM data set.....	467
Figure S21. Linear model assumptions check for baseline linear model of Luc1 response variable, from GLM data set.....	468
Figure S22. Histogram of residuals for baseline linear model of Luc1 response variable, from GLM data set.....	469
Figure S23. Linear model assumptions check for baseline linear model of Lrc2 response variable, from GLM data set.....	470
Figure S24. Histogram of residuals for baseline linear model of Lrc2 response variable, from GLM data set.....	471
Figure S25. Linear model assumptions check for baseline linear model of Luc2 response variable, from GLM data set.....	472
Figure S26. Histogram of residuals for baseline linear model of Luc2 response variable, from GLM data set.....	473
Figure S27. Linear model assumptions check for baseline linear model of Lrc3 response variable, from GLM data set.....	474

Figure S28. Histogram of residuals for baseline linear model of Lrc3 response variable, from GLM data set.....	475
Figure S29. Linear model assumptions check for baseline linear model of Luc3 response variable, from GLM data set.....	476
Figure S30. Histogram of residuals for baseline linear model of Luc3 response variable, from GLM data set.....	477
Figure S31. Linear model assumptions check for baseline linear model of Lrc4 response variable, from GLM data set.....	478
Figure S32. Histogram of residuals for baseline linear model of Lrc4 response variable, from GLM data set.....	479
Figure S33. Linear model assumptions check for baseline linear model of Luc4 response variable, from GLM data set.....	480
Figure S34. Histogram of residuals for baseline linear model of Luc4 response variable, from GLM data set.....	481
Figure S35. Linear model assumptions check for baseline linear model of Lrc5 response variable, from GLM data set.....	482
Figure S36. Histogram of residuals for baseline linear model of Lrc5 response variable, from GLM data set.....	483
Figure S37. Linear model assumptions check for baseline linear model of Luc5 response variable, from GLM data set.....	484
Figure S38. Histogram of residuals for baseline linear model of Luc5 response variable, from GLM data set.....	485
Figure S39. Linear model assumptions check for baseline linear model of Rrc1 response variable, from GLM data set.....	486
Figure S40. Histogram of residuals for baseline linear model of Rrc1 response variable, from GLM data set.....	487
Figure S41. Linear model assumptions check for baseline linear model of Ruc1 response variable, from GLM data set.....	488
Figure S42. Histogram of residuals for baseline linear model of Ruc1 response variable, from GLM data set.....	489
Figure S43. Linear model assumptions check for baseline linear model of Rrc2 response variable, from GLM data set.....	490
Figure S44. Histogram of residuals for baseline linear model of Rrc2 response variable, from GLM data set.....	491
Figure S45. Linear model assumptions check for baseline linear model of Ruc2 response variable, from GLM data set.....	492
Figure S46. Histogram of residuals for baseline linear model of Ruc2 response variable, from GLM data set.....	493
Figure S47. Linear model assumptions check for baseline linear model of Rrc3 response variable, from GLM data set.....	494
Figure S48. Histogram of residuals for baseline linear model of Rrc3 response variable, from GLM data set.....	495
Figure S49. Linear model assumptions check for baseline linear model of Ruc3 response variable, from GLM data set.....	496

Figure S50. Histogram of residuals for baseline linear model of Ruc3 response variable, from GLM data set.....	497
Figure S51. Linear model assumptions check for baseline linear model of Rrc4 response variable, from GLM data set.....	498
Figure S52. Histogram of residuals for baseline linear model of Rrc4 response variable, from GLM data set.....	499
Figure S53. Linear model assumptions check for baseline linear model of Ruc4 response variable, from GLM data set.....	500
Figure S54. Histogram of residuals for baseline linear model of Ruc4 response variable, from GLM data set.....	501
Figure S55. Linear model assumptions check for baseline linear model of Rrc5 response variable, from GLM data set.....	502
Figure S56. Histogram of residuals for baseline linear model of Rrc5 response variable, from GLM data set.....	503
Figure S57. Linear model assumptions check for baseline linear model of Ruc5 response variable, from GLM data set.....	504
Figure S58. Histogram of residuals for baseline linear model of Ruc5 response variable, from GLM data set.....	505
Figure S59. Linear model assumptions check for baseline linear model of Third response variable, from GLM data set.....	506
Figure S60. Histogram of residuals for baseline linear model of Third response variable, from GLM data set.....	507
Figure S61. Linear model assumptions check for baseline linear model of TRC response variable, from GLM data set.....	508
Figure S62. Histogram of residuals for baseline linear model of TRC response variable, from GLM data set.....	509
Figure S63. Linear model assumptions check for baseline linear model of ARC response variable, from SVR data set.....	511
Figure S64. Histogram of residuals for baseline linear model of ARC response variable, from SVR data set.....	512
Figure S65. Linear model assumptions check for baseline linear model of TRC response variable, from SVR data set.....	513
Figure S66. Histogram of residuals for baseline linear model of TRC response variable, from SVR data set.....	514
Figure S67. SVR TRC grid searches gamma vs cost.....	515
Figure S68. SVR ARC grid searches gamma vs cost.....	516
Figure S69. Map of the single significant dbMEM variable for Landmass Africa Directional Asymmetry.....	519
Figure S70. Map of the single significant dbMEM variable for Landmass Africa Fluctuating Asymmetry.....	520
Figure S71. Map of the single significant dbMEM variable for Landmass Americas Directional Asymmetry.....	521
Figure S72. Map of the single significant dbMEM variable for Landmass Asia Directional Asymmetry.....	522

Figure S73. Map of the single significant dbMEM variable for Landmass Asia Fluctuating Asymmetry.....	523
Figure S74. Map of the single significant dbMEM variable for Landmass Europe Directional Asymmetry.....	524
Figure S75. Map of the single significant dbMEM variable for Landmass Europe Fluctuating Asymmetry.....	525
Figure S76. Linear model assumptions check for baseline linear model of DA1 response variable, from GLM data set.....	526
Figure S77. Histogram of residuals for baseline linear model of DA1 response variable, from GLM data set.....	527
Figure S78. Linear model assumptions check for baseline linear model of DA2 response variable, from GLM data set.....	528
Figure S79. Histogram of residuals for baseline linear model of DA2 response variable, from GLM data set.....	529
Figure S80. Linear model assumptions check for baseline linear model of DA3 response variable, from GLM data set.....	530
Figure S81. Histogram of residuals for baseline linear model of DA3 response variable, from GLM data set.....	531
Figure S82. Linear model assumptions check for baseline linear model of DA4 response variable, from GLM data set.....	532
Figure S83. Histogram of residuals for baseline linear model of DA4 response variable, from GLM data set.....	533
Figure S84. Linear model assumptions check for baseline linear model of DA5 response variable, from GLM data set.....	534
Figure S85. Histogram of residuals for baseline linear model of DA5 response variable, from GLM data set.....	535
Figure S86. Linear model assumptions check for baseline linear model of FA1 response variable, from GLM data set.....	536
Figure S87. Histogram of residuals for baseline linear model of FA1 response variable, from GLM data set.....	537
Figure S88. Linear model assumptions check for baseline linear model of FA2 response variable, from GLM data set.....	538
Figure S89. Histogram of residuals for baseline linear model of FA2 response variable, from GLM data set.....	539
Figure S90. Linear model assumptions check for baseline linear model of FA3 response variable, from GLM data set.....	540
Figure S91. Histogram of residuals for baseline linear model of FA3 response variable, from GLM data set.....	541
Figure S92. Linear model assumptions check for baseline linear model of FA4 response variable, from GLM data set.....	542
Figure S93. Histogram of residuals for baseline linear model of FA4 response variable, from GLM data set.....	543
Figure S94. Linear model assumptions check for baseline linear model of FA5 response variable, from GLM data set.....	544

Figure S95. Histogram of residuals for baseline linear model of FA5 response variable, from GLM data set.....	545
Figure S96. Linear model assumptions check for baseline linear model of Index.Asymmetry response variable, from GLM data set.....	546
Figure S97. Histogram of residuals for baseline linear model of Index.Asymmetry response variable, from GLM data set.....	547
Figure S98. Linear model assumptions check for baseline linear model of Index.Diversity response variable, from GLM data set.....	548
Figure S99. Histogram of residuals for baseline linear model of Index.Diversity response variable, from GLM data set.....	549
Figure S100. Linear model assumptions check for baseline linear model of Index.Diversity.Holt response variable, from GLM data set.....	550
Figure S101. Histogram of residuals for baseline linear model of Index.Diversity.Holt response variable, from GLM data set.....	551
Figure S102. Linear model assumptions check for baseline linear model of DA1 response variable, from SVR data set.....	553
Figure S103. Histogram of residuals for baseline linear model of DA1 response variable, from SVR data set.....	554
Figure S104. Linear model assumptions check for baseline linear model of DA2 response variable, from SVR data set.....	555
Figure S105. Histogram of residuals for baseline linear model of DA2 response variable, from SVR data set.....	556
Figure S106. Linear model assumptions check for baseline linear model of DA3 response variable, from SVR data set.....	557
Figure S107. Histogram of residuals for baseline linear model of DA3 response variable, from SVR data set.....	558
Figure S108. Linear model assumptions check for baseline linear model of DA4 response variable, from SVR data set.....	559
Figure S109. Histogram of residuals for baseline linear model of DA4 response variable, from SVR data set.....	560
Figure S110. Linear model assumptions check for baseline linear model of DA5 response variable, from SVR data set.....	561
Figure S111. Histogram of residuals for baseline linear model of DA5 response variable, from SVR data set.....	562
Figure S112. Linear model assumptions check for baseline linear model of FA1 response variable, from SVR data set.....	563
Figure S113. Histogram of residuals for baseline linear model of FA1 response variable, from SVR data set.....	564
Figure S114. Linear model assumptions check for baseline linear model of FA2 response variable, from SVR data set.....	565
Figure S115. Histogram of residuals for baseline linear model of FA2 response variable, from SVR data set.....	566
Figure S116. Linear model assumptions check for baseline linear model of FA3 response variable, from SVR data set.....	567

Figure S117. Histogram of residuals for baseline linear model of FA3 response variable, from SVR data set.....	568
Figure S118. Linear model assumptions check for baseline linear model of FA4 response variable, from SVR data set.....	569
Figure S119. Histogram of residuals for baseline linear model of FA4 response variable, from SVR data set.....	570
Figure S120. Linear model assumptions check for baseline linear model of FA5 response variable, from SVR data set.....	571
Figure S121. Histogram of residuals for baseline linear model of FA5 response variable, from SVR data set.....	572
Figure S122. Linear model assumptions check for baseline linear model of Index.Asymmetry response variable, from SVR data set.....	573
Figure S123. Histogram of residuals for baseline linear model of Index.Asymmetry response variable, from SVR data set.....	574
Figure S124. Linear model assumptions check for baseline linear model of Index.Diversity response variable, from SVR data set.....	575
Figure S125. Histogram of residuals for baseline linear model of Index.Diversity response variable, from SVR data set.....	576
Figure S126. Linear model assumptions check for baseline linear model of Index.Diversity.Holt response variable, from SVR data set.....	577
Figure S127. Histogram of residuals for baseline linear model of Index.Diversity.Holt response variable, from SVR data set.....	578
Figure S128. SVR DA1 grid searches gamma vs cost.....	580
Figure S129. SVR FA1 grid searches gamma vs cost.....	581
Figure S130. SVR DA2 grid searches gamma vs cost.....	582
Figure S131. SVR FA2 grid searches gamma vs cost.....	583
Figure S132. SVR DA3 grid searches gamma vs cost.....	584
Figure S133. SVR FA3 grid searches gamma vs cost.....	585
Figure S134. SVR DA4 grid searches gamma vs cost.....	586
Figure S135. SVR FA4 grid searches gamma vs cost.....	587
Figure S136. SVR DA5 grid searches gamma vs cost.....	588
Figure S137. SVR FA5 grid searches gamma vs cost.....	589
Figure S138. SVR Index.Asymmetry grid searches gamma vs cost.....	590
Figure S139. SVR Index.Diversity grid searches gamma vs cost.....	591
Figure S140. SVR Index.Diversity.Holt grid searches gamma vs cost.....	592

CHAPTER ONE

INTRODUCTION, BACKGROUND, AND RESEARCH QUESTIONS

1.1 Introduction

Dermatoglyphic features have been utilized for the purposes of individual identification, clinical diagnoses, and anthropological inquiry. While the strategy of individual identification focuses on matching one image to another and clinical diagnoses using dermatoglyphic data have largely been replaced with genomics, anthropological dermatoglyphics seeks to reveal the common driving forces acting on and maintaining human variation (Meier 1980).

Dermatoglyphics, literally “skin carvings”, are found on fingers, palms, soles, and toes, and friction ridge skin maintains evolutionary function for prehensile hands and feet, mediated by the friction ridges and sweat pores, and skin flexion mediated by creases (Cummins and Midlo 1961; Holt 1968; Maceo 2011). Loesch and Martin (1984) suggested variation in finger patterns may be related to evolution of tactile sensitivity. However, within humans the wide range of variation observed across populations and data types ostensibly serves no direct functional purpose.

Unlike any other aspect of human biology, dermatoglyphic data are both discriminatory and persistent. The basis of discrimination lies in the development of friction ridges *in utero* such that identical twins do not share identical friction ridge definition. The basis of persistence begins with fixation of the friction ridge morphology *in utero* and continues with physiological maintenance of the exact definition throughout life such that only scarring or amputation can alter or remove the friction ridge definition (Wertheim 2011). As yet, no agreement exists on the nature and significance of the observed variation in dermatoglyphic features, and the

disagreement stems from a lack of consistency in selected features and methods of analysis (Jantz 1987; Crawford and Duggirala 1992).

Historical modeling approaches of dermatoglyphic data suffered from a scarcity of data, no evaluation framework, and assumption driven statistical modeling, because the properties of dermatoglyphic features do not lend well to traditional statistical analyses, since finger and palmar patterns are only qualitative, the quantitative features like finger and palmar ridge-counts are not normally distributed, dermatoglyphic features are highly correlated and hierarchical, and populations exhibit significant, unequal within-group variation (Leguebe 1982; Abraham et al. 2013). Further complicating analysis, dermatoglyphic features are spatially autocorrelated on a variety of levels such that homologous digits are correlated, adjacent digits are correlated, left and right hands are correlated, and individuals within a population within a geographic space are correlated. As a result, the field of anthropological dermatoglyphics has been abandoned in favor of genomics.

Great leaps in understanding come from comparing multiple biological data types at the same time, and research in human biological variation has incorporated the power of artificial intelligence and modern computing power. Artificial intelligence offers a means for solving problems without explicit programming. Modern computing innovations and their corresponding analysis frameworks enable a renewed attempt to document worldwide patterns of variation and to simultaneously model complex constraints. Data mining draws from statistics and artificial intelligence in the form of machine learning providing a dynamic approach flexible and powerful enough to address dermatoglyphic data that are large-scale, high-dimensional, heterogeneous, complex, and non-normally distributed, where traditional statistical methods are unsuitable.

Where researchers used to exclusively focus on a single biological data source like either dermatoglyphic data or genetic data, now systems biology approaches that focus on the complex interactions occurring within a biological system acknowledge individuals, groups, species are greater than the sum of their parts (Veenstra 2021).

The goals of this dissertation were to identify the patterns of human dermatoglyphic variation in worldwide populations by specifically addressing the existing hurdles in interpretation and analysis of dermatoglyphic features as synthesized directly from Meier (1980:165) (Figure 1). To address Factor I of the number and kinds of dermatoglyphic traits to use in this dissertation, different dermatoglyphic features were considered as response variables to contrast data types like univariate vs multivariate and summary vs raw variables. To address Factor III of the optimal sample size and composition as well as Factor V of the optimal level of population study and comparison in this dissertation, different explanatory variables were considered with different statistical modeling approaches like spatial analysis to and different forms of regression. To address Factor IV of the best methods of testing congruence in this dissertation, a data mining strategy was employed to step into model complexity with a range of approaches like description, clustering, and prediction. Dermatoglyphic features serve as the target data source for these analyses due to their abundance of differing data types within a single individual and due to their absence of susceptibility to the postnatal environment.

Dermatoglyphic data are inexpensive to collect and store, researchers can relatively easily record counts, pattern types, etc., and data collection is relatively less invasive to collect compared to genetic data or anthropometric data. Of course, dermatoglyphics are not available for skeletal individuals, but living populations can be surveyed with minimal ethical and financial concern

I.	Number and kinds of dermatoglyphic traits used – palmar versus digital, qualitative versus quantitative.
II.	Number and kinds of other biological/genetic measures used – blood groups, serum proteins, anthropometrics, etc., also nonbiological distances such as linguistic, migrational, etc.
III.	Sample size and composition – number and degree of related persons, sex ratios (in cases of lumping or analyzed separately), degree of sample overlap between data sets.
IV.	Methods of testing congruence – univariate versus multivariate, visual inspection, nonparametric measures, correlation analysis of matrix similarity, dendrogram comparisons, goodness of fit.
V.	Level of population study and comparison – local village, urban areas, tribal and racial units; all of which deal with the degree of evolutionary divergence based upon magnitude of biologic distance

Figure 1. Factors affecting congruence of dermatoglyphic distances with other biological distances. Directly from Meier (1980:165) as original Table 3

and over time can depict a suitable time-depth for additional, longitudinal analyses as well as document the living populations that may not persist due to extinction, assimilation, or oversight. Indeed many groups included in the worldwide dermatoglyphics data set have faced similar displacement, reduction/elimination of population size, and renaming altogether. Based on the results of this dissertation, the final Factor II of the number and kinds of other biological and genetic methods can be explored to compare dermatoglyphic data with other biological data sets and advance toward systems biology approaches to human biological variation.

The remainder of the chapter was organized as follows. The dermatoglyphic features were introduced as was the developmental processes required to produce dermatoglyphic morphology. Section *1.2 Previous human dermatoglyphics research* provides a brief review of the literature on the variation in human dermatoglyphic features, specifically the current perspectives for how to statistically analyze dermatoglyphic data and how to interpret the results. Section *1.3 Data mining* introduces the framework of data mining, which was selected as the strategy to expand the efforts of previous research and increase the utility of dermatoglyphic features. Finally, in Section *1.4 Research questions and dissertation organization* the research questions guiding this dissertation were specified as well as how the questions fit into the current research in human biological variation.

1.1.1 Dermatoglyphic features

Dermatoglyphic features are found on fingers, palms, soles, and toes (Cummins and Midlo 1961; Holt 1968). Few studies include the soles of the feet and the toes, and likewise here the focus of this research was the friction ridges on the hands, specifically the fingers rather than the palms.

On each hand are five fingers: digit 1 (thumb), digit 2, digit 3, digit 4, and digit 5 (little finger), and located on the palmar surface of the distal-most section of each digit are qualitative and quantitative data types within the pattern area. Type lines begin parallel to each other within the pattern area, diverge from each other, and ultimately define the pattern area, which delimits the friction ridges important for pattern classification and ridge-count analysis. Focal points, like the triradius and core, reside within the pattern area of some patterns (loops and whorls). The core represents the approximate center of the pattern area, and the triradius (also known as the delta) represents the points of divergence of the type lines on a ridge or dot (Federal Bureau of Investigation 1990). Qualitative data types include the pattern type (arch, loop, or whorl in Figure 2). A single individual may exhibit the same pattern type on all ten digits or may exhibit different pattern types on different digits, and the left and right hands may not mirror each other. Quantitative data types include the finger ridge-count and its variants and pattern summary values.

Finger ridge-counts are calculated by summing the number of ridges crossing or touching an imaginary line drawn between the triradius and the core, and the triradius and core are not included in the count (Holt 1968; Figure 2). Arches lack any triradius or the triradius is located next to the core, and thus arches have a finger ridge-count of zero. Loops can open toward the thumb side (radial loop) or toward the little-finger side (ulnar loop), and so loops can have a radial or ulnar ridge-count for each finger. For this reason, an individual's ridge-counts for each digit are divided into a radial count and an ulnar count, even if the count is zero. For example, a digit with a radial loop will have a radial finger ridge-count of some value and an ulnar finger ridge-count of zero for that digit. Whorls contain a triradius on both the radial and ulnar sides, so

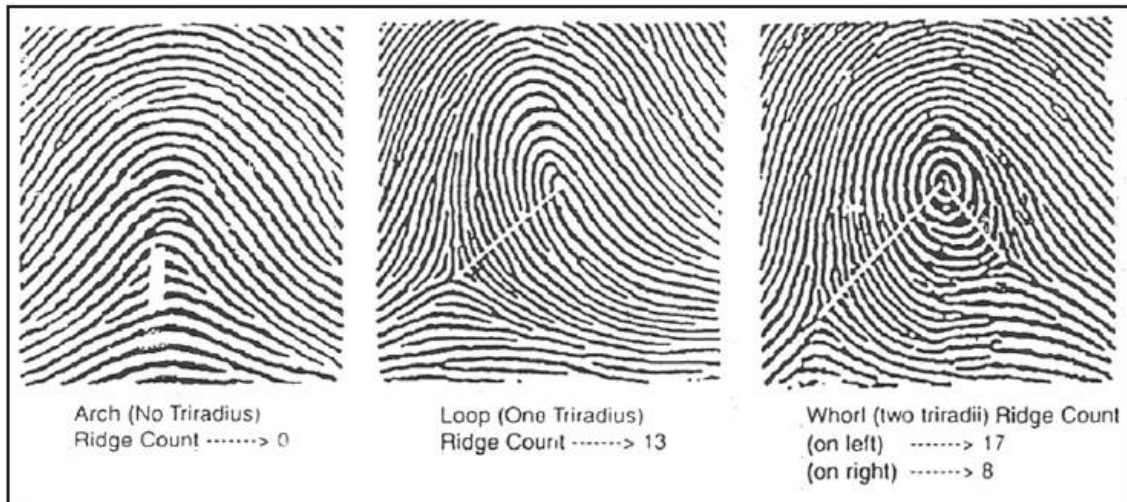


Figure 2. Finger pattern types including focal point locations (core and triradius) for finger ridge-count. Image directly from Sengupta et al. (2013), which was modified from Holt (1968).

the ridge-count of a digit will have a value for both the radial and ulnar sides. Each individual is represented by 20 finger ridge-counts for the radial and ulnar sides of each digit on both hands. The total ridge-count (TRC) represents the sum of the largest ridge-count on each finger, and usually the radial count is larger. The absolute ridge-count (ARC) represents the sum of all 20 finger ridge-counts (Meier 1980). Where the TRC reflects the pattern size, the ARC reflects the pattern intensity comprised of both pattern size and pattern type. For example, using TRC a small pattern of any type will yield a low ridge-count, and a large pattern of any type will yield a high ridge-count. But using ARC, a small whorl pattern may yield a similar ridge-count as a large loop (Holt 1968:42).

Indices of asymmetry account for the spatial autocorrelation of the homologous digits and of the left and right hands (Dipierri et al. 2018). The index of asymmetry ($\sqrt{A^2}$) accesses finger ridge-count asymmetry, where A is defined by the sum of the squared differences between finger ridge-counts on the left and right hand, for each digit. The index of diversity ($S/\sqrt{5}$) accesses the differences between non-homologous digits, where S is defined by the square root of the difference between the squared finger ridge-count minus the squared total ridge-count divided by 5 digits, summed over each digit (Jantz 1975).

The highly discriminatory and persistent nature of dermatoglyphic features stems from the variation in the developmental processes contributing to dermatoglyphic morphology. Dermatoglyphic features are specific to the individual such that even monozygotic twins do not exhibit identical friction ridge definition. Monozygotic twins share identical genes and a similar physical environment *in utero*, but at most the similarity in dermatoglyphic features is limited to a correlated TRC (Holt 1968; Wertheim 2011). Note that finger patterns and finger ridge-counts

do not encode sufficient detail to be specific to an individual. Dizygotic twins share a less similar physical environment *in utero* than monozygotic twins, and dermatoglyphic features for dizygotic twins resemble standard sibling pairs, who experience completely separate environments *in utero* (Holt 1968; Meier 1980). The results of family inheritance research suggests no two individuals exhibit identical dermatoglyphic features, and although dermatoglyphic features are partially heritable, the maternal environment *in utero* may direct some of the observed variation in dermatoglyphic features.

Dermatoglyphic features are persistent such that after fixation of the friction ridge morphology *in utero*, the dermis-epidermis physiology maintains the exact definition of the dermatoglyphic features throughout life. Only the size of the dermatoglyphic features on fingers, palms, soles, and toes increases with increasing age to a point, but the organization remains the same. Scarring or amputations are exceptional processes that can alter or remove the friction ridge definition (Wertheim 2011). Barring any destruction of the dermal or epidermal tissue throughout life, the dermatoglyphic morphology remains the same for an individual from birth to death, prior to decomposition, so the post-natal environment does not influence dermatoglyphic features.

Although heritability was not calculated here, understanding the maintenance of variation in dermatoglyphic features relies on understanding heritability of traits. Heritability can serve as an indicator for how quickly a trait can respond to evolutionary forces relative to the trait's phenotypic variance. So, given an evolutionary force of a particular strength and two traits with the same phenotypic variance, the trait with the higher heritability will evolve more quickly (Roseman and Weaver 2007). While Holt (1968) actually conducted research on inheritance of

dermatoglyphic features via twin studies, Cummins and Midlo (1961) provided a single source for the biology of human dermatoglyphics. Specifically in regard to heritability, Cummins and Midlo (1961) discussed the difficulties in interpreting the genetic inheritance of dermatoglyphic features, and Holt (1968) explained the role of dermatoglyphics in human genetics. Because dermatoglyphic features persist unchanged after formation *in utero*, any effect of genetics and any environmental perturbations will be active during embryological development. The developmental processes contributing to dermatoglyphic morphology are explained in the following section.

1.1.2 Development of discriminatory and persistent dermatoglyphic morphology

Dermatoglyphic features are considered polygenic, so multiple genes are involved in producing what we understand to be the dermatoglyphic features, and the patterning is set by about the 20th week of intra-uterine development. Dermal ridges are thought to arise from the volar pads, which are temporary expansions of mesenchyme tissue, on the palmar surfaces of the hands and the plantar surfaces of the feet. Sometime between the 5th and 10th intra-uterine week, the volar pads form and regress, and the regression of the volar pad is followed by the formation of primary, then secondary ridges that constitute the dermal ridges. An experimental study by Glover et al. (2023) utilized mouse models and human organoids and cultures to model volar skin to visualize and identify the developmental basis for finger pattern formation, and their results comparing the deeper mesenchymal layer (via mesenchymal fibroblasts) to the superficial epithelial layer (via basal epithelial keratinocytes) in ventral volar skin, minimally hairy dorsal skin, and hairy back skin were as follows.

Around 10 weeks *in utero*, the epithelium of the volar pads exhibited increased expression of lymphoid enhancer-binding factor 1 (LEF1) and enhanced ectodysplasin-A receptor (EDAR). At the sites of EDAR expression on the volar pad epithelium, follicle growth factor 20 (FGF20) and transforming growth factor alpha gene (TGFA) are also found. In the mesenchyme of the volar pads, R-spondin 2 (RSPO2) expression enhances Wnt signaling and Dickkopf Wnt signaling pathway inhibitor 2 (DKK2) expression inhibits Wnt signaling in the adjacent epithelium. Around 12 weeks *in utero*, the ventral digit epithelium exhibited a uniform distribution of EDAR until both the volar skin and the hairy skin exhibited an increased in EDAR observed in a punctuated pattern as well as an increase in expression of genes involving Wnt signaling pathways like catenin beta-1 (CTNNB1), LEF1, and low density lipoprotein receptor-related protein 4 (LRP4) around *in utero* weeks 13 and 14. In Wnt signaling specifically, CTNNB1 is a protein-coding gene for β -catenin, LEF1 is a key transducer and target of Wnt signaling, and LRP4 is an antagonist of Wnt signaling.

Around the same time of 12 weeks *in utero* in the deeper mesenchyme, hair-forming skin exhibited an increase in levels of LEF1, while the ventral volar skin mesenchyme exhibited consistently low levels of LEF1. In hairy skin, the Wnt signaling pathway stimulated mesenchymal cells and FGF20 to form a dermal condensate. However, in volar skin low levels of LEF1 contributed to low levels of Wnt activity, thus producing an unresponsive mesenchyme that produced no dermal condensate despite FGF20 activity. FGF20 normally promotes mesenchymal cell adhesion needed to form a dermal condensate, and FGF20 depends on Wnt, so if Wnt/ β -catenin activity is low no dermal condensate forms. By week 13 *in utero*, a continuous intermediate cell layer containing protein, KERATIN14, appears on the ventral digit epithelium

followed by ridge formation. Of note, ridge formation never precedes the formation of this layer containing KERATIN14, and while other areas of human fetal skin contain KERATIN14, only the ventral digit epithelium exhibits a continuous layer.

Over weeks 13 through 16 *in utero*, the primary ridges form across the rest of the volar skin continuing onto the dorsal digit, and around 15 weeks *in utero*, the primary ridges that define the overall finger pattern configuration are defined. Primary ridge formation is influenced by genes also active in early limb development, and these genes exert indirect effects on the finger pattern configurations through direct effects on digit length and proportions as well as the size and shape of the raised volar pads (Li et al. 2022). By week 16 *in utero* the sweat glands emerge from the deepest areas of the ridges, and by week 17 *in utero* primary ridge formation is complete and similar, smaller secondary ridges that flank the primary ridges begin to appear. Primary ridges only follow similar development to hair follicles in early epithelial biochemistry and morphology with strong expression of β -catenin and substantial, initial downgrowth, but the primary ridges lack the drivers of extended downgrowth and dermal condensate as in hair follicles. Secondary ridges do not exhibit strong expression of β -catenin or substantial downgrowth as in primary ridges.

Downgrowth results in a deepening of ridges, where the deepest projection hosts the greatest cellular proliferation, highest Wnt/ β -catenin activity, and upheld EDAR and LEF1 expression. Once projection of the primary ridges completes, proliferation continues to induce sweat gland downgrowth and otherwise thicken the suprabasal layer that creates the corrugations at the surface. The deepest projection of the ridge with high proliferation also exhibits active epidermal growth factor receptor (EGFR) signaling, likely as a result of the production of its

ligand, transforming growth factor- α (TGF α). TGF α triggers mitosis in the epithelial cells and is encoded by gene TGFA that is abundant in emerging ridges and is particularly strongly expressed in sweat glands as they bud off and grow down from the primary ridges (Figure 3).

Based on knockout experiments, inhibition of active Wnt signaling results in no digit ridge formation, suggesting the Wnt pathway is required for successful ridge formation but not involved in the ridge morphology patterning. Removal of EDAR signaling results in spotted ridge patterning, and heightened EDAR signaling results in thicker, widely spaced ridge formation, and the spotted-striped range is characteristic of Turing patterning systems, which suggests EDAR is required for successful ridge formation, ridge breadth, and ridge morphology patterning. Turing's reaction-diffusion model revealed how spatial patterns form from random or uniform initial interactions between two diffusible substances, and Turing patterns are common in nature like a stripe patterning on animal skins and as a mechanism governing morphogenesis (Turing 1952; Kondo 2017). Inhibiting bone morphogenic protein (BMP) activity in mice triggered expression of sonic hedgehog protein (*Shh*) in the ventral epithelium, where *Shh* is notably not normally found and is usually found in hairy skin, suggesting BMP is required to identify digit ridge areas by blocking the key of hairy skin development. BMP inhibition also results in broader ridge formation and increased Wnt signaling activity. The Turing reaction-diffusion system containing EDAR, Wnt, and BMP thus accounts for dermal ridge formation and morphological patterning. Further models of Turing reaction-diffusion systems incorporating the mechanical effect of waves of Wnt activity from various initial sites in relation to distal limb geometry reliably generate many types of finger patterns.

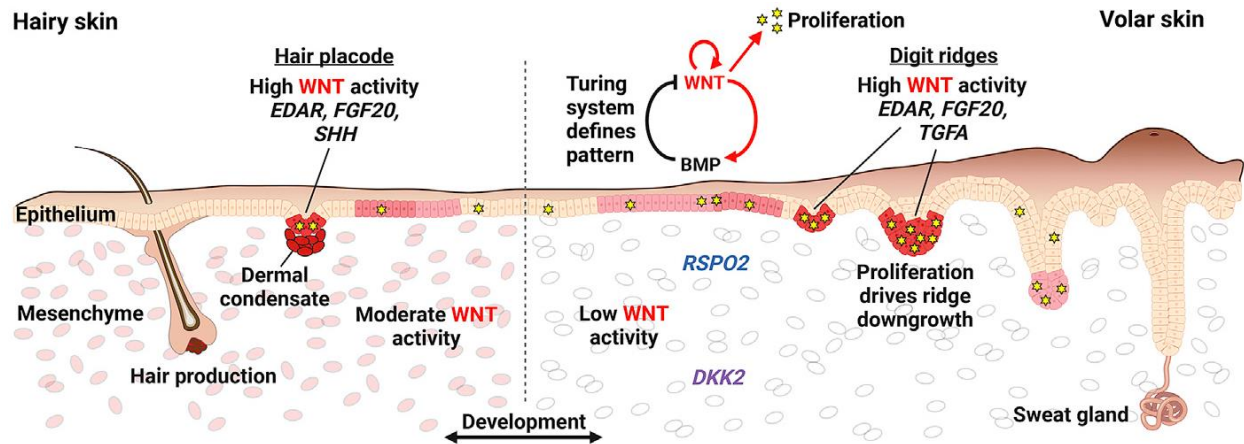


Figure 3. Biochemistry lacking mesenchymal activity and morphology lacking extended downgrowth of dermal ridge formation in volar skin vs hair follicle formation in hairy skin.

Image taken directly from Glover et al. (2023) as original Figure 6P.

Generally, volar pad symmetry theoretically relates to subsequent pattern formation, and volar pad size relates to subsequent ridge-count development (Wertheim 2011:3-21). Fournier and Ross (2016) suggested early volar pad regression and low volar pads result in an arch pattern on fingers; a late volar pad regression and high volar pads result in a whorl pattern on fingers; and an intermediate volar pad regression and height result in loops, where the highest volar pad directs the direction of the loop to the highest side. While Li et al. (2022) support the contribution of volar pad symmetry and regression to subsequent finger pattern, Glover et al. (2023) supplement these theories with experimental visualization and identification of biochemical and mechanical contributions to ridge formation and patterning. Early research in dermatoglyphics was not able to explore the developmental processes contributing to dermatoglyphic morphology in experimental animal models, and the evolution of previous trends in human dermatoglyphics research is explained in the following section.

1.2 Previous human dermatoglyphics research

Although this dissertation considers each dermatoglyphic feature of interest separately, rarely is a single dermatoglyphic feature the sole focus of dermatoglyphics research. For example, Chattopadhyay and Ganeson (1984) showed finger ridge-counts and palmar *a-b* ridge-counts – both quantitative dermatoglyphic features – did not appear correlated, but left and right finger ridge-counts were correlated. So, rather than organize the review of previous dermatoglyphics research in this section by dermatoglyphic feature, which quickly becomes redundant, this section instead focuses on the problems encountered in research that inspired the research questions of this dissertation. So, while trends in selection of explanatory variable and preference for methods of analysis are explored, the same themes pervade the dermatoglyphic

feature types like early descriptive analysis, comparison of researcher-defined groups for biomarker identification, and comparison of dermatoglyphic features to other biological data types as they became available, so are addressed.

The earliest attempts to translate the patterns of variation in dermatoglyphic features to some biologically meaningful signal rely first on knowing the influence of heredity and the environment. Holt (1955) laid the foundation for decades of subsequent researchers, because Holt addressed the genetics of dermatoglyphics by documenting the frequency distributions of TRC in families, incomplete families, siblings, twins, and unrelated persons in a British population. From Holt (1955) and related works, the organization of dermal ridges was found to be highly heritable, and females and males exhibited differences in TRC. These powerful conclusions came from simplifying areas not under scrutiny by including a single source population and a single dermatoglyphic feature, for example, and by meticulously documenting family trees and relationships. Also as a result of Holt (1955) and related works, TRC was promoted as the unit of analysis to use, which was challenged by subsequent research incorporating raw finger ridge-counts and multiple populations in which TRC appeared to distort and even mask patterns of variation.

1.2.1 How to analyze dermatoglyphic features of finger patterns and finger ridge-counts

The properties of dermatoglyphic features do not lend well to traditional statistical analyses, because finger and palmar patterns are only qualitative, the quantitative features like finger and palmar ridge-counts are not normally distributed, dermatoglyphic features are highly correlated and hierarchical, and populations exhibit significant, unequal within-group variation. Dermatoglyphics are spatially autocorrelated on a variety of levels such that homologous digits

are correlated, adjacent digits are correlated, left and right hands are correlated, and individuals within a population within a geographic space are correlated. Many dermatoglyphic researchers borrow methodological strategies used for other biological data sets that may not share the same idiosyncrasies, and Leguebe, Vrydagh, and Ducros (1981) offered a guide to specifically analyzing dermatoglyphic data. They provide experimental design tips suggesting the researcher should consider whether the sample individuals differ by time period or social class.

In terms of communicating the data, first reduce the data into relative frequency tables for qualitative variables and into summary tables (mean, standard deviation, etc.) for quantitative variables. The summary tables indicate whether the dermatoglyphic feature is normally distributed and thus can be subject to traditional hypothesis testing and statistical inference. To evaluate the relationships between dermatoglyphic features, the correlation coefficient is often used for quantitative variables even though interpretation requires the correlation coefficient has a normal bivariate distribution and even though dermatoglyphic features rarely meet this requirement. A high correlation coefficient corresponds to a strong association between the two variables, and a low correlation coefficient could still entail a strong non-linear relationship (Leguebe, Vrydagh, and Ducros 1981). For the relationships between qualitative dermatoglyphic features, independence is tested via the chi-square test, but the features are ordinal variables the appropriate measures of association are the Kendall or Spearman coefficient, unfortunately not directly comparable to each other or to the Pearson correlation coefficients. For the comparison of qualitative to quantitative dermatoglyphic features, no measure of association is particularly appropriate but the different qualitative traits can be treated as groups and then subject to regression.

To compare groups using quantitative dermatoglyphic features, the univariate features can reveal significant substructure that may bias the results overall. With two groups, a two-tailed t-test can compare the means of independent groups where the null hypothesis is the two groups originate from the same overall population and have the same mean. If the dermatoglyphic features are not normally distributed a Mann-Whitney U test or a Kolmogorov-Smirnov can test for the independence of the groups. To compare left and right sides, a Wilcoxon matched pairs signed-rank test is appropriate. An analysis of variance is appropriate to compare groups across studies where the sample size, mean, and standard deviation are known. Additional methods to compare means are with *a priori* tests or *a posteriori* tests, where multiple comparisons of means can occur with a stepwise method or with a simultaneous test procedure.

To compare groups using qualitative dermatoglyphic features, indices like the mainline index or indices of asymmetry and diversity may obscure variation. For example, populations may have significantly different frequencies of loops and whorls but the same index value. Conversely, populations with similar frequencies of loops and whorls may exhibit significantly different index values only because the relative frequencies of arches are driving the different index value. Tests of homogeneity like the chi-square test or the Shannon measure of information are appropriate for qualitative dermatoglyphic features. The Shannon measure of information represents the probabilities corresponding to categories like finger pattern. When the dermatoglyphic feature has only one category the Shannon measure of information is zero and will increase to its maximum when every category exhibits equal frequencies, so the maximum value depends on and increases with the number of categories.

Multivariate methods offer different results that are difficult to compare. Leguebe, Vrydagh, and Ducros (1981) reminded researchers to avoid introducing the same variable into a model more than once, such that the raw finger ridge-count and the summary TRC are in the same model or the raw finger ridge-count and the summary pattern intensity index (Loesch and Lafranchi 1990). They argued the multivariate methods should not be applied to the dermatoglyphic features but to information extracted from the features like a correlation matrix, dispersion matrix, or distance matrix. From here, different strategies are structure analysis, clustering, ordination, and discriminant analysis. Structure analysis is applied as multiple regression to study the asymmetry between left and right finger ridge-counts. Cluster analysis imposes a structure within complex data sets, and ordination analysis reduces the dimensionality of dermatoglyphic features. Types of ordination analysis are PCA, factor analysis, principal coordinates analysis, nonmetric multidimensional scaling, and proximity analysis, and these methods can illustrate individual-level relationships existing within population samples. Leguebe, Vrydagh, and Ducros (1981) admitted the results of PCA for dermatoglyphic features can be difficult to interpret. Discriminant analysis requires groups be defined *a priori*, assumes differences exist between them, and proceeds to calculate functions that best separate the groups. The factors that constitute the functions are not equally important to separate the groups, as illustrated in plots of discriminant functions where two groups may be spatially close to one another as defined by the first two factors but may be distant from one another when defined by the third factor. Depending on the analysis of dermatoglyphic features, the differences or similarities between groups may be highlighted, and both may be important for understanding the drivers of patterns of variation within and between populations.

1.2.2 Descriptive approaches of analyzing dermatoglyphic features

Due to the intractability of dermatoglyphic analysis, early research shied away from statistical testing and instead focused on describing select groups like small, regionally-specific populations and case-control groupings that were predefined by the researcher. Rather than an evolutionary focus with researchers suggesting the samples were representative of true biological populations, dermatoglyphics research focused on a quick, visual group identifier, and the simplest method was surveying finger patterns. In a sample of 800 (400 female, 400 male) non-related Nigerian university students, the frequencies of finger patterns were considered in the sample as a whole with 52% of the sample exhibiting ulnar loops, 27.6% with whorls, and 19.2% with arches, leaving just 1.13% with radial loops. Finger patterns were also considered within sexes with females exhibiting more arches, and bilateral finger pattern symmetry was high, above 80%, in females and males (Boroffice 1978). In a more recent sample of Sinhalese people in Sri Lanka, the frequencies of finger patterns on both hands were 60% loops, with the ulnar loop most common (58% females, 56% males), and 36% whorls (Wijerathne et al. 2013).

To directly compare sex and population differences, Jantz (1977) calculated the correlation matrices for populations selected for analysis due to their shared, or not, population history: American Whites, American Blacks, European, African, and Indian. Each correlation coefficient was then z-transformed, and the mean for each matrix was analyzed. Although American White females exceeded the mean correlation of American White and European males, in the Indian group and some African groups the males exhibited significantly higher correlations, as indicated by pairwise t-tests. Overall, males produced the most consistent patterns such that within a population the correlations were homogenous but between populations the correlations were heterogeneous. Because the patterns of correlation differed so

much by sex, Jantz (1977) supposed the sex chromosome may participate in dictating the organization of dermal ridges, and since males exhibited the most variation the Y chromosome was likely responsible through its normal task of slowing the mitotic cycle during cell division. Although not the focus of the research on the effects of parental age and birth order on dermatoglyphic asymmetry considering finger ridge-counts, palmar patterns, palmar ridge-counts, and palmar *a-b* ridge-breadth, Oliveira (1978) confirmed sex was an important factor in the variation in asymmetry for finger ridge-counts and palmar ridge-counts. Oliveira (1978) showed females exhibited increased asymmetry in finger features with increasing parental age. Jantz (2021) related finger ridge-counts to the second to the fourth digit (2D:4D) ratio, which biologically may indicate sex hormone exposure and sensitivity from intrauterine life (Kalichman, Batsevich, and Kobylansky 2019). Further, Polcerová et al. (2023) supported the importance of sex in driving the radial versus ulnar contrasts in finger ridge-counts in twenty-one, evenly-distributed human populations and encouraged future research on prenatal sex differentiation during intrauterine development.

Dennis and Sunderland (1979) compared frequencies of finger ridge-counts and palmar ridge counts in a single population subdivided into smaller and smaller groups to understand the patterns of variation in North England. The finger ridge-counts and palmar *a-b* ridge-count exhibited sex differences although not statistically significant and overall regional variation was limited suggesting scale can dictate the results and should be carefully documented for reproducibility of results.

Another strategy with descriptive techniques was to visualize the variation in dermatoglyphic features using principal component analysis or similar dimensionality reduction

techniques. Jantz (1974) used principal coordinate analysis to deal with the multivariate nature of finger patterns and finger ridge-counts, where each individual is at most represented by 10 finger pattern values and 20 finger ridge-count values. Because finger pattern numeric codes do not conform as metric variables, the finger pattern frequency on the left and right hands was pooled. Only five representative populations were selected for analysis: American Whites, American Blacks, English, Dogon, and Easter Island to represent contemporary groups, parent groups, and an outlier group. For the finger patterns, PC1 appeared to represent the frequency of whorls and loops, and PC2 appeared to separate Europeans and Africans. For the finger ridge-counts, PC1 represented TRC, and PC2 separated digits 2 and 3 from the other digits. Clearly, the finger patterns and finger ridge-counts yielded different results in terms of constituting the majority of variation. From this work, Jantz (1974) created an isosceles triangle from the principal coordinate analysis with Africans, Europeans, and Oceanian/East Asian at each vertex tip of the triangle to represent the differences in variation in dermatoglyphic features. Jantz (1974) suggested dermatoglyphics research should incorporate data from multiple dermatoglyphic features, since the features do not illustrate the same patterns in variation, and most researchers do indeed include multiple dermatoglyphic features to address a single research question. By increasing the number and breadth of populations from five to six, now including sub-Saharan Africa, Europe, West Asia, East Asia, the Pacific, and America, and by focusing only on TRC, Jantz (1977b) observed an increase in TRC from lowest in Africa, America, and Europe, to higher in West Asia, and then highest in East Asia and the Pacific.

Considering the finger patterns, finger ridge-counts, palmar patterns, palmar ridge-counts, palmar lines and creases, and palmar atd angle in just American Whites and American

Blacks, American Whites notably exhibited a higher frequency of whorls and radial loops as well as a higher finger pattern intensity index and TRC. American Blacks exhibited a higher frequency of arches and ulnar loops as well as palmar creases (Plato, Cereghino, and Steinberg 1975). The individuals included were 7 year olds, which highlighted the flexibility of sampling individuals across age groups given dermatoglyphic features become fixed *in utero*.

To explore whether multivariate techniques could surpass PCA capturing the variation in dermatoglyphic features to separate groups, Nagy and Pap (2004) considered 22 finger and 24 palmar features for comparing 1608 children in 11 Hungarian populations to 390 children in five Gypsy populations where both groups overlapped geographically but culturally remained different. The analyses differed depending on whether the variable was qualitative or quantitative and whether the analysis could accommodate univariate or multivariate data. Univariate analyses used a chi-square test for qualitative variables and a Mann-Whitney U test for quantitative variables. Multivariate analyses used homogeneity tests for qualitative variables and PCA and discriminant function analysis (DFA) for quantitative variables. Nagy and Pap (2004) did acknowledge the assumptions of a DFA were not met, so the DFA was only included to visualize the separation of groups. For the univariate analyses, Hungarian and Gypsy groups successfully separated for nine finger traits and 14 palmar traits. For the multivariate analyses, the Hungarian and Gypsy groups separated well for the homogeneity analyses for qualitative variables and the DFA for quantitative variables, but they did not separate well for the PCA for quantitative variables. Overall, Nagy and Pap (2004) suggested the palmar traits separated the two groups better than the finger traits because evolutionary changes occur more slowly in palmar traits compared to finger traits or perhaps because the organization of palmar traits depends less on

environmental factors during development than finger traits. These two suppositions can be tested by comparing dermatoglyphic traits on the palms and the fingers to other biological data types for which the evolutionary rate is known and to other biological data types for which the environmental contribution is either known or limited. The strategy of incorporating both univariate and multivariate analyses for both qualitative and quantitative variables was included in this dissertation by subjecting these different dermatoglyphic features to their appropriate analyses and comparing their performance using common metrics to all analyses and variable types.

1.2.3 Comparison of researcher-defined groups to identify biomarkers or patterns in variation

One way to inadvertently decipher the biological drivers of variation in human dermatoglyphic features was to associate specific groups with some biological difference to trends in dermatoglyphic variation and ultimately use the dermatoglyphic features as biomarkers for the case grouping, like a chromosomal abnormality. For example, Chakravathy et al. (2018) surveyed finger patterns and the palmar atd angle among hypertensive and normotensive individuals, and found individuals with hypertension had more radial loops and whorls than normotensive individuals, who displayed more ulnar loops. In the process of this kind of research, patterns of variation in dermatoglyphic data in populations are documented, but also sometimes the case grouping, like a chromosomal abnormality, may influence the expression of dermatoglyphic features and variation, thereby revealing potential developmental drivers of variation in human dermatoglyphic features.

Turner syndrome, monosomy of the X chromosome, was hypothesized to represent developmental stress that could also affect the organization of dermal ridges (Miller 1985).

Individuals with sex chromosome abnormalities like Turner syndrome reportedly featured a higher frequency of large patterns, resulting in higher TRC between 160 and 200 (100 – 150 ridge-counts otherwise) (Miller 1973). Although Turner syndrome affects tissue arising from the mesoderm and dermal ridges arise from the ectoderm, the dermal ridges do interface with the mesoderm. Considering finger ridge-counts, ARC, and palmar ridge-counts and subjecting the correlation matrices to PCA, individuals with Turner syndrome did exhibit increased total variance, increased directional asymmetry, but no changes in growth gradients as compared to individuals without Turner syndrome. The purpose of this research was not to create a dermatoglyphic biomarker for Turner syndrome but to use what was known about Turner syndrome to decipher the developmental trajectories of dermatoglyphic features (Miller 1985). Miller (1985) observed how individuals with Turner Syndrome exhibited significant differences in directional asymmetry, total variance, but not changes in the growth gradients.

A case-control approach not directed at clinical diagnosis itself but in linking dermatoglyphic features to a clinical malformation with similar developmental trajectories, Owsley (1978) concurrently studied individuals with cleft lip and cleft palate as well as their dermatoglyphic features like finger ridge-counts, palmar patterns, palmar ridge-counts, palmar ridge-breadth, and palmar maximal atd angle. Cleft lip and cleft palate malformations occur as a result of genetic abnormalities or environmental disruptions during prenatal development, and those same abnormalities and disruptions may also affect the organization of dermal ridges. By comparing the dermatoglyphic features of individuals with cleft lip and cleft palate to relatives and unrelated individuals through factor analysis, the unrelated individuals with no cleft lip or cleft palate did not differ from individuals with cleft lip or cleft palate just by looking at means.

Further, the means of relatives of individuals with cleft lip and cleft palate were similar to relatives of individuals with no cleft lip or palate. Owsley (1978) agreed descriptive studies were helpful to avoid the statistical problems of analyzing dermatoglyphic features, which are multiple variables that are highly correlated, but disagreed with descriptive studies that search for population or group differences and ignore the developmental implications of those differences. Owsley (1978) did not observe any difference in means for dermatoglyphic features of individuals with cleft lip/cleft palate compared to controls or in means of relatives of cleft-lip/palate who do not have clefts versus relatives of cleft lip/palate who do have clefts. Observing developmental abnormalities in real humans, compared to simulated models is a strength of this previous body of research.

Not every dermatoglyphic feature varies in the same way such that finger ridge-counts and palmar *a-b* ridge-counts did not appear correlated, but left and right finger ridge-counts were correlated (Chattopadhyay and Ganeson 1984). To add to the confusion of selecting an appropriate dermatoglyphic feature, Jantz and Chopra (1983) tried analyzing many dermatoglyphic features (finger ridge-counts, TRC, finger pattern intensity, palmar mainline endings, palmar patterns, palmar atd angle, palmar mainline index, frequency of missing *c* triradii, numbers of combinations of palmar loops, and palmar *a-b* ridge-count) with different methodological approaches for only males from four endogamous groups in India to ultimately detect intergroup variation, with the result of no clear superior dermatoglyphic feature emerged. TRC performed least well, which suggests retention of both the radial and ulnar counts is important for distinguishing populations, and palmar features overall exhibited greater intergroup differences. None of the patterns of intergroup variation exhibited by the many dermatoglyphic

features agreed with patterns of variation seen with anthropometric or serological data types. This difficulty in comparison with other data types may suggest selection of dermatoglyphic feature alone may not define success but also the appropriate statistical model must be applied, and this idea was explicitly explored in this dissertation by subjecting both summary and raw dermatoglyphic features to specific analyses that can accommodate the mathematical quirks of dermatoglyphic features. Jantz and Chopra (1983) did show the dermatoglyphic features with the lowest heritability produce the greatest intergroup differences, suggesting much of dermatoglyphic variation arises from non-genetic processes and thus may be difficult to consistently interpret.

Binkley (1985) leveraged the ability of dermatoglyphic features to highlight intergroup differences and population structure and evaluated the introduction of agriculture into Europe via the demic expansion hypothesis versus the cultural diffusionist hypothesis. The cultural diffusionist hypothesis suggests agriculture spread through ideas and not genes. The demic diffusion hypothesis suggests Neolithic farmers spread through migrations and gene flow into Europe from Southwest Asia, and dermatoglyphics research could directly address the intergroup differences between populations in Europe (47 female, 48 male populations) and whether the group differences mimic a directional trend. Binkley (1985) thoroughly highlighted the limitations in dermatoglyphics research like how studies of heritability were not totally consistent, like how comparison of dermatoglyphic features to other biological data types ranged from excellent concurrence to nil so the biological meaning of dermatoglyphic features remained unclear, like how methods for how to include finger ridge-counts as variables was inconsistent – as a single summary variable, as a vector of means for each of 10 digits, or as 20 raw variables –,

and like how contemporary populations exhibited unexpected finger ridge-counts compared to their parent population. The inconsistent heredity evidence showed TRC as more heritable than finger ridge-counts, and finger ridge-counts from each hand more heritable than individual digits for monozygotic and dizygotic twin pairs (Machado et al. 2010). Despite these admitted limitations and consciously ignoring the assumptions of normal distributions for multiple regression, which were all commonly accepted admissions in dermatoglyphics research, Binkley (1985) performed PCA, multiple regression, and spatial analysis. PCA addressed the patterns of variation between groups in Europe, and multiple regression addressed the biological distances between groups compared to the spatial distances between groups, interpreted with latitude and transformed longitude so an increase in each represented an increase in distance to the North and West, respectively. The spatial analysis used Moran's I criterion for whether two groups were considered connected or not, and the Moran's I values were regressed on different distance classes.

To evaluate whether females and males needed to be analyzed separately Binkley (1985) compared the variance/covariance matrices of all females to all males as well to the variance/covariance matrices of each population, and the chi-square values of the all-female and all-male groups were much higher than for each population, and Binkley (1985) suggested the all-female-and-all-male groups contained larger sample sizes thus allowing for greater representation of variation. Despite this, Binkley (1985) analyzed the females and males of a population separately as though they were separate subpopulations. Because Binkley (1985) separated females and males throughout his analysis as others suggested despite compromising

the amount of variation contained within a grouping, this dissertation explored the importance of sex in directing patterns of variation in dermatoglyphic features.

The results of the multiple regression suggest the variation accounted for by geography was 19% for both males and females. Within Europe, Binkley (1985) found high correlations of Moran's I values up to 2100 km and even for some values up to 2800 km, which appears to identify the limiting scale for Europe. This dissertation sought to validate the results of Binkley (1985) by using an updated method of spatial analysis and by including more Landmasses in addition to Europe (Borcard and Legendre 2002; Brind'Amour, Mahévas, and Legendre 2018).

Just looking at the finger ridge-count means, groups in Southeast Europe exhibited higher variation than groups in Northwest Europe, which would appear to support a demic diffusion hypothesis. Although the results of Binkley's (1985) analyses tend to support a demic diffusion hypothesis, he admits other aspects of his results do not strongly support the hypothesis like how PC4 was significantly associated with longitude exhibiting an increase in the magnitude of contrast between digit 3 and digits 2 and 5 from east to west. Not specific to dermatoglyphics research, Novembre and Stephens (2008) argued against interpreting gradients in PCA biplots as evidence of migration, because the gradients may be the results of mathematical artifacts of genetic similarity that decreased with distance. An untestable assumption of Binkley's (1985) research was evidence of a migrating population would be an increase in finger ridge-count contrast, and by comparing these results to known patterns of variation in other biological data types like genetic data this assumption could be explored with additional modeling approaches (Sokal and Menozzi 1982; Binkley 1985; Sokal, Harding, and Oden 1989). An overlooked assumption of this work was that the variation captured by dermatoglyphic features in

contemporary populations resembled the variation during the Neolithic period. The evolutionary timing has yet to be sufficiently addressed in dermatoglyphics research. Just like with genetic analyses, future dermatoglyphics research would benefit from comparison of dermatoglyphic features in modern populations to those in historic populations to estimate evolutionary rates in large phylogenies (Tao et al. 2019).

This dissertation addressed the limitations thoughtfully tabulated by Binkley (1985) for dermatoglyphics research by including multiple dermatoglyphic features and explicitly evaluating the performance using consistent metrics. Further, models that do not assume a normal distribution were evaluated to specifically address whether the non-normality of dermatoglyphic features could be ignored. By addressing methodological limitations, the theoretical hurdles like lack of clarity on heritability and how it is population-specific and the lack of understanding in comparison to other data types can thus be addressed in subsequent research and were not explored in this dissertation. Similar to Binkley (1985), the geographic coordinates in this dissertation were estimated for the groups with the understanding a single latitude and longitude value cannot encompass the settlement distribution of an entire population, and the estimation of variables in Binkley (1985) inspired the creation of artificial groups in this dissertation like Landmass to facilitate spatial analysis and to explore the ability of these artificial grouping to capture patterns of variation in dermatoglyphic features (Reddy et al. 2004).

Another approach testing hypotheses of human migration focused on the origin of modern humans in Africa supported by genetic analyses of modern populations such that individuals in sub-Saharan Africa exhibited the highest within-population variation which decreased with increasing geographic distance from Africa. Daly, Ousley, and Jantz (2010)

utilized radial finger ridge-counts from only males in 81 populations, containing a minimum sample size of 50 individuals. The exclusion of females negated any question of sex-specific differences, the inclusion of worldwide populations increased the resolution of any gradient that could be observed with increasing distance from Africa, the requirement of a minimum sample size prevented small populations containing only a subset of variation from skewing the results since within-population variation was the focus. The radial ridge-count distance matrices were generated to compare with geographic distances through linear regression of an R matrix of intra-population variation and geographic distance. Daly et al. (2010) observed finger ridge-counts only from males in multiple populations around the world and showed that North and South American populations exhibited the highest variance and not sub-Saharan Africa as other researchers show with different data types. Daly et al. (2010) did use finger ridge-counts as solely neutral features. Dermatoglyphics were accepted as polygenic features that were likely influenced by a complex gene network linked to a functional ‘trait(s)’ (that is not related to dermatoglyphics) rather than the finger ridge-counts being directly influenced by evolutionary forces. Jantz (1987) reminds us we still do not have a concept of how evolutionary forces act on dermatoglyphic features that do show inter-population relationships, and it is more likely selection is acting on development itself rather than the features produced by development. This harkens to Pigliucci’s (2010) suggestion to include developmental encoding as an integral piece to understand genotype and phenotype variation. Other researchers using different data types have identified development as a key source of variation, explicitly hormone exposure and sensitivity during intra-uterine growth (Sengupta and Karmakar 2006; Dipierri et al. 2018; Kalichman, Batsevich, and Kobylansky 2019; Jantz 2021; Polcerová et al. 2023). Heritability is

also disparate across dermatoglyphic features and still requires testing with family pedigrees (Machado et al. 2010). Also, sex differences are common for finger ridge-counts, so including a model for females may also shift the results of Daly et al. (2010; Jantz 1977; Mbaka et al. 2016).

1.2.4 Comparison of dermatoglyphic features to other biological data types

Comparing populations to reveal relationships usually relies on tethering a hypothesis to a well-documented pattern, like across data types comparing dermatoglyphic features to tried and true blood groups or like across populations comparing contemporary groups to their likely pedigreed parental groups (Kobyliansky and Livshits 1986; Nagy and Pap 2004). For example, Harich et al. (2002) conducted pairwise population comparison using finger patterns, TRC, palmar *a-b* ridge-counts, and palmar mainline D terminations. The population of interest was Berbers (103 females, 120 males) from Morocco, contrasted with Arab-speakers, another dominant North African group but differing in ethno-linguistic history, as well as other populations around the Mediterranean, accessed through the literature. Females and males were analyzed separately within populations, with a chi-square test and visualization via PCA for qualitative data like finger patterns and pattern frequencies and t-test for quantitative data like TRC and pattern intensity index, which reflects the number of triradii and thus the complexity of a finger pattern. The Berber males exhibited more arches than individuals from North Africa or the Middle East, and Berber females exhibited more radial loops but fewer whorls. Principal components analysis illustrated Berber males proximal to North African groups, while Berber females related to Iberian groups. Multivariate analysis of quantitative features suggested the female and male Berbers were the most differentiated group considered, which differed from genetic analyses.

Another comparison of populations to tether dermatoglyphics research to well-understood biological data types and population histories was the investigation of the Lapps to discern their origin, which was debated as from Asia with some European contribution or as solely from Europe. Jantz, Eriksson, and Brehme (1993) converted finger ridge-counts and palmar ridge-counts into F_{ST} values and concluded Lapps more likely originated from Asia with some European contribution. The results agreed with previous work using genetic distances, but blood markers suggested a solely European origin.

Jantz, Brehme, and Eriksson (1992) used finger patterns, finger ridge-counts, a vector of 10 finger ridge-counts, a vector of 20 finger ridge-counts, palmar patterns and mainlines, and palmar ridge-counts to situate males of the Finno-Ugric speaking populations between populations to the West and the East using a canonical variates analysis with a Wilk's lambda. The relationships between populations depended on which dermatoglyphic feature was considered, so the populations were all most differentiated when considering palmar mainlines and least differentiated when considering finger ridge-counts. The results of analyses including finger patterns and finger ridge-counts were consistent with language distances. And summary dermatoglyphic features like TRC, ARC, pattern intensities, and the mainline index may distort the relationships between populations (Jantz, Brehme, and Eriksson 1992). The argument that summary features collapse variation and mask differences between populations is explicitly explored in this dissertation by subjecting summary features like TRC, ARC, indices of asymmetry and diversity, directional asymmetry, and fluctuating asymmetry to raw features like finger patterns and finger ridge-counts to the same analyses and comparing their performance.

Jantz and Hawkinson (1980) compare the finger ridge-counts separately for females and males of American Whites, American Blacks, and African Blacks using PCA as well as the Mahalanobis distance of the principal component scores as a proxy for some genetic component. PC1 represented size and was akin to the ARC, PC2 represented radial versus ulnar counts, and PC3 represented both the contrast between the ulnar count on digit 5 with the ulnar counts on digits 1 and 2 as well as the radial and ulnar counts of digit 1. PC4 represented the radial and ulnar counts of digit 1 versus all other digits, and PC5 represented the gradient of strongly negative loadings on digit 5 to strongly positive loadings on digit 1 for the radial counts, and ulnar counts were not represented in PC5. PC6 represented the contrast of the radial and ulnar counts on digits 1 and 2 as well as the digits themselves since the direction of the contrasts for the two digits is reversed (Jantz and Hawkinson 1980). Only PC6 and PC12 exhibited significant heterogeneity in both females and males. From this work, the importance of including both radial and ulnar finger ridge-counts in analysis became clear. Further, while PC6 exhibited little population variation, once the principal component scores were converted into Mahalanobis distances, PC6 then became the most significant contributor to interpopulation variation, accounting for more than 45% of the heterogeneity. The heterogeneity differed for females and males within the selected populations, so Jantz and Hawkinson (1980) support the strategy of analyzing females and males separately, and their results could form the basis for subsequent research into the effect of sex-specific settlement patterns or the effect of sex chromosomes on the patterns of dermatoglyphic variation. For example, Mbaka et al. (2016) found females exhibited lower F_{ST} differentiation based on finger ridge-counts, TRC, ARC, and finger pattern types, and could be considered more migratory than their male counterparts for three ethnic

groups in Liberia (209 females, 218 males). And Jantz and Parham (1978) found the variation in palmar *a-b* ridge breadth, defined by the sum of the distance in mm on the left hand and the distance in mm on the right hand divided by the sum of the two plus the ridge-counts on the left hand and the ridge-counts on the ridge hand, was associated with hand size, body size, and sex. The greatest separations between females and males arose from palmar *a-b* ridge-counts and palmar *a-b* ridge breadths, and for population grouping the Africans exhibited the lower palmar *a-b* ridge-counts due to wider ridges than the population groupings for the English and Jews, which exhibited little interpopulation variation.

When genetic analyses became accessible, especially in publishing patterns of global human variation, subsequent research sought to quell the confusion in dermatoglyphics research through comparison with genetic data and also followed by encompassing multiregional data sets (Rosenberg et al. 2002). For example, Zhang et al. (2010) were able to detect geographic patterning in Chinese ethnic groups when considering TRC, finger patterns, palmar *a-b* ridge-counts, and palmar hypothenar and thenar patterns as a result of including 122 population samples, 31 combined samples, as well as Africans, American Whites, and Gin Vietnamese. The Han Chinese, who actually represented 1.2 billion out of China's 1.3 billion total population, clustered with local groupings, suggesting the Han Chinese contain lots of diversity. The power of these dermatoglyphic data could distinguish between North and South clusters with the Yangtze River as a natural boundary, and this pattern is similarly found with anthropometric, human leukocyte antigen (HLA), and immunoglobulin data types. The Mang were not assigned to any ethnic group and did not cluster with any groups. The Minnan Han, the largest population in Taiwan, were more closely similar to mainland groups in the North cluster. Although all nine

Tibetan samples geographically reside in Southwestern China, the samples all clustered more similarly to the North cluster. Multiregional data sets can capture variation in broader scales but can also identify nonconforming subpopulations for whom further research may uncover untold population histories. While not every scale was exhaustively tested, this dissertation explored the issue of scale by including a worldwide dermatoglyphics data set not yet analyzed in its entirety as well as an explicit spatial analysis to identify spatially structured components of variation in dermatoglyphic features.

Previous spatial analyses of dermatoglyphic variation highlighted the difference in dermatoglyphic features to reflect patterns in variation, like how distances from palmar patterns did not faithfully reflect genetic structure and how finger patterns and finger ridge-counts were highly correlated but palmar features and finger features were minimally correlated (Sokal, Jantz, and Thomson 1996). This previous research focused on populations within Europe and unexpectedly observed no autocorrelation between short distances. For populations in Siberia, finger ridge-counts exhibited evidence of spatial connections along north-to-south and east-to-west clines, while craniometrics exhibited east-to-west connections, and serological data exhibited north-to-south connections (Yokota 1997), so dermatoglyphic data did not agree with other biological data types at this scale.

1.2.5 How to analyze dermatoglyphic features of asymmetry and diversity

One byproduct of consistent results indicating summary dermatoglyphic features mask variation was the increase in research focus on radial and ulnar differences and left and right digit and hand differences. But the research into asymmetry suffered from the same theoretical and analytical issues as summary and raw dermatoglyphic features like population-specific

heritability, different results compared to other biological data types, and no agreement upon modeling approach. Asymmetry does delve into the developmental underpinnings of dermatoglyphic variation as many suggested greater asymmetry reflected greater developmental stress or instability.

Jantz, Fohl, and Zahler (1979) compared the variation in finger ridge-counts to self-reported left or right handedness in a sample of American White university students (250 females, 226 males) and found left-hand dominant individuals exhibited lower ridge-counts and lower variances as well as lower within-population variation and high within-individual variation. Within-individual variation, measured as the correlation between homologous structures, has been considered an indicator of developmental stability. Left and right handed females overall exhibited higher correlations than males, suggestive that females were more canalized during development. Not examined here, palms have demonstrated a lower heritability than fingers, and palms have demonstrated a higher association with handedness. This research did not codify the etiology of handedness as genetic or environmental or how handedness can inform heritability of dermatoglyphic features, but even with a small sample from a single population this simple research design yielded results useful for subsequent research. Jantz, Fohl, and Zahler (1979) showed how finger ridge-counts exhibited high within-individual variation and comparatively low within-population variation, where females were the most canalized or groups with the highest robusticity.

In addition to observing the raw differences between left and right sides for finger ridge-counts as an example, more formal calculations of asymmetry and diversity arose with specific guidelines for analysis. Fluctuating asymmetry (FA) represents a measure of developmental

instability, developmental noise, and robustness on a population-level and only weakly infers individual-level developmental instability. FA refers to the variance of the difference between left and right sides and is calculated by subtracting the left from the right side and taking the absolute value. Directional asymmetry (DA) similarly subtracts the right side from the left side and retains the sign, but the mean is not equal to zero. Antisymmetry does have a mean of zero or close to zero, like when half of lobsters have a larger left claw and half of lobsters have a larger right claw. The least powerful way to observe asymmetry is through the correlation coefficient or coefficient of determination, because the range of observations influences the values such that a sample may appear asymmetric merely because the range of variation is small. Indices that adequately capture asymmetry are the mean absolute deviation, a Procrustes distance of geometric morphometric data, dense surface modeling, and continuous symmetry measures. When considering asymmetry research questions, size scaling may not always be necessary, and because the corrections can conceal asymmetry, the reasons should always be documented. Depending on the data type, error may be multiplicative as in actively growing tissue like bone with age as a covariate, or error may be additive as in inert data types like dermatoglyphic features and measurement error. For data types with additive error, size scaling is not appropriate, because it should be absent. For dermatoglyphic features like finger patterns, individual digits do not achieve statistical independence, so the statistical unit of observation is minimally the individual and thus maintains the correct degrees of freedom during analysis. However, Graham and Özener (2016) argued finger ridge-counts can be treated finger-by-finger, homologous digits may be analyzed as separate traits, or all five digits on a hand can be considered a single variable. Presumably, the difference in rules of analysis for finger patterns

compared to finger ridge-counts referred to some developmental understanding that was not explained, even though finger ridge-counts depend on the finger pattern. Of the types of asymmetry, directional asymmetry, fluctuating asymmetry, and an index of asymmetry were included in this dissertation. Although the suggested methods of analysis were not applied in this dissertation, following Graham and Özener (2016), no size scaling was performed for the majority of analyses, especially because most dermatoglyphic features were in the same units: counts, unless explicitly documented how and why.

As an example of early adoption of formal calculation of asymmetry and diversity, Holt (1958) used a variable for diversity: $S^2 = \sum q^2 - Q^2/10$, where S indicated the measure of diversity, q indicated the finger ridge-counts on individual digits, and Q indicated the TRC, and updated the measure of diversity as $S/\sqrt{10}$, which accounts to the standard deviation of the ten digits. Applying these two approaches to American Whites, American Blacks, African Blacks, and part of Holt's (1955) English sample, the finger ridge-count means decreased in order from digit 1, digit 4, digit 5, digit 3, to digit 2, and from finger to finger American Whites were more similar to the English sample and American Blacks were more similar to the African sample (Jantz 1974b). From this proof of concept, $S/\sqrt{10}$ as a measure of inter-finger variation could distinguish between populations, even contemporary populations from their putative contemporary parent population.

By increasing the number and breadth of populations from four to five, now including American Whites, American Blacks, African Blacks, English, and Easter Island, and by including both indices of asymmetry and diversity, Jantz (1975) observed patterns of variation that agreed more with shared population history than indicated environmental instability during

development. Here, the index of asymmetry: $\sqrt{A^2}$ was introduced which accessed finger ridge-count asymmetry as well as an updated measure of diversity: $S/\sqrt{5}$, which accessed the sum of ridge-counts of homologous digits where higher values indicated greater asymmetry. Overall, $\sqrt{A^2}$ and $S/\sqrt{5}$ exhibited different results than how TRC reflected intergroup variation Jantz (1974). For example, the American Blacks exhibited the lowest values for $\sqrt{A^2}$ and $S/\sqrt{5}$ with the overall greatest homogeneity between digits and overall the lowest finger ridge-counts. The Easter Island group exhibited only slightly higher values of $\sqrt{A^2}$ and $S/\sqrt{5}$ than the American Blacks, but overall the Easter Island group exhibited the highest finger ridge-count values. The American Whites and the English group exhibited the overall greatest heterogeneity between digits and asymmetry. Jantz (1975) showed African populations exhibited the lowest index of asymmetry and index of diversity, where Easter Island populations were just higher, and European populations exhibited the highest heterogeneity across all the studied population groups. The agreement of the index of asymmetry and index of diversity with population differences in Jantz (1975) challenged the core concept of asymmetry and diversity as a biological indicator of developmental stress with a high environmental component. The relationship between finger ridge-counts and population structure was not direct, which suggested finger ridge-counts represent multiple modules that differ in evolvability, or our heritability calculations are not sufficient to actually express how quickly features can respond to an evolutionary force relative to the phenotypic variance.

Hawkinson (1981) specifically addressed the concept of developmental stress in driving dermatoglyphic variation by comparing an Armenian population located in West Australia that experienced stress *in utero* compared to an Armenian population located in East Australia that

was secure as children born from the stressed parents. Finger ridge-counts, palmar ridge-counts, and palmar ridge-breadth were converted into variables of directional asymmetry, with a bias to one side, and fluctuating asymmetry, with a mean of zero, and greater asymmetry was hypothesized to indicate a lower ability to withstand stress or even perhaps a greater overall amount of stress. The stress experienced *in utero* for the West group would have had mothers undergoing physical stress and malnutrition while pregnant with this West group sample population. Using PCA and a multiple analysis of variance, no differences were detected between the West and East group for any of the dermatoglyphic features. Furthermore, directional asymmetry performed least well in detecting stress in these sample groups. Hawkinson (1981) investigated the effect of stress during intra-uterine development as a producer of directional asymmetry in fingers and palms but did not find any relationship.

If directional asymmetry and fluctuating asymmetry do not represent developmental stress in Hawkinson (1981), then perhaps subdividing the population by females and males may reveal evidence of stress. Jantz and Brehme (1993) considered the influence of sex in interpreting directional asymmetry and fluctuating asymmetry as well as whether palmar ridge-counts exhibited directional asymmetry in a German group. For directional asymmetry in Germans, females exhibited higher left palmar ridge-counts than males, and the radial and ulnar margins of the palm exhibited greater fluctuating asymmetry than the center of the palm. Overall, sex differences did not clarify any developmental component to directional asymmetry and fluctuating asymmetry. To validate the results, a simulated distribution of a bivariate normal with the same mean and standard deviation as the left subtracted from the right side was created and evaluated via linear regression in the same way as the sample data. In this dissertation, rather

than studying simulations for comparison of methods, cross-validation on the sample data was performed as proven appropriate. “When the true data generation mechanism is unknown or too complex, it is preferable to evaluate the performance of methods by cross validation on real data” (Xing, Lesperance, and Zhang 2020:68; Leisch and Dimitriadou 2021).

To continue the search for clarity in heritability, familial relationships were accounted for when analyzing directional asymmetry and fluctuating asymmetry for finger and palmar dermatoglyphic features from both hands. In a sample of two generations of parents (200 females, 200 males) and their offspring (195 females, 229 males) in West Bengal, India, the directional asymmetry and fluctuating asymmetry dermatoglyphic features were z-transformed and subject to analysis of familial correlation, to PCA, and to pedigree analysis that calculated the maximum likelihood estimates of heritability (Sengupta and Karmakar 2006). Intergeneration correlation was calculated using Pearson’s correlation, and intra-generation correlation as between two siblings was evaluated using a t-test on the correlation coefficient. The directional asymmetry and fluctuating asymmetry correlations overall were much lower than the theoretical values, suggesting an environmental component to asymmetry. For the PCA, the first five principal components accounted for 74% of the variation, and PC1 indicated the fluctuating asymmetry of finger ridge-counts for 24% of the variation; PC2 indicated the directional asymmetry of finger ridge-counts for 19% of the variation, and PC2 exhibited the highest heritability (0.19); PC3 indicated the asymmetry of digits 2 and 3 ridge-counts for 12% of the variation; PC4 indicated the asymmetry of the palmar atd angle and the palmar pattern intensity index for 10% of the variation; and PC5 indicated the asymmetry of the palmar *a-b* ridge-counts for 7% of the variation, and PC5 exhibited the lowest heritability (0.08). The lower heritability of

palmar ridge-counts suggested the asymmetry of the palm served as a more reliable indicator of developmental homeostasis than asymmetry of the fingers. The z-transformation of the asymmetry variables, which are already transformed from raw dermatoglyphic features, as well as the selection of correlation coefficient did not conform to the guidelines subsequently offered by Graham and Özener (2016). Subsequent dimensionality reduction via PCA on z-transformed asymmetry variables may have masked the variation between groups as well as dermatoglyphic features, so the results found in Sengupta and Karmakar (2006) likely represent a conservative picture of the patterns of variation in these groups as well as the heritability estimates.

In addition to interpopulation variation and population histories, human adaptation to stressful environments remains an active research avenue, and Dipierri et al. (2018) focused on comparing dermatoglyphic asymmetry from finger patterns and finger ridge-counts in groups living in high, mid, and low altitude environments in Jujuy Province, Argentina, (170 females, 140 males) that undergoes an altitudinal gradient from 500 m up to 4000 m. In these groups, fluctuating asymmetry reflected the small, random deviations from perfect symmetry of the average individual after all direct effects of the genotype and the environment on the character of interest have been removed. When considering total phenotypic variation as a combination of genetic and environmental components, the fluctuating asymmetry translates as the random aspect of the environmental component and also as a measure of developmental precision. Dipierri et al. (2018) defined developmental stability as the ability of an individual to correct for the random perturbations caused by developmental noise, and increased developmental noise lowers developmental precision, resulting in an increase in fluctuating asymmetry. Directional asymmetry, then, reflects the amplified development on one side more than the other side and

can artificially inflate values of certain fluctuating asymmetry indices, so calculating directional asymmetry usually follows calculating fluctuating asymmetry, because traits that exhibit directional asymmetry should be excluded from fluctuating asymmetry analyses. The contribution of the genotype tends to be larger in directional asymmetry than in fluctuating asymmetry, and fluctuating asymmetry does not change much as a result of inbreeding since fluctuating asymmetry is not highly heritable. Here, Dipierri et al. (2018) used the amount of variation in fluctuating asymmetry to represent population fitness indirectly, and three goals defined this research. First, all along the altitudinal gradient the variation of ridge-counts were considered separated by digit, left or right side, sex, and population. Second, all along the altitudinal gradient the pattern of bilateral asymmetry was established by digit, left or right side, sex, and population. Third, all along the altitudinal gradient the asymmetry and diversity indices were calculated separately by sex and by population in Jujuy Province, compared to other populations like Amerindians from Chile, Europeans (including a population from Ohio, USA), Western Asian, African, Polynesian, and Indian. Females and males were analyzed separately within populations, and in addition to directional asymmetry and fluctuating asymmetry variables, the raw finger ridge-counts, TRC, the index of asymmetry ($\sqrt{A^2}$), the index of diversity ($S/\sqrt{5}$), the sum of bilateral radial finger ridge-counts, and the sum of bilateral ulnar finger ridge-counts were also calculated. These dermatoglyphic features did not meet the requirements of parametric tests of normal distributions, so non-parametric Kruskal-Wallis tests were employed to compare populations, the Mann-Whitney U test compared sex, the Wilcoxon matched pairs signed-rank test compared radial to ulnar sides, the Friedman's test compared digits, the Spearman's rho correlated the TRC, index of asymmetry, and index of diversity by

sex, and PCA visualized the patterns of variation between asymmetry, diversity, sex, and the populations. Finger ridge-counts exhibited higher homogeneity overall between populations and females and males as well as for non-homologous digits in the high altitude population. Despite different degrees of stress due to hypoxia due to altitudinal differences, populations with the same ethnic origin exhibited no significant differences in asymmetry of finger ridge-counts, except for the index of diversity in males. No directional asymmetry was detected, but fluctuating asymmetry was present and differed by digit and by sex, and overall males appeared to have less developmental stability or be susceptible to developmental noise. The frustrations of researchers in dermatoglyphics pertaining to lack of consistent and satisfying methodology were addressed here using the foundation of data mining that incorporates interdisciplinary statistical tools to tackle difficult data sets, as described in the next section.

1.3 Data mining

Historical modeling approaches of dermatoglyphic data suffered from a scarcity of data, no evaluation framework, and simplistic statistical modeling that was assumption driven (Abraham et al. 2013). Both the commitment to data-driven discovery and the structure provided by data mining guided the theme and structure of this dissertation in the choice of types of methods to include, since data mining draws from statistics and artificial intelligence in the form of machine learning providing a dynamic approach flexible and powerful enough to address dermatoglyphic data that are large-scale, high-dimensional, heterogeneous, complex, and non-normally distributed, where traditional statistical methods are unsuitable. While this worldwide dermatoglyphics data set contains many individuals and many variables, the data set does not fall under big data principles where data are scraped from multiple sources continually generating

observations. Data mining offers a means for extracting previously unknown but potentially useful information and patterns implicit to a data set through exploration and testing (Tan et al. 2019a). In the workflow of data analysis, data mining occurs after the input data are preprocessed, and data mining occurs before post processing to extract information (Figure 4).

Before preprocessing, the researcher must acknowledge the variable types present in the data set, like whether variables are nominal, ordinal, or interval/ratio numeric, and whether the variables are multivariate or even collinear. Real data do not always neatly fit into a single category, so sometimes researchers may treat a variable type as another to simplify data analysis, and this may be approximately correct and may not unduly bias the results. How to treat the data also relies on the researcher's interest in the dimensionality, sparsity, resolution, and size of the data set. High dimensionality can overload a model or can reduce interpretability when high dimensions are reduced. Sparsity is the focus on a variable only when present, even when the absence of an observation for that variable may hold information of interest. Resolution considers how patterns in data depend on the scale of the data, and size considers how a method of analysis can depend on the size of the data set, like with the law of large numbers suggesting as a sample size grows the mean approaches the population mean so variation may be constricted and standard error calculation may be biased (Tan et al. 2019b). Real data sets, however, may be compositional, contain many zeros, contain binary presence/absence variables, and contain clustering of observations, for example. When observations cluster into groupings, mixed models and multilevel models may be appropriate that fit a different regression line, called a random effect, for each grouping and estimate a typical slope, called a fixed effect, across all groupings. Linear-based methods can be approximately correctly applied to these nonlinear response

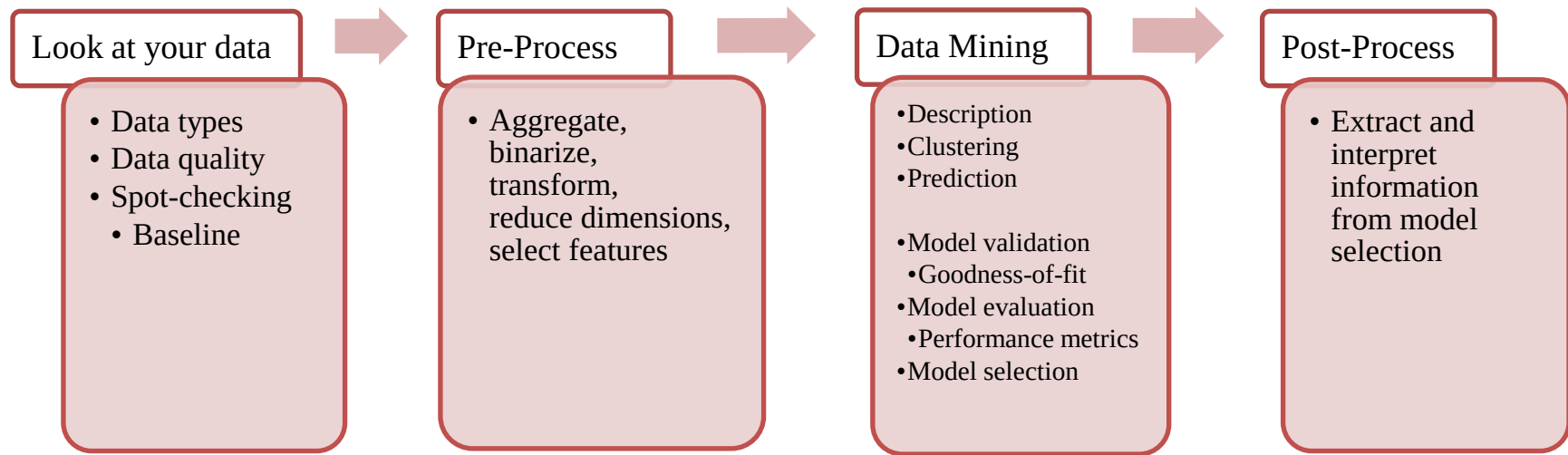


Figure 4. Data analysis process using data mining approach

variables if the variables can be transformed to linearity, or if within a more narrow resolution of the environmental gradient the distribution of the response variables approximates a linear function (Paliy and Shankar 2016:1041).

Further, the data quality should be identified and considered prior to even preprocessing, because poor data quality may restrict subsequent data mining and post processing steps. Issues of data quality include noise, outliers, wrong data, missing observations, and duplicate variables. For variables, noise refers to the modification of original values so their magnitude and shape are distorted and can mask patterns. Outliers are data observations with characteristics different from the majority of other observations, and sometimes outliers are noise obscuring the data patterns, and sometimes outliers are the target of the analysis. The presence of outliers may bias the calculation of standard errors and model coefficients. The treatment of wrong data and missing observations should be consistent and clearly documented for reproducibility. The definition of duplicate variables should be clearly outlined, and any exclusion of these variables from analysis as a result should be rationalized, because while duplicate variables dilute models not all researchers may agree on definition or treatment of duplicate variables (Tan et al. 2019b).

Data may also be expressed as dissimilarity or similarity measures, where dissimilarity measures with lower values indicate high similarity and where similarity measures with higher values indicate high similarity. Each variable type (nominal, ordinal, interval/ratio numeric) hosts specific dissimilarity and similarity measures, with their own drawbacks. For example, correlation represents a similarity measure for numeric data, but a drawback is correlation relies on a linear similarity, and correlation may be valued at zero for variables with a non-linear relationship like a mirrored or hyperbolic relationship. Dissimilarity and similarity measures

become flexible when weights are applied to the different variable measures as a way to combine similarities and to treat some variables differently from others (Tan et al. 2019b).

Given a data set, the methods to use, the data variables that will be useful, and whether the research question can be effectively modeled, are all unknown. Prior to preprocessing, multiple methods using multiple, different variables should undergo a spot-checking to perform a preliminary assessment whether the research problem can be addressed (Brownlee 2014a). Spot checking incorporates a diverse set of methods and calculates baseline models as a jumping off point for subsequent data mining (Wu et al. 2008). The preprocessing, data mining, and post processing steps depend on the baseline results and thus direct these steps rather than data analysis proceeding along an arbitrary path (Figure 4).

Preprocessing represents the first step in the data analysis workflow and proceeds once the data have been defined in terms of data types and quality, already discussed, and assumes the algorithms to use are known. Aggregation, binarization, transformation, dimensionality reduction, feature subset selection, and feature creation all occur during preprocessing to prepare the data for the data mining of exploration and testing and for the post processing of interpretation, which requires constant consideration of the research questions. Aggregation reduces data for simpler data analysis like aggregating groups into regions and landmasses or possibly reduces variation for easier data interpretation like aggregating finger ridge-counts into a single TRC measure. Binarization converts nominal or numeric variables into one or more binary variables, which often permits direct comparison of nominal to numeric variables. Transformation replaces old observations with new observations that are a function of the old. Standardization by subtracting by the mean and dividing by the standard deviation represents a

transformation strategy, and with dermatoglyphic features the reason for transformation should always be documented (Graham and Özener 2016). In biological data sets like dermatoglyphic data, standardization reduces the impact of the differences in the means of the variables between groupings thereby reducing the evidence of biological difference or distance (Paliy and Shankar 2016:1051). Dimensionality reduction reduces the complexity of an algorithm and primarily permits visualization of complex data sets. By reducing the dimensions, the variables important for producing some pattern can be identified as well as the variables not contributing and even actively producing noise. Feature subset selection also reduces complexity of an algorithm by reducing the number of variables but through different ways of statistical testing to remove redundant variables or variables not improving the performance of a model via interpretability and reducing overfitting. Feature creation relies on the researcher's ability to recognize how variables in the existing data set can be combined or transformed into new variables that capture more information more efficiently from the data set for analysis (Tan et al. 2019a).

Data mining includes techniques of description, clustering, and prediction through classification and regression, to identify patterns implicit to the data, group similar observations in the data, and predict unknown or future variable values (Wu et al. 2008). Clustering represents a descriptive technique that can generate new variables important for driving patterns in the data set but were not initially measured in the data set and important for reducing the size of a large data set. Through various measures, observations are placed into groupings where observations that are more similar cluster together and observations that are more dissimilar are placed into other groupings. The new cluster assignment variables can be included as explanatory variables in subsequent predictive modeling (Tan et al. 2019c).

In the classification prediction task, models are constructed to relate a class grouping variable as a function of other variable values, and the model that performs the best by correctly classifying the most observations based on previously known class membership is selected to represent the data (Tan et al. 2019d). No classification methods were included in this dissertation, because the classification of individuals into groups was not the goal of this dissertation. The goal of this dissertation was to illustrate patterns in worldwide dermatoglyphics data of finger ridge-counts and to identify the dermatoglyphic features, explanatory variables, and methods that best highlight the patterns of variation. A regression prediction task performs a prediction of a continuous numeric variable value based on the values of other variables through the assumption of some model of dependency, like a linear relationship. The model that performs best by predicting values closest to a known set of values is selected to represent the data.

Model building involves three steps: validation, evaluation, and selection (Shmueli 2010). Model validation encompasses model fit and relays how well the estimated model fits the real world data and how closely the estimated statistic compares to the population parameters (Zhang 2016). In explanatory models, detecting multicollinearity is part of analyzing model fit. Prior to using any regression model, a goodness-of-fit test is required to indicate whether the model is appropriate for the data. Measures of fit are solely within-sample approaches, because they measure how well the model can make inferences about the past data used to fit the current model. Informal ways of assessing model fit involve comparing the residuals against various values. The residuals encompass the variation in the response variable not captured by the model, calculated by subtracting the fitted response variable values from the observed response variable values. Residuals are not the same as the model error but are calculated from it. Comparing the

residuals to the fitted response values allows detection of heteroscedasticity, or non-constant variance in the error, and comparing the residuals against the explanatory variables allows detection of curvature or a horseshoe shape. Comparing the residuals to the sequence of data collection allows detection for autocorrelation, or temporal correlation between the observations, and comparing the residuals against the standard normal derivatives allows for detection of non-normality of the errors (Crawley 2013a:404).

Many goodness-of-fit measures appropriate for nonlinear models focus on this comparison of a null model to a fitted or saturated model (Chen and Onnela 2019). Description of traditional goodness-of-fit measure R^2 is found in *Appendix A Supplementary information and table, Chapter One*. The other two traditional goodness-of-fit tests, the deviance test and the likelihood ratio test, compare models to one another to measure the model with the better fit. Information criteria fit tests similarly compare models to one another using an information criterion, and the model with the lower score indicates a relatively better model fit (Grünwald 2007; Shmueli 2010; Bonifay 2022). The Akaike information criterion (AIC) uses a log likelihood approach for the goodness-of-fit: $-2(model \log likelihood) + 2p$, where p represents the number of model parameters including the intercept.

Like the R^2 value, the deviance value improves as the model increases in complexity, and a smaller deviance indicates a better model fit (McElreath 2020:184). Deviance is calculated as two multiplied by the log likelihood of the fitted model subtracted from the log likelihood of the saturated model, and the null deviance is calculated as two multiplied by the log likelihood of the null model subtracted from the log likelihood of the saturated model. While the null model contains no explanatory variables, the saturated model is the full model containing all

explanatory variables without removing any through feature selection or regularization. The Kullback-Leibler R^2 value is calculated as one minus the deviance divided by the null deviance and represents an R^2 value appropriate for nonlinear models and for count data (Cameron and Windmeijer 1997). The Kullback-Leibler R^2 value is not the same as but is related to the Kullback-Leibler divergence, also known as the relative entropy or the information divergence, and leverages the concept of learning through data compression, that the regularity in data can be used to compress data to the minimum description length, or how many bits of information in a binary string are required to code the data for transmission with some given precision (Rissanen 1978; Rissanen 1983). The best hypothesis for a data set is the one that results in the largest compression of the data, and as such the minimum description length is a generalization of the maximum likelihood where maximum likelihood estimates for model selection are not appropriate (Hansen and Yu 2001). If a model fits the data well, then instead of storing and transmitting the data, sending and storing the model description satisfies the minimum description length and is preferred, and simpler models are preferred over complex models (Grünwald 2007).

Accordingly, a benefit of the Kullback-Leibler R^2 value compared to traditional goodness-of-fit measures is that the Kullback-Leibler R^2 value does not inflate simply because explanatory variables are added to the model (Kvalseth 1985; Cameron and Windmeijer 1997). A higher Kullback-Leibler R^2 value calculated as $1 - (\text{model deviance} \div \text{null model deviance})$ reflects a lower model deviance in the numerator, and a smaller model deviance indicates a better model fit. A higher null deviance suggests the null model fits the data poorly compared to the fitted or reduced model represented in the numerator as the model deviance, so a low-valued numerator and high-valued denominator create a small number; one minus a small number is

closer to one. Thus, the model deviance is optimally a low value, and the Kullback-Leibler R^2 value is optimally a high value closer to one. Finally, although these goodness-of-fit measures reduce overfitting, due to the bias against complexity, model selection should not be based on goodness-of-fit measures and should be based on out-of-sample prediction (McElreath 2020:168).

Model evaluation and selection do not differ in any substantive way but can be differentiated by considering model selection the end process of model evaluation by selecting the best performing model for the specific research question. Model evaluation, then, is the process of calculating the performance of the models and improving the models based on their performance so they can ultimately be selected as the best answer to the research question. Model evaluation focuses on how accurately the model predicts future data rather than accurately inferring the true population distribution. Out-of-sample predictions on future or unseen data depend on the underlying data-generation process and how much data can be supplied to the models. Consequently, just adding new explanatory variables does not improve model evaluation and selection as happens with goodness-of-fit measures (McElreath 2020:185). When the underlying data generation process is unknown or too complex, cross-validation represents a common method of model evaluation and selection compared to simulation studies (Xing, Lesperance, and Zhang 2020; Leisch and Dimitriadou 2021). Performance metrics that compare the difference between observed and predicted response variable values represent another approach to model evaluation and selection, and while no single metric can adequately represent overall performance, including multiple performance metrics that each have their own strengths and flexibilities is recommended for optimal interpretation. Model evaluation procedures include

training for feature selection via regularization and optimization via hyperparameter tuning. Model evaluation continues until researcher-defined goals are met, and the optimal model is selected via model selection that addresses the research question. Model selection as the final step in data analysis leads into post-processing to extract and interpret information, the final step in data mining (Shmueli 2010). The results of model selection that feed into post-processing will always be relative to the variables and methods considered and to the research question. Although the differences are clearly noted here, because model evaluation and model selection often are interchanged terms in the literature and do not reflect mutually exclusive processes, in this dissertation the terms model evaluation and model selection were used interchangeably to represent training and improvement of the model and the calculation of model performance metrics for model comparison. Overall, data mining is a powerful tool that can help researchers make sense of the vast amounts of data they collect, and use that data to improve their decision-making processes. Given the difficulties in previous dermatoglyphics research and the flexibility but intentional progression of data mining to accommodate and explore the statistical peculiarities of dermatoglyphic data, the research questions in this dissertation were constructed on these foundations and are specified in the following section.

1.4 Research questions and dissertation organization

In the current study, introduced for the first time in entirety was the worldwide dermatoglyphics data set of more than 51,000 individuals to explore how to identify patterns of variation in these dermatoglyphic features. The contributions of this work rely on the consideration of the worldwide data set in three ways. First, different dermatoglyphic features were considered as response variables using perspectives of data comparison like univariate vs

multivariate and summary vs raw. Summary variables like total ridge-count mask the variation and covariation between sides of digits, between digits, and between hands of a single individual (Jantz 1987). Thus, I hypothesized even considering raw finger ridge-counts as univariate response variables, the digits primarily responsible for driving variation in summary variables like total ridge-count can be identified as well as covariant relationships between sides of digits and between digits. Further, I hypothesized multivariate directional asymmetry and fluctuating asymmetry variables that were transformations of raw finger ridge-counts would permit identification of covariant relationships between hands of a single individual but also perform better in statistical models due to the transformation and thus compression of variation. Univariate indices of asymmetry and diversity were included to contest the hypothesis for multivariate directional asymmetry and fluctuating asymmetry.

Second, different explanatory variables were considered using perspectives of sample size and composition as well as levels of population study and comparison. Because individuals in the worldwide dermatoglyphics data set were structured as part of larger groups like females and males, Ethnic Groups, larger Regions, and Landmasses, the appropriate level of population study has yet to be confirmed for different dermatoglyphic features, because many previous researchers focused on small, regional data sets (Meier 1980). I hypothesized that given the fixation of dermatoglyphic morphology during intrauterine development and thus limits of environmental contribution, the debated population-specific heritability of dermatoglyphic features, and the lack of observed population differences even between populations of shared history, that the resolution of groups via dermatoglyphic features was coarse (Dennis and Sunderland 1979; Ousley 1997; Glover et al. 2023). This means levels of population study

representing a fine scale like with Ethnic Group territories would exhibit significant overlap in variation of dermatoglyphic features. The extent of the effect of spatial autocorrelation of individuals within the worldwide dermatoglyphics data set was explicitly addressed with the hypothesis that removing spatially structured variation in dermatoglyphic features would clarify the remaining biologically significant sources of variation in the data.

Finally, different methods of descriptive modeling and analysis were considered using perspectives of unsupervised vs supervised and distribution-dependent vs distribution-free. Here, my overall goal was to create transparent methodological workflows where the contributions of researcher assumptions were clear as well as the statistical assumptions perhaps constraining model performance and interpretation. The unsupervised methods would permit clustering of dermatoglyphic features without the constraints of researcher assumptions of appropriate sample compositions and level of population study so any novel patterns of variation in dermatoglyphic features could arise and be interpreted through the lens of the assumptions of the statistical distribution assumptions but with little else. I hypothesized the distribution-dependent supervised methods would perform worst in terms of capturing the variation in dermatoglyphic features, because both statistical distribution assumptions and researcher assumptions were present and interacting in unknown ways. I further hypothesized supervised methods with limited statistical distribution assumptions would better capture the variation in dermatoglyphic features and also be more interpretable, especially since dermatoglyphic features tend to contain many zeros and to not to follow a normal statistical distribution (Holt 1955; Binkley 1985).

By addressing the three research questions of this dissertation, the dermatoglyphic data can be integrated and compared with other biological and genetic data toward a systems biology approach to understanding patterns in human biological variation (Veenstra 2021).

The rest of the dissertation was organized as follows. In *Chapter Two* the impetus for selecting each method of analysis was explained in detail. *Chapter Three* describes the experiment design including the data set, ethical considerations, and experimental procedures including model selection and evaluation of model performance and other implementation details. In *Chapter Four* experimental results were discussed and were separated by dermatoglyphic feature, such that results of the analyses applied to finger ridge-counts and summaries of ridge-counts were discussed before the results of the analyses applied to directional asymmetry, fluctuating asymmetry, and indices of asymmetry and diversity. Finally, conclusions were drawn and outlined in *Chapter Five*. All abbreviations used in this paper can be found in Table S1.

CHAPTER TWO

METHODOLOGY

Each method of analysis in this study specifically addressed an opportunity to visualize or measure the observed variation in the human dermatoglyphic data, as detailed in this chapter. The contribution of each method of analysis to accommodate different response data types, explanatory data types, and analytical constraints is found in (Table 1). For clarification, the response variables may also be referred throughout as output attributes, features, or variables or target attributes, features, or variables, and the explanatory variables may also be referred to as input attributes, features or variables. Examples of the response variables are finger ridge-counts, and examples of explanatory variables are Ethnic Group, Latitude, Sex, etc.

2.1 k-means clustering

Clustering techniques describe structure in data and aid visualizing data by assigning data points to a group or cluster without necessarily assuming the data derive from any statistical probability distribution, and k-means clustering represents a method for reducing the complexity of multivariate quantitative data. Since k-means clustering is an unsupervised method, groups are assigned based on the evaluation of high intra-group similarity and low inter-group similarity between different input data points rather than any relationship to a predicted output variable. The input data can either be cases (rows) or variables (columns) depending on which dimensions of the data set should be reduced (Morissette and Chartier 2013). Each group is defined by its centroid, which is the group mean in k-means clustering; thus, the group assignments are sensitive to outliers, and often the preprocessing step to normalize the features is necessary if the original data were measured in different units or scales. The clusters are assumed to be sampled

Table 1. Strategy of dissertation divided into response variable type, explanatory variable type, and method. The methods included in this study are listed across the top row. GLM = generalized linear model; LM = linear model; PCA = principal component analysis;

SVR = support vector regression

		k-means	PCA	LM	GLM	SVR
Response variables	Univariate vs multivariate	Multivariate	Multivariate	Univariate	Univariate	Univariate
	Summary vs raw	Either	Either	Either	Either	Summary
Explanatory variables	Sample size and composition	✓	✓	✓	✓	✓
	Level of population study	✓	✓	✓	✓	✓
Methods	Unsupervised vs supervised	Unsupervised	Unsupervised	Supervised	Supervised	Supervised
	Distribution-dependent vs free	Dependent	Dependent	Dependent	Dependent	Free

with equal probability from some normal distribution and to have spherical covariance matrices. Instead of Euclidean distance, other distance metrics can be used that allow for clusters of different shapes, like the Mahalanobis distance, which accounts for the covariance between the vectors (Paliy and Shankar 2016:1040). Although few statistical assumptions are required for k-means clustering, some K number of clusters is required information for the function. If no value of K is realistically known, then many different values should be tested to find the optimum separation of the data and the lowest within-cluster variation.

To estimate the optimal number of clusters before performing k-means clustering, use a direct method to optimize a criterion like within cluster sums of squares or the average silhouette, or use a statistical testing method to compare observations against a null hypothesis. One direct method, nicknamed the elbow method, measures the compactness of clusters by comparing the total within cluster sums of squares to a range of numbers of clusters. When assigning clusters, one algorithm of k-means clustering minimizes the total within cluster sum of squares, or the total intra-cluster variation to create compact clusters. Evaluating the plot of the elbow method, the smallest total within cluster sums of squares at the bend in the plot is optimal but is sometimes ambiguous. A second direct method, called the average silhouette, measures the quality of clustering by how similar a point is to its own cluster compared to other clusters. Ranging from -1 to 1, a high silhouette value indicates a point matches better with its own cluster and worse to a neighboring cluster. If the average silhouette is low or negative, then the clustering configuration is not appropriate by having too few or too many clusters. The plot of the average silhouette compares the average silhouette to a range of numbers of clusters (Kassambara 2017). Evaluating the plot of the average silhouette approach, the largest average silhouette is optimal. A statistical testing method, called the gap statistic, measures the

organization of the clusters by comparing the total intra-cluster variation of a range of numbers of clusters with simulated expected values under a random uniform reference distribution. Evaluating the plot of the gap statistic is performed computationally as the smallest cluster value k whose gap statistic is within one standard deviation of the gap at $k+1$, and roughly the largest gap between the clustering of the simulated values and the observed values is optimal (Tibshirani, Walther, and Hastie 2001).

The most widely used k-means algorithms are Forgy/Lloyd, MacQueen, and Hartigan and Wong, and the choice of algorithm depends on the research question and goals. The Forgy/Lloyd algorithm implements a batch, or offline, centroid approach that implements a transformative step to all cases at the same time, and thus performs well on large data sets. Incremental, or online, algorithms store the cluster membership of each case after one or two steps and thus require more computation the larger the data set. The Forgy algorithm itself considers the data to be derived from a continuous distribution, while the Lloyd algorithm considers a discrete distribution; otherwise, the two algorithms perform the same procedure (Forgy 1965; Lloyd 1982). After normalizing the data and selecting the number of K clusters, some number of group centroids is randomly initialized in the feature space. The Euclidean distance between the input data and the centroids is calculated so data points are assigned to the group whose centroid is closest. Next, each group centroid is re-calculated based on the mean of all data points assigned to that group, and if the new centroid is different from the old centroid, the process of group assignment according to shortest Euclidean distance begins anew, and some data points may end up in a different group than in the beginning. This iterative process of adjusting the centroid and reassigning group membership based on similarity continues until the old and new centroids are identical, until the difference between centroids is small, or until a

researcher-set maximum number of iterations is reached. This is the Forgy/Lloyd algorithm.

Problems arise if all cases in a cluster are reassigned at the same time, because an empty cluster may result (Hamerly and Drake 2015).

The MacQueen algorithm implements an incremental, or online, centroid approach, where the centroid is recalculated after any case changes group membership as well as after each time the algorithm passes through all cases. In this way, the MacQueen algorithm updates the centroids more often and thus may only require one complete pass through the cases to converge, meaning the old and new centroids are the same (MacQueen 1967; Hamerly and Drake 2015).

The Hartigan and Wong algorithm implements a batch, or offline, approach that targets a local optimum and assigns clusters by minimizing the within cluster sums of squares. As long as the total within cluster sums of squares is minimized where compactness of the clusters is favored, a case may be reassigned to a new cluster even if the old centroid is closer. Just like the Forgy/Lloyd algorithm, the centroids are randomly initialized, and cases are assigned to the closest cluster. When the group centroids are re-calculated, if the new centroid is different, the within cluster sums of squares is calculated for every case in another cluster. If one of the other cluster sums of squares is smaller than the current cluster, the case is assigned to the new cluster. The algorithm continues until no cases change cluster assignments, and to continue would either increase the within cluster variation or make the clusters more similar to each other (Morissette and Chartier 2013:17).

Problems arise in the general performance of k-means clustering when the optimal cluster shapes are not spherical, so points lying in a straight line may not cluster together, and points on a different line may cluster together, based on distance alone. Kernel methods like support vector machines, where the data are mapped into a higher dimensional space to make the cases linearly

separable, would be appropriate to correctly classify cases into non-spherical clusters. Further, because clusters are assigned by minimizing within-cluster variation, clusters with more cases are given greater weight than clusters with fewer cases, and this may beget incorrect classification.

2.2 Spatial analysis via distance-based Moran's eigenvector maps (dbMEM)

Beta diversity denotes the spatial patterning in community composition and may account for much of the observed variation in data rather than the alpha diversity, which addresses the within-sample diversity of the number of groups (richness) and different sizes of the groups (evenness) (Jost 2006; Jost 2007; Chao, Chiu, and Jost 2014). Initial identification and discrimination between alpha and beta diversity as different sources of variation allows the relevant sources to be subsequently modeled separately and in richer detail. Examples of spatial patterning in human populations are large-scale dynamics of climate and altitude that can produce gradients of change in human variation but also patchy structures with distinct interfaces as well as regional-scale dynamics of gene flow, response to local disease, and even social organization. Various environmental processes generate the response variables of human variation that have observable spatial patterns (Legendre and Legendre 2012a). Examples of structures of variation may reflect a combination of spatial patterning and non-spatial patterning to include the overall mean of the response variable (i.e. finger patterns or finger ridge-counts); a gradient change in the mean of the response variable operating across the entire sampling area; a regional-scale spatial structure; local deterministic structures that contain no easily recognizable spatial structures; and random error as the residual or stochastic component of the observed variation (Borcard, Gillet, and Legendre 2018:300). In this way, although space itself does not explain ecological variation, spatial variation is vital to the successful functioning of an

ecosystem or population and can be studied as non-random patterns (Legendre 1993; Legendre and Legendre 2012a:860).

Spatial structure in the response variable originates either from external environmental factors as induced spatial dependence or from internal community factors as spatial autocorrelation. The model for induced spatial dependence indicates the observed response variable can acquire a gradient-like deterministic spatial structure merely as a function of dependence on spatially-structured environmental explanatory variables, and importantly the error terms from the external variables are not autocorrelated (Legendre and Legendre 2012b). Autocorrelation refers to the nonrandom value of a variable depending on the position of the value in a sequence or, generally, on other nearby values (Legendre 1993). Spatial autocorrelation, then, refers to the nonrandom relationship between the value of a variable at a particular location and another value of a variable at another particular location, and here each variable value must have an associated location (Legendre 1993). The model for spatial autocorrelation, then, uses a weighted sum of the centered response variable to account for the influence of location on the individual response variable values. Spatial autocorrelation may mimic a gradient structure if the underlying process driving the variation influences a larger area than the sampling area, and this is deemed a false gradient (Legendre 1993:1662). No statistical approaches can distinguish the false gradient from a true gradient whose spatial structure arises from spatial dependence, and thus the underlying process of a false gradient must be tested as a biological hypothesis (Legendre and Legendre 2012b).

Further complicating statistical analysis, spatial autocorrelation suggests values at one location can be predicted, even in part, from values at another location if the underlying biological process and the location are known. For models of spatial autocorrelation, each data

value is not stochastically independent so the error terms are not independent, and because the data values are not free to vary, each data value does not merit a full degree of freedom but rather some unknown fraction. Statistical bias when the null hypothesis is rejected too often really only occurs with a parametric test where both response and environmental explanatory variables are spatially correlated (Borcard, Gillet, and Legendre 2018:301; Legendre 1993:1660). To evaluate the hypothesis of induced spatial dependence, multiple regression or canonical analyses are appropriate. To determine whether the entire spatial structure detected in the response variable is explained by the environmental explanatory variables or whether a spatial autocorrelation hypothesis is more realistic such that some spatial variation remains unexplained in the response variable, variation partitioning and multiscale ordination are appropriate analyses (Legendre and Legendre 2012b:15).

To identify the spatial structures within the sampling area, spatial variables are constructed at relevant scales using distance-based Moran's eigenvector maps (dbMEM), a newer iteration of the principal coordinates of neighbor matrices (PCNM) method (Borcard and Legendre 2002). The spatial variables are eigenvectors of truncated matrices of geographic distances among sampling sites and are constructed following initial steps of the PCNM method and then final changes to the distance matrix to be considered dbMEM. First, construct a distance matrix among sites by applying a Euclidean distance function to a set of Cartesian geographic coordinates, and geographic distances like latitude and longitude can be transformed into Cartesian coordinates. Second, truncate the matrix to include only close neighbors within a threshold distance. The threshold distance is selected by the researcher or software as a small distance such that any remaining points must be connected by links smaller than or equal to the threshold distance. Any pairs of points above the threshold distance are assigned an arbitrarily

large distance value, like 4 times the threshold distance value. Third, compute a principal coordinates analysis of the truncated distance matrix. Fourth, compute Moran's I for all eigenvectors, and retain the resulting eigenvectors that are modeling positive spatial correlation, meaning the Moran's I is larger than the expected I. Fifth, use the resulting eigenvectors as spatial explanatory variables in subsequent regression analyses (Legendre and Legendre 2012a:861; Borcard, Gillet, and Legendre 2018:315).

Initially, PCNM methods produce diagonal values of 0 on the truncated distance matrices confirming a sampling site is connected to itself with zero distance (Borcard and Legendre 2002). However, replacing the diagonal values of 0 with the arbitrarily large 4 times the threshold value now implies that a site is not connected to itself and creates dbMEMs, or eigenvalues of the resulting spatial eigenvectors, that are proportional to Moran's I coefficient calculated in step 4, and any eigenvalues larger than the Moran's I expectation are identified as those modeling positive spatial correlation. Keeping values of zero on the diagonal of the truncated distance matrix as in PCNM artificially results in more positive eigenvalues than in dbMEM. Further of value, the eigenvectors calculated using PCNM methods and dbMEM are the same, and the spatial descriptions resulting from the use of dbMEM are orthogonal, or linearly independent (Legendre and Legendre 2012a:862).

The dbMEM method operates best on regular sampling designs, because with an irregular design the researcher must select a larger threshold distance so points more distant can remain connected on the map at the expense of fine scale spatial details. For irregular sampling designs, data interpolation techniques are available, and for sampling designs with large gaps a nested approach may be warranted (Borcard, Gillet, and Legendre 2018). Because the threshold distance is selected such that all other links must be smaller than or equal to the threshold,

including any large distance like two locations separated by an ocean across continents will limit the interpretation of the MEM eigenfunctions to that cross-continental scale so is not recommended (Legendre and Legendre 2012b:864). For studies including locations spanning multiple continents or where locations are similarly separated by large distances, the recommended action is to divide the sample locations into relevant within-landmass groups and analyze the within-landmass groups separately. Trends in data, which are broad-scale spatial structures larger than the extent of the observed data, will be detected in spatial analyses whether or not finer-scale patterns of spatial correlation exist, so should be removed by detrending the response variables (Legendre and Legendre 2012b:16).

2.3 Principal component analysis (PCA)

Dimensionality reduction compresses data and aids visualizing data, and principal component analysis (PCA) represents a means of dimensionality reduction for multivariate quantitative data. Principal component analysis projects data onto a lower-dimensional surface that maximizes the data separation and minimizes the projection error, or the sum of squared distances between the original variables and the new projected dimension line (Dimensionality Reduction n.d.). PCA uses Euclidean distance by default to measure the dissimilarity between objects like variables or individuals, and when the biological data set contains many zeros PCA can exhibit a horseshoe visualization artefact such that objects at the edges of an environmental gradient appear close to each other in the feature space, like the tips of a horseshoe (Novembre and Stephens 2008; Legendre and Legendre 2012b; Paliy and Shankar 2016:1037). Often, a preprocessing step to scale or normalize the variables is necessary so the variables all have a comparable range of values and so no one variable exerts high influence just because the variable numeric values are high.

After feature scaling or mean normalization, a singular value decomposition is performed on the covariance matrix to compute the eigenvectors, which are now uncorrelated components. Eigenvectors indicate the direction, or angle, of projection. These eigenvectors can serve as input into subsequent machine learning algorithms or other types of analyses, and alone the contribution of the eigenvectors, indicated by the eigenvalues, reveals which reduced features are responsible for a proportion of the overall input variable variance (Krzanowski 2000:23).

Although commonly confused, linear regression and PCA fundamentally differ in the goals and the process. The goal of a researcher using linear regression is to predict an output variable based on some input variable so is a supervised method, whereas the goal of a researcher using PCA is dimensionality reduction to compress and visualize data but not to predict anything and does not consider both input and output variables, so is an unsupervised method. Further, linear regression proceeds to minimize the square error vertical line distance between any input variable value and the line predicted by the hypothesis, whereas the process of PCA proceeds to minimize the distance between the variables and the projected dimension line at an angle, or an orthogonal straight-line distance (Alpaydin 2014; Dimensionality Reduction n.d.). Problems arise as PCA only relies on the similarity between variables and intentionally does not capture the variation within a single variable. PCA maximizes the variation in data on the axes for the purpose of visualization, so has limits because the important components of the variation that allow you to discriminate between groupings that covary with some explanatory variable are not considered (Marion, Fordyce, and Fitzpatrick 2015).

2.4 Baseline models

Prior to preprocessing in data mining, data sets undergo spot checking, subjecting the multiple, different variables to a diverse set of methods, called baseline models to establish a

base performance (Wu et al. 2008). The baseline models provide a lens through which to interpret the results of methods during formal modeling and data mining steps. Baseline methods can be rule-based methods or traditional methods. Rule-based methods invoke a common rule like for classification to create baseline models that predict the most frequent class of the training data response variable or for regression to predict the mean, median, constant value, or quantile of the training data response variable. Traditional methods focus on minimizing complexity by using a simple algorithm like linear regression (Brownlee 2014b).

Baseline modeling follows the same steps for both rule-based and traditional methods: specify the data set for the target response variables and the explanatory variables; split the data into partitions like training and test sets; create the regressor like the mean estimation; fit the regressor to the training response variable; predict new samples using the fitted regressor and test explanatory variables; and score the performance using the predictions and the test response variable (Pedregosa et al. 2011). After the baseline models are created, the researcher can add or change the data variables, the analytical methods, and the parameters of the methods, and a baseline result provides a reference for whether these new changes and results add value and improve the solution to the research question. The baseline results direct the subsequent preprocessing, data mining, and post processing steps, rather than data analysis proceeding along an arbitrary path.

Baseline linear regression models (also just called linear models) must abide by the assumptions about the residuals and errors to be interpretable and thus useful as a baseline for subsequent regression modeling (Paliy and Shankar 2016:1041). First, the observations must represent random samples from a population. Second, the explanatory variables must exhibit no multicollinearity, and the errors must be independent of each other, which is sometimes referred

to as the independent error assumption. Multicollinearity reduces the precision of the coefficient estimates in the linear regression. Third, no multivariate outliers should be present, such that no one residual is visibly away from the random pattern of residuals, because outliers can unduly influence the regression trends.

Fourth, the residual errors for the explanatory variables must be normally distributed with equal variance, and this is sometimes referred to as the normality assumption. Non-normally distributed residual errors may indicate a useful explanatory variable was not included in the model or that a polynomial term to model a curve or an interaction term should be added as explanatory components to the regression model. Residual errors that do not follow a normal distribution may also indicate autocorrelation where one adjacent residual can predict another, which results in artificially smaller p values, because although the coefficient estimates truly contain higher variance, the model does not detect that. Autocorrelated residuals may suggest the use of robust standard errors to return more accurate coefficient estimates or even that a mixed model with error term for the clustering groups should be used (Lüdtke, Makowski, Ben-Shachar, et al. 2023). An Anderson-Darling test can test the normality assumption of the normality of residuals, and a Durbin-Watson test can test for autocorrelation of residuals.

Finally, the relationship between the explanatory variables and the response variable is assumed to be linear with a mean zero and an equal variance. Here, to exhibit linearity, the residuals should be randomly spread around the mean zero line. To exhibit homogeneity of variance, the residuals should form an approximately horizontal, even band around the mean zero line, and this is sometimes referred to as the homoscedasticity assumption. Violation of the homoscedasticity assumption, or heteroscedasticity, reduces the precision of the coefficient estimates in the linear regression. In residual plots, a distinctive cone or fan shape appears,

because as the fitted values increase along the x-axis, the variance of the residuals on the y-axis also increases. Heteroscedasticity may occur when the observed values encompass a large range between the smallest value and the largest value. Pure heteroscedasticity refers to a correct specification of the model but still observing a non-constant variance in the residuals. Impure heteroscedasticity applies when the model is incorrectly specified by leaving an important explanatory variable out of the model.

2.5 Generalized linear modeling (GLM)

Generalized linear models (GLM) represent a supervised learning algorithm that focuses on which explanatory variables have an effect on the response variables, where the response variables do not follow a normal distribution like with binary, count, and continuous variables. Because these response variables may not then range from $-\infty$ to $+\infty$, the relationship between the response variable and the explanatory variables may not be linear. Often, the response variables are categorical like multiple-choice responses to a survey or affiliation with a grouping. GLM may be appropriate if violation occurs of the assumptions of standard linear regression.

In generalized linear modeling, the response variable is modeled as a linear function of the explanatory variables and an error term. The GLM uses a link function to perform a transformation of the expected response values, like the mean, into a linear form by some unknown parameters (Paliy and Shankar 2016:1040). Another way to conceptualize, the link function links the explanatory variables to the distribution of the response variables, and the output of the GLM is a set of regression coefficients that define the relationship between explanatory and response variables captured by the model (Paliy and Shankar 2016:1045).

Regression models are fundamentally based on the underlying probability distribution function, like the Gaussian model is based on the Gaussian probability distribution function. The

Gaussian (normal) distribution reflects a symmetric continuous probability distribution defined from $-\infty$ to $+\infty$ and can handle response variable values of zero. To more closely follow a Gaussian distribution, data are often transformed like taking a logarithm of positive values to increase the symmetry of the distribution. Skewness, outliers, and other data characteristics can reduce the appropriateness of modeling data with the Gaussian distribution. The Poisson distribution reflects rate data or count data, like the probability an event occurs during a set timeframe or like finger ridge-counts in a population grouping. An assumption is the occurrence of the event, as indicated by the response variable, does not depend on previous occurrences of the event. The parameter of the Poisson distribution represents the average number of times an event occurred during a set unit of exposure. Exposure may be throughout a period of time, within physical area of space, over some distance, or in a set population size. The Poisson distribution is bounded at zero, so an event cannot occur negative times. Further, the mean of the event parameter is assumed to equal the variance of the event parameter, called equidispersion.

A Poisson probability distribution function maintains the assumption of equidispersion where the greater the mean value the greater the variation in the data, a phenomenon often absent in real-world data. Also assumed is the counts are independent of one another, and when they are not independent due to excessive correlation in the data the underlying distribution properties are disrupted, resulting in extra-dispersion. Overdispersion, a type of extra-dispersion, indicates the response variance term is greater than the mean, so there is excessive unexplained variation in the response (Hilbe 2012a; Crawley 2013b:580). Other causes of overdispersion may be correlation between the response variables or excess variation between response variable probabilities or count values. The consequences of overdispersion are the underestimation of the standard error of the estimates so variables may appear to significantly affect the response

variables when they actually do not (Hilbe 2012b:141). Other consequences of overdispersion are the model focuses on the effects on the variables from the dispersion, like differences between groups in the variables' variances, rather than the desired effects on the variables from the location, like differences between groups in the variable means (Paliy and Shankar 2016:1051). To recognize dispersion in a model, the dispersion statistic, or the Pearson (X^2) statistic divided by the residual degrees of freedom, exceeds the value of one. A dispersion value exceeding 1.25 merits correction of the dispersion, whereas a dispersion value around 1.05 does not require correction and usually represents a model with a large sample size (Hilbe 2012b:142). Problems arise with deviant dispersion, and immediate solutions are that the response variables do not outnumber the explanatory variables and that the explanatory variables are informative and not redundant.

Correcting overdispersion begins with fixing apparent overdispersion like by transforming the data or accounting for interactions between predictors, and should real overdispersion persist the correction is based on the structure of the overdispersion, like how enhanced Poisson models correct Poisson overdispersion and enhanced negative binomial models correct negative binomial overdispersion (Hilbe 2012b). Specifically, correcting overdispersion can occur by standardizing the within-group variance for each variable and will eliminate the unequal mean-variance relationship, but this may become computationally untenable for methods that use permutation to separate the groups. A second way of correcting overdispersion is to use multivariate GLM where different mean-variance functions can correct the unequal mean-variance relationship (Paliy and Shankar 2016:1051). Assuming an underlying Poisson distribution, the link function is the log, and a Poisson regression model can also be referred to as a log-linear model, for example. Should overdispersion persist, the Poisson

distribution should be substituted for a different statistical probability distribution like a negative binomial (Espinoza et al. 2021). The negative binomial model is appropriate to use when overdispersion in a Poisson model takes a gamma distribution. The negative binomial model is based on a mixture of Poisson and gamma distributions (Hilbe 2012a:1). Further, when the overdispersion takes no discernible shape or distribution, many researchers choose a negative binomial model. Negative binomial regression allows for overdispersion in count data, because the negative binomial probability distribution models the number of failures occurring before a specified number of successes, like a set count value.

2.6 Support vector regression (SVR)

Artificial intelligence offers a means for solving problems without explicit programming. Machine learning supplies the statistical tools to explore and learn trends in existing data to solve problems using new, unseen data. Machine learning approaches consist of supervised learning, unsupervised learning, and reinforcement learning. Reinforcement learning, also deemed semi-supervised learning, trains an agent to reach a goal by reinforcing interactions with the environment that move the agent closer to the goal, and a self-driving car avoiding accidents learns through reinforcement. Unsupervised learning is appropriate when data do not have output or target labels like when group membership is unknown, and clustering techniques and dimensionality reduction extract underlying trends in data. Supervised learning is appropriate when the output or target data are labeled like for a group membership, and classification and regression techniques train a model to predict the output data given existing input data so the model can then solve problems using new, unseen input data (Mouloudi et al. 2021). A main assumption is that the training data as well as unseen examples are independently generated by the same but unknown probability distribution. Whereas classification techniques predict discrete

categories of output data like disease status or flower species, regression techniques predict continuous output data like toxic chemical concentrations or human stature (Schrider and Kern 2018). The supervised machine learning methods primarily differ in their approach to structuring and training the function that predicts the output or target data from the input data (Schrider and Kern 2018:304).

Support vector machine represents a supervised machine learning algorithm that focuses on the objective of identifying the optimal hyperplane with the maximum margin, or distance, to separate different groups in the transformed higher-dimensional feature space (Cortes and Vapnik 1995). Support vector regression (SVR) focuses on the objective of identifying a regression function approximating the relationship between the input data and the target variable by finding the best fit line or curve passing as close as possible to the support vectors and while maximizing the margin. Specifically, SVR refers to the problem to predict a scalar output $\hat{y}$ given a d -dimensional input vector x and is formulated as $\hat{y} = w^T \varphi(x) + b$, where w is the weight vector of each dimension of the input variable in the kernel feature space, $\varphi(x)$, and b is a constant. When input data can linearly be divided into two groups a simple linear plane is a sufficient boundary, and when input data can linearly be divided into three groups, the dividing plane becomes two-dimensional like a flat plate (Krzanowski 2000:16). Hyperplanes represent multidimensional boundaries between data groups when input data can be divided into more than three groups. When input data cannot linearly be divided, support vector machine nonlinearly maps the input data into a higher dimensional feature space where a linear best fit is then created. To reduce the computational burden of creating the optimal hyperplane in a high dimensional feature space, only data points positioned closest to the best-fit curve and thus defining the

maximum margin are considered in the model. These select data points are called the support vectors, and the nonlinear transformation or mapping strategy is known as the kernel function.

To perform the kernel trick, which is used in other algorithms, a kernel function is selected to be applied between each pair of data points in the original feature space. Using the kernel function, the dot product between every pair of data points is calculated, and this calculation is equivalent to projecting the data points into the higher-dimensional space without actually calculating their coordinates (Sanchez-Fernandez et al. 2004:2301). The significance of the kernel trick is the input feature inner product calculations approximate the high-dimensional relationship, but the kernel function does not actually perform the transformation to a higher dimension. The final predicted output equals a constant term as well as the sum of the inner products added using weight variables (Smola and Schölkopf 2004:206). Each support vector is associated with a weight coefficient that reflects the importance of the support vector in defining the best-fit curve. The support vectors with large negative or positive non-zero weight coefficient values of the kernel function have the greatest impact on the definition of the best-fit curve. After training, the support vector machine model can be used to make predictions for new, previously unseen data points. The kernel function computes the distance of the new data points to the support vectors, which are the data points that lie closest to or along the best-fit curve, and the support vector machine model can compare the new data to the values originally predicted by the model.

The choice of kernel function can affect the final model output, and either a linear function kernel or one of many non-linear kernels like polynomial function, radial basis function, or sigmoid function, should be considered (Paliy and Shankar 2016:1040). The radial basis function is flexible over other kernels, because the radial basis function kernel can approximate a

linear function kernel and behave similarly to a sigmoid function kernel, and the polynomial function kernel contains more parameters to complicate model selection (Xu et al. 2013:1081; Zhang, Ma, and Lv 2021). The radial basis function permits a rounded boundary shape in the feature space and is defined by $e^{-\gamma\|x_i-x_j\|^2}$ where x_i indicates an input feature data point and x_j indicates another input feature data point, and the squared difference in their lengths is multiplied by gamma as the overall negative exponential.

Two parameters require tuning when using the radial basis function: the solely positive kernel width (gamma, γ) that determines the shape of the optimal hyperplane and the cost parameter (C). The kernel width value represents a scaling factor for the distance between data points in the feature space by determining the influence of a single training data point on the prediction of the model. A higher kernel width value indicates each training data point has a small influence on the prediction of the model resulting in increased complexity of the best-fit curve and may be prone to overfitting when applied to test data. In lieu of the kernel width parameter gamma, sigma (σ) can be used as equivalent to gamma by $1/2\sigma^2$. The loss function represents the difference between the prediction of the model and the actual output values, and a risk function is the average value of the loss function across the input values. Successful learning occurs when the risk function is minimized. The cost parameter (C) is a solely positive penalty value and represents a regularization parameter to balance maximizing the margin and a large value of the loss function (Chapelle et al. 2002). A higher cost value penalizes training errors more, potentially leading to a closer fit of the best-fit curve so fits the training model more closely but may be prone to overfitting when applied to test data, and a lower cost value is more tolerant to errors on the training data and allows a larger margin between the support vectors so may generalize better but may under fit the training data. A cost value of zero effectively

removes the regularization so the SVR model solely focuses on minimizing the error on the training data, which is not recommended since overfitting often results. The type of support vector regression called epsilon-regression does not stipulate the number of data points that become support vectors. In epsilon-regression, the researcher completely controls the amount of error permitted by the model, and these points lie within an epsilon-insensitive loss function area where the predefined value of epsilon indicates the radius of the loss function area, or the error. Any error within the epsilon-insensitive loss function area is disregarded as within an acceptable margin of error, and any error beyond the epsilon-insensitive loss function area is penalized according to the cost parameter and contributes to the loss function. The value of the epsilon parameter controls the trade-off between model accuracy and complexity by adjusting the regression function. Model evaluation strategies like grid searching and gradient descent perform well for single output SVR where a single performance metric for prediction error is evaluated unlike data sets with multiple output variables and thus the problem of reconciling a performance metric for each, interrelated output variable into some final model (Mao et al. 2014). Support vector regression addresses nonlinear problems but cannot accommodate multiple, dependent outputs, which are common in many real-world data sets (Glowacz 2021). Therefore, it can be beneficial to treat model selection as a multi-objective optimization problem, like in multiple output support vector regression.

2.7 Model selection and evaluation of model performance

The strategies of model evaluation for training and final model selection differed by method in this dissertation and are outlined below.

2.7.1 k-means clustering internal performance using the Dunn Index

Two metrics, internal and external, can evaluate the solution of a k-means clustering algorithm. The internal metric is solely based on the data set to which the algorithm was applied and measures the compactness, connectedness, and separation of the k-means clusters.

Compactness refers to how close cases are within a cluster, so the within-cluster variation based on distance measures between cases should be low. Connectedness refers to the extent the nearest case is in the same cluster, and although this value can range from zero to infinity, connectedness should be minimized. Separation refers to how clusters are separated, which can be measured by distances between cluster centroids or by pairwise distances between cases in different clusters. Many internal validation methods combine compactness and separation, and the Dunn Index generally reports the ratio of the inter-cluster similarity, or minimum separation, measured by the distance between clusters to the intra-cluster similarity, or maximum compactness, as measured by the diameter of the clusters. The minimum separation should be large, and the maximum diameter should be small, so the Dunn Index should be maximized, and the solution with the highest Dunn Index is preferred (Dunn 1974; Morissette and Chartier 2013:18). External metrics are appropriate when the k-means clusters are used as explanatory variables in subsequent classification or regression model frameworks, and the appropriate external performance metric depends on the model framework.

2.7.2 Generalized linear modeling with LASSO regularization

To optimize GLM, model evaluation through feature selection and resulting parameter estimation is performed. The least absolute shrinkage and selection operator (LASSO) regularization technique simultaneously performs feature selection and parameter estimation for GLM regression. By adding a penalty term to the likelihood function, the coefficients for some

variables shrink to zero, thereby reducing overfitting. One parameter requires tuning when using LASSO regularization; the solely positive lambda value controls the strength of the regularization by determining the amount of shrinkage of the coefficients and can range up to $+\infty$ but can commonly be restricted to range between zero and one. The lambda value represents the ratio of the penalty to the maximum penalty. A lambda value of zero represents no penalty and becomes standard linear regression, and a lambda value of one represents the maximum penalty, so a higher lambda value results in more coefficients set to zero and removed from the model. The lambda value controls the regularization, and the optimal lambda value can be identified through cross-validation. Using the optimal lambda value, the coefficients that correspond to the selected variables in the model are estimated, and only those variables with a non-zero coefficient remain in the model as those most informative or relevant. From the coefficients, the magnitude and sign indicates the importance and direction of effect. To evaluate performance metrics on new test data, apply the LASSO regularization technique on the test data, apply the method containing only the best variables on the test data to generate predictions, and compare the predictions to the observed values to compute performance metrics.

2.7.3 Support vector grid search model evaluation via hyperparameter optimization

To optimize SVR hyperparameters, model evaluation through grid searching is performed. The two hyperparameters to optimize in SVR are gamma, or the kernel width, and the cost, which is a regularization parameter to balance maximizing the margin, which is the objective of SVR, with large values of the loss function. Grid searching proceeds in multiple steps from a broad or coarse range of possible hyperparameters to a fine range of possible hyperparameters, and the number of steps and resolution of the search depends on the data set.

For example, the search range for gamma may include a sequence of 10^{-8} to 10^8 , incremented by 1, and the search range for the cost may include a sequence of 2^{-16} to 2^{16} , incremented by 1. The coarse grid search applies a single pair of hyperparameter gamma and cost values in the SVR model with the training data with k-folds cross-validation, and this continues until all gridded pairs of hyperparameter gamma and cost values are applied in the SVR model with the training data with k-folds cross-validation. Next, the resulting models from the coarse grid search are applied to explanatory variables from a validation data set to generate predicted response values. The predicted response values are compared to actual response variable values from a validation data set to generate relevant performance metrics like MAE, RMSE, and MAPE.

The more narrow range of hyperparameters gamma and cost from the best performing coarse grid models is selected for a subsequent medium grid search, encompassing a smaller incremental search. For example, the search range for gamma may now include a sequence of 10^{-6} to 10^{-3} , incremented by 0.25, and the search range for the cost may include a sequence of 2^{10} to 2^{16} , incremented by 0.25. The medium grid search proceeds the same way as the coarse grid search, and the more narrow range of hyperparameters gamma and cost from the best performing medium grid models is selected for a subsequent fine grid search, encompassing a smaller incremental search. For example, the search range for gamma may now include a sequence of 10^{-6} to 10^{-4} , incremented by 0.125, and the search range for the cost may include a sequence of 2^{10} to 2^{14} , incremented by 0.125. The fine grid search proceeds the same way as the coarse and medium grid searches, except at the conclusion of the fine grid search the hyperparameters gamma and cost from the best performing fine grid model are used for the final best SVR model calculation using the training data. The decision of best performance is based on some

performance metrics like RMSE. The final model formulation is applied to explanatory variables from an unseen test data set to generate predicted response values. The predicted response values are compared to actual response values from the unseen test data set to generate relevant performance metrics like MAE, RMSE, and MAPE.

2.7.4 Regression performance metrics

To examine model performance across regression methods, different metrics compare the residuals, or individual differences, between predicted and observed values (Tuia et al. 2011). Despite the use of regression performance metrics, they cannot convey the extent to which the formal error term contributes to the discrepancy between the observed and predicted values, and any discussion of error here refers to the residuals not the true error.

The mean absolute error (MAE) equally weights the individual sample differences in the typicality by calculating the absolute value difference in the predicted values from the observed values for each sample and averaging that absolute value difference. A smaller MAE value indicates the closer the predicted values are to the observed values, on average (Pan et al. 2017:94). MAE is calculated as

$$\frac{1}{n} \sum_{i=1}^n |o_i - p_i|$$

where n is the number of samples, o_i is the observed value and p_i is the predicted value for each sample i . The MAE value is in the same units as the response variable. Because the MAE represents the typical magnitude of the residuals, no interpretation of the direction of performance such as underperformance or over-performance is possible. To compare different model performance using MAE, different models may exhibit different scales or ranges depending on the data set and the specific modeling strategy, so comparing the MAE values to

some baseline model like linear regression may improve interpretation; nevertheless, given these considerations comparing different models with different response variables is possible.

The root mean square error (RMSE) measures the average magnitude of the error by calculating the difference in the predicted from the observed values for each sample and summing all the differences. The summed differences are squared, and the average value is computed. Finally, the square root of the average value is calculated. The RMSE more highly weights larger errors, and the RMSE will increase if outliers are present, so this metric is most useful when large errors are important to avoid. RMSE is calculated as

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (o_i - p_i)^2}$$

where n is the number of samples, o_i is the observed value and p_i is the predicted value for each sample i . A smaller RMSE value indicates a better predictive ability for the model (Pan et al. 2017). The RMSE is scale-dependent, because the units of the RMSE value are the same units as the response variable. The RMSE is the almost the same as the standard deviation of the regression, except the RMSE does not adjust for the number of explanatory variables, and the standard error of the regression does.

The RMSE and MAE can range between zero and infinity, and the RMSE and MAE are commonly applied to continuous response variables, so when applied to discrete, non-negative count data like finger ridge-counts the RMSE and MAE evaluate the performance of models predicting counts. For zero-inflated models where the excess zeros are modeled separately from the non-zero counts, RMSE and MAE may not capture the zero-inflation component.

To compare different model performance using RMSE, the models should be trained on the same data sets and tested on the same data sets, and the training and test sets should be split

the same with the same preprocessing steps. If these similar conditions were not present, the RMSE may not be directly comparable across models, because the absolute values of the errors may differ due to the data types and analytical methods rather than any differences in model performance. For example, comparing the RMSE for GLM and SVR models may reflect the different assumptions and characteristics of the modeling approaches, like how GLM assumes the response variable follows a probability function and uses a link function to relate the response variables to the explanatory variables and like how SVR does not assume any probability function and instead maximizes the margin between support vectors to find the best-fit curve for the data. If the target variable has a large scale, RMSE values will be large and thus another metric may be more appropriate.

The mean absolute percentage error (MAPE) measures the average absolute percentage difference between predicted values and observed values to provide an average magnitude of error relative to observed values, and the MAPE is most similar to the MAE in this formulation. Thus, the MAPE is not sensitive to outliers and can be considered a ratio of the residual over the actual observed values. A smaller MAPE value indicates the closer the predicted values are to the observed values. Percentage errors, in general, calculate absolute errors. MAPE is calculated as

$$\frac{1}{n} \sum_{i=1}^n \left| \frac{o_i - p_i}{o_i} \right|$$

where n is the number of samples, o_i is the observed value and p_i is the predicted value for each sample i .

Although the MAPE is beneficial regarding the ease of interpreting a percentage, the MAPE does have drawbacks and is more limited than the MAE. First, because the formula includes scaling everything by the actual observed value, if the observed value is zero the MAPE

is undefined, given by Not a Number (NaN) in R Statistical Program (R Core Team 2022). The zero value for a variable is assumed to be meaningful, so for variables that contain positive and negative values including zero as an arbitrary point, the MAPE may perform poorly. The scaling inherent in MAPE also produces model-specific metrics, so the MAPE should not be compared across models with different response variables.

Next, if the observed values are very small the MAPE can become exceptionally large. For example, a predicted value 3 and observed value 2 results in an error of 50%, but a predicted value 101 and observed value 100 results in an error of 1%, even though both scenarios include an absolute error of one. This behavior can make MAPE a useful metric when the difference between values of 3 and 2 are meaningful. Another drawback, the MAPE exhibits bias toward predictions smaller than the observed values, so that the MAPE would be lower if the prediction is lower than the observed value than if the prediction is higher by the same amount (Sefidian 2022).

CHAPTER THREE

MATERIALS AND METHODS

3.1 Data set

The worldwide dermatoglyphics data set is a combined effort of 100+ researchers who collected actual inked prints of individuals' hands and feet across the globe from 328 ethnic groups, including 57,478 individuals, between the 1930's and 1980, most before 1970. From the inked prints, the late Mr. Heinz Brehme recorded the finger pattern classification and counted the ridges for most of the individuals in the worldwide data set (Figure 5). Presumably, all ridge counting was performed following guidelines from Brodhage and Wendt (1951; Jantz, Eriksson, and Brehme 1993). Dr. Miyo Yokota counted the ridges for the Ainu Ethnic Group. In addition to his academic pursuits, the late Mr. Brehme worked for the Institute for Human Genetics and Anthropology at the University of Freiburg, Germany, to establish paternity using detailed fingerprint comparison based on the assumption of high heritability of fingerprints, on behalf of the legal system. The general finger patterns and finger ridge-counts of the German samples in the worldwide dermatoglyphics data set largely came from these paternity tests. After the death of Mr. Brehme, the Institute for Human Genetics and Anthropology transferred curation responsibility of the worldwide dermatoglyphics data set to Dr. Richard Jantz, Professor Emeritus, University of Tennessee, Knoxville. Individual researchers have published the raw data of finger patterns and finger ridge-counts as well as the results of analysis for some of the ethnic groups in this data set. However, the entirety of the worldwide dermatoglyphics data set has not yet been published or made publicly available.

Hands of individuals were printed in entirety onto a single sheet of paper by the trained investigator (Figure 6A). Feet were not considered in this dissertation so were not described here.

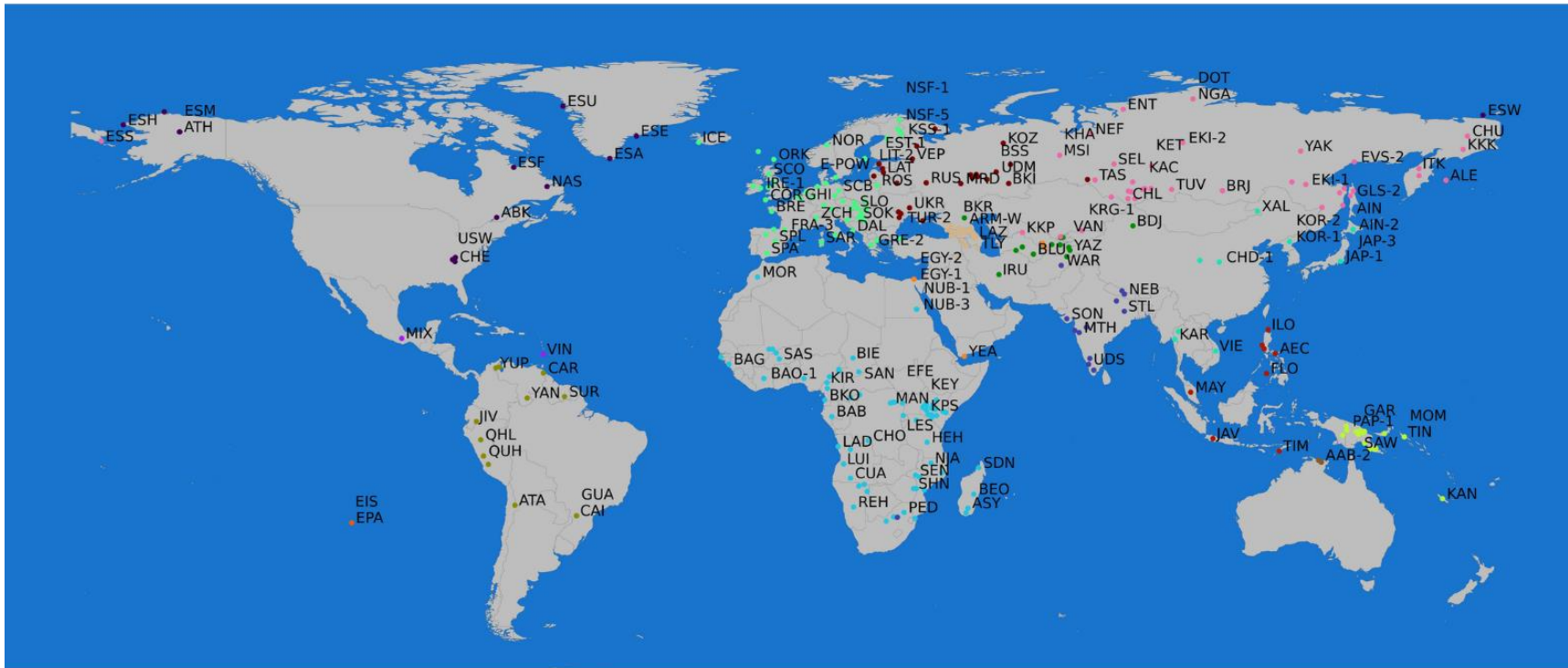


Figure 5. Map of Ethnic Group locations, colored by Region. Not all Ethnic Groups labelled here due to overlapping locations

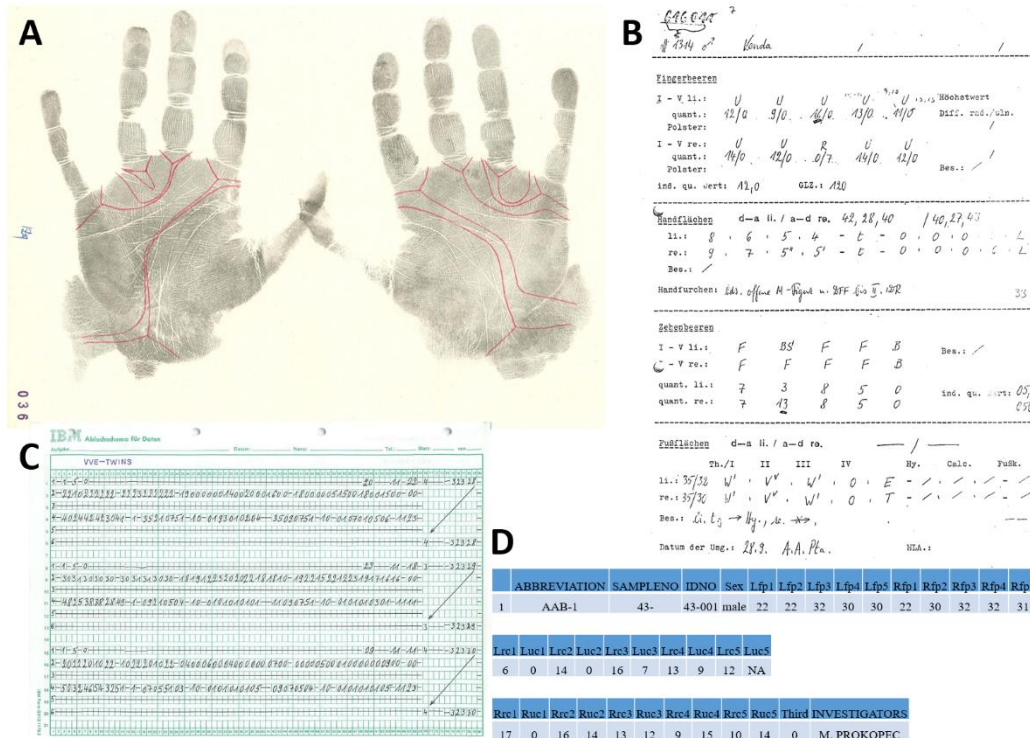


Figure 6. Process from handprint to data file. (A) Hand printed by trained investigator, (B) data recorded on data sheet, (C) numeric codes recorded on IBM card for ID number, finger patterns, and finger ridge-counts, and (D) transcription into digital data file. The illustrations of the four example steps do not depict the same individual throughout and were intended to demonstrate the structure of the data recoding steps rather than the transcription of the same data points across media.

The variables were recorded onto a data sheet (Figure 6B) and then transcribed into codes for an IBM data card (Figure 6C). Finally, the ID number, finger patterns, and finger ridge-counts were transcribed into a digital format (Figure 6D). Only two researchers, the late Mr. Heinz Brehme and Dr. Miyo Yokota, assessed the handprint image (Figure 6A) to identify finger patterns and count the ridges on fingers as recorded in the data sheets (Figure 6B), so the inter-observer error for the worldwide human dermatoglyphics data set was minimized. All physical handprints and data sheets are archived at the University of Tennessee, Knoxville, for long-term preservation.

3.1.1 Data set, original

The original data set consisted of 57478 observations of 45 attributes. Three main error types appeared in the original data set and doubtless more exist that are currently undetected, but every attempt was made to identify, fix, and report the errors in this data set. 1) Column errors in the original text file erroneously shift data values into another column, thus offsetting subsequent column values. For example, the finger pattern codes for some ID Numbers suggest no Third ridge-count where the original data set contained non-zero values for the Third ridge-count, so the errors in Third ridge-counts containing non-zero counts were changed to zero. 2) Keystroke errors resulted in incorrect values entered into the original data file. For example, finger ridge-count values unusually high (ex. 81) were identified through outlier detection and fixed based on reference to the respective finger pattern codes. So, a sequence of 01, 81, 71 was likely 0, 18, 17, 1, etc. Transposition errors resulted in incorrect values entered into the original data file. For example, a finger ridge-count value of 40 was identified as an error through outlier detection and fixed based on reference to the respective finger pattern codes. The pattern type for this digit was

20, which suggests a ridge-count of 4 or less, so the correct finger ridge-count value was likely 04.

The metadata file consisted of 508 observations of 18 attributes. Five attributes from the original data set were removed to facilitate merging: Latitude, Longitude, Phylum, Subphylum, Family, and Language. In this way, the merged data used the Latitude and Longitude columns from the metadata file. To merge, metadata and data set were matched by metadata column ABBREVIATI and data set column Abbreviation; then metadata columns were added to data set. Column classes were converted for the attributes. Any duplicate and mislabeled Sample Numbers and ID Numbers were corrected. Any rows where finger pattern columns contained 100% NA values were removed. ID Numbers can contain finger pattern codes but no ridge-counts, and these data were retained, because this situation occurs when the original print photo was not clear enough to assess ridge-count but finger patterns were clear. A problem arises at this step, because three Ethnic Groups with no finger pattern data but data for ridge-counts were excluded. So, the rows where more than 50% of the ridge-counts were present for AIN-2, GRE-2, and JAP-3 were retained. One new attribute of character class was created: Landmass. Fifteen new attributes were created, all of numeric class: total ridge-count, absolute ridge-count, index of asymmetry, index of diversity, index of diversity by Holt's definition, five attributes for Directional Asymmetry of digits 1-5, and five attributes for Fluctuating Asymmetry of digits 1-5.

For a quality check to ensure data manipulation did not compromise the association of a finger pattern code or ridge-count to an ID Number, a value of a finger pattern or a ridge-count from an arbitrary ID Number was compared from the original data file to the value at each data manipulation step. So, if the two values were not the same, the problem occurred at that specific data manipulation step, which should be fixed before moving on to the analysis of data. Where

new attributes were created, the "correct" values were calculated by-hand and then compared to the computer-generated value in the final data set to test for computational errors.

3.1.2 Data set, final

The final, cleaned data set consists of 51699 observations of 50 attributes. Attributes can exist as five types or classes: logical or boolean, integer or 32-bit integer, numeric or double precision, complex or double complex, and character or string. Data represented by named groups were considered categorical, aka nominal and character, like Landmass encompassing Africa, Americas, Asia, and Europe. Instead of representing categorical data types by single scalar values like one or five, which implies group two is closer to group one than to group five, categorical data types were represented by vectors. For example, Landmass Africa may be represented by {1, 0, 0, 0}, and Landmass Americas may be represented by {0, 1, 0, 0}. This allows groups with multiple categories to be represented without implying distance or order (Kruschke 2015:429). Specifically in R, character groups can be classified as factors, with both character and numeric types where the character is the group name, and the numeric value is the vector level representing the group (R Core Team 2022). For example, group Landmass Americas would have a character group name "Americas" and a vector level of 2 for the second group where Africa was the first group level. Categorical, aka nominal and character, and factor data are qualitative data.

Data represented solely as numbers are either interval or ratio type. Interval data have equal distances between values but no real or true zero. A real or true zero indicates a complete absence of some variable. For example but not in this dissertation, Celsius and Fahrenheit temperature scales represent interval data, because they have equal distances between values but no true zero, while the Kelvin temperature scale represents ratio data, because there is an equal

distance between values as well as a true zero. Further, Latitude and Longitude variables do have equal distances between values but no true zero, so are interval data. Directional Asymmetry variables were considered numeric ratio data, because they have equal distances between values, a true zero of no asymmetry, and the ability to calculate meaningful ratios. For Directional Asymmetry, the positive or negative sign merely indicates the direction of asymmetry, whether the left or right side has the greater number of ridges, and the positive or negative sign does not invalidate the calculation of meaningful ratios between values. Fluctuating Asymmetry variables, represented by the absolute value of Directional Asymmetry, were considered numeric ratio count data with equal distances between values, a true zero of no asymmetry, and all values were positive discrete integer ridge-counts. The indices of asymmetry and diversity variables represent numeric continuous ratio data types, because there were equal distances between values, they had a true zero value, and all values were continuous. Numeric, interval and ratio, discrete count and continuous data are quantitative data.

The original attributes were overwritten with the new attributes of new class (Table 2, Table S2). Phylum, Subphylum, Family, and Language are all empty columns and were not considered further in this dissertation. The ten digits per individual were referenced by values one through five on either the left or the right hand. Digit 1 indicated the thumb; digit 2 indicated the pointer finger; digit 3 indicated the middle finger; digit 4 indicated the ring finger; and digit 5 indicated the little or pinky finger.

If the attributes were part of this dissertation as response or explanatory variables, their description is found in the respective following section, and if the attributes were not response or explanatory variables in this dissertation, their description immediately follows.

Table 2. New classes and descriptions for attributes from worldwide dermatoglyphics data set.

Attribute	New Attribute Class	Description
Abbreviation	character	unique abbreviation of Ethnic Group
Sample Number	character	numbers, symbols, and spaces to indicate Ethnic Group
Ethnic Group	character	grouping name
Region	character	grouping containing multiple Ethnic Groups
Investigator Location	character	location where investigator obtained prints
ID Number	character	Sample Number followed by a sequential number
Sex	character	female or male
Lfp1	factor	left finger pattern on digit 1 (thumb)
Lfp2	factor	left finger pattern on digit 2 (pointer)
Lfp3	factor	left finger pattern on digit 3 (middle)
Lfp4	factor	left finger pattern on digit 4 (ring)
Lfp5	factor	left finger pattern on digit 5 (little/pinky)
Rfp1	factor	right finger pattern on digit 1 (thumb)
Rfp2	factor	right finger pattern on digit 2 (pointer)
Rfp3	factor	right finger pattern on digit 3 (middle)
Rfp4	factor	right finger pattern on digit 4 (ring)
Rfp5	factor	right finger pattern on digit 5 (little/pinky)
Lrc1	numeric	left radial ridge-count on digit 1 (thumb)
Luc1	numeric	left ulnar ridge-count on digit 1 (thumb)
Lrc2	numeric	left radial ridge-count on digit 2 (pointer)
Luc2	numeric	left ulnar ridge-count on digit 2 (pointer)
Lrc3	numeric	left radial ridge-count on digit 3 (middle)
Luc3	numeric	left ulnar ridge-count on digit 3 (middle)
Lrc4	numeric	left radial ridge-count on digit 4 (ring)
Luc4	numeric	left ulnar ridge-count on digit 4 (ring)
Lrc5	numeric	left radial ridge-count on digit 5 (little/pinky)
Luc5	numeric	left ulnar ridge-count on digit 5 (little/pinky)
Rrc1	numeric	right radial ridge-count on digit 1 (thumb)
Ruc1	numeric	right ulnar ridge-count on digit 1 (thumb)
Rrc2	numeric	right radial ridge-count on digit 2 (pointer)
Ruc2	numeric	right ulnar ridge-count on digit 2 (pointer)
Rrc3	numeric	right radial ridge-count on digit 3 (middle)
Ruc3	numeric	right ulnar ridge-count on digit 3 (middle)
Rrc4	numeric	right radial ridge-count on digit 4 (ring)
Ruc4	numeric	right ulnar ridge-count on digit 4 (ring)
Rrc5	numeric	right radial ridge-count on digit 5 (little/pinky)
Ruc5	numeric	right ulnar ridge-count on digit 5 (little/pinky)
Third	numeric	ridge-count for a third triradius

Table 2 Continued

Attribute	New Attribute Class	Description
Investigators	character	named researcher who collected the original prints
Sample Declaration	character	type of group
City for Lat/Long	character	city and country associated with updated Latitude and Longitude
Latitude	character	decimal degrees
Longitude	character	decimal degrees
Phylum	character	Language phylum
Subphylum	character	Language subphylum
Family	character	Language family
Language	character	name of Language
Landmass	character	grouping not spanning oceans
TRC	numeric	sum of largest ridge-counts on each digit
ARC	numeric	sum of all ridge-counts on all digits

Abbreviation is a string of characters that is a unique identifier of each group. For example, Ethnic Group Germans of Sample Number 335 has Abbreviation E-DU (for European Dusseldorf), Sample Number 286 has Abbreviation E-KL (for European Kiel), and Sample Number 2 has an Abbreviation E-LZ (for European Leipzig). In the data set, 373 unique Abbreviations were present. Dr. Miyo Yokota performed the finger ridge-counts of these Abbreviations (Sample Number): AIN-2 (385), BUS-2 (387), GKU (390), JAP-3 (386), KWA (388), MBK (392), MBU (389), and SYU (391), for these eight groups the Third ridge-count was not recorded for most individuals (Yokota 1997).

Sample Number is a string of characters (numbers, symbols, and spaces) that is not a unique identifier. For example, Sample Number -90 refers to Ethnic Group British with Abbreviation ELD, and ID Number -90229 refers to a male with specific finger patterns and ridge-counts but also ID Number -90229 refers to a female with different finger patterns and ridge-counts. Multiple Sample Numbers contain duplicated ID Numbers, where one duplicated ID Number is a female and the other duplicated ID Number is a male: -59, 31-, -90, 135, -53, -46, 1-1, 1-2, 103, 104, 105, 44-, -43, -84, -45. So, even though some of the ID Numbers within these Sample Numbers were duplicated, the information is not redundant. In data, 367 unique Sample Numbers are present.

Investigator Location is a string of characters representing the location where the investigator obtained the prints. This may be a major city and may not represent the homeland of the group.

ID Number is a string of characters that is not a unique identifier of each individual. The beginning of the ID Number is always the Sample Number followed by a sequential number. For example, ID Number -01233 indicates the 233rd individual of Sample Number -01. In data, 745

ID Numbers are appear twice, once for a male, once for a female, and are listed in *Appendix B Supplementary information and tables, Chapter Three*.

Finger patterns are of factor class indicating the pattern type and side of the digit. Pattern codes include 10, 11, 12, 13, 15, 16, 19, 20, 21, 22, 23, 24, 25, 26, 27, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 45, 46, 47, and NA. Pattern codes beginning with number 1 indicated arches, and the second number of the pattern code indicated a type of arch finger pattern. Pattern codes beginning with number 2 indicated loops, like radial and ulnar loops. Pattern codes beginning with number 3 indicated whorls and their variations, and pattern codes beginning with number 4 indicated complex and uncommon patterns.

Investigators is a string of characters indicating the name of the researcher who collected the original prints onto paper in the field (Figure 6A). Investigators can take a variety of forms: first-initial full-last-name (e.g. W Lehmann); first-initial middle-initial full-last-name (e.g. HJ Hammer); full-first-name middle-initial full-last-name (e.g. Derek F Roberts); full-first-name full-last-name (e.g. Heinz Brehme); and full-first-name full-middle-name full-last-name (e.g. Lidiya Ivanovna Tegako). Multiple authors are separated by a semicolon. "Et al." indicates additional unnamed investigators collected the prints. "No Name" indicates no known information about the investigators. The approximately 116 named Investigators are listed in *Appendix B Supplementary information and tables, Chapter Three* in plain text in alphabetical order by last name.

Sample Declaration is a string of characters representing the type of group, like SCHCH (schoolchildren) and VOLS (volunteers). In data, 31 unique Sample Declarations are present, and many unique Sample Declarations are merely different spellings or abbreviations of the

same kind of group. For example, at least 14 Sample Declaration's refer to students and/or schoolchildren.

City for Lat/Long is a string of characters similar to Investigator Location representing the city and country associated with the updated Latitude and Longitude coordinates. This may be a city that does not represent the broad homeland of the group. This variable was arbitrarily set by the researcher based on Ethnic Group history and Investigator Location to provide a single geographic location via designation of Latitude and Longitude for each Ethnic Group.

3.1.2.1 Response variables, dermatoglyphic features

The response variables were discussed in order of their place in the worldwide dermatoglyphics data set, outlined in *3.1.2 Data set, final* by their attribute type. In this way, the raw finger ridge-counts were discussed before summary variables like TRC and ARC and derived variables like Directional Asymmetry, Fluctuating Asymmetry, and indices of asymmetry and diversity. Directional Asymmetry and Fluctuating Asymmetry variables were capitalized in the text when reflecting the specific formulation in this dissertation, rather than some universal concept. For evaluating performance metrics of regression models containing these response variables, specific characteristics of the response variables can affect the application and interpretation of the performance metrics. So, a common list of characteristics was included here: presence of outliers; discrete, non-negative counts; large values; zero values; and arbitrary zeros.

Finger ridge-counts are of numeric class indicating the number of ridges from the triradius (delta) to the core of the print pattern (Table 3). Third is of numeric class indicating the third ridge-count when an individual exhibits a third (other than radial and ulnar) triradius (Table S1). For an individual to have a Third ridge-count, the pattern for that digit must be a loop or

Table 3. Response variables and their variable type. Variables in alphabetical order

Response variable	Variable type	Qualitative or quantitative
ARC	discrete count	quantitative
Directional Asymmetry: DA1, DA2, DA3, DA4, DA5	discrete ratio	quantitative
Finger ridge-counts: Lrc1, Luc1, Lrc2, Luc2, Lrc3, Luc3, Lrc4, Luc4, Lrc5, Luc5, Rrc1, Ruc1, Rrc2, Ruc2, Rrc3, Ruc3, Rrc4, Ruc4, Rrc5, Ruc5, Third	discrete count	quantitative
Fluctuating Asymmetry: FA1, FA2, FA3, FA4, FA5	discrete count	quantitative
Index.Asymmetry	continuous ratio	quantitative
Index.Diversity	continuous ratio	quantitative
Index.Diversity.Holt	continuous ratio	quantitative
TRC	discrete count	quantitative

whorl. Individuals do not commonly have Third ridge-counts so were converted from NA to 0 so can be used in analyses that require no missing data.

Finger ridge-counts contained many outliers, particularly for the ulnar ridge-counts where zero counts dominate, so any deviance by a positive count can represent an outlier (Figure 7). Finger ridge-counts were in units of discrete, non-negative counts. The finger ridge-counts did not encompass large values, with a maximum of 53 counts, and the finger ridge-counts did contain zeros but no arbitrary zeros due to the nature of counts (Table 4).

TRC is of numeric class of the total ridge-count, which was calculated by taking the largest radial or ulnar ridge-count for each of the 10 digits and summing that count (Table 3). So, each individual may have 10 ridge-count values contributing to the TRC. Individuals with any finger ridge-count containing an NA will have NA for TRC, as well.

Total ridge-count contains many outliers and was in units of discrete, non-negative counts (Figure 8). The total ridge-counts did encompass large values, with minimum of 0 and maximum of 320, within the possible upper limit of 350 ridges hypothesized by Jantz (1977). Total ridge-count values did contain zeros but no arbitrary zeros due to the nature of counts (Table 4).

ARC is of numeric class of the absolute ridge-count, which was calculated by summing both the radial and ulnar ridge-count for each of the 10 digits (Table 3). So, each individual may have up to 20 ridge-count values contributing to the ARC. Individuals with any finger ridge-count containing an NA will have NA for ARC, as well.

Absolute ridge-count contained many outliers, relegated to the higher values in the range, and was in units of discrete, non-negative counts (Figure 9). The absolute ridge-counts did

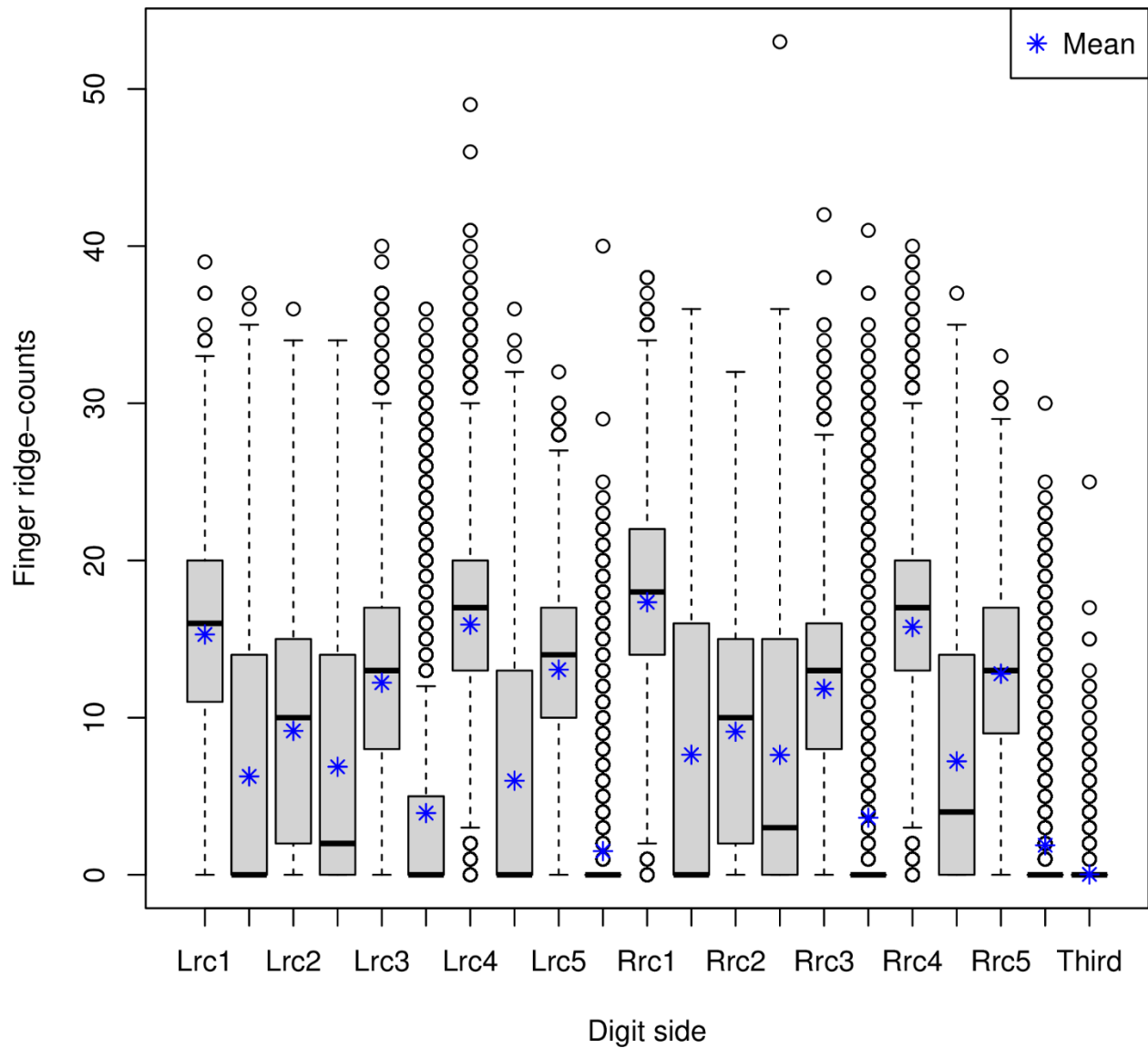


Figure 7. Boxplots of finger ridge-counts separated by digit and side. Luc for digits 1-5 and Ruc for digits 1-5 are indicated by the unlabeled hash marks on the x-axis. Open circles are outliers, horizontal lines equal maximum ($Q3 + 1.5 \cdot IQR$) and minimum ($Q1 - 1.5 \cdot IQR$) values, and bold horizontal lines are the median.

Table 4. Performance metric evaluation of TRC, ARC, and finger ridge-counts

Response variable	Outliers	Discrete, non-negative counts	Large values	Zeros	Arbitrary zeros	Best performance metric(s)
	yes MAE, no RMSE, yes MAPE	no MAE, no RMSE	no RMSE	undefined MAPE	no MAPE	
TRC	✓	✓	✓	✓		MAPE
ARC	✓	✓	✓	✓		MAPE
Lrc1	✓	✓		✓		MAPE
Luc1	✓	✓		✓		MAPE
Lrc2	✓	✓		✓		MAPE
Luc2		✓		✓		MAPE
Lrc3	✓	✓		✓		MAPE
Luc3	✓	✓		✓		MAPE
Lrc4	✓	✓		✓		MAPE
Luc4	✓	✓		✓		MAPE
Lrc5	✓	✓		✓		MAPE
Luc5	✓	✓		✓		MAPE
Rrc1	✓	✓		✓		MAPE
Ruc1		✓		✓		MAPE
Rrc2		✓		✓		MAPE
Ruc2		✓		✓		MAPE
Rrc3	✓	✓		✓		MAPE
Ruc3	✓	✓		✓		MAPE
Rrc4	✓	✓		✓		MAPE
Ruc4	✓	✓		✓		MAPE
Rrc5	✓	✓		✓		MAPE
Ruc5	✓	✓		✓		MAPE
Third	✓	✓		✓		MAPE

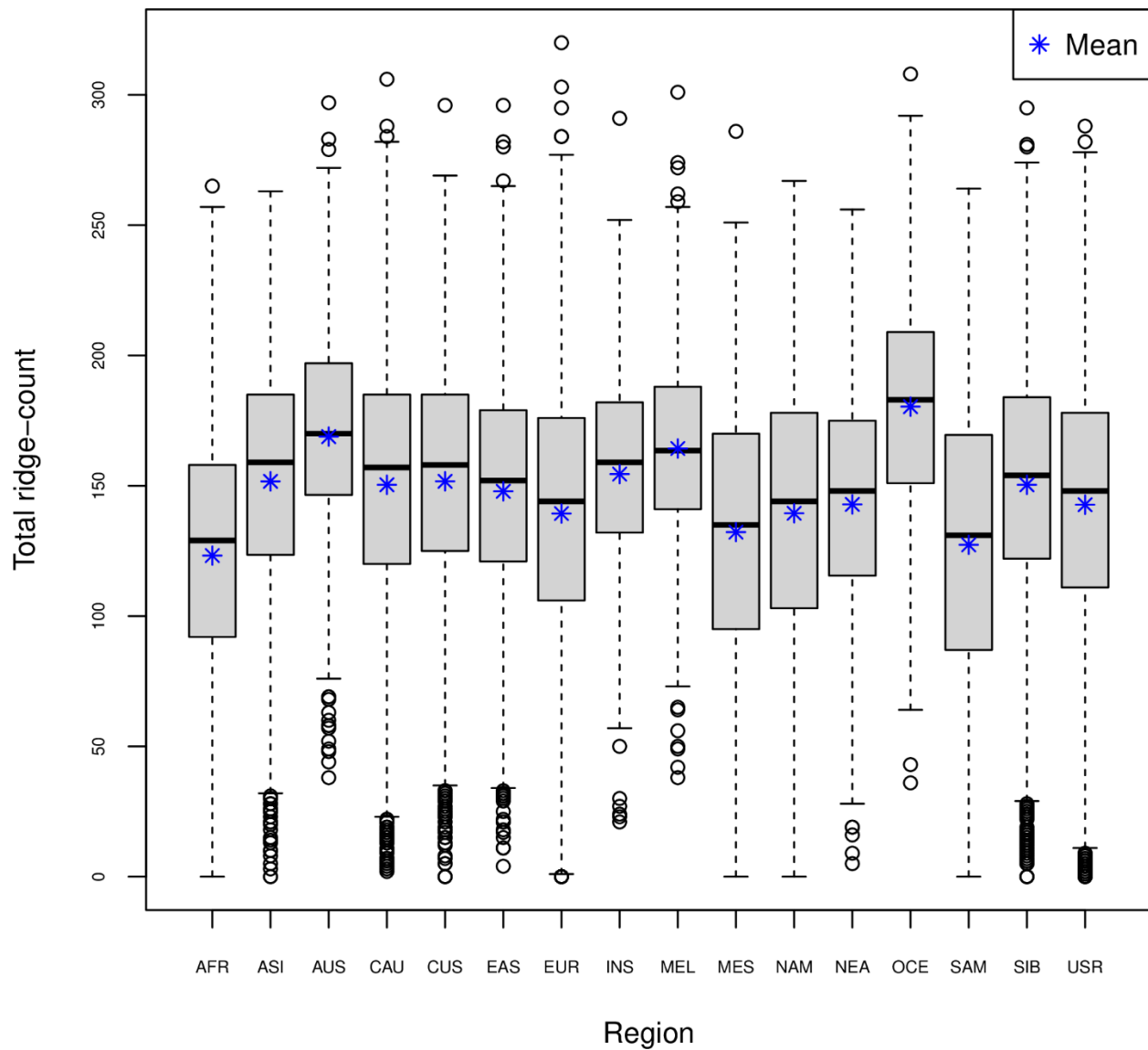


Figure 8. Boxplots of total ridge-count separated by Region. Open circles are outliers, horizontal lines equal maximum ($Q3 + 1.5 \cdot IQR$) and minimum ($Q1 - 1.5 \cdot IQR$) values, and bold horizontal lines are the median.

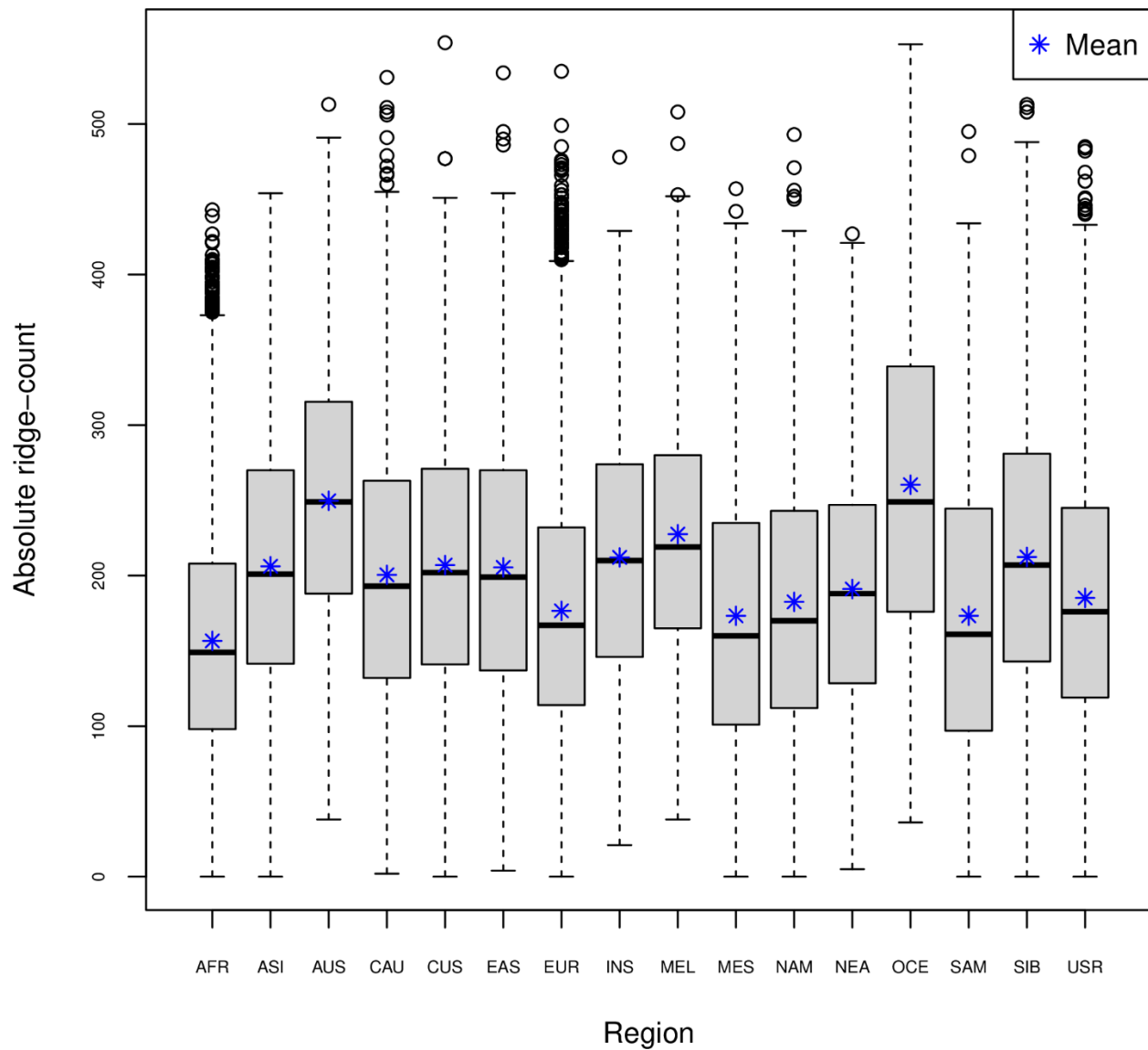


Figure 9. Boxplots of absolute ridge-count separated by Region. Open circles are outliers, horizontal lines equal maximum ($Q3 + 1.5 \cdot IQR$) and minimum ($Q1 - 1.5 \cdot IQR$) values, and bold horizontal lines are the median.

encompass large values, with a minimum of 0 and maximum of 554. Absolute ridge-count values did contain zeros but no arbitrary zeros due to the nature of counts (Table 4).

Directional Asymmetry variables for each digit were of numeric class of the directional asymmetry, which was calculated for each individual for each of the five digits by the subtracting the larger of the radial or ulnar ridge-count on the right side from the larger of the radial or ulnar ridge-count on the left side (Table 3). Individuals with any finger ridge-count containing an NA for each digit will have an NA for the respective Directional Asymmetry variable, as well. Directional Asymmetry indicated the direction of asymmetry by retaining a positive or negative sign:

$$d_k = \text{left}_k - \text{right}_k$$

where d is the difference between homologous k digits.

Directional Asymmetry variables for each digit contain many outliers and were in units of negative and positive integers (Figure 10). Directional Asymmetry does not encompass large values, with a minimum of -45 and a maximum of 33. The Direction Asymmetry variables do not contain zeros that are arbitrary, because zero does represent no asymmetry, and the negative or positive Directional Asymmetry values additionally indicate greater ridge-counts on the right or left side for that digit (Table 5).

Fluctuating Asymmetry variables for each digit were of numeric class of the fluctuating asymmetry, which was calculated for each individual for each of the five digits by the absolute value of the Directional Asymmetry: $abs(d_k)$ (Table 3). Individuals with any finger ridge-count containing an NA for each digit will have an NA for the respective Fluctuating Asymmetry variable, as well.

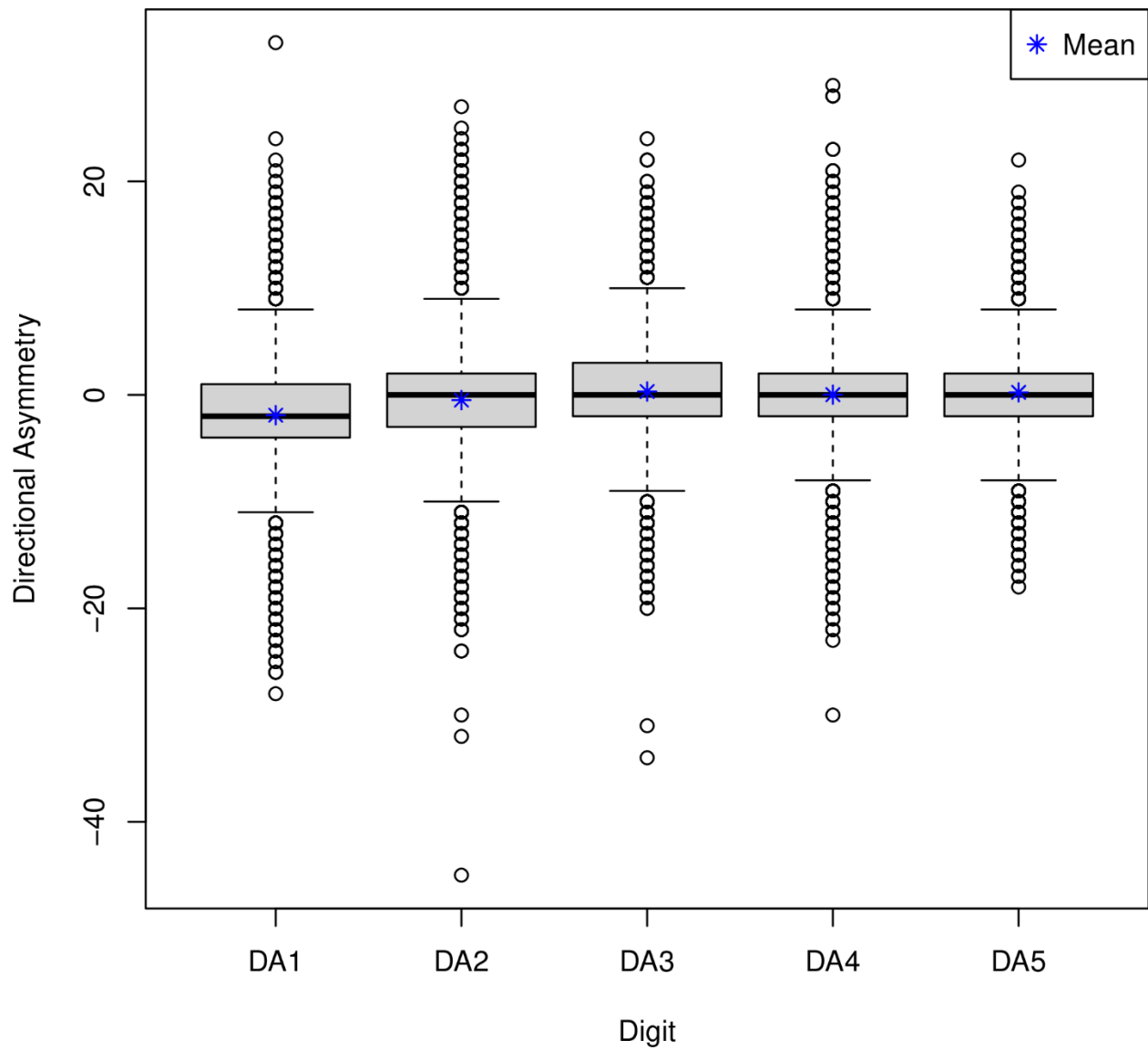


Figure 10. Boxplots of Directional Asymmetry variables separated by digit in Data. Open circles are outliers, horizontal lines equal maximum ($Q3 + 1.5 \cdot IQR$) and minimum ($Q1 - 1.5 \cdot IQR$) values, and bold horizontal lines are the median.

Table 5. Performance metric evaluation of Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry and diversity

Response variable	Outliers	Discrete, non-negative counts	Large values	Zeros	Arbitrary zeros	Best performance metric(s)
	yes MAE, no RMSE, yes MAPE	no MAE, no RMSE	no RMSE	undefined MAPE	no MAPE	
DA1	✓			✓	✓	MAE
DA2	✓			✓	✓	MAE
DA3	✓			✓	✓	MAE
DA4	✓			✓	✓	MAE
DA5	✓			✓	✓	MAE
FA1	✓			✓		MAE, MAPE
FA2	✓			✓		MAE, MAPE
FA3	✓			✓		MAE, MAPE
FA4	✓			✓		MAE, MAPE
FA5	✓			✓		MAE, MAPE
Index.Asymmetry	✓			✓		MAE, MAPE
Index.Diversity	✓			✓		MAE, MAPE
Index.Diversity.Holt	✓			✓		MAE, MAPE

Fluctuating Asymmetry variables for each digit contain many outliers, relegated to the higher values in the range, and were in units of positive integer counts (Figure 11). Fluctuating Asymmetry does not encompass large values, with a maximum of 45. The Fluctuating Asymmetry variables do contain zeros but no arbitrary zeros due to the absolute value function (Table 5).

Index.Asymmetry is of numeric class of the Index of Asymmetry, which accesses the ridge-count asymmetry between the right and left hands on homologous digits (Table 3). Index.Asymmetry was calculated by selecting the larger of the radial or ulnar ridge-count on the right for each digit, and selecting the larger of the radial or ulnar ridge-count on the left for each digit; this step yielded 10 values. Then, subtract the left larger ridge-count for each digit from the right larger ridge-count for the same digit; this step yielded five values. Next, square the difference, and sum all five squared differences, which yielded a single value. Finally, the square root was taken of the sum of squared differences, which yielded a single value for the Index of Asymmetry:

$$\sqrt{A^2} = \sqrt{\sum_{i=1}^5 (R_i - L_i)^2}.$$

Individuals with any finger ridge-count containing an NA will have an NA for Index.Asymmetry, as well.

Index.Asymmetry contains many outliers, relegated to the higher values in the range, and was in units of positive continuous ratios (Figure 12). The Index.Asymmetry does not encompass large values, with a maximum of 43.0465, and Index.Asymmetry does contain zeros but no arbitrary zeros (Table 5).

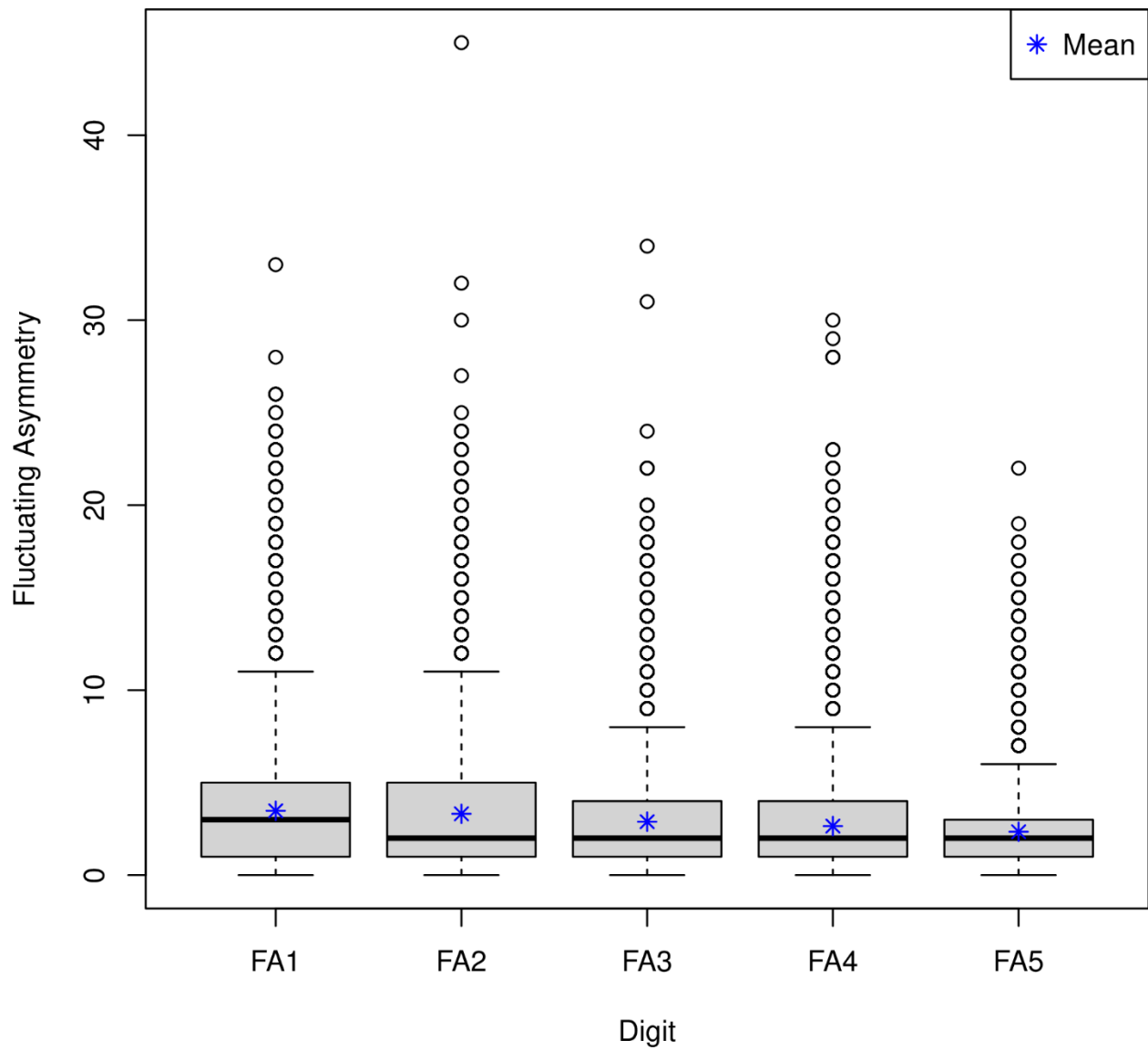


Figure 11. Boxplots of Fluctuating Asymmetry variables separated by digit in Data. Open circles are outliers, horizontal lines equal maximum ($Q3 + 1.5 \cdot IQR$) and minimum ($Q1 - 1.5 \cdot IQR$) values, and bold horizontal lines are the median.

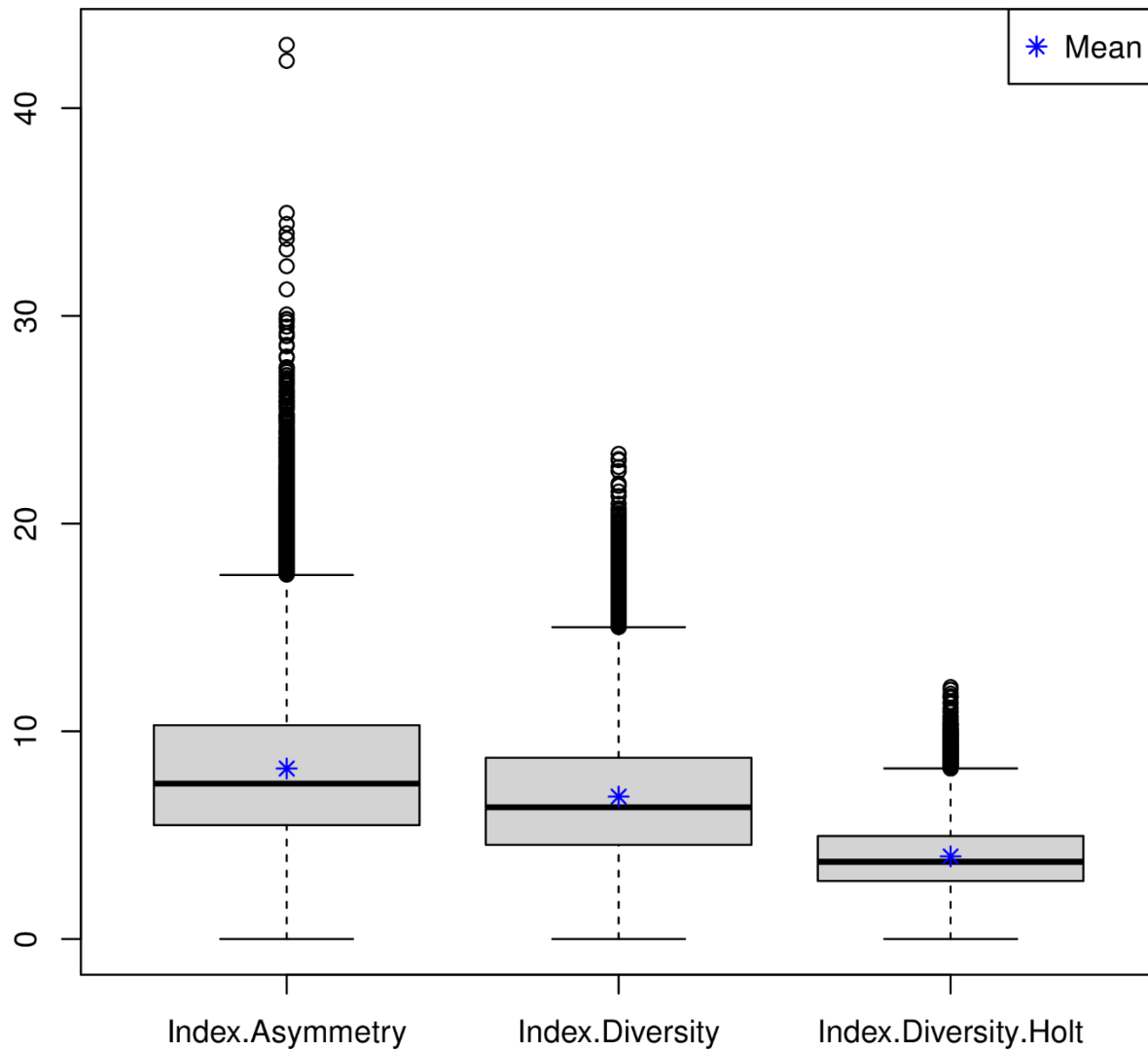


Figure 12. Boxplots of Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt in Data. Open circles are outliers, horizontal lines equal maximum ($Q3 + 1.5 \cdot IQR$) and minimum ($Q1 - 1.5 \cdot IQR$) values, and bold horizontal lines are the median.

Index.Diversity is of numeric class of the Index of Diversity, which is the sum of the ridge-counts of homologous digits (Table 3) (Jantz 1975). The goal of this index was to quantify the differences between non-homologous digits. Index.Diversity is calculated by selecting the larger of the radial or ulnar ridge-count on the right for each digit, and selecting the larger of the radial or ulnar ridge-count on the left for each digit. Next, sum the right larger ridge-count for each digit with the left larger ridge-count for the same digit, which yielded five values. Then, square the right plus left larger ridge-count for each digit, and sum those squared values to yield a single value. Separately, square the TRC for each individual, and divide that single value by 5. Further, subtract the 'squared TRC divided by 5' from the sum of squared digit ridge-counts, and that value is designated S. Finally, the square root of 'S divided by 5' yielded a single value for the Index of Diversity:

$$\frac{S}{\sqrt{5}} = \sqrt{\frac{\sum_{i=1}^5 q_i^2 - \frac{Q^2}{5}}{5}}$$

where q_i is the ridge-count for the i -th digit, and $Q = \sum q_i$, or the TRC. Individuals with any finger ridge-count containing an NA will have an NA for Index.Diversity, as well.

Index.Diversity contains many outliers, relegated to the higher values in the range, and was in units of positive continuous ratios (Figure 12). The Index.Diversity does not encompass large values, with a maximum of 23.3581, and Index.Diversity does contain zeros but no arbitrary zeros (Table 5).

Index.Diversity.Holt is of numeric class of the Index of Diversity from Holt (Table 3) (Holt 1960; Jantz 1975). Index.Diversity.Holt is calculated by selecting the larger of the radial or ulnar ridge-count for each digit, which yielded 10 values. Next, square the larger ridge-count for each digit, and sum those 10 squared values to yield a single value. Separately, square the TRC

for each individual, and divide that single value by 10. Further, subtract the 'squared TRC divided by 10' from the sum of squared digit ridge-counts, and that value was designated S. Finally, the square root of 'S divided by 10' yielded a single value for the Index of Diversity from Holt:

$$\frac{S}{\sqrt{10}} = \sqrt{\frac{\sum_{i=1}^{10} q_i^2 - \frac{Q^2}{10}}{10}}$$

where q_i is the ridge-count for the i -th digit, and $Q = \sum q_i$, or the TRC. Individuals with any finger ridge-count containing an NA will have an NA for Index.Diversity.Holt, as well.

Index.Diversity.Holt contains many outliers, relegated to the higher values in the range, and was in units of positive continuous ratios (Figure 12). The Index.Diversity.Holt does not encompass large values, with a maximum of 12.1264, and Index.Diversity.Holt does contain zeros but no arbitrary zeros (Table 5).

3.1.2.2 Explanatory variables

The Explanatory variables were discussed in order of their place in the worldwide dermatoglyphics data set, outlined in 3.1.2 *Data set, final* by their attribute type. Thus, the explanatory variables were not discussed in alphabetical order as they were found in Table 6.

Ethnic Group is a string of characters representing the ethnic group that is not unique. For example, Ethnic Group Germans refers to Sample Number 335, 286, 2, -01, and -4. In data, 328 unique Ethnic Groups are present (Figure 5). Thirty-five Ethnic Groups contained finger patterns but not finger ridge-counts and are listed in *Appendix B Supplementary information and tables, Chapter Three*.

The original Ethnic Group designations reflected the existing geopolitical climate of the time the handprints were recorded. Because populations change over time, many of the ethnic

Table 6. Explanatory variables and their variable type. Variables in alphabetical order;
qualitative in italics

Explanatory variable	Variable type	Qualitative or quantitative
Ethnic Group	character, categorical	<i>qualitative</i>
k-means cluster: DA.kmeans.2, DA.kmeans.4, DA.kmeans.10, DA.kmeans.322, FA.kmeans.2, FA.kmeans.4, FA.kmeans.8, FA.kmeans.322, frc.kmeans.2, frc.kmeans.4, frc.kmeans.10, frc.kmeans.16, frc.kmeans.322	categorical, nominal, factor	<i>qualitative</i>
Landmass	character, categorical	<i>qualitative</i>
Latitude	numeric, interval	quantitative
Longitude	numeric, interval	quantitative
Moran's Eigenvector Maps	numeric, continuous ratio	quantitative
Region	character, categorical	<i>qualitative</i>
Sex	character, categorical, factor	<i>qualitative</i>

groups in this worldwide dermatoglyphics data set do not exist in the same form today through assimilation or eradication. Further, the original Ethnic Group designations were arbitrary social constructs often filtered through the lens of the investigator collecting the handprints, and for many reasons some Ethnic Group names were changed, so only the updated Ethnic Group names were included in this dissertation, but the association of the updated Ethnic Group names to the original Ethnic Group names to the original data were maintained in the data files. The updated Ethnic Group names consist of the group endonym (autonym), or the name used inside a particular group to designate themselves, their homeland, or their language, if possible. If not possible, the updated Ethnic Group names consist of the group exonym (xenonym), the name used by those outside the particular group to designate the people (ethnonym), the place names or nationality (toponym), or the predominate language (glossonym). Every attempt was made to maintain searchability of the groups using other resources through a common name or spelling as well as to respect the group by removing clearly derogatory group names. As a result, the updated Ethnic Groups were not truly ethnic groups and may refer to a group by its toponym, because no other suitable designator was found, for example. Ultimately, the Ethnic Group was arbitrarily set by me to provide a local designation for each group and may not reflect truly local understanding of groupings.

Region is a string of characters that is not unique. Regions include Africa (AFR), Asia (ASI), Australia (AUS), the Caucasus (CAU), Central Soviet Union (CUS), East Asia (EAS), Europe (EUR), Indonesia (INS), Melanesia (MEL), Mesoamerica (MES), North America (NAM), Northeastern Africa (NEA), Oceania (OCE), South America (SAM), Siberia (SIB), and the Soviet Union (USR). In data, 16 unique Regions were present (Figure 5).

Sex is a string of characters, either female or male.

Latitude is a string of characters representing the latitude of the group in decimal North or South to the tenth's place.

Longitude is a string of characters representing the longitude of the group in decimal West or East to the tenth's place. Initial latitude and longitude assignments in the original data set included whole numbers with no decimal precision based on original researcher use of printed atlas and so encompassed large areas. The limits of precision for the initial geographic locations sometimes placed an Ethnic Group's location primarily in the middle of a body of water with little nearby land. Updated Latitude and Longitude variables were arbitrary assignments by me based on the Ethnic group history and Investigator Location to assign a single geographic location to each Ethnic Group with greater precision while maintaining an area realistic to population spread.

Landmass is a string of characters that is not unique. Landmasses include Africa, Americas, Asia, and Europe. In the worldwide dermatoglyphics data set, four distinct Landmasses are present. Landmass Africa contained two Regions: Africa (AFR) and Northeastern Africa (NEA). Landmass Americas contained three Regions: Mesoamerica (MES), North America (NAM), and South America (SAM). Landmass Asia contained nine Regions: Asia (ASI), Australia (AUS), the Caucasus (CAU), Central Soviet Union (CUS), East Asia (EAS), Indonesia (INS), Melanesia (MEL), Oceania (OCE), and Siberia (SIB). Landmass Europe contained two Regions: Europe (EUR) and the Soviet Union (USR). The Landmass grouping was arbitrarily set by me to provide a large grouping that did not span oceans for spatial analysis. As such, the Landmasses are not the same as true continents and were created by dividing the existing Regions into larger groups that did not cross major bodies of water.

The k-means cluster and Moran's Eigenvector Maps explanatory variables resulted from the k-means clustering and the spatial analysis via distance-based Moran's eigenvector maps (dbMEM) (Table 6). The k-means cluster objects were specific to the response variable type and to the set optimal K . For example, explanatory variable DA.kmeans.2 included the cluster assignments for all individuals from k-means clustering including only Directional Asymmetry variables for all digits and an optimal $K = 2$; so all individuals within this explanatory variable exhibit a cluster assignment of 1 or 2. FA.kmeans.2 included Fluctuating Asymmetry variables for all digits as the response variables, and frc.kmeans.2 included the finger ridge-count variables as the response variables.

The Moran's Eigenvector Maps explanatory variables resulted from the spatial analysis performed within Landmasses so remain relevant at this level. Each Landmass exhibited a different number of significant Moran's Eigenvector Maps variables. For example, the spatial analysis of finger ridge-count response variables within Landmass Africa resulted in significant Moran's Eigenvector Maps 3, 7, 8, 9, 13, and 15, for which all individuals exhibited a specific value for Moran's Eigenvector Map 3, for Moran's Eigenvector Map 7, etc. Further, the spatial analysis of finger ridge-count response variables within Landmass Americas resulted in significant Moran's Eigenvector Maps 2, 3, 4, 5, and 6. Of note, the number signifier does not indicate a set meaning, so Moran's Eigenvector Map 3 for Landmass Africa and for Landmass Americas does not mean the same thing; Moran's Eigenvector Map 3 is just the third Moran's Eigenvector Maps that was specific to that Landmass grouping and those response variables.

3.2 Ethics, privacy, and data management

The scope of this dissertation was pre-approved by the Human Research Protection Program at the University of Tennessee, Knoxville, with their conclusion the project does not

meet the definition of private identifiable information so was excluded from requiring review by the institutional research board at the University of Tennessee, Knoxville.

While the descendants of the individuals sampled between the 1930's and 1980, most before 1970, for the worldwide dermatoglyphics data set inherit responsibility as stakeholders of the data, the resulting research, and the findings, this dissertation and other publications arising from this data set will provide open access to generate additional meaningful research questions (Huttenhower, Finn, and McHardy 2023). The level of informed consent for individuals to originally participate in the sampling was unknown due to the lengthy and collaborative nature of the data collection. However, collecting hand and foot prints is relatively non-invasive, and no additional interviews to elicit private identifiable information took place, to the knowledge of the current researcher. The question of informed consent may arise in the general finger patterns and finger ridge-counts of the German samples generated from legally compelled paternity tests, but this concern is isolated to this group and can be excluded from the worldwide dermatoglyphics data set without great effect.

With any research in human biological variation, the potential for misuse and abuse arises toward research perpetuating harmful stereotypes. In this dissertation, the groupings like Ethnic Group were based on the original data collection strategies and may not represent any true biological population. Further, an explicit goal of this work was to explore the appropriate level of population study, so additional groupings were arbitrarily created that also may not represent any true biological population. The methodology was structured to elucidate patterns of variation in dermatoglyphic features and not to classify individuals into the arbitrary groupings created during this work. Any attempts to extrapolate the research in the appropriate level of population

study to support racial classification are misguided and do not reflect sound application of dermatoglyphic data to understanding patterns in human biological variation.

To be identifiable, dermatoglyphic data must be highly discriminatory and linked with a known identity. In terms of individual privacy, no identifiable information like names of individuals in the populations was included in this dissertation. Dermatoglyphic traits contain three levels of detail, such that Level 1 detail encompasses finger patterns and finger ridge-counts that do not contain discriminatory information enough for a positive identification, whereas Level 2 detail of the shape, direction, and orientation of traits forming friction ridges and Level 3 detail of the individual friction ridge structure contains discriminatory information commonly used by forensic fingerprint examiners (Fournier and Ross 2016). Thus, finger ridge-counts and any summary or derived variables are not highly discriminatory. Further, even if an individual seeks to discover potential inclusion in the data set, human variation patterns exhibit overlap across populations and individuals so no one individual can exclusively be identified (Oh et al. 2021).

The goals of data management were to reproducibly document the analyses in this dissertation from the original data to the presented figures and results for subsequent open access of the data and R code (Alpaslan-Roodenberg et al. 2021; R Core Team 2022; Kowal et al. 2023). The original data files were maintained in separate folders not used for data analysis and were not modified. Only copies of the original data files were used as working data files for data analysis and were renamed so as not to confuse the original with the working file. The data files were saved as comma-delimited files (.csv) and read into R as .csv files without opening and modifying in Microsoft Excel, which erroneously modified content and shifted column information. All data cleaning and analyses occurred in R to make these steps reproducible, so

any bias as a result of removing rows (individuals) can be recreated and explored, and both the original, raw data and the cleaned data together are publicly accessible with digital object identifier (DOI) for persistent reference. The computing environment including program versions, package versions, etc., in which the analyses occur, was preserved in *3.3.1 Computing environment*. Only the finger ridge-counts as dermatoglyphic data were originally available, so any summary or derived variables were also calculated in R for reproducibility. Only the metadata files were manually updated by me. Specifically, the designations for Ethnic Group, Latitude, Longitude, Landmass, and unused in this dissertation City for Lat/Long, were set by me for reasons outlined in *3.1.2.2 Explanatory variables* of this dissertation. Finally, all figures in this dissertation were created and formatted in R and are fully reproducible, so changing the underlying data input will automatically change the figure.

3.3 Experimental procedures

In this section, the procedures for applying the methodology to the worldwide dermatoglyphics data set were discussed in detail.

3.3.1 Computing environment

All computation was performed in R: A Language and Environment for Statistical Computing (R Core Team 2022), and when additional computing power and memory were required, the Open OnDemand RStudio Server was remotely accessed on Infrastructure for Scientific Applications and Advanced Computing (ISAAC), which is sponsored by the High Performance and Scientific Computing group at the University of Tennessee, Knoxville. All software and computing package algorithms are listed with versions and sources in Table S3. The sample sizes were specific to each method of analysis and then further refined based on the specific response variable.

3.3.2 *k-means clustering*

The method of *k-means* clustering was included in this dissertation, because the method can accommodate quantitative, multivariate response variables for either raw finger ridge-counts or summary variables, and *k-means* clustering represents an unsupervised but distribution-dependent method. Although *k-means* clustering represents an unsupervised method without response and explanatory variables, the *k-means* cluster assignment results can be used in subsequent regression analyses as explanatory variables so the perspectives of sample size and composition as well as level of population study can be analyzed in this way (Table 1). For *k-means* clustering, any individuals with NA for response variables finger ridge-counts, Directional Asymmetry variables, and Fluctuating Asymmetry variables were removed, then the first step to estimate the optimal value *K* (number of clusters) was performed on a smaller 10% subset of the data set for the finger ridge-counts and also for Directional Asymmetry variables and Fluctuating Asymmetry variables to ease computational burden. The distribution of the dermatoglyphic features in the 10% subset data set was confirmed to be similar to the distribution of dermatoglyphic features in the overall data set: finger ridge-counts (Table S4, Figure 13), Directional Asymmetry variables (Table S5, Figure 14), and Fluctuating Asymmetry variables (Table S6, Figure 15). The tables were created with the *table1* package (Rich 2023).

The optimal value of *K* was estimated using three approaches: gap statistic, within-cluster sum of squares, and the silhouette method. The gap statistic was calculated with four initial random configurations, 50 maximum iterations, and 500 bootstrap samplings using the *clusGap* function in the *cluster* package and came with a warning “Quick-TRANSfer stage steps exceeded maximum (= 232750)” for finger ridge-counts and came with warning “Quick-TRANSfer stage steps exceeded maximum (= 232800)” for Directional Asymmetry and Fluctuating Asymmetry variables (Maechler et al. 2022). The within-cluster sum of squares approach targets the “elbow”

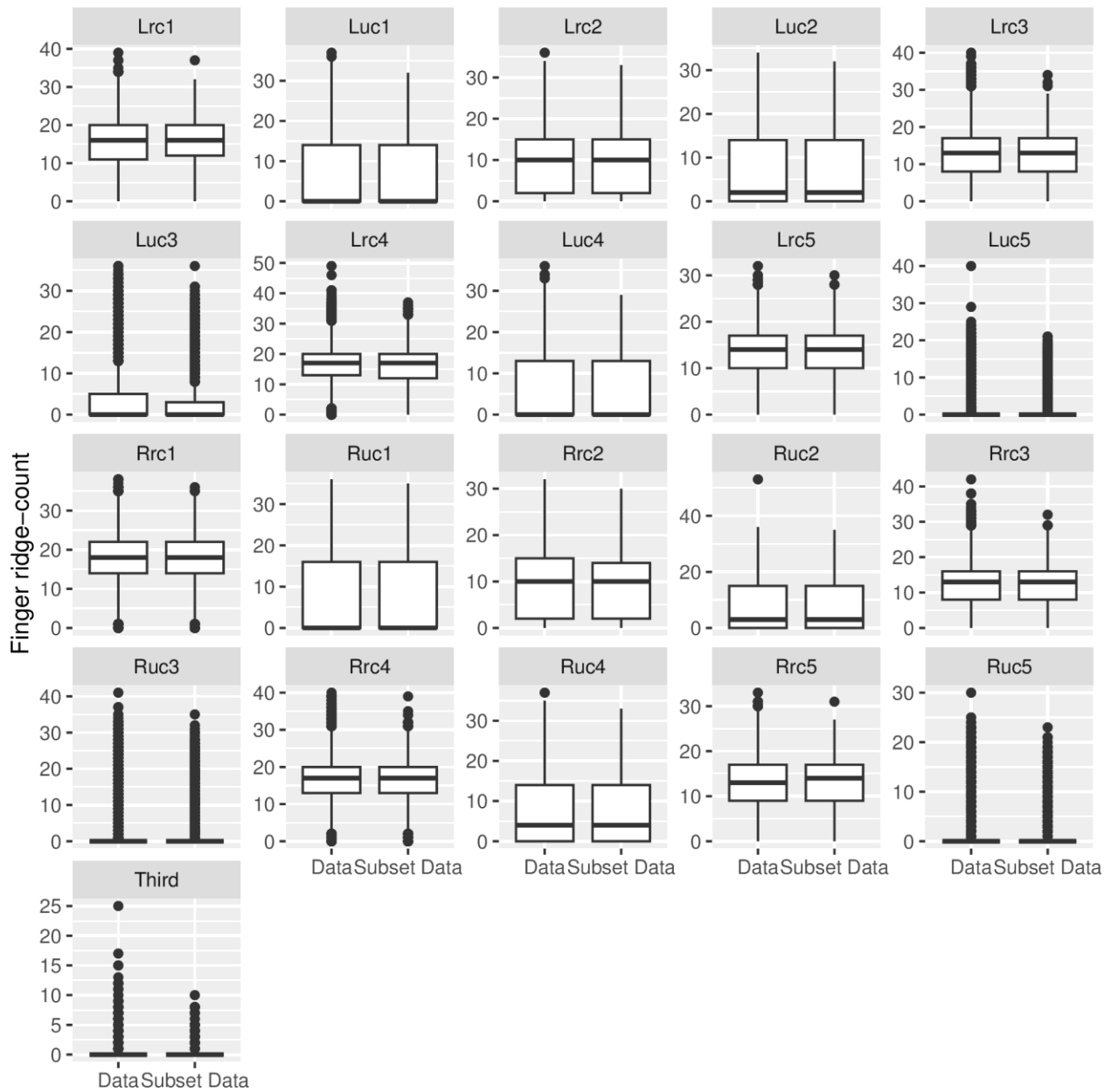


Figure 13. Boxplot of distribution of finger ridge-counts in overall data set compared to 10% subset. Closed black circles are outliers, horizontal lines equal maximum ($Q3 + 1.5 \cdot IQR$) and minimum ($Q1 - 1.5 \cdot IQR$) values, and bold horizontal lines are the median.

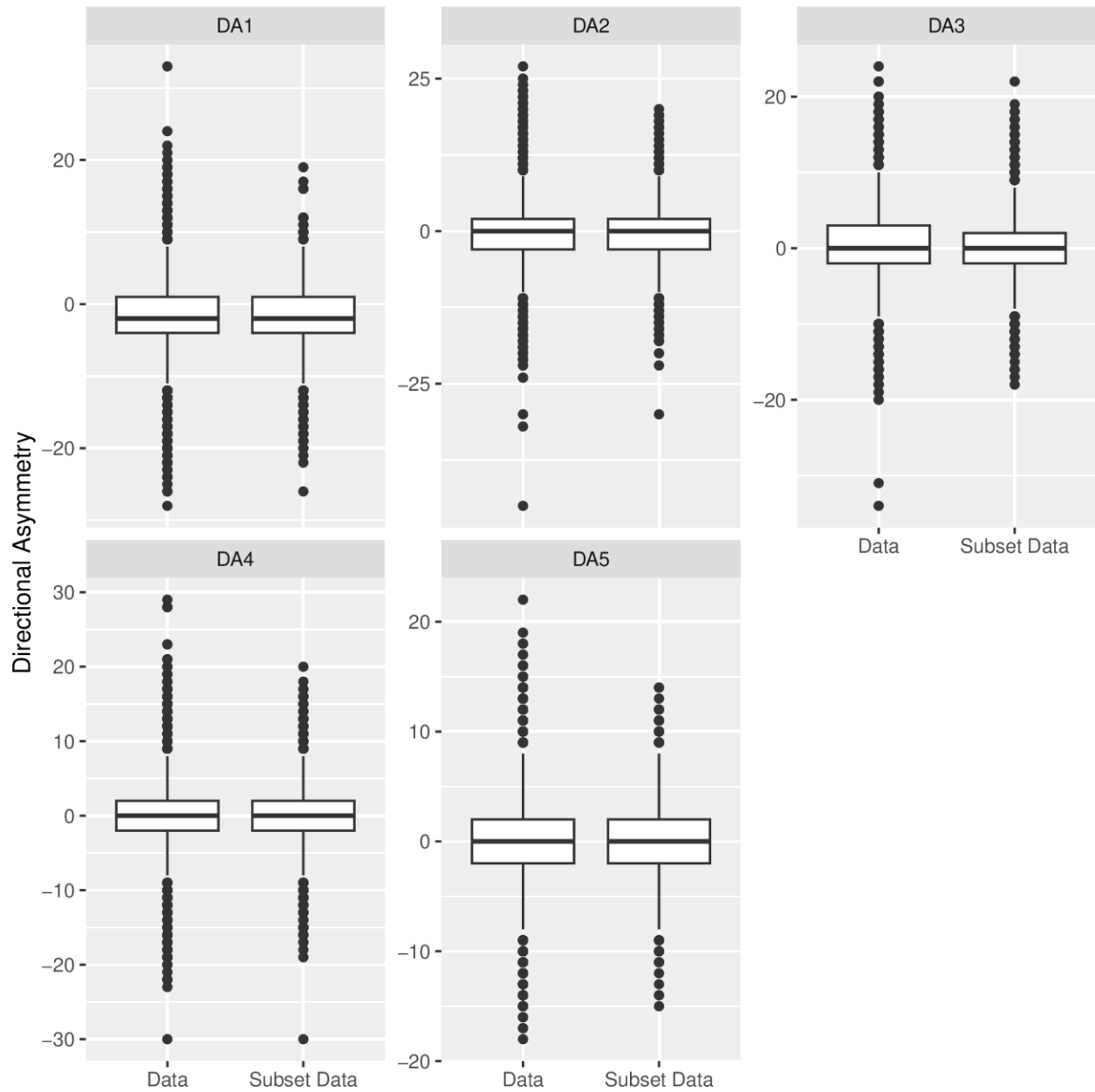


Figure 14. Boxplot of distribution of Directional Asymmetry variables separated by digit in overall data set compared to 10% subset. Closed black circles are outliers, horizontal lines equal maximum ($Q3 + 1.5 \cdot IQR$) and minimum ($Q1 - 1.5 \cdot IQR$) values, and bold horizontal lines are the median.

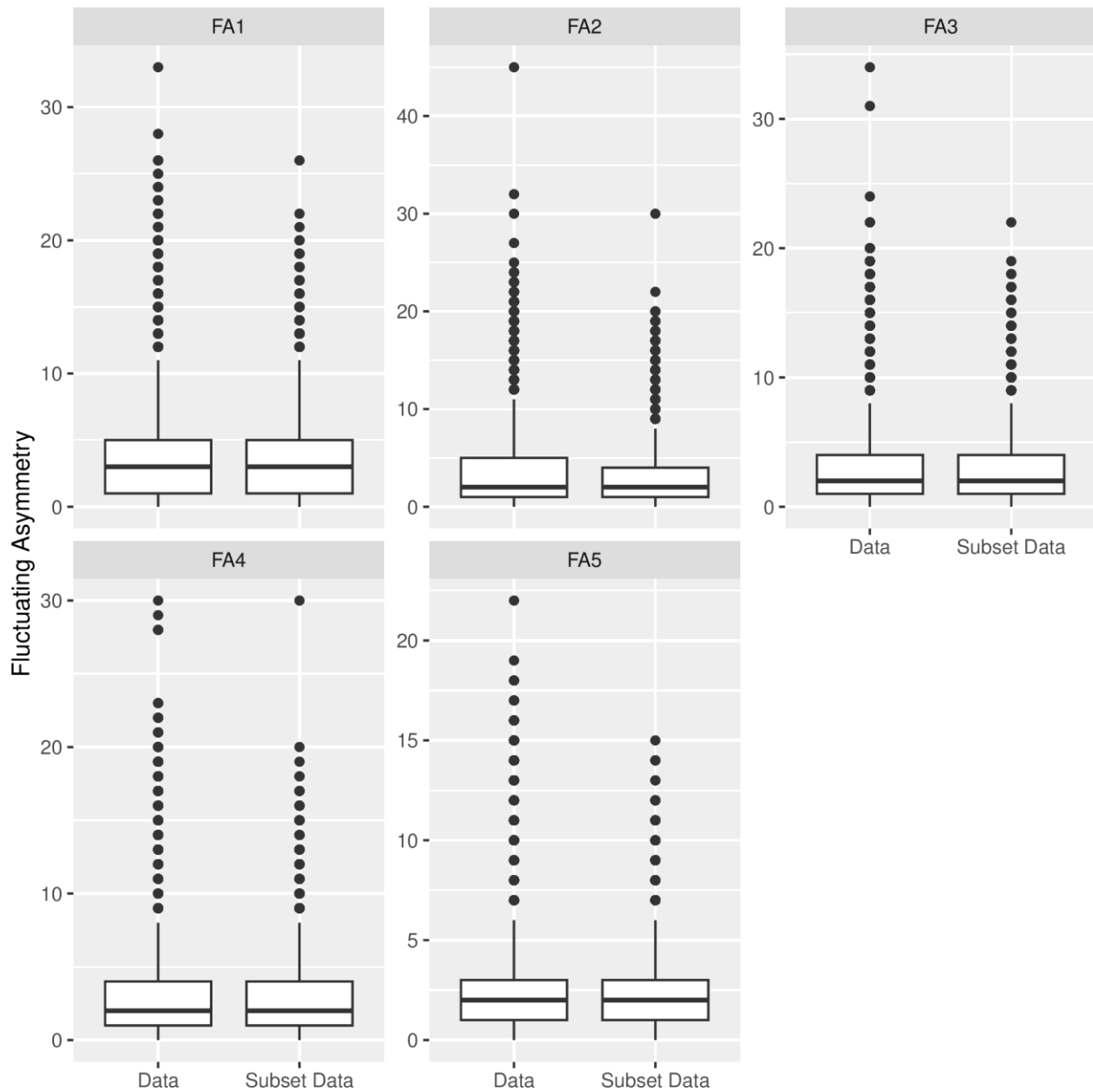


Figure 15. Boxplot of distribution of Fluctuating Asymmetry variables separated by digit in overall data set compared to 10% subset. Closed black circles are outliers, horizontal lines equal maximum ($Q3 + 1.5 \cdot IQR$) and minimum ($Q1 - 1.5 \cdot IQR$) values, and bold horizontal lines are the median.

or leveling-off in the line plotting the number of clusters against the total within sum of squares. The within-cluster sum of squares approach and the average silhouette approach both used the `fviz_nbclust` function in the `factoextra` package (Kassambara and Mundt 2020). Each of the three approaches purported a different ideal K value, so each of the K values was considered further.

Second, each observation was assigned to an initial cluster, from one to K , and k-means clustering was then performed on the full data set with a maximum of 50 iterations and 25 initial configurations. The default Hartigan-Wong k-means algorithm was used to find the initial configurations leading to the smallest within-cluster variation. Visualizing the three resulting cluster plots did not reveal an obvious interpretation of the K value generated by the three statistics. So, two additional K values were chosen for additional k-means clustering analysis, and the K values were dictated by the number of Landmasses and number of Ethnic Groups in this data set. The second step of assigning each observation to an initial cluster, from one to K clusters and performing k-means clustering analysis was repeated with optimal K values of 4 and 322, respectively.

At the conclusion of k-means clustering for finger ridge-counts, the objects containing the cluster column for each analysis with different optimal K contain only 46,574 rows. At the conclusion of k-means clustering for Directional Asymmetry and Fluctuating Asymmetry variables, the objects containing the cluster column for each analysis with different optimal K contain only 46,582 rows. For any subsequent analysis with k-means clusters as explanatory variables, the explanatory variables must come from these data sets, because many analyses require the response variables to have the same number of rows as the explanatory variables.

The internal performance of k-means clustering was evaluated by calculating the Dunn Index for the cluster assignments of different optimal K values using the `dunn` function in the

ciValid package (Brock et al. 2008). The external performance of each clustering attempt based on a different optimal K was evaluated when the cluster assignments were used as explanatory variables in subsequent GLM and SVR, and each modeling approach calculated its own performance metrics like MAE, RMSE, and MAPE.

3.3.3 Spatial analysis via distance-based Moran's eigenvector maps (dbMEM)

An explicit spatial analysis was included in this dissertation to supply a mathematical evaluation of the importance of sample size and composition as well as level of population study. To identify spatial structures within the sampling area, spatial variables were constructed at relevant scales using distance-based Moran's eigenvector maps (dbMEM). Spatial analysis was performed separately for each Landmass: Africa, Americas, Asia, and Europe, because spatial structures spanning continents separated by oceans limit the fineness of the MEM eigenfunctions produced by the dbMEM functions. The fineness is determined by the largest link in the minimum spanning tree connecting all data points. The Latitude and Longitude values in decimal form with cardinal direction were first converted into positive and negative numbers to eliminate the character indicator of the cardinal direction. The Longitude East values remained positive numbers, while the Longitude West values were negative numbers. The Latitude North values remained positive numbers, while the Longitude South values were negative numbers.

Spatial analysis was performed separately for finger ridge-counts and finger patterns. Not all individuals in the data set contain records for both finger patterns and finger ridge-counts, so while separating the individuals into data sets according to Landmass different sample sizes and compositions result (Table 7). The same procedures were applied to the finger patterns as well as the finger ridge-counts, so the remainder of the explanation of the spatial analysis refers to the response variables, which were either the finger patterns or the finger ridge-counts.

Table 7. Different sample sizes in different spatial analyses; finger patterns = 10 variables; finger ridge-counts = 21 variables

Variables for spatial analysis	Sample size
finger patterns Africa	11675
finger ridge-counts Africa	10806
finger patterns Americas	3887
finger ridge-counts Americas	3522
finger patterns Asia	18648
finger ridge-counts Asia	15813
finger patterns Europe	17056
finger ridge-counts Europe	16400

The response variables, Latitude, Longitude, Abbreviation, Region, Ethnic Group, and Sex were split into different data sets by Landmass. The explanatory variables of Region, Ethnic Group, and Sex were converted from character into factor class. Any rows containing NA values were removed. To ease computational burden, the data sets were reduced by the unique Abbreviations, so a data set containing just the Latitude and Longitude would just contain coordinates for each unique Abbreviation. Because some groups share the same Latitude and Longitude coordinates, selecting for the unique Latitude and Longitude would exclude some groups, so this method was not used. Using the reduced data sets of coordinates, the Vincenty ellipsoid distance matrices were calculated using the default arguments using the `distVincentyEllipsoid` function in the `geosphere` package, to account for curvature of Earth, resulting in distances in meters, and then the matrices of dbMEM variables were created with automatic calculation of the truncation levels using the `dbmem` function in the `adespatial` package (Dray et al. 2022; Hijmans 2022). Tables of truncation distances for each data set were created using the `kable` function in the `kableExtra` package (Zhu et al. 2021). From this, the dbMEM eigenvectors with positive spatial correlation were recorded. The spatial analysis followed R code and functions `plot.links`, `quickMEM`, `scalogram`, and `sr.value` provided by Borcard, Gillet, and Legendre (2018), except the response variables were not Hellinger-transformed as suggested, because the dermatoglyphic response variables were in the same scale so do not need transformation (Graham and Özener 2016).

The R packages `ade4`, `adegraphics`, `ape`, and `spdep` and were also used to perform spatial analysis (Paradis and Schliep 2019; Bivand et al. 2023; Dray et al. 2023; Siberchicot and Julien-Laferrière 2023). To assess whether a linear trend was present in the data, constrained ordination was performed on the response variables with the coordinates then an ANOVA to test for

significance of the model. Unless otherwise indicated, all ANOVA ran for 99999 permutations, and the model was significant (p value < 0.05). Since a linear trend was present with each response variable type and within each Landmass, the data were detrended by regressing the response variables on the coordinates, and the residuals were retained. After detrending, the dbMEM variables were repeated to match the size of the original data set and not the reduced data set containing only unique Abbreviations. A global dbMEM analysis was performed on the detrended response variable and the full dbMEM variables to identify further spatial trends. If the R^2 value was significant, then the adjusted R^2 value was calculated, and forward selection was performed of the detrended response variables and the full dbMEM variables using the adjusted R^2 value as the threshold for selection (Blanchet, Legendre, and Borcard 2008).

From this reduced model, a number of significant dbMEMs was identified. A new dbMEM analysis was conducted on the detrended response variables and the reduced models' significant dbMEM variables, and a new adjusted R^2 value was calculated. An ANOVA was performed to test for significance of trends, and a separate ANOVA was performed by-axis to test for the significance of canonical axes. To calculate the percent of total variance explained by the significant canonical axes, multiply the total adjusted explained variance, as represented by the adjusted R^2 values of the significant dbMEM variables retained by forward selection, by the accumulated constrained eigenvalues of the redundancy analysis (RDA) output for the proportion of variance explained by the significant canonical axes with respect to the total explained variance. To assess whether the variation in the significant canonical axes was related to any of the explanatory environmental variables, simple linear regression was performed with the fitted site scores of the canonical axes as response variables, and the environmental variables as explanatory variables. To verify the residuals of the regressions were normally distributed,

which is an assumption of linear regression, the Anderson-Darling normality test was performed with the `ad.test` function in the `nortest` package (Gross and Ligges 2015).

To interpret the spatial variation, regression of the significant canonical axes was performed on the environmental variables: Region, Ethnic Group, Sex, Latitude, and Longitude, and an Anderson-Darling normality test was performed on the residuals. To discriminate between different scales, a map of the significant dbMEM variables was created, and a scalogram was constructed with the significant dbMEMs versus the unadjusted R^2 values using 999 permutations.

For each significant dbMEM variable from the reduced model, a dbMEM analysis was performed with the detrended response variable. This ends the focus on the global spatial structure of the response variables where all significant dbMEM variables are combined, which masks any broad, medium, or fine-scaled spatial structuring. Because dbMEM analysis was performed by-Landmass and not for the entire worldwide data set, the term global here refers to the spatial structure at the maximum level of the data set within this analysis rather than an actual global, worldwide scale. Maps for significant dbMEMs or canonical axes were created using the `cowplot` and `gridGraphics` packages (Murrell and Wen 2020; Wilke 2020).

Variation partitioning was performed as a way to explicitly incorporate any present linear trend and consider spatial structuring on multiple levels not only on a global scale. The variation partitioning function, `varpart` from the `vegan` package, only accepts numeric variables, so the environmental variables that were factor class: Region, Ethnic Group, and Sex, were recoded as dummy binary variables (Oksanen et al. 2022). For example, Sex female may be represented by two variables {1, 0}, and Sex male may be represented by {0, 1}. Similar procedures were performed to identify spatial trends. Constrained ordination was performed on the response

variables with the environmental variables: Region, Ethnic Group, Sex, Latitude, and Longitude. The adjusted R^2 value was calculated, and forward selection was performed of the response variables and the environmental variables using the adjusted R^2 value as the threshold for selection. From this reduced model with the environmental variables, a number of significant environmental variables were identified.

The significant dbMEM variables from the reduced model with the detrended response variables were arbitrarily split and assigned to a broad or fine scale, and this specifically included the detrended response variables to exclude the linear trend from influencing the detection of broad or fine scale spatial structures, whereas the significant environmental variables specifically include the response variables to include any linear trend. Scale is a relative term used here only by comparing broad to fine spatial patterns, but scale also contains inherent values of grain size, sampling intervals, and extents. Grain size is the resolution of the study, like the surface supporting the measurements, and sampling intervals are the average distance between neighboring sampling units but were irregular here because the sampling design was effectively irregular. The extent is the total area of the study such that grain size and extent are expressed in the same units (Legendre and Legendre 2012c:786). Variation partitioning was performed by including the response variables, the significant environmental variables, the Latitude and Longitude coordinates, the significant dbMEM variables assigned to the broad scale, and the significant dbMEM variables assigned to the fine scale. Finally, the fractions of variation were tested for significance with an ANOVA: pure environmental, pure trend, pure broad scale, and pure fine scale.

The significant dbMEM variables represent the significant spatial variation and can be divided into multiple levels of spatial structure, so rather than Latitude and Longitude

coordinates, the significant dbMEM variables for each Landmass can be included as explanatory variables in subsequent model frameworks at the Landmass scale. Because the significance of the dbMEM variables accounted for the variation in the response variable like finger ridge-counts or finger patterns, only analyses with the respective response variable should include the significant dbMEM variables as explanatory variables. For example, any derived variables like TRC or ARC have their own patterns of variation that were not accounted for in identifying the significance of dbMEM variables, so no analyses with TRC or ARC as the response variables should include dbMEM explanatory variables and should only use Latitude and Longitude where appropriate. Because the residuals from variation partitioning were higher than 90% for all undetrended response variables across Landmasses, the source of variation in the undetrended response variables was possibly not accounted for, so no subsequent modeling included significant dbMEM variables as explanatory variables to minimize the over-parameterization of models.

The methods for spatial analysis of finger ridge-counts and finger patterns by-Landmass were also applied to Directional Asymmetry for digits 1 through 5 and Fluctuating Asymmetry for digits 1 through 5.

3.3.4 Principal component analysis (PCA)

The method of principal component analysis (PCA) was included in this dissertation, because the method can accommodate quantitative, multivariate response variables for either raw finger ridge-counts or summary variables, and PCA represents an unsupervised but distribution-dependent method. Although PCA provided similar content for response variables as k-means clustering, the results of PCA are often difficult to interpret, so both k-means clustering and PCA were included to evaluate any different results (Table 1). For each data type, any individuals with

NA for response variables finger ridge-counts, Directional Asymmetry variables, and Fluctuating Asymmetry variables, were removed from subsequent analysis. Next, the individuals with finger ridge-counts, Directional Asymmetry variables, and Fluctuating Asymmetry variables were split by-Sex so females and males were analyzed separately. Also, the finger ridge-counts, Directional Asymmetry variables, and Fluctuating Asymmetry variables containing no NA were split by-Landmass and then further split by-Sex so females and males within each Landmass were analyzed separately: Africa, Americas, Asia, and Europe (Table S7). Only finger ridge-counts, Directional Asymmetry variables, and Fluctuating Asymmetry variables were included in the rda function from the vegan package, because PCA only analyzes quantitative variables. Without specifying any explanatory variables, the rda function performs a PCA (Oksanen et al. 2022). No variables were scaled, so the PCA was performed on the covariance matrix of the finger ridge-counts, Directional Asymmetry variables, and Fluctuating Asymmetry variables. Different explanatory variables like Sex, Region, and Landmass can be included by coloring the points in a PCA biplot by these variables to search for any patterns. Biplots with accessible color palettes were created using cowplot, ggpubr, scales, and tidyverse packages (Wickham et al. 2019; Wilke 2020; Wickham, Seidel, and RStudio 2022; Kassambara 2023; Wickham and RStudio 2023).

3.3.5 Baseline models

During spot-checking to preliminarily assess how to approach the research questions, baseline models served as an initial reference for how subsequent regression models could improve upon capturing the variation in finger ridge-counts or other response variables. The mean regressor, median regressor, and linear model were selected as the baseline models, and ultimately the goodness-of-fit and performance of the regression models was compared to the baseline models. For comparison, the same data sets were required for the baselines as for the

experimental models. Because the generalized linear models (GLM) were partitioned into a training and test set and the support vector regression (SVR) models were partitioned into a training, validation, and test set, the baselines were calculated separately for GLM training and test sets and for SVR training and test sets. For example, the mean baseline was calculated for GLM training set and compared to the GLM test set, so the performance of the GLMs have a baseline containing the same observations. Each baseline model underwent the same steps no matter for the GLMs or the SVR models, so subsequent discussion of methods reflects the baseline modeling steps for either GLM or SVR.

The Latitude and Longitude decimal values with character cardinal direction were first converted into positive and negative numbers to eliminate the character indicator of the cardinal direction. The Longitude East values remained positive numbers, while the Longitude West values were negative numbers. The Latitude North values remained positive numbers, while the Longitude South values were negative numbers.

The data set was then split into training data (70%) and test data (30%) for GLM or into training data (50%), validation data (25%), and test data (25%) for SVR. All rows containing NA were removed from each data set, and any Ethnic Group factors that were present in the test data but not in the training data were removed to facilitate the comparison of the split data sets.

To calculate the mean regressor, the training input was specified as the same explanatory variables as the experimental regression models, and the training output was specified as a single response variable. Using the `dummy_regressor` function in the `basemodels` package, the training output was supplied with the mean strategy (Chen et al. 2023). Next, a data matrix of the test input was specified as the same explanatory variables as the training input but with the test data, and the test output was specified as the same response variable but with the test data. To use the

mean regressor, the `predict_dummy_regressor` function in the `basemodels` package compared the dummy regressor to the test input to generate predictions of the response variable. The observed test output was compared to the predicted values to standard regression performance metrics using the MAE function in the `caret` package, the RMSE function in the `caret` package, and the MAPE function in the `MLmetrics` package (Yan 2016; Kuhn 2022). The mean and median regressors were included as baselines, because poor performance of a model is equivalent to just estimating the central tendency, the mean or median depending on the data set. The dermatoglyphic features may not follow a normal distribution, so including the median regressor provides a buffer should the mean not be informative. The median regressor followed the same steps as the mean regressor but with the median strategy in the `dummy_regressor` function.

The linear model was included as a baseline model for its limited complexity and thus its ease of interpretation to direct subsequent experimental regression approaches. The full model of the training data included all explanatory variables and the single response variable in a single object. To evaluate the goodness-of-fit of the full model, the AIC value of the full model was calculated using the `ols_aic` function in the `olsrr` package by supplying the full model, and the default method was “R”. To create a linear regression model, the `lm` function was used with the traditional formula syntax ($\text{response} \sim \text{explanatory}_a + \text{explanatory}_b \dots \text{explanatory}_n$, `data`), and to perform feature selection the `ols_step_backward_aic` function in the `olsrr` package was used to perform backward stepwise selection starting from a full model with all explanatory variables and then removing one in each iteration based on their AIC value until no variables should be removed due to a diminishing improvement (reduction) in the AIC value (Hebbali 2020).

For the three baseline linear models using the SVR training and test sets with response variable Directional Asymmetry of digit 2, response variable left ulnar count of digit 2, and

response variable Third ridge-count, the `ols_step_forward_aic` function in the `olsrr` package was used to perform forward stepwise selection, starting with an intercept-only model with no explanatory variables and then adding one in each iteration based on their AIC value until no variables should be added due to a diminishing improvement in the AIC value. Each of these models exhibited a final model with a single explanatory variable, which the `ols_step_backward_aic` function was evidently unable to accommodate.

After feature selection, a new fitted final linear model was created with the `lm` function, the same response variable, and the smaller subset of explanatory variables selected from the stepwise feature selection. To evaluate the goodness-of-fit of the final fitted model, the AIC value of the final model was calculated using the `ols_aic` function in the `olsrr` package by supplying the final reduced model, and the default method was “R”. The goodness-of-fit AIC value from the full model was compared to the AIC value from the final reduced model to demonstrate the improvement in model fit due to feature selection. Because linear models assume a constant error in the variance, the heteroscedasticity was formally evaluated using the `check_heteroscedasticity` function in the `easystats` package by supplying the final model, and heteroscedasticity was detected based on a p value < 0.05 from the Breusch-Pagan test (Breusch and Pagan 1979; Lüdecke et al. 2022). Because linear models assume the normality of residuals, a formal evaluation by the Anderson-Darling test, which gives more weight to the tails of the distribution, used the `ad.test` function in the `nortest` package, supplying the residuals of the final model (Gross and Ligges 2015). Plotting the residual values offered a visual assessment of their potential normality and used the `ols_plot_resid_hist` function in the `olsrr` package by supplying the final model. Next, linear models assume no autocorrelation between adjacent residuals, meaning one adjacent residual value does not partially predict or inform the value of another

residual, so the autocorrelation was formally evaluated using the `check_autocorrelation` function in the `easystats` package by supplying the final model as well as setting the seed to ensure reproducibility, since the autocorrelation check performs a Durbin-Watson test and relies on simulations comprising the iterative sampling with replacement of a residual value, setting that value as a response variable with the existing explanatory variables, and then calculating the residuals of that model to check for patterns. In this way, the p value was a bootstrapped p value. Extracting the non-iterative `durbWats` function from the `check_autocorrelation` function that performed the Durbin-Watson test, to calculate the test statistic, any value equal to 2 indicated no autocorrelation, any value between zero and less than 2 indicated positive autocorrelation, and any value greater than 2 and less than 4 indicated negative autocorrelation. In practical use, any test statistic value between 1.5 and 2.5 could be approximated as no autocorrelation (Bedre 2019; SE n.d.). The other assumptions of linear models were visually assessed using the `check_models` function and the `see` package (Lüdecke, Makowski, Patil, et al. 2023).

The performance of the linear model was assessed on the yet unseen test set, where the test input contained the reduced explanatory variables that passed feature selection and made the final model, and the test output was the same response variable as the final model. The final model formulation was applied to the test input to generate predictions of the response variable, and the predictions were compared to the observed test output values to measure regression performance metrics using the `MAE` function in the `caret` package, the `RMSE` function in the `caret` package, and the `MAPE` function in the `MLmetrics` package (Yan 2016; Kuhn 2022). The `RMSE` was used for model evaluation during training of `SVR` models with grid searching for hyperparameter optimization, so although the `RMSE` was not suitable for use with final model

selection based on the properties of the response variables, the RMSE was still reported alongside MAE and MAPE as final performance metrics (Table 4, Table 5).

3.3.6 Generalized linear modeling (GLM)

The method of generalized linear modeling (GLM) was included in this dissertation, because the method can accommodate quantitative, univariate response variables for either raw finger ridge-counts or summary variables, and GLM represents a supervised and distribution-dependent method (Table 1). Since dermatoglyphic features like TRC do not exhibit a normal distribution but instead were negatively skewed and flattened for both females and males from a British population, the GLM provided a flexible modeling framework to specify a different statistical distribution depending on the properties of the response variables (Holt 1955).

Generalized linear modeling was performed only for the single output quantitative response variables like TRC, ARC, Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt. Because GLM via LASSO regularization assesses which explanatory variables have an effect on the response variables, GLM was not performed on each Landmass or other grouping separately, since that would presume an effect of Landmass or other grouping on the response variables. At the conclusion of k-means clustering on finger ridge-counts, the objects containing the cluster column for each analysis with different optimal K contained only 46,574 rows. For any subsequent analysis of finger ridge-counts, TRC, and ARC with k-means cluster as explanatory variable like GLM, the explanatory variables come from this data set, because the analysis requires the response data to have the same number of rows as the explanatory data (Table S8). Although k-means clustering was not performed for TRC and ARC, the k-means cluster assignments from the finger ridge-counts were used as explanatory variables,

because TRC and ARC represent summary measures of the finger ridge-counts. At the conclusion of k-means clustering on Directional asymmetry variables for each digit, Fluctuating asymmetry variables for each digit, Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt, the objects containing the cluster column for each analysis with different optimal K contained only 46,582 rows. For any subsequent analysis of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt, with k-means cluster as explanatory variable like GLM, the explanatory variables came from this data set, because the analysis requires the response data to have the same number of rows as the explanatory data (Table S9). Although k-means clustering was not performed for Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt, the k-means cluster assignments from the Directional Asymmetry variables were used as explanatory variables, because the optimal K for Directional Asymmetry variables was similar to the optimal K for Fluctuating Asymmetry variables, and including both would likely overload the model with too many redundant explanatory variables.

Next, the Latitude and Longitude decimal values with cardinal direction were first converted into positive and negative numbers to eliminate the character indicator of the cardinal direction. The Longitude East values remained positive numbers, while the Longitude West values were negative numbers. The Latitude North values remained positive numbers, while the Longitude South values were negative numbers. The data set was then split into training data (70%) and test data (30%). All rows containing NA were removed from each data set, and any Ethnic Group factors present in the test data but not in the training data were removed to facilitate the comparison of the two data sets.

GLM was selected as a modeling strategy, because the dermatoglyphic features did not meet the assumptions of standard linear regression and overdispersion was present such that the mean of the response variable exceeded the variance for all response variables, except for Index.Diversity.Holt where the mean was 3.99, and the variance was lower at 2.53. To detect dispersion across different groupings, from the vegan package the betadisper function was applied to the vegdist function of the response variable and the grouping, like Ethnic Group, Region, Sex, and Landmass (Oksanen et al. 2022). The betadisper function represents the absolute value distance between groups from a multivariate centroid, and the anova function was applied to the resulting betadisper object to test for significance in dispersion across groups.

Each training and test set was specified to only include the required variables and was divided into input and output objects for modeling. For example, the training data set with TRC as the response variable was divided into training output that contained 32580 observations of only the TRC response variable column and into a training input data matrix that contained the same number of 32580 observations of the 11 explanatory variables: Ethnic Group, Region, Landmass, Sex, Latitude, Longitude, frc.kmeans.2, frc.kmeans.4, frc.kmeans.10, frc.kmeans.16, and frc.kmeans.322 (Table S1). To evaluate the goodness-of-fit of the full model containing the single response variable as a function of all 11 explanatory variables from the training data set, the Kullback-Leibler R^2 value was calculated. The final test data set with TRC as the response variables was divided into test output that contained 13961 previously unseen observations of only the TRC response variable column and into a test input data matrix that contained the same number of 13961 observations but only the best explanatory variables: Landmass, Sex, frc.kmeans.2, frc.kmeans.4, frc.kmeans.10, and frc.kmeans.16, for example.

The Kullback-Leibler R^2 value was used as the goodness-of-fit metric for the nonlinear generalized linear models in this dissertation, because the traditional R^2 value is not an appropriate goodness-of-fit measure for nonlinear models, and pseudo R^2 values can exhibit strange mathematical behaviors that limit comparability and interpretability (Cameron and Windmeijer 1997). In the performance package in R, the `r2_kullback` function takes a base R glm model as the only input, and because the argument `adjust` was specified as `FALSE` the Kullback-Leibler R^2 value was not adjusted by the degrees of freedom in the model (Lüdtke, Makowski, Ben-Shachar, et al. 2023). Without adjustment, the `r2_kullback` function actually calculates $1 - (\text{model deviance} \div \text{null model deviance})$.

All GLM were processed using the `glmnet` function in the `glmnet` package, so these models were not compatible with the `r2_kullback` function that requires a base R glm model, so the official function in the package was not used (Friedman et al. 2023). The source formula $1 - (\text{model deviance} \div \text{null model deviance})$ was used, which is equivalent to the `dev.ratio` function in the `glmnet` package. Because the `glmnet` models perform LASSO regularization, and base R glm does not, setting `lambda` to zero in the `glmnet` models to remove the LASSO regularization produced the same coefficients as in base R glm and the same calculated Kullback-Leibler R^2 value as with the `r2_kullback` function (results of quality check not included in this dissertation).

To identify the best explanatory variables for each response variable, a regression model with a probability distribution family and associated link function was fit with LASSO regularization with k-folds cross-validation for feature selection. Using the `cv.glmnet` function in the `glmnet` package the training input and training output objects were specified (Friedman et al. 2023). For response variables TRC, ARC, the finger ridge-counts, and Fluctuating Asymmetry

variables for each digit, a negative binomial family ($\theta = 5$) loaded by the MASS package and log link function was specified, because these variables were all positive, discrete counts (Ripley et al. 2023b). For response variables Directional Asymmetry variables for each digit, Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt, a Gaussian family was specified, because these variables included negative integer values like Directional Asymmetry variables for each digit or were positive continuous including zeros like Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt. The alpha value was set to 1 for all models, specifying LASSO regression versus ridge regression or elastic-net regularization. By default, cross-validation progressed for 10-folds. The minimum lambda value was saved as the best lambda value for that model.

The final model was constructed with the same training output object, but the training input object contained only the explanatory variables that did not have zero coefficients as a result of LASSO regularization, using the coefplot package to plot coefficients and the viridis package for color accessibility (Garnier et al. 2022; Lander 2022). The same family, link function, and alpha were included in the final model, and now the best lambda value was specified in the model. To evaluate the goodness-of-fit of the final fitted model, the Kullback-Leibler R^2 value was calculated. The goodness-of-fit Kullback-Leibler R^2 value from the full model was compared to the Kullback-Leibler R^2 value from the final reduced model to demonstrate the improvement in model fit due to feature selection.

The performance of the final model was evaluated by generating predictions using the final model with the yet unseen test input data that contained the same best explanatory variables that did not have zero coefficients as a result of LASSO regularization. Those predictions were compared to the yet unseen test output variable values, and standard regression performance

metric quantified the performance using the MAE function, the RMSE function in the caret package, and the MAPE function in the MLmetrics package (Yan 2016; Kuhn 2022)

3.3.7 Support vector regression (SVR)

The method of support vector regression (SVR) was included in this dissertation, because the method can accommodate quantitative, univariate response variables for either raw finger ridge-counts or summary variables (although only summary variables were included in this dissertation), and SVR represents a supervised but distribution-free method as compared to GLM, which is supervised, distribution-dependent (Table 1). Since dermatoglyphic features like TRC do not exhibit a normal distribution but instead were negatively skewed and flattened for both females and males from a British population, the SVR provided a free modeling framework that does not require specifying a statistical distribution of the response variables (Holt 1955). Support vector regression (SVR) was performed only for the single output quantitative response variables like TRC, ARC, Directional Asymmetry variables for each digit, and Fluctuating Asymmetry variables for each digit. Because SVR assesses which explanatory variables have an effect on the response variables, SVR was not performed on each Landmass separately, since that would presume an effect of Landmass on the response variables. From the overall data set, the Latitude and Longitude decimal values with cardinal direction were first converted into positive and negative numbers to eliminate the character indicator of the cardinal direction. The Longitude East values remained positive numbers, while the Longitude West values were negative numbers. The Latitude North values remained positive numbers, while the Longitude South values were negative numbers. The overall data set was split into training data (50%), validation data (25%), and test data (25%) (Table S10, Table S11). All rows containing NA were removed from each data set, and any Ethnic Group factors that were present in the validation

data or test data but not in the training data were removed to facilitate the comparison of the three data sets. Each training, validation, and test set was specified to only include the required variables for modeling. For example, the final validation data set with TRC as the response variable contained 11625 observations of seven features: TRC, Sex, Ethnic Group, Region, Latitude, Longitude, and Landmass, and the final test data set that was reserved for performance evaluation with TRC as the response variable contained 11604 observations of those same seven features.

First, a baseline support vector machine model was created using the `svm` function in the `e1071` package with `eps-regression` type, a radial kernel and standard parameter values (Meyer et al. 2022). Gamma was set as $1/\text{number of column variables in the training data}$, for all SVR models was $1/7$, and the cost was set at 1. The baseline model trained for 200-fold cross-validation. For the baseline model, the training data were divided into training input for the explanatory variables: Sex, Ethnic Group, Region, Latitude, Longitude, and Landmass; and training output for the response variable, like TRC, ARC, Directional Asymmetry variables for each digit, and Fluctuating Asymmetry variables for each digit. The training input variables were applied to the baseline model to generate predictions of the response variable, and the predictions were compared to the observed training output variable, so a baseline model performance was recorded.

Next, grid searching was performed to choose the optimal hyperparameters of gamma and the cost. Grid searching proceeded from broad preset ranges of each hyperparameter. For example, the search range for gamma was a sequence of 10^{-8} to 10^8 , incremented by 1, and the search range for the cost was a sequence of 2^{-16} to 2^{16} , incremented by 1. This coarse grid search performed model tuning with the training data using the `train` function in the `caret` package and

method `svmRadial` to use the radial basis function kernel from the `kernlab` package (Karatzoglou et al. 2004; Kuhn 2022).

Using the `trainControl` function in the `caret` package, the model was trained over a 5-fold cross validation (Kuhn 2022). The validation input variables: Sex, Ethnic Group, Region, Latitude, Longitude, and Landmass were applied to the models produced with each possible hyperparameter value to generate predictions of the response variable, and the predictions were compared to the observed validation output variable for which performance metrics were calculated and recorded. Based on the smaller range of hyperparameters that yielded the best performance via lowest RMSE, a subsequent grid search was performed with narrower search ranges incremented by smaller values. To visualize the RMSE value of each gamma and cost value in the grid search, a grid plot was constructed using the `ggplot2` and `ggthemes` packages and the `viridis` and `viridisLite` packages for accessibility of the color palettes in the tuning graph (Wickham 2016; Arnold 2021; Garnier et al. 2022). When a large range performed well or if the lower/upper boundary of the optimal hyperparameter remained unclear, a medium grid search was performed again with the training data and evaluated on the validation data and did not contain small increment values. For example, a medium grid search range for gamma was a sequence of 10^{-4} to 10^{-2} , incremented by 0.25, and the search range for the cost was a sequence of 2^3 to 2^{10} , incremented by 0.25. Only when the lower boundaries of the optimal hyperparameters were clear and a more narrow range of optimal hyperparameters were evident would the fine grid search be appropriate. The fine grid search on the training data and evaluated on the validation data contained the most narrow range of hyperparameter values like the search range for gamma was a sequence of 10^{-4} to 10^{-3} , incremented by 0.125, and the search range for the cost was a sequence of 2^2 to 2^{11} , incremented by 0.125. The range of hyperparameter values

for coarse, medium, and fine grid searching differed among single output quantitative response variables. When tuning the hyperparameters, often no single value yields the best performance, rather a range for each hyperparameter yields a similar, best performance, and so as long as values within the range of best values are used the model performance is upheld. At the conclusion of each coarse, medium, and fine grid search, the range of hyperparameters to search was selected based on the performance of the models with each possible hyperparameter value.

After grid searching, the final model contained a single gamma value and a single cost value and was performed on the training data. The performance of the final model was evaluated by generating predictions using the final model with the yet unseen test input data using the caret package (Kuhn 2022). The test input variables contained the same explanatory variables Sex, Ethnic Group, Region, Latitude, Longitude, and Landmass. The predicted response variables were compared to the yet unseen test output variable values, and standard regression performance metrics quantified the performance using the MAE function, the RMSE function in the caret package, and the MAPE function in the MLmetrics package (Yan 2016; Kuhn 2022).

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Results for finger ridge-counts and summaries of ridge-counts

The results for the finger ridge-count response variables and TRC and ARC summary response variables were separated by the method of analysis.

4.1.1 *k-means clustering*

The optimal K for the finger ridge-counts using the gap statistic was 16 clusters, around 10 clusters for the within-cluster sum of squares approach, and 2 clusters for the average silhouette approach (Figure 16). The internal performance of k -means clustering evaluated by calculating the Dunn Index for the cluster assignments of different optimal K values ranged from 0.017 to 0.053. Because the Dunn Index should be maximized, the optimal K value of two clusters from the average silhouette approach with Dunn Index of 0.053 represents the most appropriate number of clusters of those evaluated (Table 8).

The cluster assignments from k -means clustering of finger ridge-counts were included in subsequent GLM and SVR as additional explanatory variables. To interpret the cluster assignments, stacked bar charts were created for $K = 2, 4$, and 16 and compared to known groupings for the finger ridge-counts using packages tidyverse, viridis, viridisLite, scales to set the viridis color palette for accessibility, cowplot for converting base R plots into ggplot-friendly objects, and ggpubr for multipanel plots (Wickham et al. 2019; Wilke 2020; Garnier et al. 2022; Wickham, Seidel, and RStudio 2022; Kassambara 2023; Wickham and RStudio 2023). If the cluster assignments matched the known groupings for finger ridge-counts, then a known grouping would contain the predominately one cluster. For example, if $K = 2$ corresponded to Sex, then females would primarily contain one cluster, and males would primarily contain the

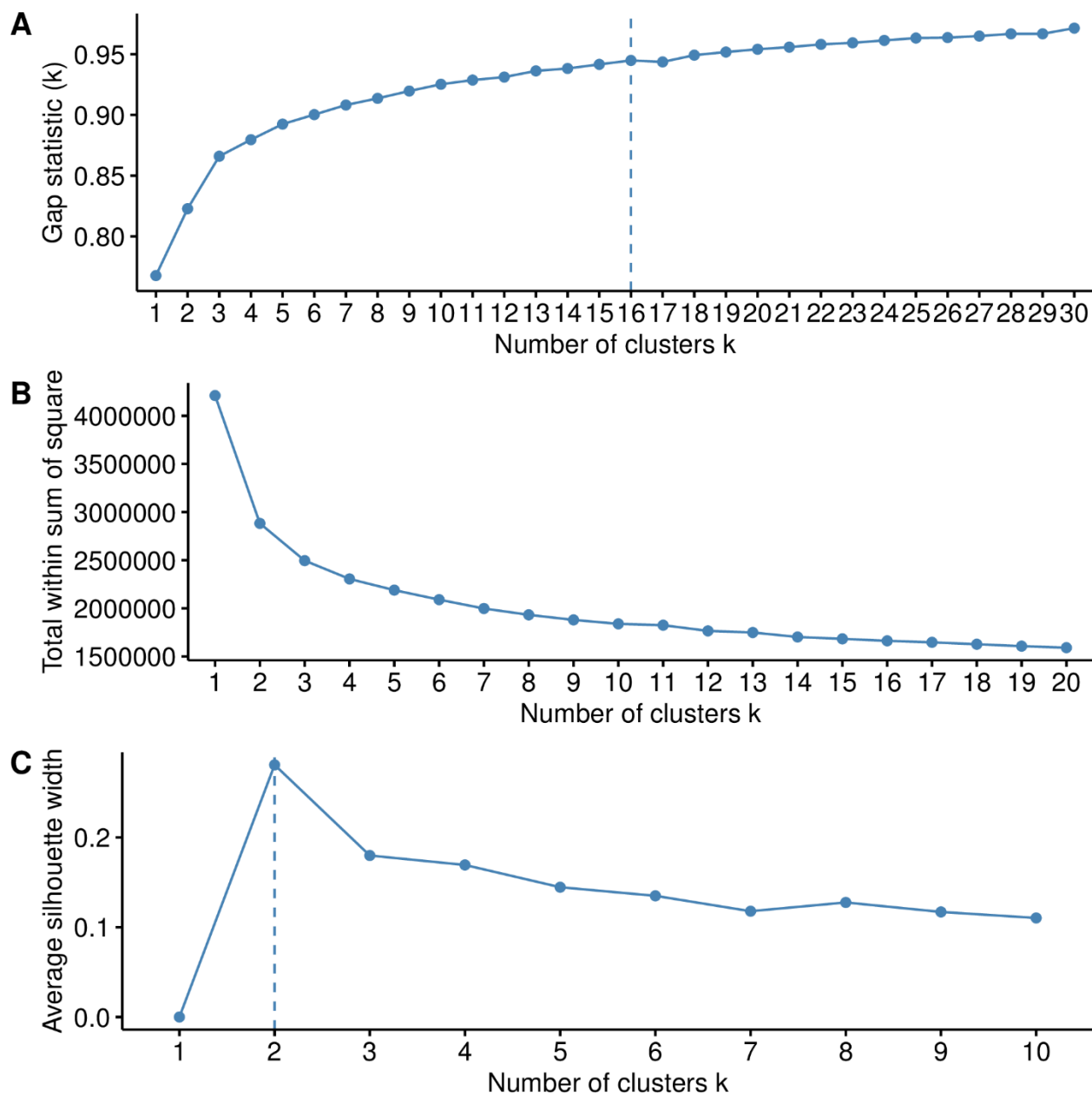


Figure 16. Finger ridge-counts optimal K for three approaches: (A) the gap statistic, (B) within-cluster sum of squares, and (C) average silhouette.

Table 8. Dunn Index values for optimal K values for finger ridge-counts. Dunn Index values in decreasing order; the Dunn Index should be maximized.

Number of Clusters	Approach for K selection	Dunn Index
2	Average silhouette	0.0525
16	Gap statistic	0.0302
4	Number of Landmasses	0.0292
10	Within sum of squares	0.0284
322	Number of Ethnic Groups	0.0171

second cluster (Figure 17). If $K = 4$ corresponded to Landmass, then each Landmass would primarily contain one cluster assignment each (Figure 18). If $K = 16$ corresponded to Ethnic Group, then each Ethnic Group would primarily contain one cluster assignment each (Figure 19). The cluster assignments for $K = 10$ did not readily match the number of levels in any known grouping for finger ridge-counts, and the closest number of levels in a known grouping for finger ridge-counts was 16, for which the gap statistic already suggested as the optimal K and so were not included in a stacked bar chart. The cluster assignments for $K = 322$ were not compared to the number of Ethnic Groups in a stacked bar chart, because the chart would not be readable with 322 groups on the x-axis.

The stacked bar charts for $K = 2, 4$, and 16 exhibited evenly distributed clusters for Landmass (Figure 17A, Figure 18A, and Figure 19A), Region (Figure 17B, Figure 18B, and Figure 19B), and Sex (Figure 17C, Figure 18C, and Figure 19C) known groupings for finger ridge-counts, which suggests $K = 2$ does not represent Sex, $K = 4$ does not represent Landmass, and $K = 16$ does not represent Ethnic Group. Thus, the cluster assignments from the various k-means clustering for finger ridge-counts with different optimal K represent explanatory variables presently unknown.

4.1.2 Spatial analysis via distance-based Moran's eigenvector maps (dbMEM)

The truncation distance for the finger ridge-counts and the finger patterns for each Landmass was the same, except for the Americas (Table 9). For finger ridge-counts in the Americas, the truncation distance was 2182080 m greater than the truncation distance for the finger patterns, and the sample sizes in the Americas overall represent roughly one-third the size of samples in other Landmasses.

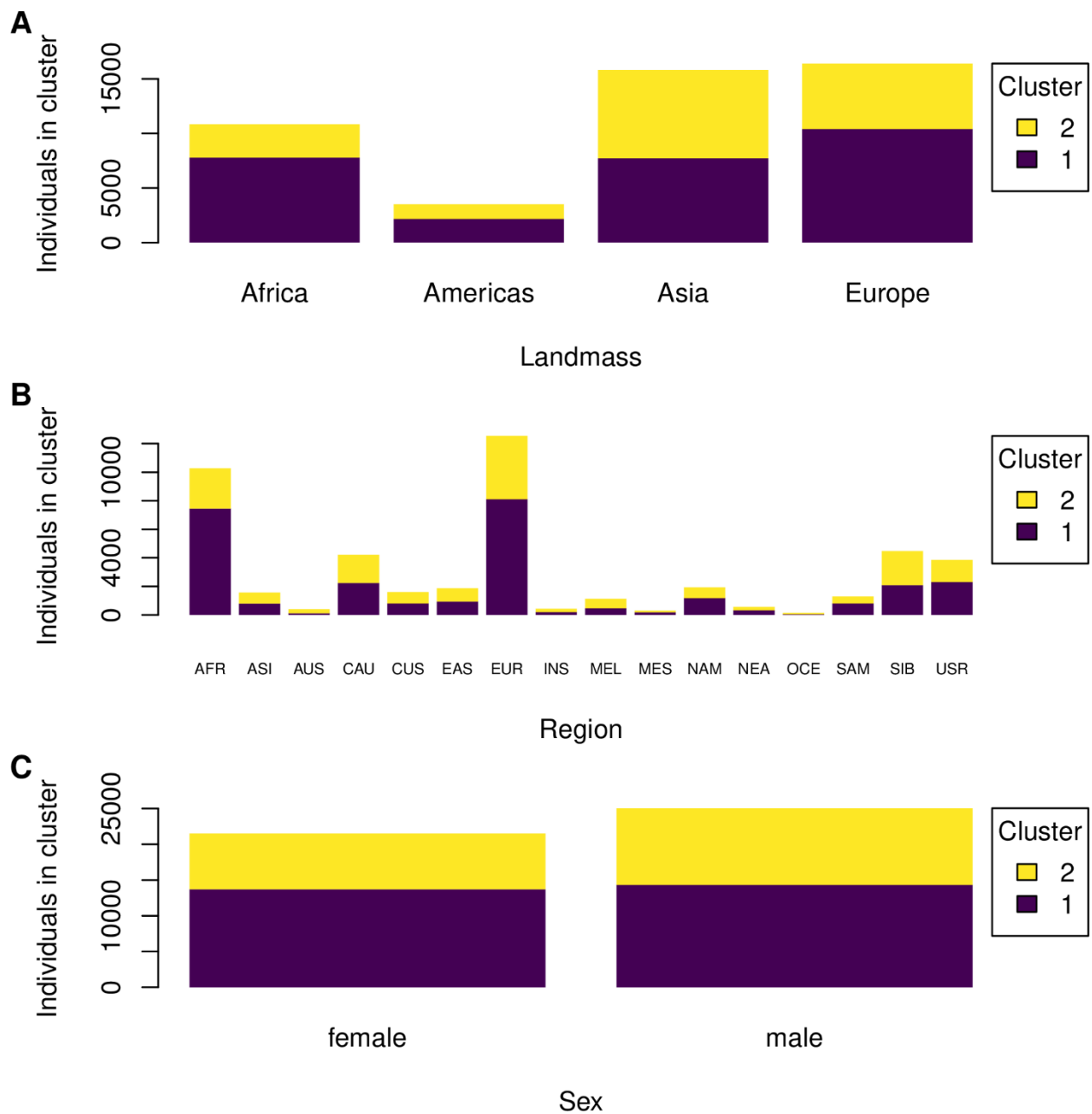


Figure 17. Stacked bar chart for $K = 2$ from the average silhouette method for finger ridge-counts grouped (A) by-Landmass, (B) by-Region, and (C) by-Sex

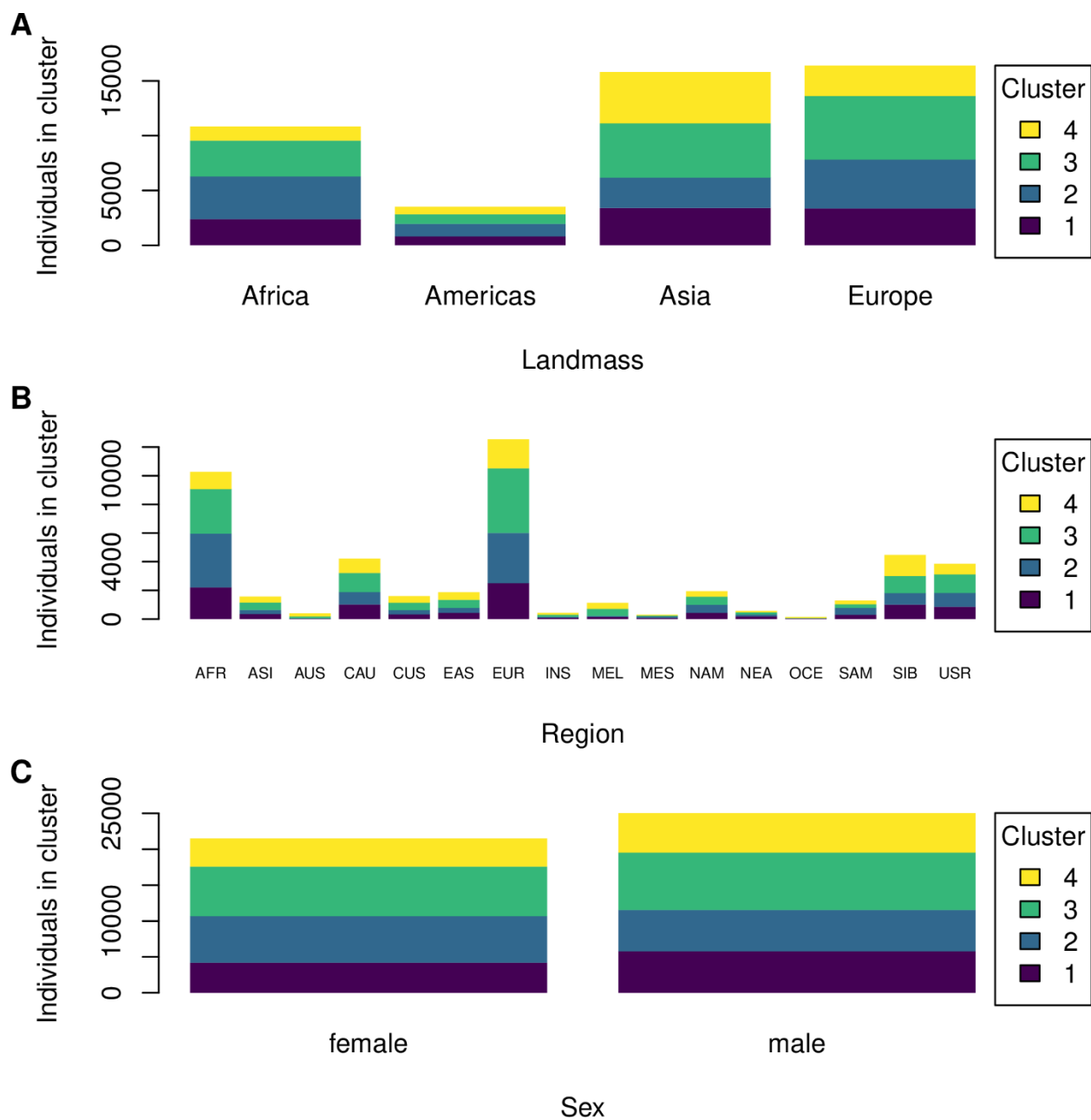


Figure 18. Stacked bar chart for $K = 4$ for finger ridge-counts grouped (A) by-Landmass, (B) by-Region, and (C) by-Sex

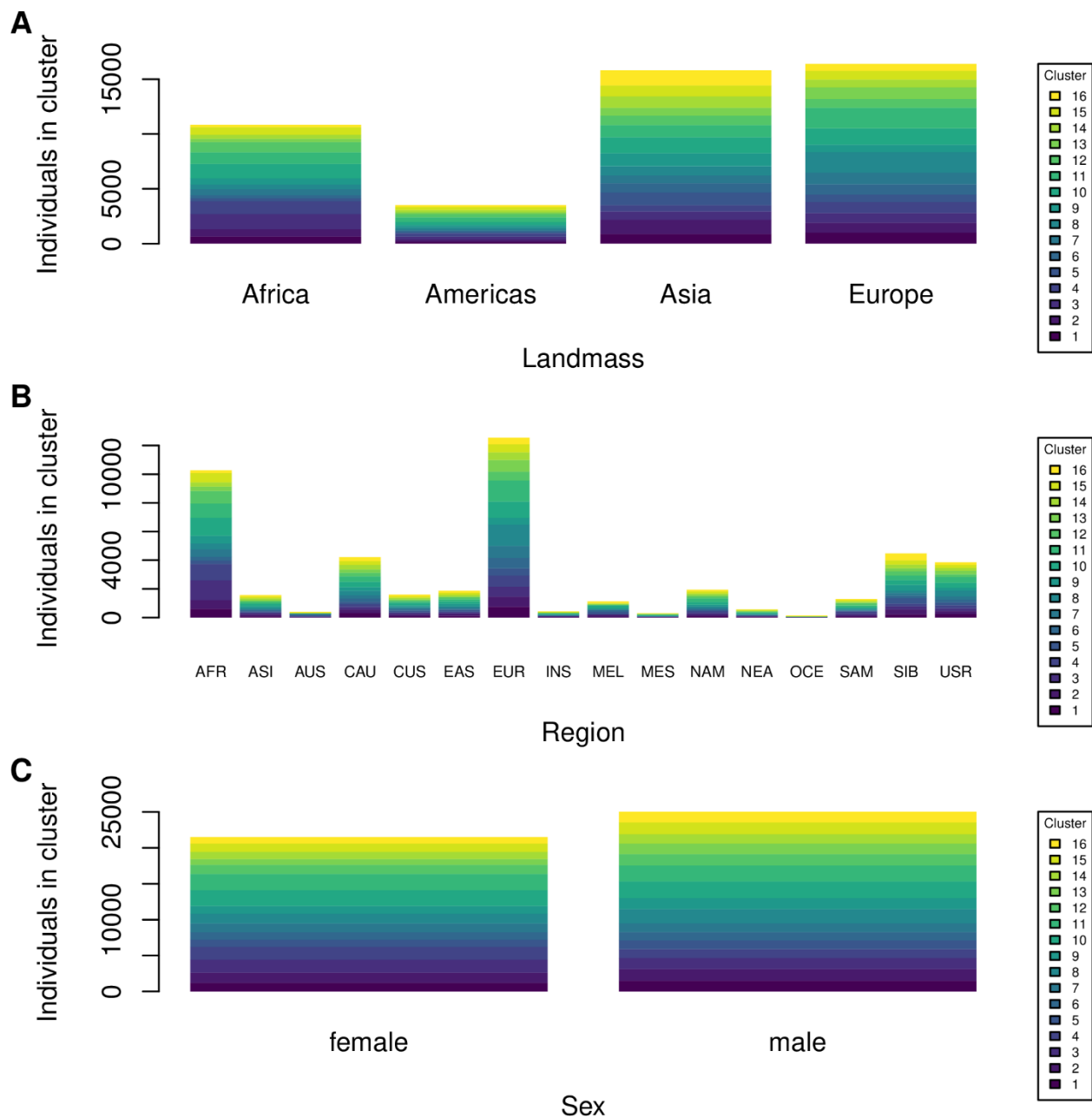


Figure 19. Stacked bar chart for $K = 16$ from the gap statistic method for finger ridge-counts grouped (A) by-Landmass, (B) by-Region, and (C) by-Sex

Table 9. Truncation distances for finger patterns and finger ridge-counts within Landmass; m = meters

Dermatoglyphic feature	Landmass	Sample Size	Truncation distance (m)
Finger patterns	Africa	11675	3067568
Finger ridge-counts	Africa	10806	3067568
Finger patterns	Americas	3887	8520334
Finger ridge-counts	Americas	3522	10702414
Finger patterns	Asia	18648	3051902
Finger ridge-counts	Asia	15813	3051902
Finger patterns	Europe	17056	1240025
Finger ridge-counts	Europe	16400	1240025

The dbMEM eigenvectors with positive spatial correlation were recorded in Table 10 as number of dbMEMs (distance-based Moran's eigenvector maps). Here, only the Americas shared the same number of dbMEMs at six for both finger patterns and finger ridge-counts, while the finger patterns for Landmasses Africa, Asia, and Europe had a higher number of dbMEMs than the finger ridge-counts for the same Landmass. The number of dbMEMs for Landmasses Africa and Europe were one higher for the finger patterns compared to the finger ridge-counts, but the Landmass Asia had 10 more dbMEMs for finger patterns than for finger ridge-counts. Despite the difference in number of dbMEMs across finger patterns and finger ridge-counts among Landmasses the adjusted R^2 values for these dbMEMs were similar by orders of magnitude, except for higher adjusted R^2 value (0.014) in Africa finger ridge-counts than the adjusted R^2 value (0.008) in Africa finger patterns.

The number of significant dbMEMs was identified through forward selection of the detrended response variables and the full dbMEM variables using the adjusted R^2 value as the threshold for selection, and Landmasses Americas and Europe contained the same number of significant dbMEMs and similar adjusted R^2 values for finger ridge-counts and finger patterns (Table 10). The adjusted R^2 values represent the percent variance explained by the dbMEM analysis of the detrended response variables. For Landmass Africa, the finger ridge-counts yielded a higher number of significant dbMEMs and higher adjusted R^2 value, while for Landmass Asia the finger patterns yielded a higher number of significant dbMEMs but similar adjusted R^2 value. Overall, the adjusted R^2 values for the significant dbMEMs were lower than the adjusted R^2 values for the dbMEMs before selection. For Landmass Europe the adjusted R^2 values reduced by over half after forward selection. The cumulative variance explained by the significant canonical axes out of the total variance is found in Table 11, and the plots

Table 10. Number of dbMEMs and the adjusted R^2 and the number of significant dbMEMs and the adjusted R^2 for finger patterns and finger ridge-counts within Landmass; dbMEMs = distance-based Moran's eigenvector maps, R^2_{adj} = adjusted R squared

Dermatoglyphic feature	Landmass	Number of dbMEMs	R^2_{adj}	Number of significant dbMEMs	R^2_{adj}	Significant dbMEMs
Finger patterns	Africa	18	0.008	2	0.004	8, 16
Finger ridge-counts	Africa	17	0.014	6	0.010	3, 7, 8, 9, 13, 15
Finger patterns	Americas	6	0.022	5	0.022	2, 3, 4, 5, 6
Finger ridge-counts	Americas	6	0.027	5	0.026	2, 3, 4, 5, 6
Finger patterns	Asia	16	0.008	3	0.004	1, 3, 5
Finger ridge-counts	Asia	6	0.006	2	0.004	1, 6
Finger patterns	Europe	23	0.001	1	0.001	20
Finger ridge-counts	Europe	22	0.002	1	0.001	2

Table 11. Variance explained by the significant canonical axes for finger patterns and finger ridge-counts. Where the R^2_{adj} is the variance explained by the dbMEM analysis of the detrended response variables, the proportion of variance in the significant axes includes only the significant axes with respect to the total explained variance. The variance explained by the significant axes is the product of the proportion of variance in significant axes and the R^2_{adj} ; R^2_{adj} = adjusted R squared

Dermatoglyphic feature	Landmass	Number of significant axes	Proportion of variance in significant axes	R^2_{adj}	Variance explained by significant axes
Finger patterns	Africa	1	0.947	0.004	0.003
Finger ridge-counts	Africa	2	0.949	0.011	0.010
Finger patterns	Americas	2	0.951	0.022	0.021
Finger ridge-counts	Americas	2	0.941	0.0262	0.024
Finger patterns	Asia	2	0.992	0.004	0.004
Finger ridge-counts	Asia	2	1.000	0.004	0.004
Finger patterns	Europe	1	1.000	0.001	0.001
Finger ridge-counts	Europe	1	1.000	0.001	0.001

representing the spatially structured variation in the detrended response variables are found in Figure S1 through Figure S8.

The significant canonical axes represent a combination of dbMEM variables ranging from broad to fine scale but without any way to discriminate between scales. Evaluating the relation of the significant canonical axes to the environmental explanatory variables through linear modeling created normally-distributed residuals for all dermatoglyphic feature types and Landmasses as indicated by the Anderson-Darling normality test with an $\alpha = 0.05$ threshold.

For the single significant axis for Landmass Africa finger patterns, only Sex male was not a significant explanatory variable of the spatially structured variation in the detrended finger patterns, and Ethnic Group Egyptians-Arab exhibited NA results (Figure S1). For the first and second significant axes for Landmass Africa finger ridge-counts, Ethnic Group Egyptians and Sex male were not significant explanatory variables of the spatially structured variation in the detrended finger ridge-counts, and Ethnic Group Egyptians-Arab exhibited NA results (Figure S2).

Both finger patterns and finger ridge-counts for Landmass Americas shared the same results. For the first and second significant axes, only Sex male was not a significant explanatory variable of the spatially structured variation in the detrended dermatoglyphic features, and Ethnic Groups Mixtecs, Yupa-Indians, and Longitude exhibited NA results (Figure S3, Figure S4).

For the first significant axis for Landmass Asia finger patterns, Ethnic Groups Balkars, Easter-Island, Garmend-Ayom, Hagahai, Ingush, Kabardinians, Swaiab-Ayom, and Sex male were not significant explanatory variables of the spatially structured variation in the detrended finger patterns. For the second significant axis for Landmass Asia finger patterns, Ethnic Groups Easter-Island, Garmend-Ayom, Hagahai, Karachay, Swaiab-Ayom, and Sex male were not

significant explanatory variables of the spatially structured variation in the detrended finger patterns. For both significant axes, Ethnic Groups Easter-Island-Polynesians, Rembranga, Timorese, Vietnamese, Yakuts, Yazgulyami, and Yuwamb-Ayom exhibited NA results (Figure S5). For the first significant axis for Landmass Asia finger ridge-counts, Ethnic Groups Easter-Island, Evens, and Sex male were not significant explanatory variables of the spatially structured variation in the detrended finger ridge-counts. For the second significant axis for Landmass Asia finger ridge-counts, Ethnic Groups Cherkess, Easter-Island, Hagahai, Kobon, and Sex male were not significant explanatory variables of the spatially structured variation in the detrended finger ridge-counts. For both significant axes, Ethnic Groups Easter-Island-Polynesians, Pinai, Rembranga, Timorese, Ulchis, Vietnamese, and Yaghnobis exhibited NA results (Figure S6).

For the single significant axis for Landmass Europe finger patterns, Region USSR, Ethnic Groups Aland-Mariehamm, Cornish, and Sex male were not significant explanatory variables of the spatially structured variation in the detrended finger patterns (Figure S7). For the single significant axis for Landmass Europe finger ridge-counts, Ethnic Groups Aland-Mariehamm, Poles, Romani-Gemer, Romani-Vinodol, Serbians, and Sex male were not significant explanatory variables of the spatially structured variation in the detrended finger ridge-counts (Figure S8).

Because the combination of the dbMEM variables plotted via the significant canonical axes (Figure S1 through Figure S8) does not allow discrimination between broad and fine scales of spatial patterning, maps of the significant dbMEM variables were created (Figure S9 through Figure S16). Based on the map of significant dbMEM variables for Landmass Africa finger patterns, dbMEM 8 was defined as broad scale patterning, and dbMEM 16 was defined as medium/fine scale patterning (Figure S9). These delineations were arbitrary as evidenced by the

similar appearance in the map for dbMEM 8 and 16, but for subsequent analyses some separation of scale was necessary. For Landmass Americas finger patterns, dbMEM 2 and 3 were defined at broad scale patterning, with dbMEM 4 and 5 as medium scale, and dbMEM 6 as fine scale (Figure S10). Despite outlying groups compressing the map, the patterning is still evident. For Landmass Asia finger patterns, dbMEM 1 and 3 were defined as broad scale patterning, and dbMEM 5 was defined as medium/fine scale patterning (Figure S11). Here, the outlying groups compressing the map do obscure the discrimination between scales. For Landmass Europe finger patterns, the single significant dbMEM 20 was assigned to represent broad scale patterning, so subsequent analyses for Landmass Europe finger patterns had no representation of fine scale patterning (Figure S12).

For the map of Landmass Africa finger ridge-counts, dbMEM 3 was defined as broad scale patterning, dbMEM 7, 8, and 9 as medium scale, and dbMEM 13 and 15 as fine scale (Figure S13). The map of Landmass Americas finger ridge-counts lead to dbMEM 2, 3, and 4 defined as broad scale patterning and dbMEM 5 and 6 as fine scale (Figure S14). Similar to the map of finger patterns for Landmass Americas, outlying groups appear to compress the map, but patterns of scale are still evident. For the Landmass Asia finger ridge-counts, dbMEM was defined as broad scale patterning, and dbMEM 6 was defined as fine scale (Figure S15). Similar to the map of finger patterns for Landmass Asia, the outlying groups compress the map and obscure the discrimination between the scales. For the map of Landmass Europe finger ridge-counts, the single significant dbMEM 2 was assigned to represent broad scale patterning, so subsequent variation-partitioning analyses for Landmass Europe finger ridge-counts had no representation of fine scale patterning (Figure S16). Compared to Landmass Europe finger

patterns, the map of Landmass Europe finger ridge-counts does appear to actually represent broad scale patterning with clear delineation between groupings.

The results of variation partitioning should only be interpreted if the pure environmental, pure linear trend, pure broad spatial scale, and pure fine spatial scale components were significant at the $\alpha = 0.05$ threshold. The four components were significant at the $\alpha = 0.05$ threshold for the majority of Landmasses and dermatoglyphic features. However, the Landmass Europe finger patterns and Landmass Africa, Americas, and Europe finger ridge-counts pure linear trend component contained no constrained component so did not calculate any significance, and Landmass Americas finger ridge-counts pure broad scale component was not significant at the $\alpha = 0.05$ threshold. The majority of variation of the undetrended response variables was not captured by the fractions considered during variation partitioning, as evidenced by the high residual values (Figure 20). Note the total of the partitions including the residual value sometimes exceeds one, due to the variation partitioning protocol subtracting of R^2 values that were adjusted based on a different number of explanatory variables. Any sum of partitions was reached by summing the actual values in each partition rather than subtracting other partitions from the value one.

The amount of spatially-structured variation captured by the variation partitioning was 0.7% for Landmass Africa finger patterns, 1.9% for Landmass Africa finger ridge-counts, 3% for Landmass Americas finger patterns, 4.6% for Landmass Americas finger ridge-counts, 1.6% for Landmass Asia finger patterns, 1.3% for Landmass Asia finger ridge-counts, 0.8% for Landmass Europe finger patterns, and 0.6% for Landmass Europe finger ridge-counts. Very little variation captured by variation partitioning was spatially structured but not environmental: 0.3% for

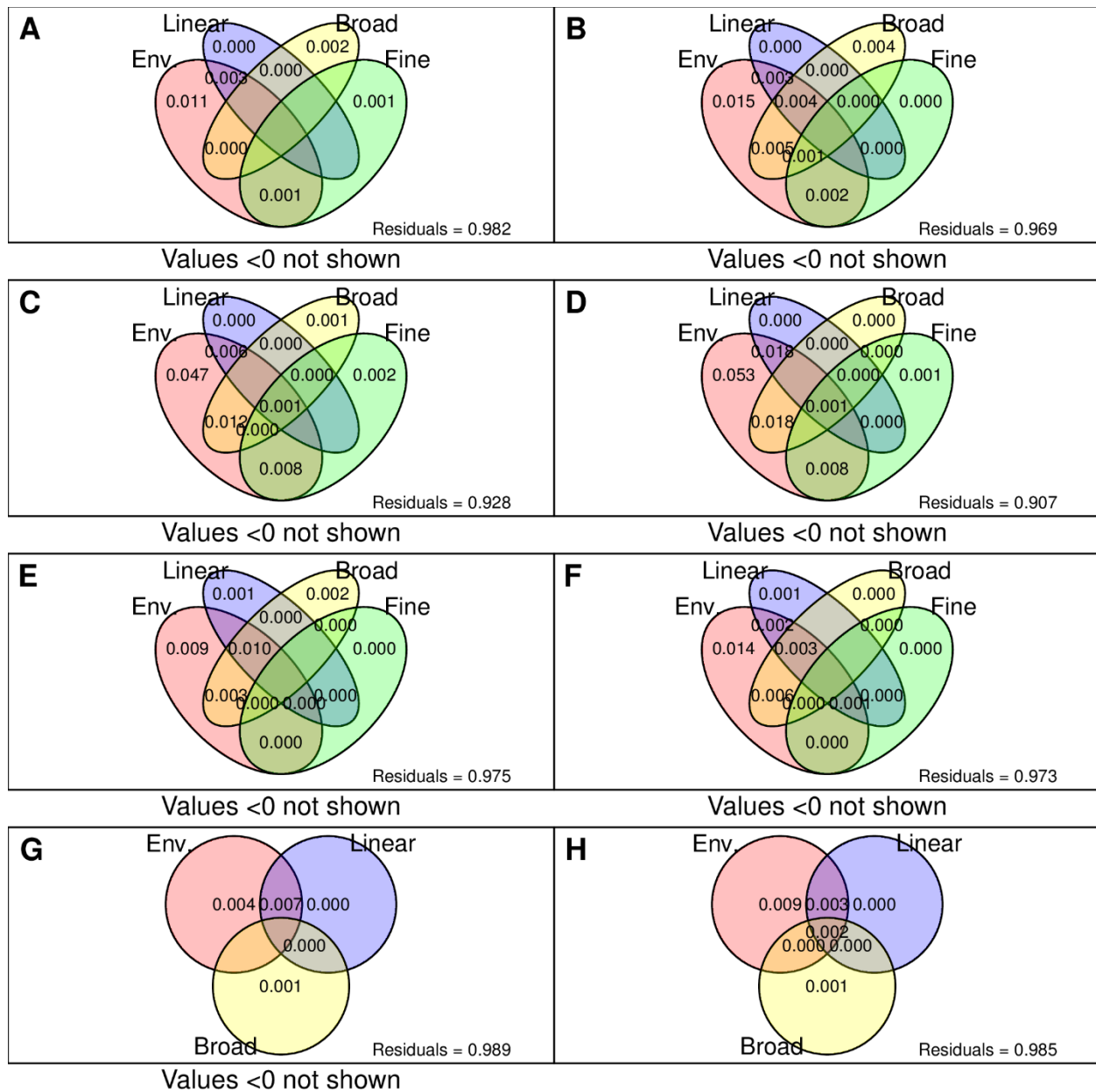


Figure 20. Variation partitioning of undetrended finger pattern and finger ridge-counts response variables into an environmental component, a linear trend, a broad scale, and a fine scale for (A) Landmass Africa finger patterns, (B) Landmass Africa finger ridge-counts, (C) Landmass Americas finger patterns, (D) Landmass Americas finger ridge-counts, (E) Landmass Asia finger patterns, (F) Landmass Asia finger ridge-counts, (G) Landmass Europe for finger patterns, and (H) Landmass Europe finger ridge-counts; Env. = environmental

Landmass Africa finger patterns, 0.4% for Landmass Africa finger ridge-counts, 0.3% for Landmass Americas finger patterns, 0.1% for Landmass Americas finger ridge-counts, 0.3% for Landmass Asia finger patterns, and 0.1% for Landmass Asia finger patterns as well as Landmass Europe finger patterns and finger ridge-counts.

The majority of variation for both Landmass Africa and Americas finger patterns and finger ridge-counts considered in the variation partitioning was not spatially structured, that is the purely environmental component. For example, 1.1% of the variation captured in the undetrended Landmass Africa finger patterns is just due to the relationship between individuals and the environment (Figure 20A), and 1.5% of the variation in the undetrended Landmass Africa finger ridge-counts is purely environmental (Figure 20B). The highest portion of variation captured by the variation partitioning at all and as purely environmental was in the Landmass Americas with finger ridge-counts 5.3% (Figure 20D) and finger patterns 4.7% (Figure 20C). The highest portion of spatially structured environmental variation was also found in Landmass Americas in finger ridge-counts as 1.8%, 1.8%, and 0.8% for a total of 4.4%. The percent of variation explained by environmental and linear trends is 1.8%, by environmental and broad scale trends is 1.8%, and by environmental and fine scale trends is 0.8%. Similarly high are the portions of spatially structured environmental variation in Landmass Americas finger patterns as 0.8%, 1.2%, and 0.8% for a total of 2.8%, where the percent of variation explained by environmental and linear trends is 0.8%, by environmental and broad scale trends is 1.2%, and by environmental and fine scale trends is 0.8%.

4.1.3 Principal component analysis (PCA)

The loadings of the principal components illustrated the importance of the finger ridge-counts, and by visualizing the loadings of the first six principal components the finger ridge-

count variables that contributed the most to each component were apparent (Figure 21). To interpret the loadings that ranged from -1 to 1, a higher absolute value indicated the finger ridge-count variable more heavily influenced the component than an absolute value close to zero. For PC1, all finger ridge-counts exhibit negative values indicating a negative correlation of the finger ridge-counts with PC1 (Figure 21A). PC2 captured the positive correlation of digit 1 against all other finger ridge-counts (Figure 21B). PC3 captured the positive correlation of ulnar ridge-counts against the radial ridge-counts (Figure 21C). PC4 exhibited a higher positive correlation for the ulnar counts of digit 2 on both left and right sides and equally negative correlation for the left ulnar count of digit 4 (Figure 21D). PC5 exhibited higher positive correlation for the radial ridge-counts on the left and right sides for digits 4 and 5 and a negative correlation for the right radial count of digit 2 (Figure 21E). PC6 captured a higher positive correlation for the left and right ridge radial counts of digit 1 with a negative correlation for the right radial count of digit 2 (Figure 21F). The Third ridge-count exhibited loadings at value zero for all six principal components, indicating no correlation.

The first principal component explained 46.3% of the variation, and the second principal component explained 9.2% of the variation in the finger ridge-counts for the full data set. Visualizing the biplot of PC1 vs PC2 for only the centroids of different groupings illustrated Landmass Asia as separated from the other three Landmasses but only Regions Australia, Melanesia, and Oceania, which were in Landmass Asia, were separated from all other Regions across all other Landmasses (Figure 22). Landmass Americas appeared to exhibit more variation than the other three Landmasses (Figure 22A). The centroids for finger ridge-counts of females and males in the full data set were entirely separate, as were the error bars representing the standard error on the PC1 axis and the PC2 axis (Figure 22C). Although separate, females and

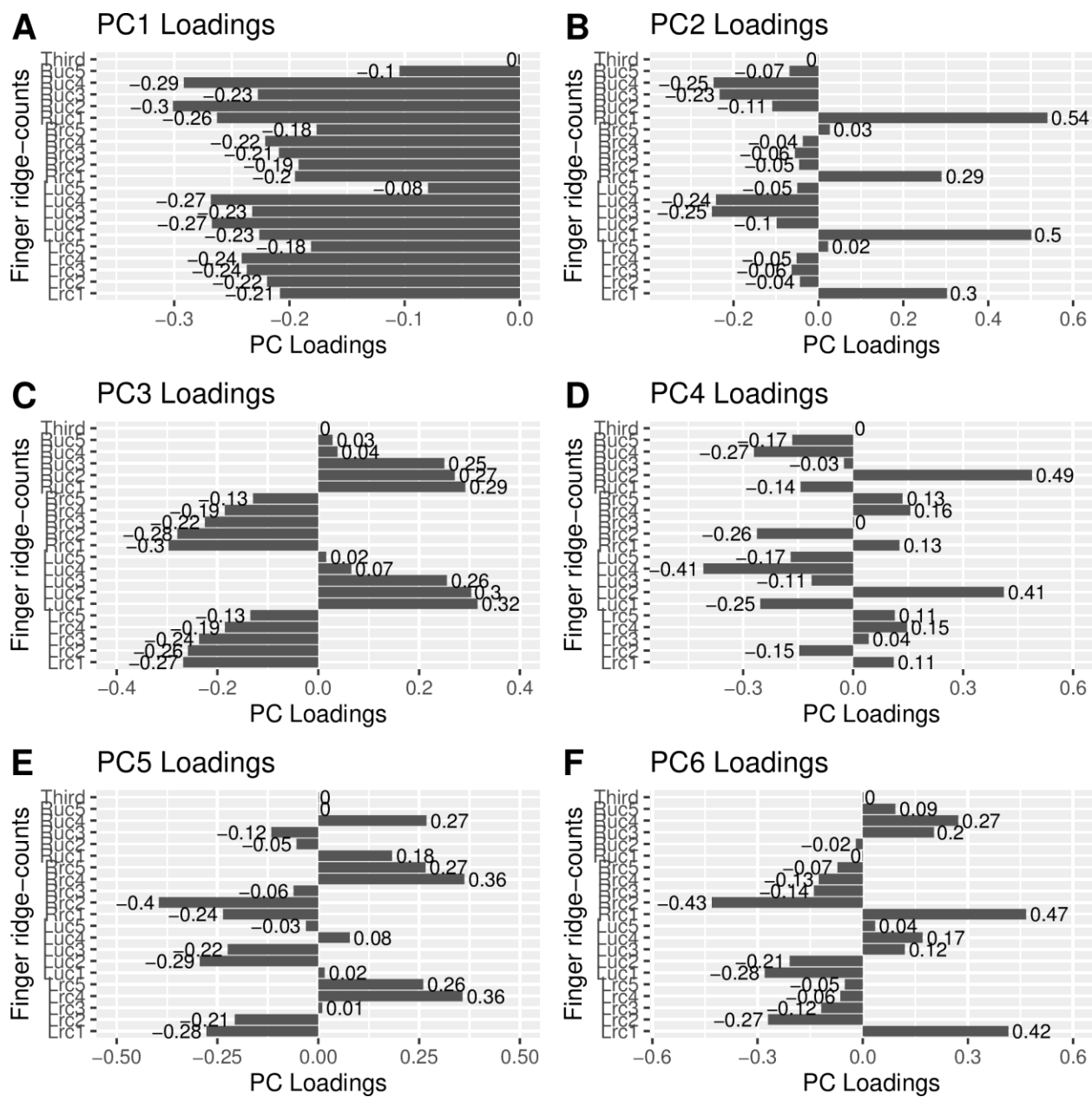


Figure 21. PC 1-6 loadings from PCA of covariance matrix of finger ridge-counts. (N = 46541)

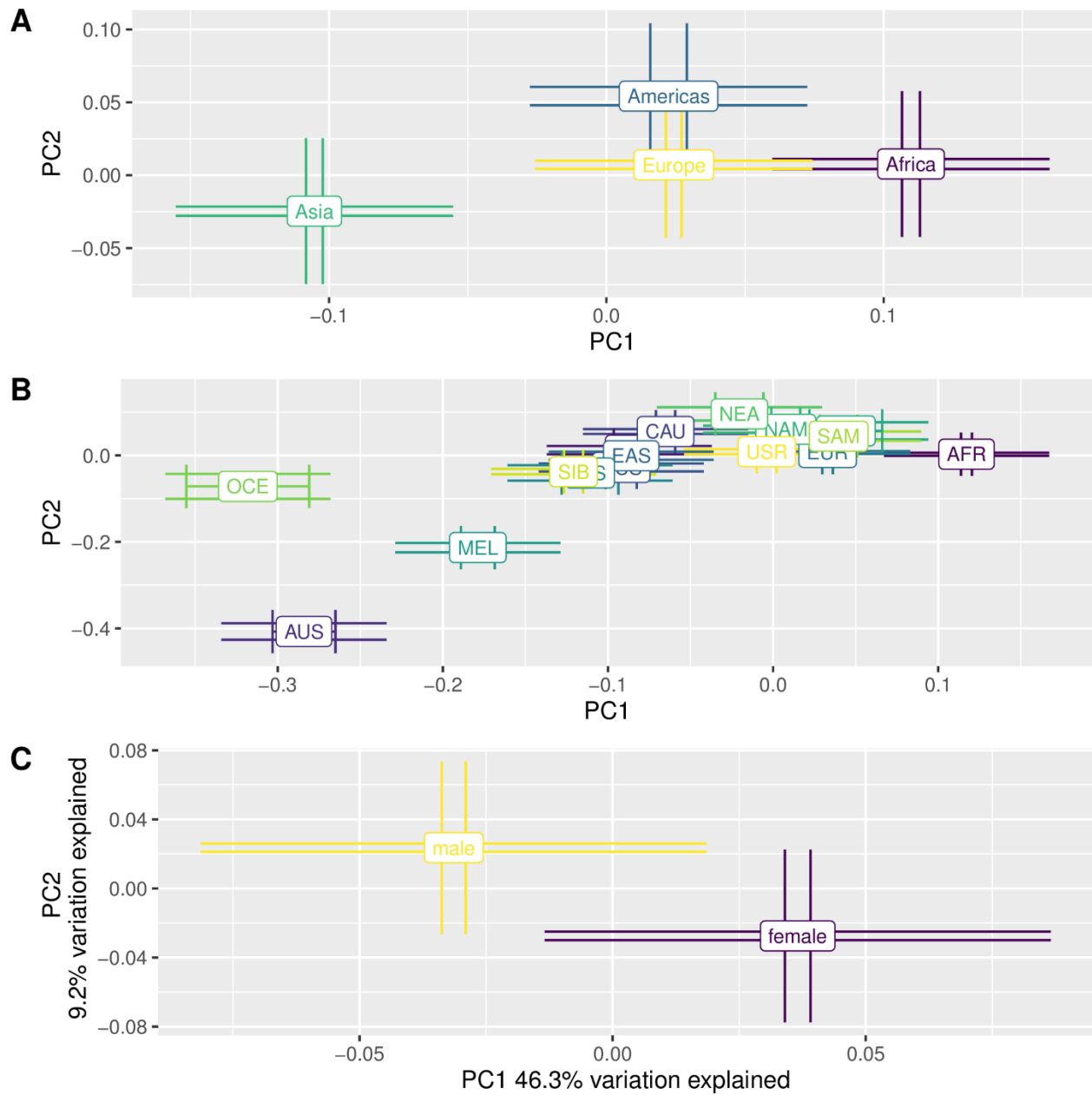


Figure 22. Principal component analysis of covariance matrix of finger ridge-counts. Viewed by different grouping centroids: (A) by-Landmass, (B) by-Region, and (C) by-Sex (N = 46541).

Error bars indicate the standard error.

males appeared to share a similar amount of variation.

When the full data set was split by-Sex, the loadings of the principal components illustrated the differences between females and males of finger ridge-counts (Figure 23, Figure 24). For PC1, the finger ridge-counts of both females and males exhibited negative values indicating a negative correlation of the finger ridge-counts of females and males with PC1 (Figure 23A, Figure 24A). PC2-6 exhibited the same trends but opposite values of similar value for females and males. For example, PC2 captured the correlation of digit 1 against all other finger ridge-counts, and in females the digit 1 values were negative and the other digits were positive, and in males the digit 1 values were positive and the other digits were negative (Figure 23B, Figure 24B). The magnitude and direction of loadings of PC1-6 for male finger ridge-counts matched the pooled finger ridge-counts containing both females and males (Figure 21, Figure 24). As further illustration of the opposing PC loading values for female and male finger ridge-counts, the biplot of finger ridge-counts by-Landmass for females and males exhibited similar organization along PC1 but opposite organization along PC2 (Figure 25A, Figure 25C). For the finger ridge-counts by-Region for females and males, the organization along PC1 was similar but along PC2 was opposite (Figure 25B, Figure 25D). In females, the first principal component explained 46.2% of the variation in the finger ridge-counts, and the second principal component explained 9.1% of the variation in finger ridge-counts (Figure 25A, Figure 25B). In males, the first principal component explained 46.4% of the variation in finger ridge-counts, and the second principal component explained 9.3% of the variation in finger ridge-counts (Figure 25C, Figure 25D).

When the full data set was split by-Landmass, the loadings of the principal components illustrated the importance of the finger ridge-counts in the each Landmass separately.

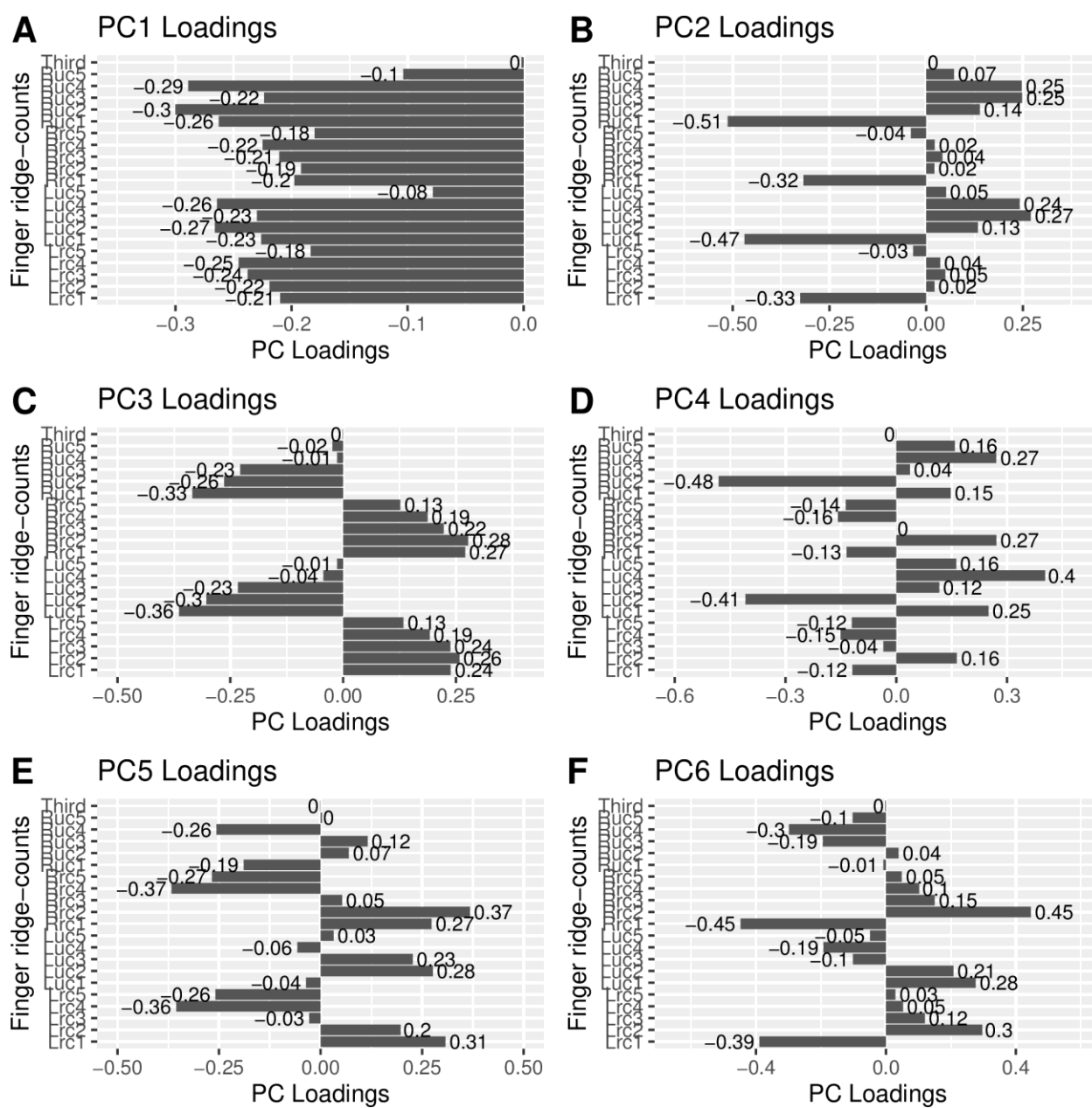


Figure 23. PC 1-6 loadings from PCA of covariance matrix of female finger ridge-counts. (N = 21641)

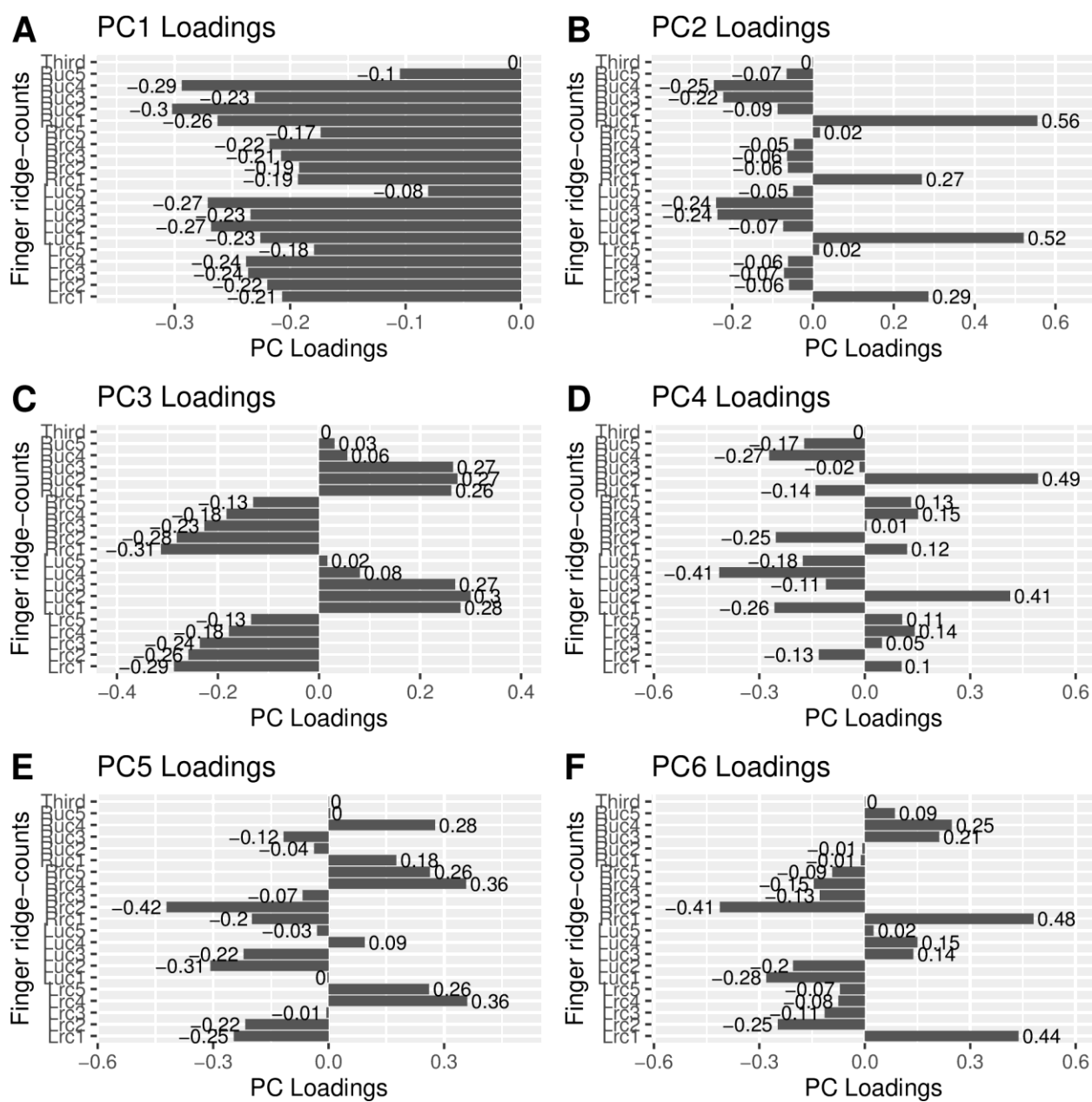


Figure 24. PC 1-6 loadings from PCA of covariance matrix of male finger ridge-counts. (N = 24864)

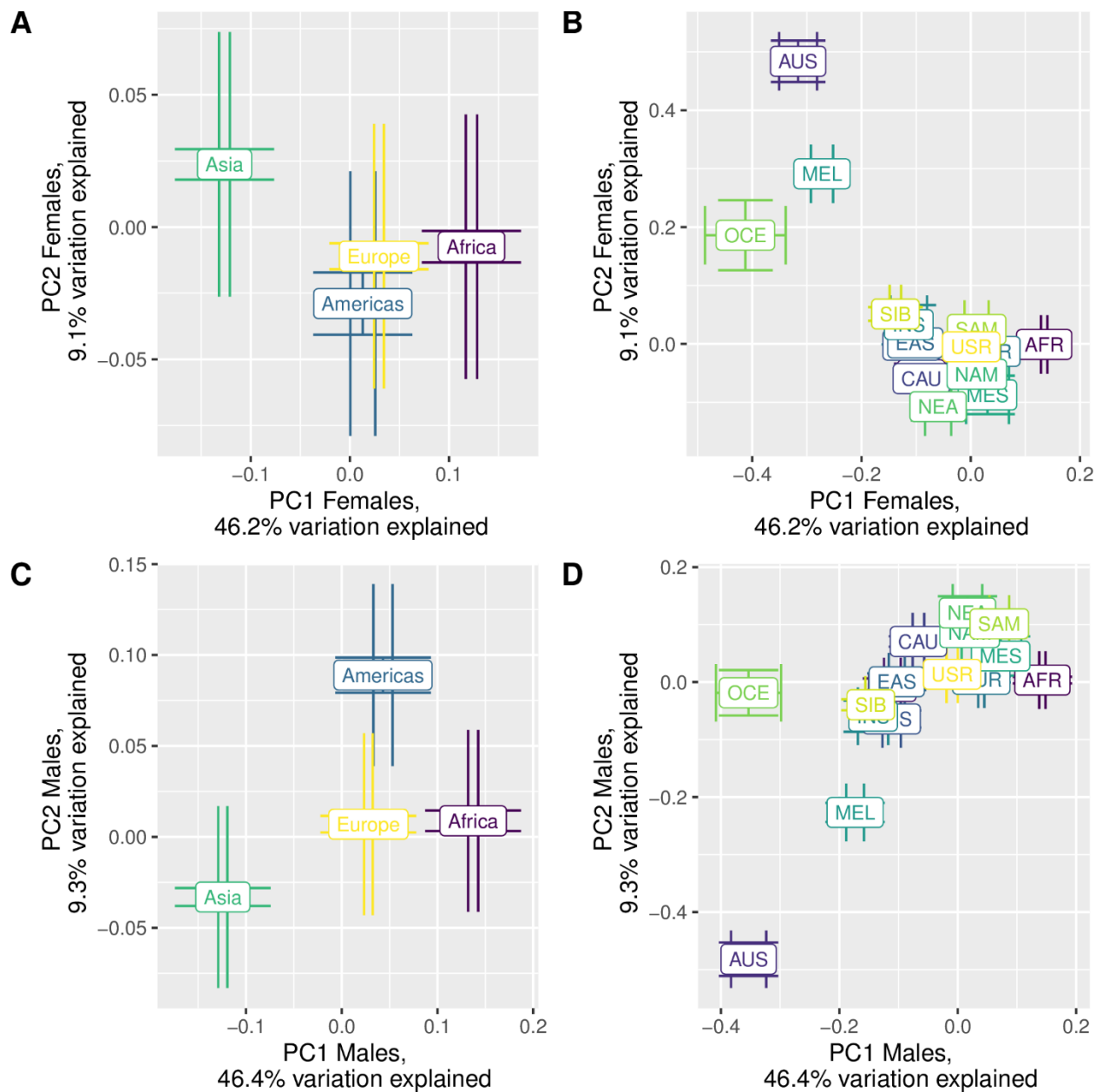


Figure 25. Principal component analysis of covariance matrix of female and male finger ridge-counts. Viewed by different grouping centroids: (A) females by-Landmass, (B) females by-Region, (C) males by-Landmass, and (D) males by-Region (N = 21641 females, 24864 males).

Error bars indicate the standard error.

For PC1, the finger ridge-counts of Landmass Africa exhibited positive values indicating a positive correlation of the finger ridge-counts of Landmass Africa with PC1 (Figure 26A). PC2 captured the negative correlation of digit 1 against all other finger ridge-counts of Landmass Africa (Figure 26B). PC3 captured the positive correlation of radial ridge-counts against the ulnar ridge-counts of Landmass Africa (Figure 26C). PC4 exhibited a higher positive correlation for the ulnar ridge-counts of digit 4 on both left and right sides and equally negative correlation for the left ulnar count of digit 2 (Figure 26D). PC5 exhibited higher positive correlation for the radial ridge-counts on the left and right sides for digit 1 and a negative correlation for the radial ridge-counts on the left and right sides for digits 4 and 5 (Figure 26E). PC6 captured a higher positive correlation for the left and right radial counts for digit 2 with a negative correlation for the right ulnar count for digit 4 (Figure 26F). The Third ridge-count exhibited loadings at value zero for all six principal components, indicating no correlation. The loadings for PC1-4 were similar in magnitude but opposite in direction of correlation between the finger ridge-counts in the full data set versus in Landmass Africa only (Figure 21, Figure 26).

The first principal component explained 46% of the variation, and the second principal component explained 10% of the variation in the finger ridge-counts of Landmass Africa. Visualizing the biplot of PC1 versus PC2 for only the centroids of different groupings illustrated separation of Region Africa from Region Northeast Africa along the PC1 axis but not the PC2 axis (Figure 27). Region Northeast Africa appeared to exhibit more variation than Region Africa. While the centroids for finger ridge-counts of females and males in Landmass Africa were separate, the error bars representing the standard error still overlapped on the PC2 axis (Figure 27B). Females and males of Landmass Africa appeared to exhibit a similar amount of variation.

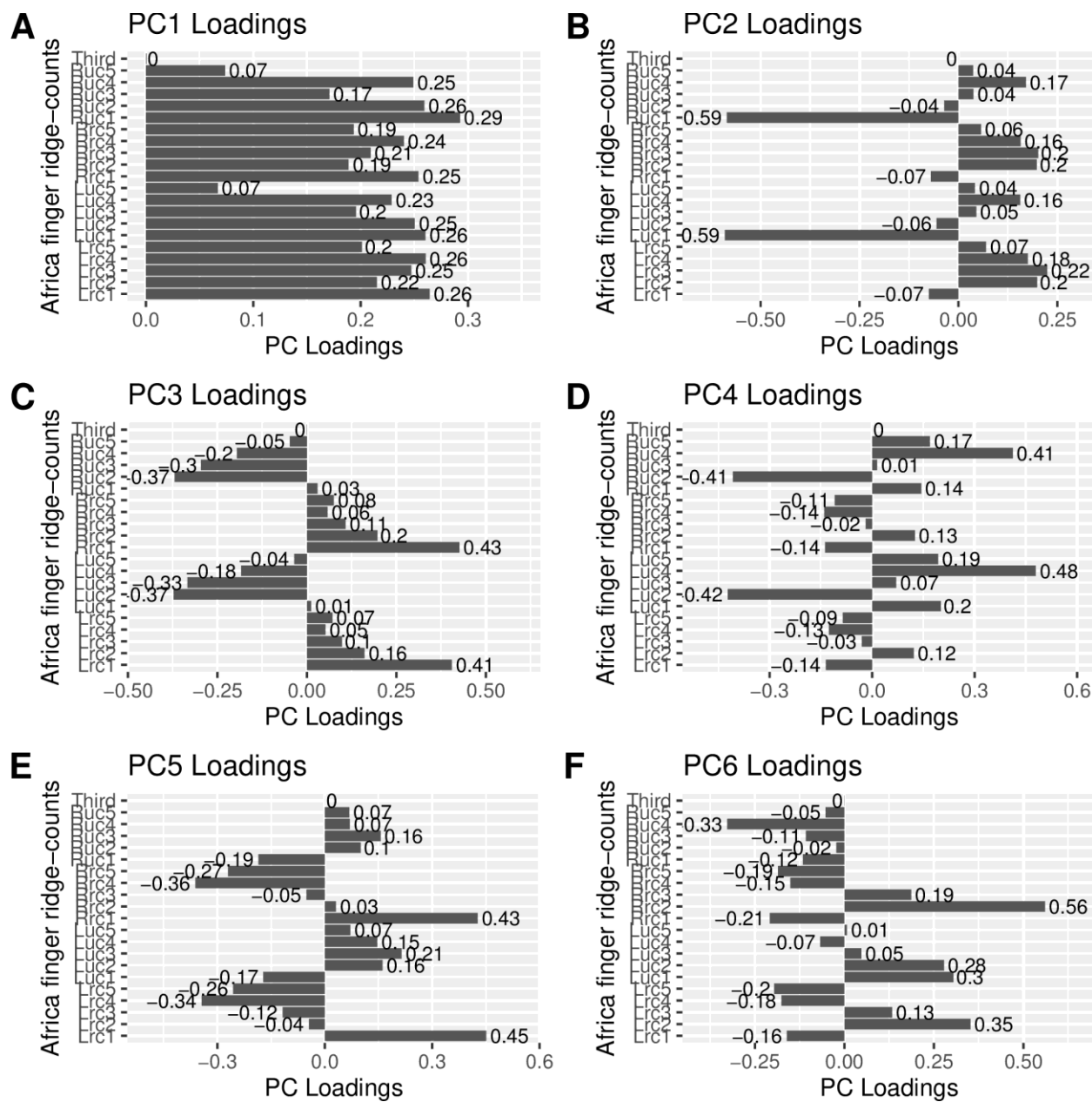


Figure 26. PC 1-6 loadings from PCA of covariance matrix of Landmass Africa finger ridge-counts. (N = 10806)

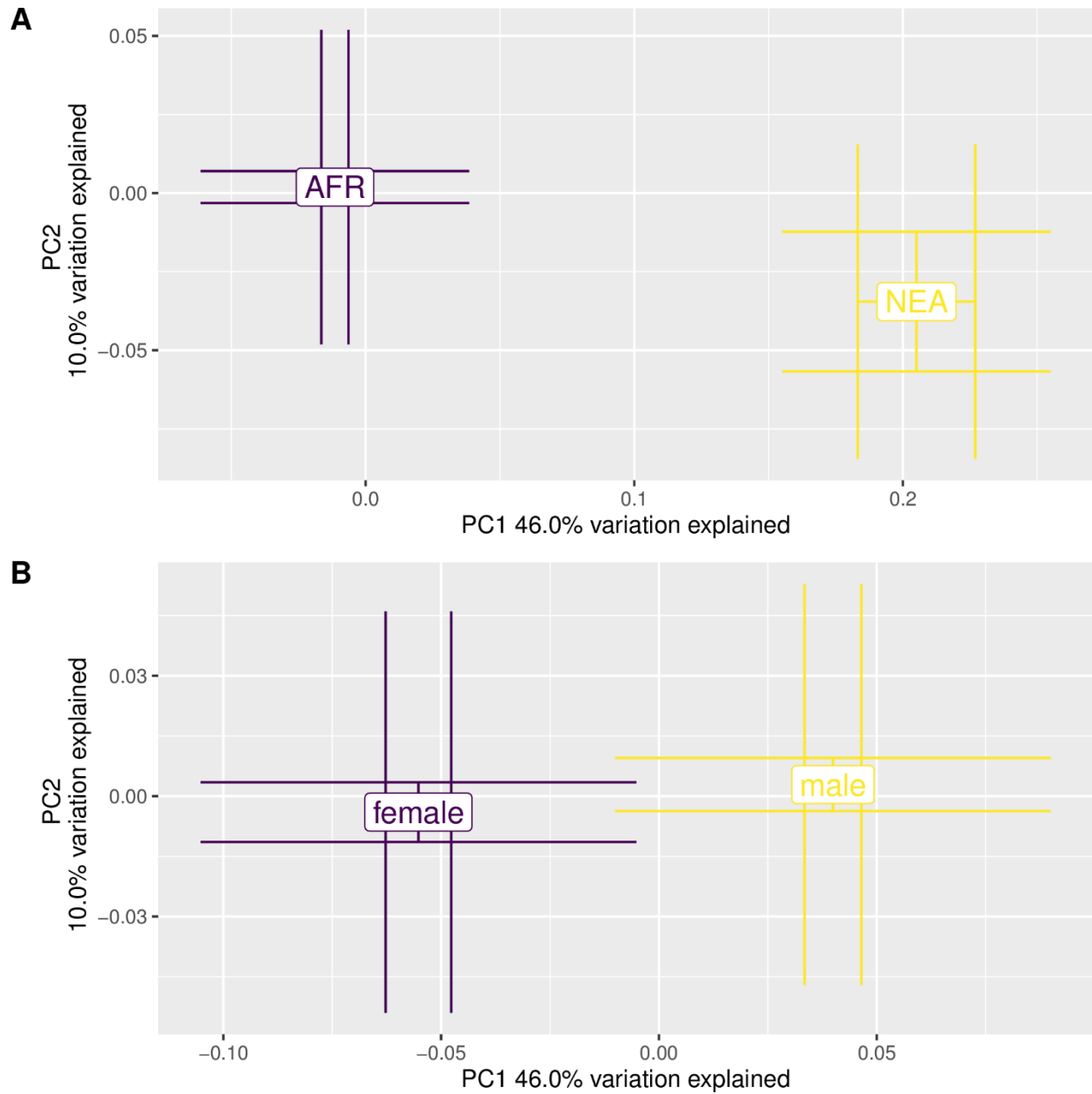


Figure 27. Principal component analysis of covariance matrix of Landmass Africa finger ridge-counts. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 10806).

Error bars indicate the standard error.

Further splitting the finger ridge-counts of Landmass Africa by-Sex, the two Regions exhibited opposing positions and non-overlapping separation along the PC1 axis in the biplots for females and males. The first principal component for females exhibited 46.5% of the variation in finger ridge-counts of Landmass Africa, and the second principal component for females exhibited 9.9% of the variation in finger ridge-counts of Landmass Africa (Figure 28A). The first principal component for males exhibited 45.5% of the variation in finger ridge-counts of Landmass Africa, and the second principal component for males exhibited 10.2% of the variation in finger ridge-counts of Landmass Africa (Figure 28B).

For PC1, the finger ridge-counts of Landmass Americas exhibited positive values indicating a positive correlation of the finger ridge-counts of Landmass Americas with PC1 (Figure 29A). PC2 captured the positive correlation of digit 1 against all other finger ridge-counts of Landmass Americas (Figure 29B). PC3 captured the positive correlation of ulnar ridge-counts against the radial ridge-counts of Landmass Americas (Figure 29C). PC4 exhibited a higher positive correlation for the ulnar counts of digit 2 on both left and right sides and similarly negative correlation for the left ulnar count of digit 4 and for the right radial count of digit 2 and to a lesser extent a negative correlation for the right ulnar count of digit 4 (Figure 29D). PC5 exhibited higher positive correlation radial counts on the left and right sides for digits 4 and 5 and a negative correlation for the radial count on the left and right sides for digit 1, the ulnar ridge-counts on the left and right sides for digit 2, and the right radial count of digit 2 (Figure 29E). PC6 captured a higher positive correlation for the left and right radial counts of digit 1 as well as the right ulnar count for digit 4 with a negative correlation for the right radial count of

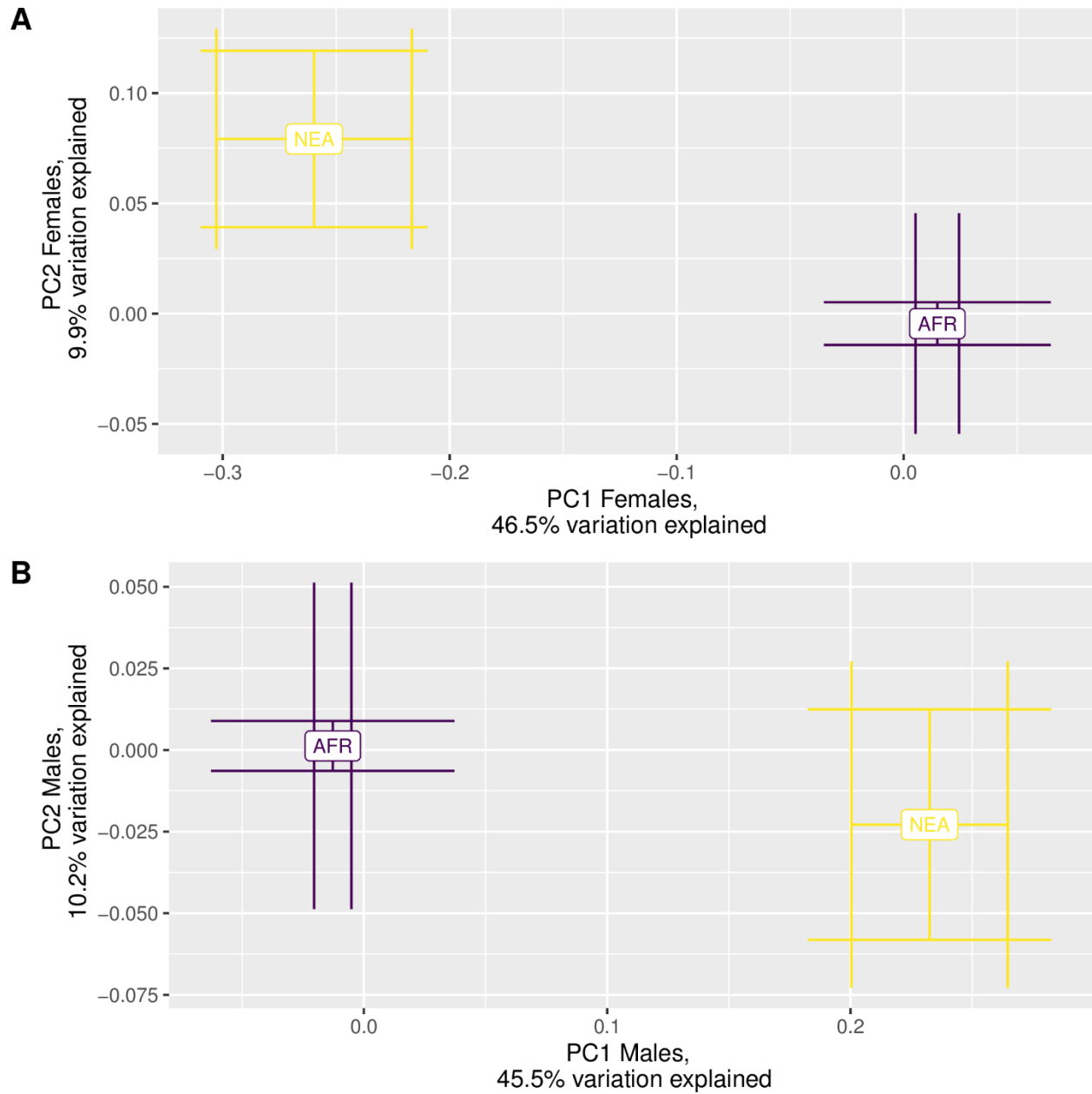


Figure 28. Principal component analysis of covariance matrix of female and male Landmass Africa finger ridge-counts. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 4537 females, 6269 males). Error bars indicate the standard error.

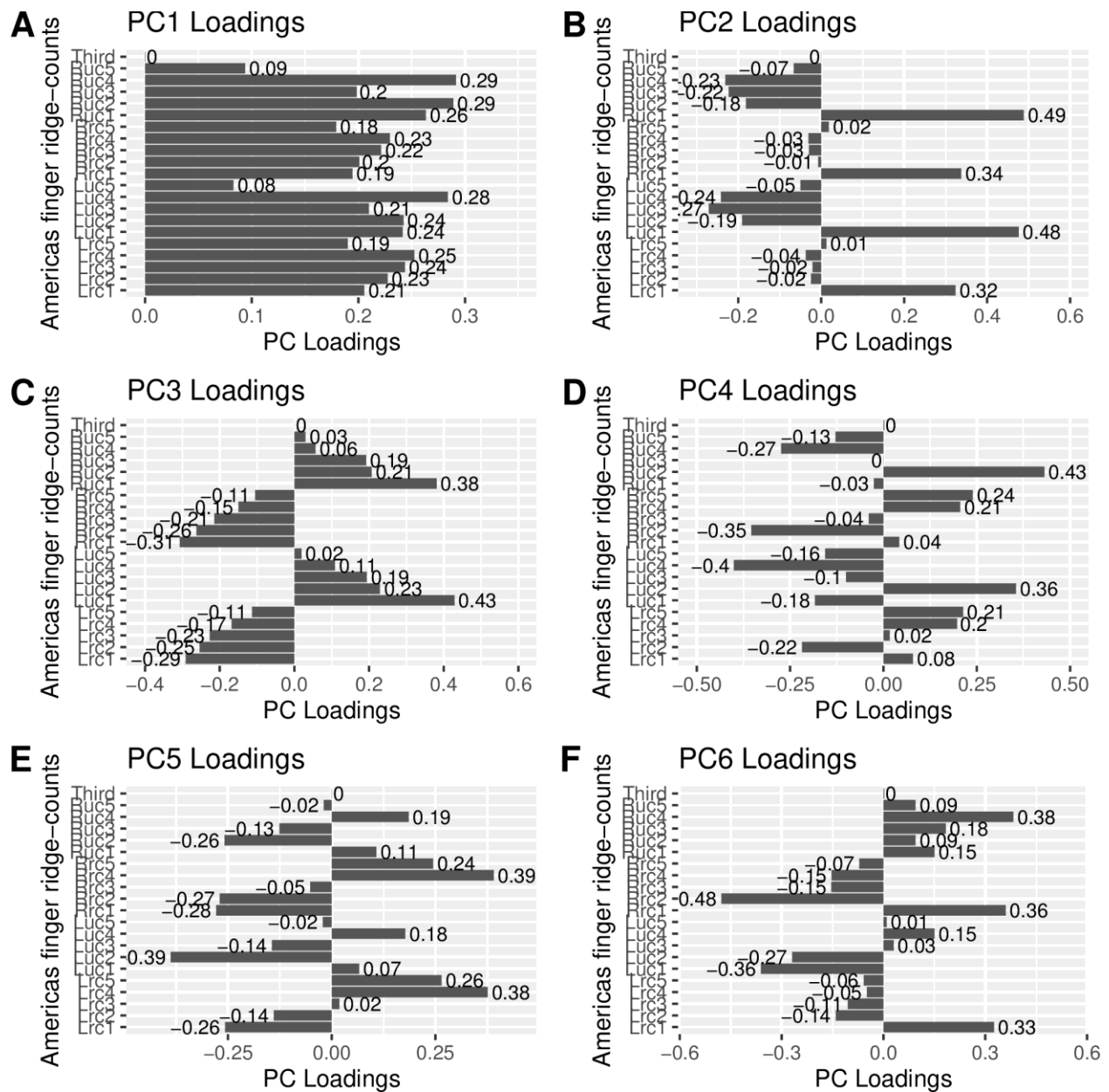


Figure 29. PC 1-6 loadings from PCA of covariance matrix of Landmass Americas finger ridge-counts. (N = 3522)

digit 2 and the left ulnar counts for digits 1 and 2 (Figure 29F). The Third ridge-count exhibited loadings at value zero for all six principal components, indicating no correlation. The loadings for PC1-3 were similar in magnitude and direction of correlation between the finger ridge-counts in the full data set and Landmass Americas only (Figure 21, Figure 29).

The first principal component explained 47.9% of the variation, and the second principal component explained 8.7% of the variation in the finger ridge-counts of Landmass Americas. Visualizing the biplot of PC1 vs PC2 for only the centroids of different groupings illustrated little separation of Region Mesoamerica from Region North America or Region South America along the PC1 and PC2 axes (Figure 30). Region Mesoamerica appeared to exhibit more variation with Region North America exhibiting less variation. The centroids for finger ridge-counts of females and males in the full data set were separate with no overlap on the PC1 or PC2 axes (Figure 30B).

Further splitting the finger ridge-counts of Landmass Americas by-Sex, the Regions North America and South America exhibited similar positions in the biplots for females and males along PC1 but opposing along PC2. Region Mesoamerica exhibited the greatest difference between females and males along PC1 with positive values for females and negative values for males. The first principal component for females exhibited 49.0% of the variation in finger ridge-counts of Landmass Americas, and the second principal component for females exhibited 8.3% of the variation in finger ridge-counts of Landmass Americas (Figure 31A). The first principal component for males exhibited 46.7% of the variation in finger ridge-counts of Landmass Americas, and the second principal component for males exhibited 9.1% of the variation in finger ridge-counts of Landmass Americas (Figure 31B).

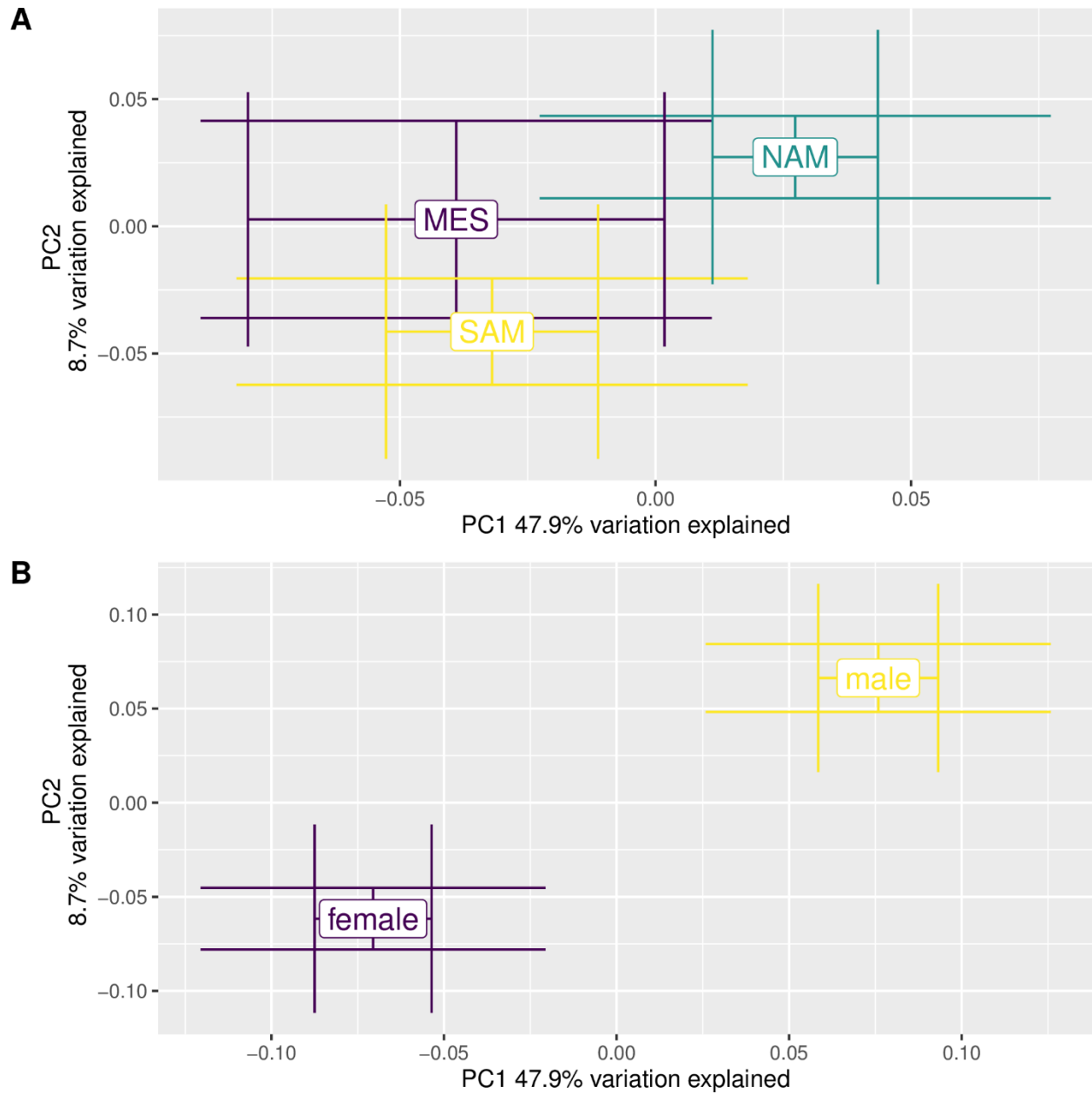


Figure 30. Principal component analysis of covariance matrix of Landmass Americas finger ridge-counts. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 3522). Error bars indicate the standard error.

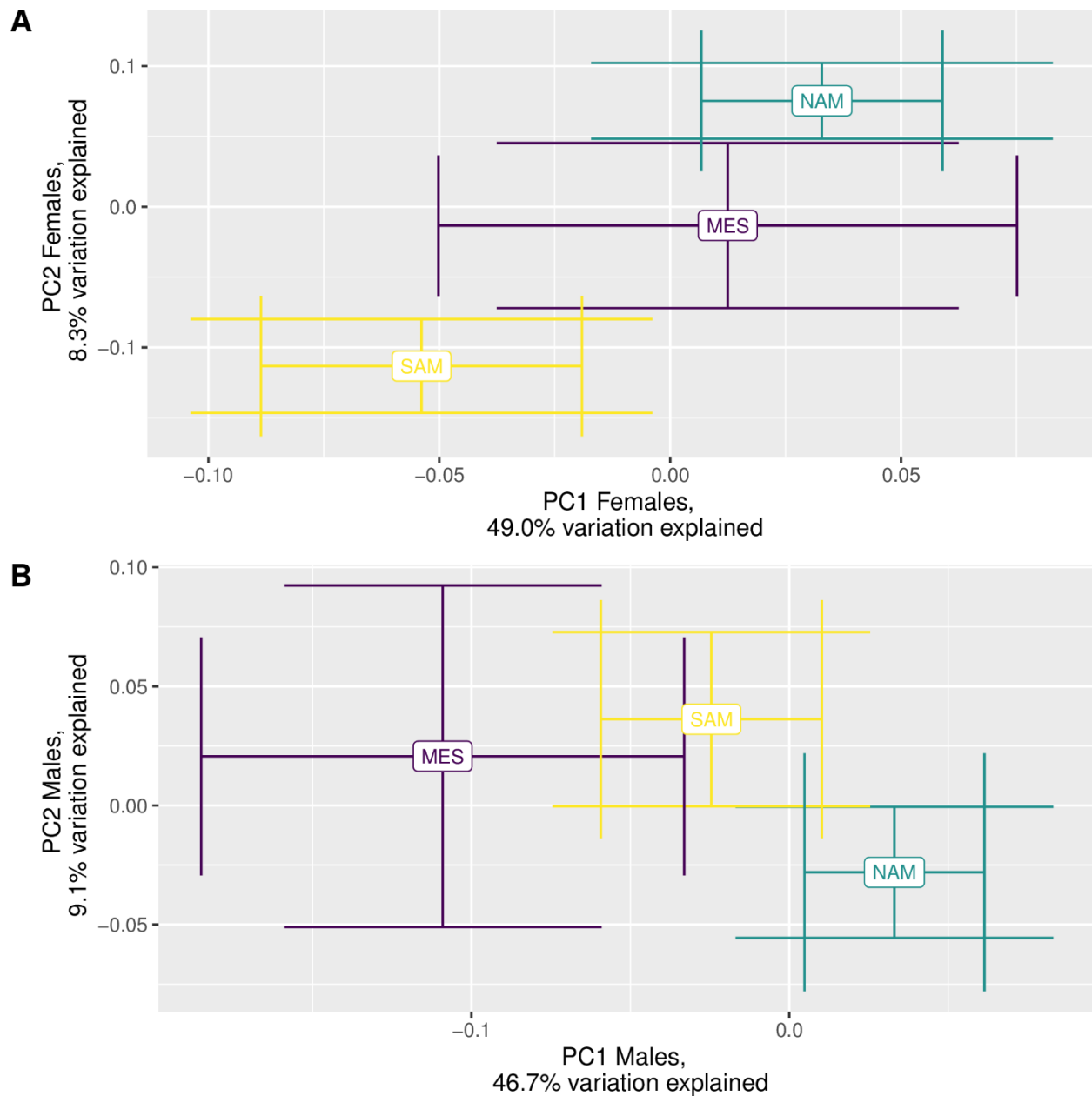


Figure 31. Principal component analysis of covariance matrix of female and male Landmass Americas finger ridge-counts. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 1825 females, 1697 males). Error bars indicate the standard error.

For PC1, the finger ridge-counts of Landmass Asia exhibited negative values indicating a negative correlation of the finger ridge-counts of Landmass Asia with PC1 (Figure 32A). PC2 captured the positive correlation of digit 1 against all other finger ridge-counts of Landmass Asia (Figure 32B). PC3 captured the positive correlation of ulnar ridge-counts against the radial ridge-counts of Landmass Asia (Figure 32C). PC4 exhibited a higher positive correlation for the ulnar counts of digit 2 on both left and right sides and similarly negative correlation for the left ulnar count of digit 4 (Figure 32D). PC5 exhibited higher positive correlation for radial ridge-counts on the left and right sides for digits 4 and 5 and to a lesser extent the ulnar count for digit 4 on the left and right sides and a negative correlation for the right radial count for digit 2 (Figure 32E). PC6 captured a higher positive correlation for the right ulnar counts of digit 3 with a negative correlation for the left ulnar count of digit 2 (Figure 32F). The Third ridge-count exhibited loadings at value zero for all six principal components, indicating no correlation. The loadings for PC1-4 were similar in magnitude and direction of correlation between the finger ridge-counts in the full data set and Landmass Asia only (Figure 21, Figure 32).

The first principal component explained 46.4% of the variation, and the second principal component explained 9.9% of the variation in the finger ridge-counts of Landmass Asia. Visualizing the biplot of PC1 vs PC2 for only the centroids of different groupings illustrated separation of Regions Australia, Melanesia, and Oceania from the remaining six Regions in Landmass Asia, and the range of variation in Region Melanesia does not overlap with either outlying Regions Australia or Oceania or with the remaining six clustered Regions in Landmass Asia (Figure 33). Region Oceania appeared to exhibit more variation than other Regions in Landmass Asia. The centroids for finger ridge-counts of females and males in Landmass Asia were separate. (Figure 33B). Females and males of Landmass Asia appeared to exhibit a similar

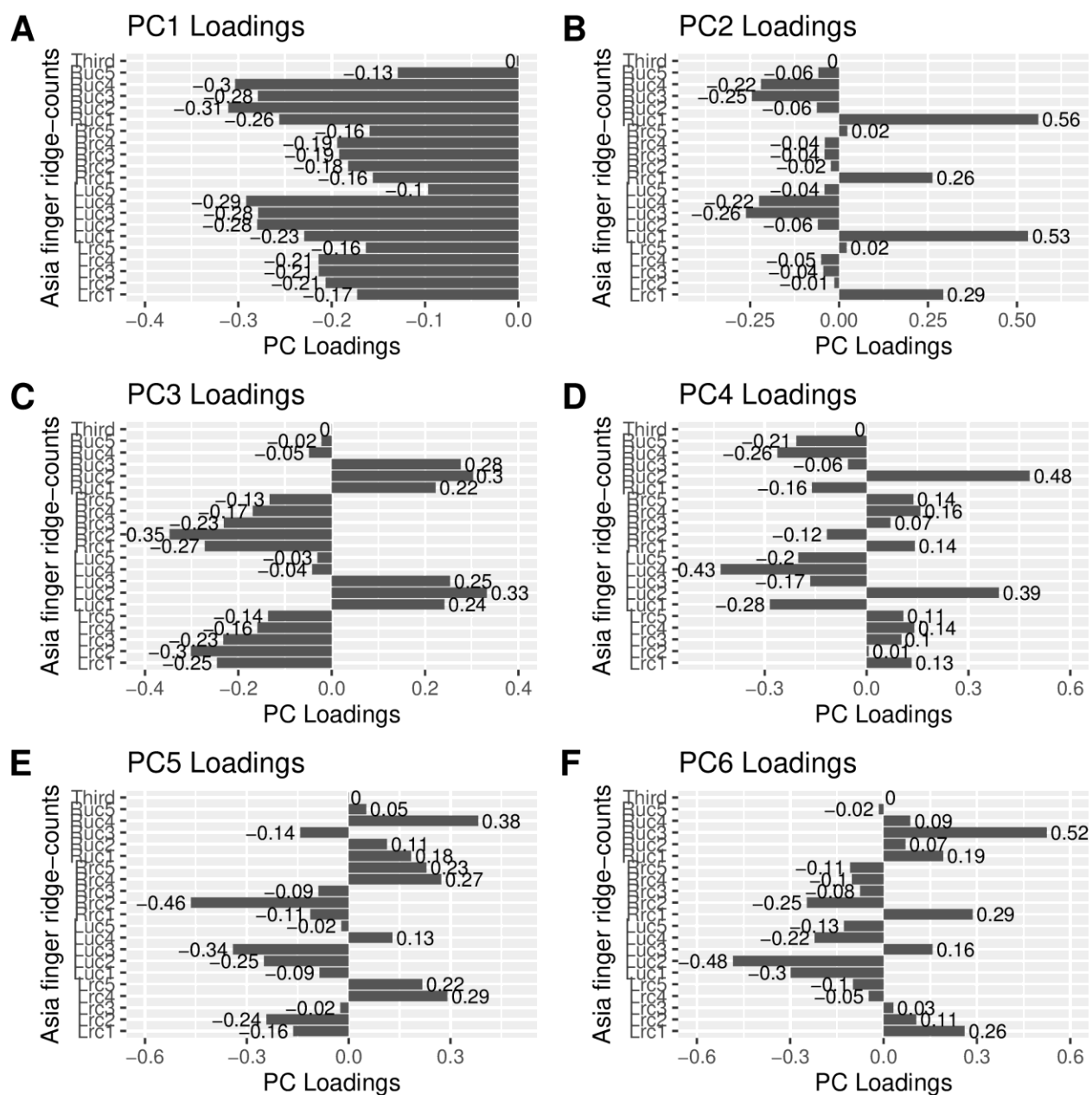


Figure 32. PC 1-6 loadings from PCA of covariance matrix of Landmass Asia finger ridge-counts. (N = 15813)

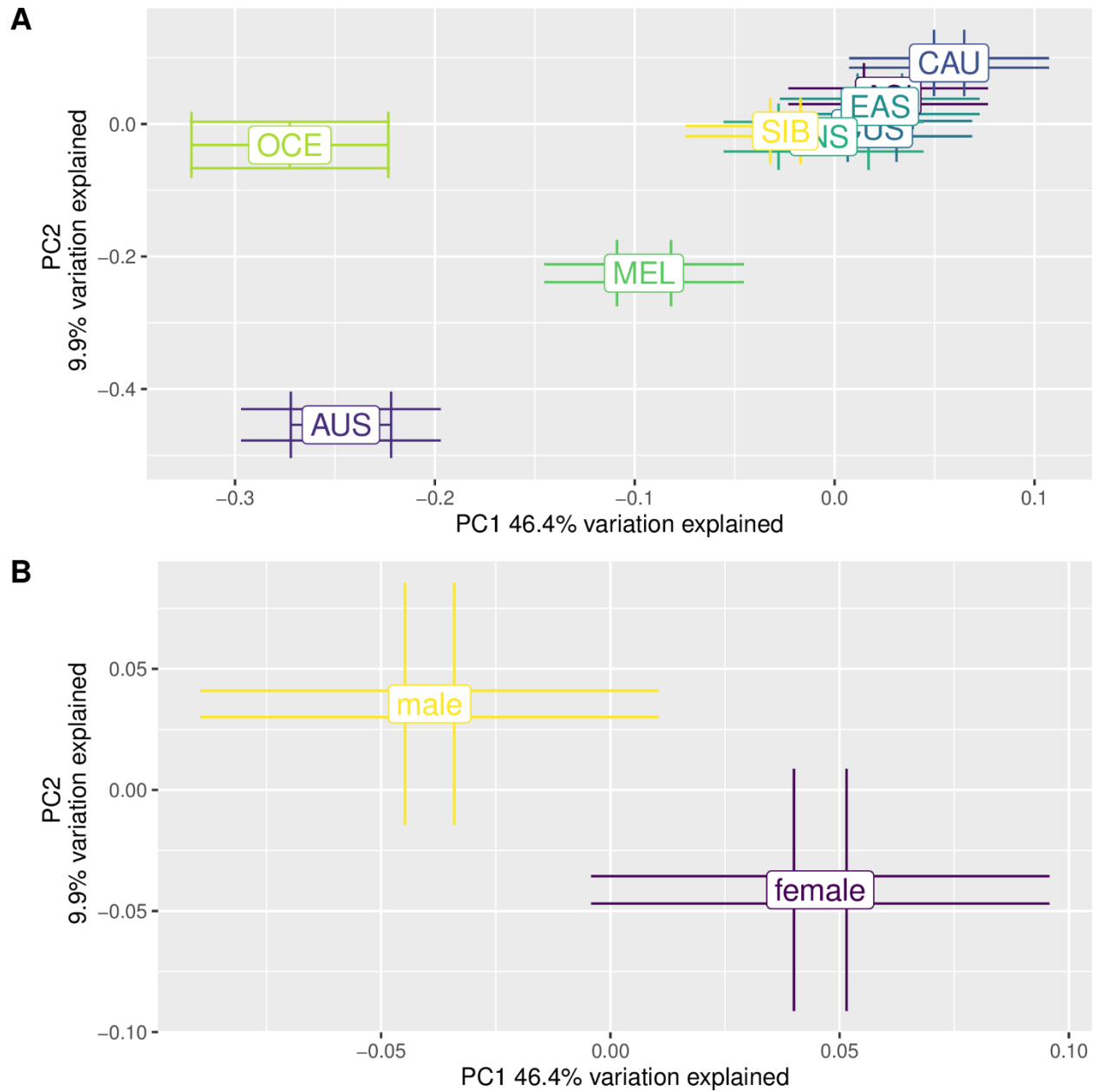


Figure 33. Principal component analysis of covariance matrix of Landmass Asia finger ridge-counts. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 15813).

Error bars indicate the standard error.

amount of variation.

Further splitting the finger ridge-counts of Landmass Asia by-Sex, the Regions exhibited opposing positions in the biplots for females and males on both PC1 and PC2. The first principal component for females exhibited 46.8% of the variation in finger ridge-counts of Landmass Asia, and the second principal component for females exhibited 9.7% of the variation in finger ridge-counts of Landmass Asia (Figure 34A). The first principal component for males exhibited 46.1% of the variation in finger ridge-counts of Landmass Asia, and the second principal component for males exhibited 10.1% of the variation in finger ridge-counts of Landmass Asia (Figure 34B).

The Landmass Asia Regions Australia, Melanesia, and Oceania appeared as outliers that compress the remaining Regions, so the three Regions were excluded from visualization (Figure 34, Figure 35). In Landmass Asia females Region Indonesia appeared with the highest loadings on PC1 and PC2 (Figure 35A). But in Landmass Asia males, Region Indonesia appeared with the other Regions along PC1 and with the lowest loadings on PC2 (Figure 35B). Females in the Caucasus Region exhibited negative loadings on PC1 and negative loadings on PC2, but males in the Caucasus Region exhibited positive loadings on PC1 and positive loadings on PC2. Other Regions for females and males exhibited different loadings along PC1 but similar loadings along PC2. Females in Region Siberia exhibited positive loadings on PC1 and loadings close to zero on PC2, but males in Region Siberia exhibited negative loadings on PC1 and loadings still close to zero on PC2. Females in Region Asia exhibited slightly negative loadings on PC1 and slightly positive loadings on PC2, but males in Region East Asia exhibited slightly positive loadings on PC1 and loadings still slightly positive on PC2. Females in Region Asia exhibited negative

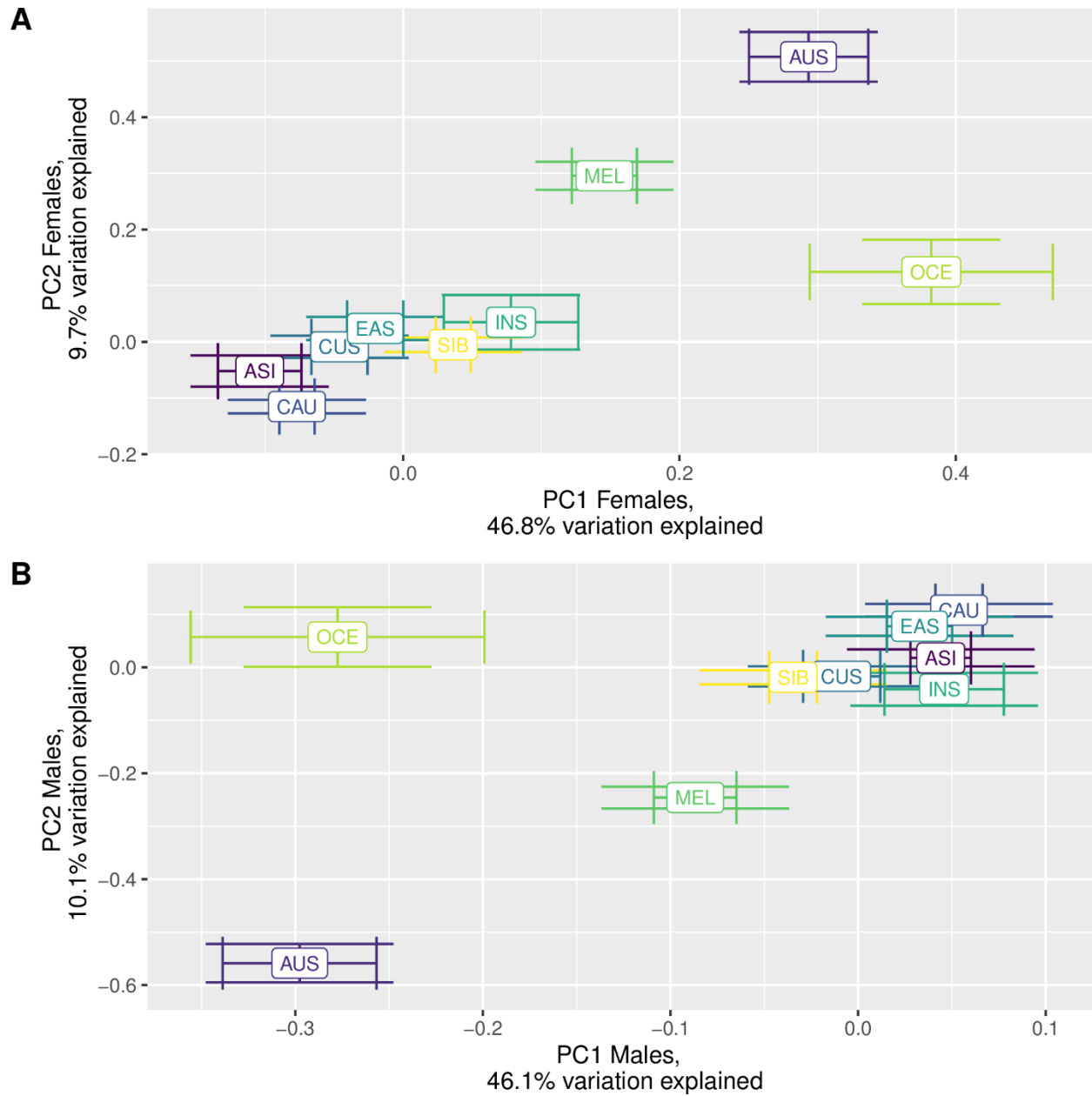


Figure 34. Principal component analysis of covariance matrix of female and male Landmass Asia finger ridge-counts. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 7316 females, 8497 males). Error bars indicate the standard error.

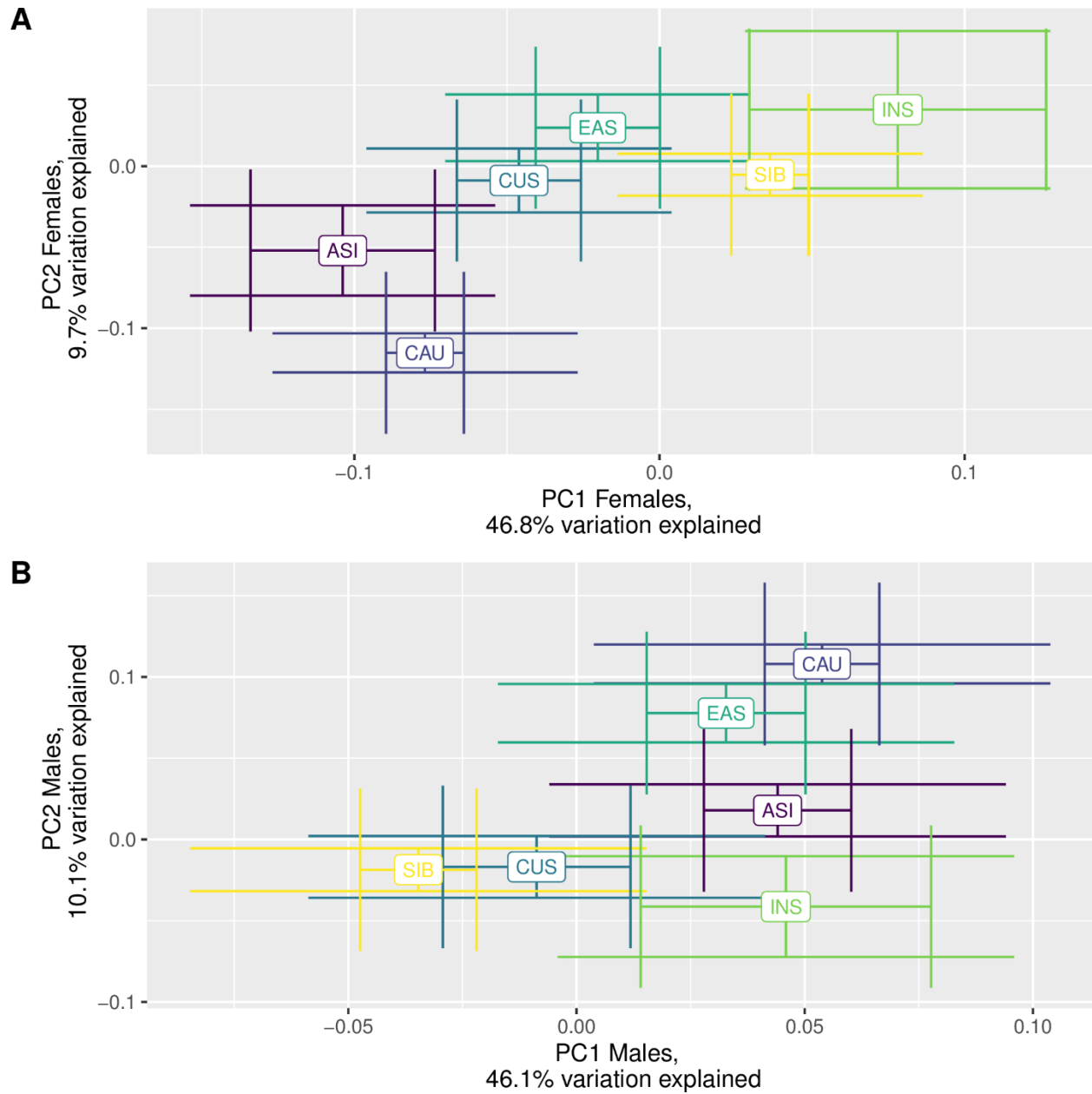


Figure 35. Principal component analysis of covariance matrix of female and male Landmass Asia finger ridge-counts with Regions Australia, Melanesia, and Oceania removed. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 7316 females, 8497 males). Error bars indicate the standard error.

loadings on PC1 and slightly negative loadings on PC2, but males in Region Asia exhibited slightly positive loadings on PC1 and loadings close to zero on PC2. Both females and males in Region Central Soviet Union exhibited slightly negative loadings on PC1 and loadings close to zero on PC2.

For PC1, the finger ridge-counts of Landmass Europe exhibited positive values indicating a positive correlation of the finger ridge-counts of Landmass Europe with PC1 (Figure 36A). PC2 captured the positive correlation of digit 1 against all other finger ridge-counts of Landmass Europe (Figure 36B). PC3 captured the positive correlation of ulnar ridge-counts against the radial ridge-counts of Landmass Europe (Figure 36C). PC4 exhibited a higher positive correlation for the ulnar ridge-counts of digit 2 on both left and right sides and negative correlation for the ulnar ridge-counts of digit 4 on both left and right sides and to a lesser extent right radial count for digit 2 and left ulnar count for digit 1 (Figure 36D). PC5 exhibited higher positive correlation for the radial ridge-counts on the left and right sides for digit 4 and for the right ulnar count for digit 4 and a negative correlation for the radial ridge-counts on the left and right sides for digit 2 and for the left ulnar count for digit 2 (Figure 36E). PC6 captured a higher positive correlation for the left radial count for digit 2 and the right ulnar count for digit 2 with a negative correlation for the left ulnar count for digit 2 (Figure 36F). The Third ridge-count exhibited loadings at value zero for all six principal components, indicating no correlation. The loadings for PC1 were similar in magnitude but opposite in direction of correlation between the finger ridge-counts in the full data set versus in Landmass Europe only, and the loadings for PC2 and PC3 were similar in magnitude and direction of correlation between finger ridge-counts in the full data set and Landmass Europe (Figure 21, Figure 36).

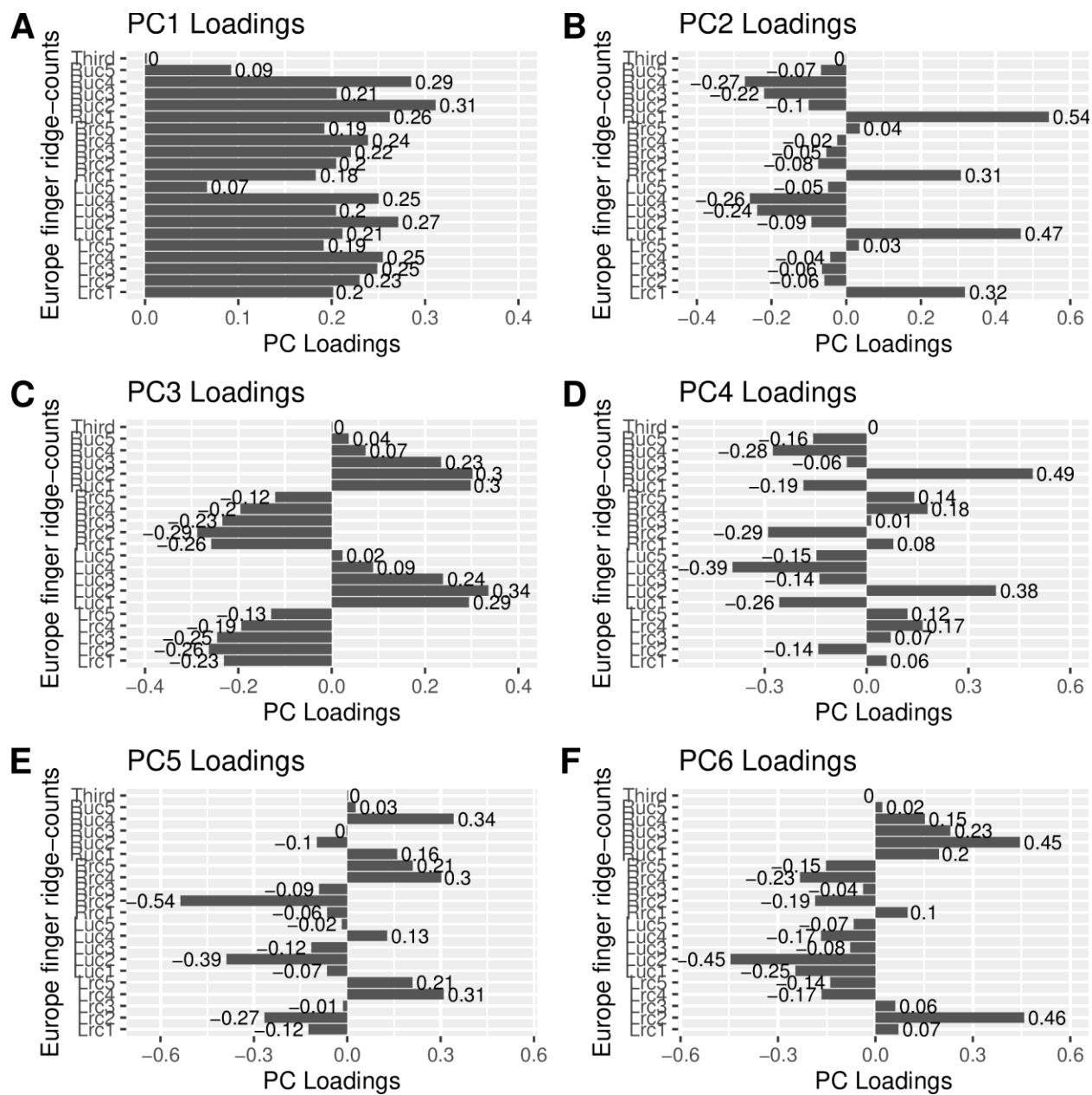


Figure 36. PC 1-6 loadings from PCA of covariance matrix of Landmass Europe finger ridge-counts. (N = 16400)

The first principal component explained 44.4% of the variation, and the second principal component explained 8.7% of the variation in the finger ridge-counts of Landmass Europe. Visualizing the biplot of PC1 versus PC2 for only the centroids of different groupings illustrated minimal separation of Region Europe from Region Soviet Union along the PC1 and PC2 axes (Figure 37). Region Soviet Union appeared to exhibit more variation than Region Europe. The centroids for finger ridge-counts of females and males in Landmass Europe were separate (Figure 37B). Females and males of Landmass Europe appeared to exhibit a similar amount of variation.

Further splitting the finger ridge-counts of Landmass Europe by-Sex, the two Regions exhibited similar positions and loading values in the biplots for females and males, even though the amounts of variation in finger ridge-counts of Landmass Europe were different for females and males. The first principal component for females exhibited 45.4% of the variation in finger ridge-counts of Landmass Europe, and the second principal component for females exhibited 8.5% of the variation in finger ridge-counts of Landmass Europe (Figure 38A). The first principal component for males exhibited 43.5% of the variation in finger ridge-counts of Landmass Europe, and the second principal component for males exhibited 9.1% of the variation in finger ridge-counts of Landmass Europe (Figure 38B).

4.1.4 Baseline models for generalized linear modeling

For the LASSO regularization of the GLM and thus baseline models for the GLM of TRC, ARC, and finger ridge-counts the final training set contained 32580 observations of 11 variables: Ethnic Group, Region, Landmass, Sex, Latitude, Longitude, frc.kmeans.2,

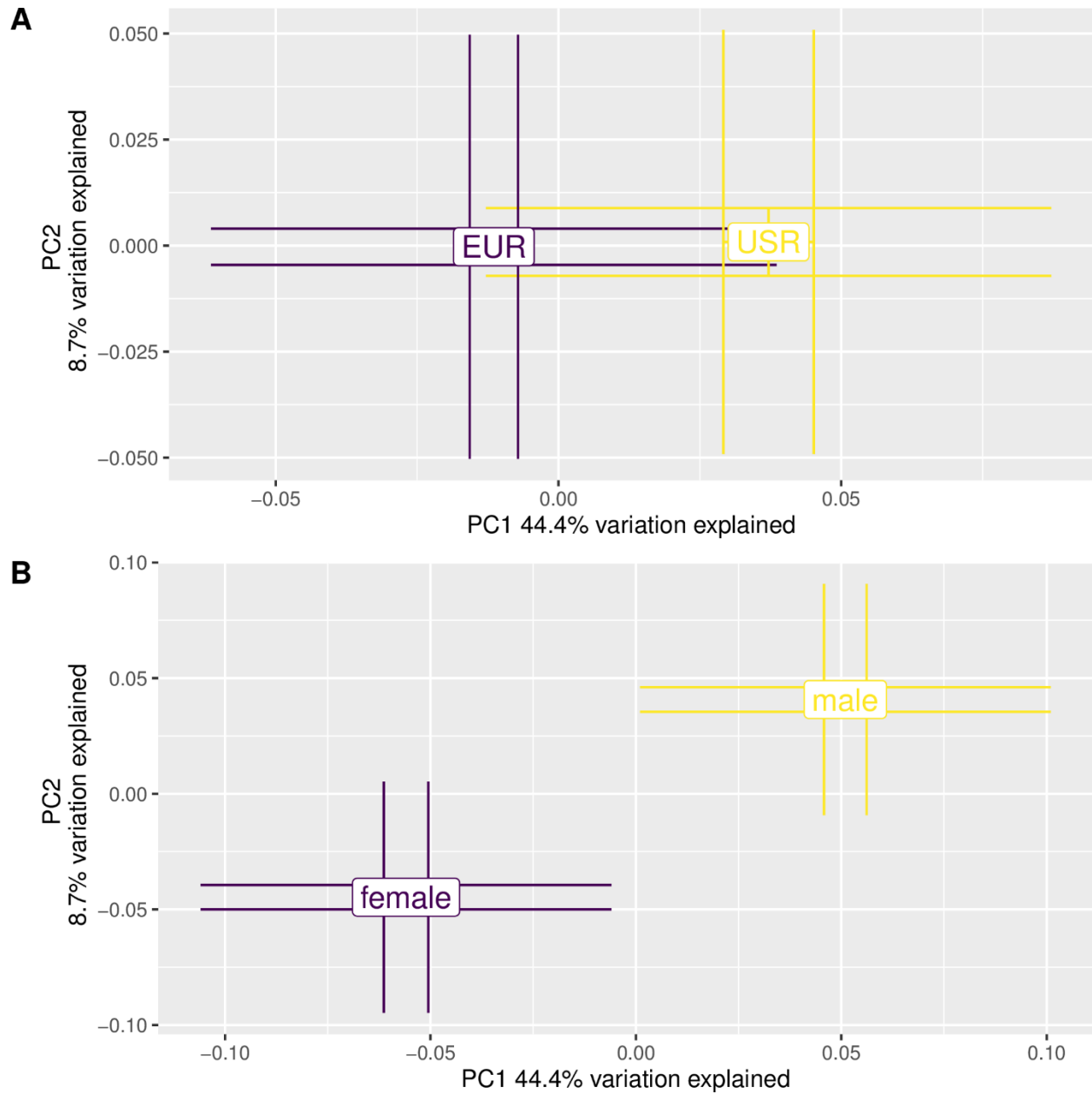


Figure 37. Principal component analysis of covariance matrix of Landmass Europe finger ridge-counts. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 16400).

Error bars indicate the standard error.

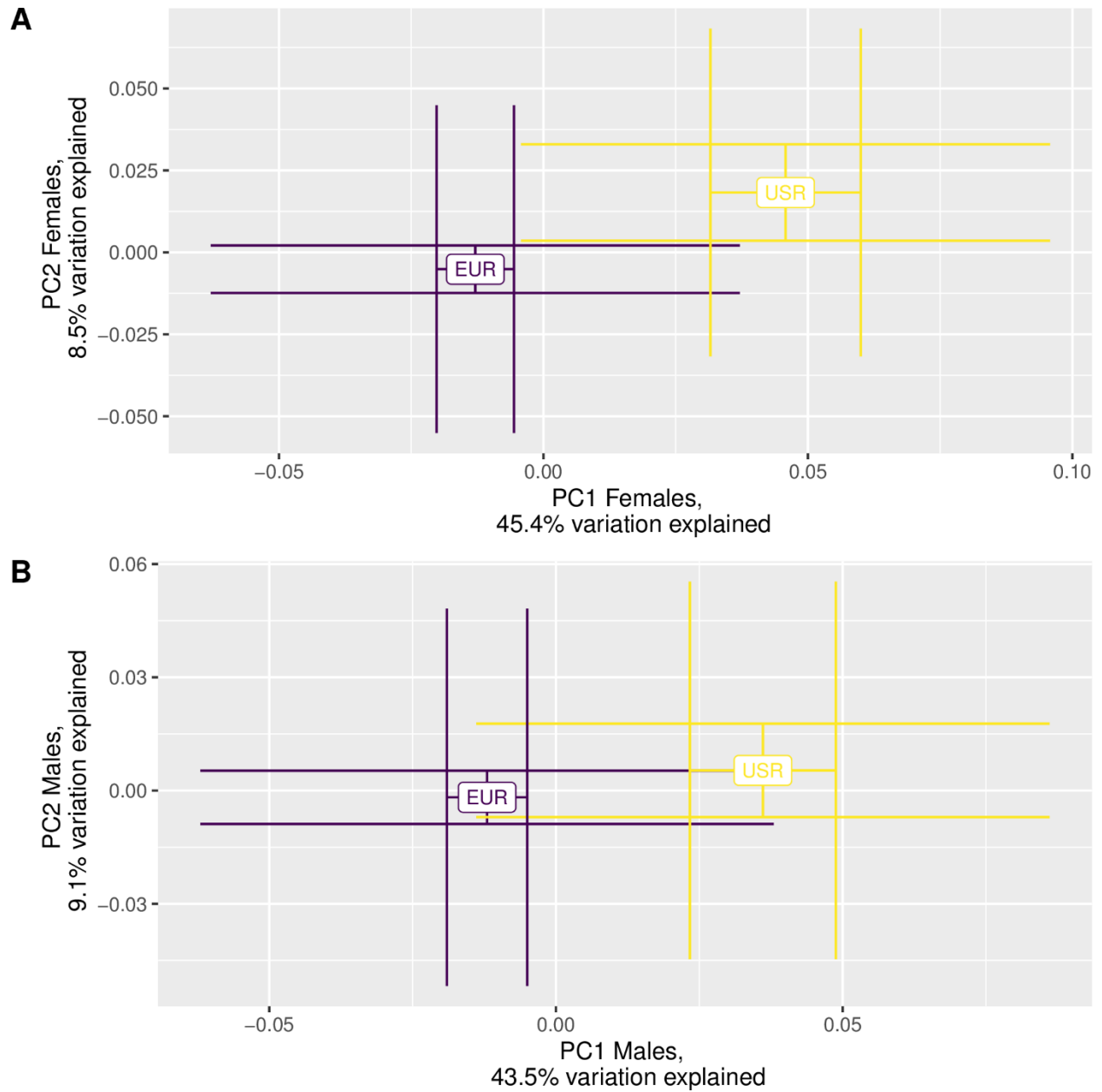


Figure 38. Principal component analysis of covariance matrix of female and male Landmass Europe finger ridge-counts. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 7821 females, 8579 males). Error bars indicate the standard error.

frc.kmeans.4, frc.kmeans.10, frc.kmeans.16, and frc.kmeans.322, and for evaluating the final model on a separate test set to calculate performance metrics, the final test set contained 13961 observations of the same 11 variables.

To evaluate the assumptions of the mean regressor as an approximation of the central tendency of the response variables, the similarity to the median regressor was visualized, and for the TRC, ARC, and finger ridge-counts response variables the mean of each appeared similar to the median of each (Figure 7, Figure 8, Figure 9).

To evaluate whether the baseline linear model could even be interpreted in terms of goodness-of-fit or performance, initial model checking was performed to see whether the assumptions of linear models were met or could be approximately met, since dermatoglyphic features like TRC do not exhibit a normal distribution but instead were negatively skewed and flattened for both females and males from a British population (Holt 1955).

Ten baseline linear models for the GLM exhibited multicollinearity (high > 10 , moderate 5-10, low < 5) in the explanatory variables as computed via the variance inflation factor, which assesses how much the linear correlation of an explanatory variable with other explanatory variables increases the variance of the coefficient estimate compared to theoretically no correlation. For brevity, only the moderate or high explanatory variables were reported here, and the remaining explanatory variables not specified in that model exhibited low collinearity. The linear model with left ulnar count of digit 1 as the response variable exhibited high collinearity (> 1000) for explanatory variables Ethnic Group and Latitude (Figure S22). The linear model with left ulnar count of digit 2 as the response variable exhibited moderate collinearity for explanatory variables Latitude and Region (Figure S25). The linear model with left ulnar count of digit 3 as the response variable exhibited moderate collinearity for explanatory variables

Longitude and Region (Figure S29). The linear model with left radial count of digit 4 as the response variable exhibited high collinearity (> 1000) for explanatory variables Ethnic Group and Region (Figure S31). The linear model with left ulnar count of digit 5 as the response variable exhibited moderate collinearity for explanatory variables Latitude and Region (Figure S37). The linear model with right ulnar count of digit 2 as the response variable exhibited moderate collinearity for explanatory variables Latitude and Region (Figure S45). The linear model with right radial count of digit 4 as the response variable exhibited high collinearity (> 1000) for explanatory variables Ethnic Group and Latitude (Figure S51). The linear model with right ulnar count of digit 4 as the response variable exhibited high collinearity (> 1000) for explanatory variables Ethnic Group and Longitude (Figure S53). The linear model with right ulnar count of digit 5 as the response variable exhibited high collinearity (> 1000) for explanatory variables Ethnic Group and Latitude (Figure S57). The linear model with Third ridge-count as the response variable exhibited moderate collinearity for explanatory variables Longitude and Region (Figure S59). The consequences of continuing to use a model known to exhibit high collinearity between explanatory variables are primarily relegated to interpretation of coefficients and resulting p values, which were not the focus of these models in this dissertation. Further, when collinearity is present, the predictions, precision of the predictions, and the goodness-of-fit statistics do not change (Neter 1996:289). Solutions to collinearity between explanatory variables are to remove the highly correlated explanatory variables, to perform LASSO regression that can handle collinearity, etc. (Frost 2017).

For all linear models with response variables of TRC, ARC, and finger ridge-counts, the Anderson-Darling normality test p value was < 0.05 , so reject the null hypotheses the errors were from a normal distribution (Table 12). For a visual assessment of normality, the normality of

Table 12. Violation of standard regression assumptions of residuals for baseline linear models for GLM of TRC, ARC, and finger ridge-counts; ARC = absolute ridge-count; GLM = generalized linear model; TRC = total ridge-count; p value < 0.05 in bold

GLM	Multicollinearity	Anderson-Darling Normality		Durbin-Watson Autocorrelation		Breusch-Pagan Heteroscedasticity
Response variable	Variance Inflation Factor > 10	A test statistic	p value < 0.001	d test statistic	p value	p value < 0.001
TRC		28.828	✓	1.984	0.002	✓
ARC		36.556	✓	1.988	0.012	✓
Lrc1		63.245	✓	1.980	< 0.001	✓
Luc1	✓	320.45	✓	2.011	0.552	✓
Lrc2		107.97	✓	2.005	0.274	✓
Luc2		382.13	✓	1.986	0.176	✓
Lrc3		118.56	✓	1.979	< 0.001	✓
Luc3		496.92	✓	1.994	0.514	✓
Lrc4	✓	58.697	✓	1.998	0.076	✓
Luc4		353.58	✓	2.000	0.082	✓
Lrc5		18.924	✓	2.008	0.396	✓
Luc5		2275.4	✓	1.980	0.038	✓
Rrc1		73.042	✓	1.988	0.006	✓
Ruc1		233.49	✓	2.006	0.258	✓
Rrc2		97.967	✓	1.996	0.064	✓
Ruc2		373.92	✓	1.972	0.008	✓
Rrc3		130.5	✓	1.989	0.012	✓
Ruc3		500.41	✓	1.987	0.002	✓
Rrc4	✓	65.999	✓	2.032	0.168	✓
Ruc4	✓	352.86	✓	2.005	0.256	✓
Rrc5		5.9647	✓	2.008	0.378	✓
Ruc5	✓	1269.3	✓	2.003	0.166	✓
Third		10302	✓	1.974	0.046	✓

residuals plots exhibit an almost perfect one-to-one line toward the center comparing the standard normal distribution quantiles on the x-axis as the residuals from a perfect model to the sample quantiles on the y-axis as the real residuals from our models, but the relationship breaks down as seen from the dots falling away from the line particularly when the standard normal distribution quantiles on the x-axis dip below -2.5 and exceed 2.5 at the tails, and for the histograms of residuals the lower and upper tails were unequal (Figure S17 through Figure S62). The subjectivity of the visual assessment of normality was clear when the normality of residuals plots for left and right radial count of digit 5 as the response variables appeared to show all dots falling along the line, suggesting the errors were normally distributed, but the Anderson-Darling normality test still revealed a p value < 0.05 indicating the errors were indeed not normally-distributed (Table 12, Figure S35, Figure S55). Thus, all linear models with response variables of TRC, ARC, and finger ridge-counts exhibited non-normal distribution of errors, which could be attributed to autocorrelation. Other reasons for non-normal distribution of errors could be an important explanatory variable not included in the model or some polynomial or interaction term not included in the model. The consequences of continuing to use a model known to violate the normality of residuals assumption are primarily biased estimate of standard errors affecting interpretation of confidence intervals and the p values, which were not the focus of these models in this dissertation, and with large sample sizes where the number of observations per variable exceeds 10 the impact of the negative consequences diminishes (Schmidt and Finan 2018). Therefore, no solutions were necessarily sought, and transformation as a common solution was inherently considered with summary response variables for finger ridge-counts like TRC and ARC (discussed here) as well as with transformation response variables like Directional Asymmetry, Fluctuating Asymmetry, and indices of asymmetry and diversity, but transformation

was otherwise avoided because it would more drastically bias the results than it would reduce the bias attributed to non-normality of residuals (Graham and Özener 2016).

One potential source of the non-normally distributed errors from linear models with response variables of TRC, ARC, and finger ridge-counts was evaluated with the Durbin-Watson test for autocorrelation of residuals, indicated by a p value < 0.05 (Table 12). Autocorrelation of residuals was detected by the Durbin-Watson test for response variables: TRC, ARC, left radial count of digit 1, left radial count of digit 3, left ulnar count of digit 5, right radial count of digit 1, right ulnar count of digit 2, right radial count of digit 3, right ulnar count of digit 3, and Third ridge-count (Table 12). However, in practical use any Durbin-Watson test statistic value equal to 2 indicated no autocorrelation and between 1.5 and 2.5 could be approximated as no autocorrelation (Bedre 2019; SE n.d.). The highest Durbin-Watson test statistic was 2.010865 for response variable left ulnar count of digit 2, and the p value was 0.552 indicating the null hypothesis of no autocorrelation of residuals was supported. The lowest Durbin-Watson test statistic with a p value > 0.05 (actual p value = 0.176) indicating no autocorrelation of residuals was 1.986 for response variable left ulnar count of digit 2. This test statistic was less than 2. The lowest Durbin-Watson test statistic with a p value < 0.05 (actual p value = 0.008) indicating autocorrelation of residuals was 1.972 for response variable right ulnar count of digit 2. This test statistic where the p value indicated autocorrelation was very similar to the previously stated test statistic where the p value indicated no autocorrelation. While non-normality of residuals does not pose a serious threat to the interpretation of the baseline linear models, the consequences of continuing to use a model known to exhibit autocorrelation of residuals, even with large sample sizes, are great. Autocorrelation of residuals reduces the interpretability of the estimates of the coefficients but also their standard errors. As a result, the p values would be biased, and the

standard error of the regression coefficients may underestimate the true standard deviation of the estimated regression coefficients, so the mean squared error performance metric may greatly underestimate the true variance of the errors. Similarly, the RMSE would be unreliable (Broersen 2005). Given no Durbin-Watson test statistics exceeded the bounds of the practical use of the Durbin-Watson test, no baseline linear models with response variables of TRC, ARC, and finger ridge-counts were treated as autocorrelated (Bedre 2019; SE n.d.).

Other violations of linear regression assumptions that can seriously bias the interpretation of the model even with a large sample size are the presence of extreme observations that unduly influence the regression and also the heteroscedasticity that reduces the interpretation of the coefficients. For all linear models with response variables of TRC, ARC, and finger ridge-counts, the results of the heteroscedasticity check exhibited a p value < 0.001 , so reject the null hypothesis the error variance was constant. Further, the homogeneity of variance plots for ARC, finger ridge-counts, and TRC as response variables did not show a flat and horizontal reference line comparing the fitted values on the x-axis to the square root of the standardized residuals on the y-axis (Figure S17, Figure S19, Figure S21, Figure S23, Figure S25, Figure S27, Figure S29, Figure S31, Figure S33, Figure S35, Figure S37, Figure S39, Figure S41, Figure S43, Figure S45, Figure S47, Figure S49, Figure S51, Figure S53, Figure S55, Figure S57, Figure S59, Figure S61). Heteroscedasticity can occur when a large range separates the smallest and largest observed values, like with TRC and ARC response variables. The consequences of continuing to use a model known to exhibit heteroscedasticity rely on identifying the purity of the heteroscedasticity and so do the solutions. For impure heteroscedasticity where the model was not correctly specified by leaving an important variable out, and the generation of the coefficient values was not biased, so opportunities exist to solve the problem with additional explanatory

variables. For pure heteroscedasticity where the model is correctly specified and yet non-constant variance in the residuals persists, the issues with interpretation are the same as with autocorrelation, and the prediction error significantly changes over the range of prediction of the model, so the performance metrics, which rely on the prediction error likely represent an underestimate of the performance error, or a minimum. For the baseline linear models exhibiting significant heteroscedasticity, the performance metrics were interpreted as a minimum value, meaning if the comparison in performance between the baseline linear model that exhibits heteroscedasticity and the mean regressor were close in value, the mean regressor would likely be identified as the better performing baseline. Solutions to pure heteroscedasticity are the use of weighted least squares, obtaining standard errors corrected for heteroscedasticity, or retooling the model. Weighted least squares divides each variable in the regression model by a proportionality factor that turns the residuals into a white noise process after which the regression is rerun so the second regression model will be free from heteroscedasticity (Bucevska 2011).

Calculating a goodness-of-fit for each model serves to indicate how well the explanatory variables in the model explain the observed variation in the response variable. However, comparing goodness-of-fit values for multiple models with different response variables is not appropriate. So, only the AIC of the full model was compared to the reduced model containing the same response variable and model formulation to indicate how the feature selection process improved the fit of the model (Schneider, Hommel, and Blettner 2010). For the mean regressor, median regressor, and linear regression model as baseline comparisons for the GLM, only the linear model incorporated explanatory variables in the model construction, and so the goodness-of-fit was only assessed for the linear model as a baseline via the AIC value, for which lower values indicate a better model fit (Table 13).

Table 13. Goodness-of-fit for full and reduced baseline linear models of TRC, ARC, and finger ridge-counts, of same GLM data set; AIC = Akaike information criterion; ARC = absolute ridge-count; GLM = generalized linear model; LM = linear model; TRC = total ridge-count

Analytical Method	Response Variable	Model	AIC
LM	TRC	full	313640.3
		reduced	313604.8
LM	ARC	full	335389.9
		reduced	335354.7
LM	Lrc1	full	201217.4
		reduced	201182.7
LM	Luc1	full	212579.6
		reduced	212540
LM	Lrc2	full	200966.4
		reduced	200928.1
LM	Luc2	full	209297.4
		reduced	209201.8
LM	Lrc3	full	192545.2
		reduced	192506.3
LM	Luc3	full	196838.5
		reduced	196680
LM	Lrc4	full	193925.3
		reduced	193894.1
LM	Luc4	full	200027.6
		reduced	199995
LM	Lrc5	full	183332.8
		reduced	183310.9
LM	Luc5	full	173109.7
		reduced	173032.6
LM	Rrc1	full	201229.1
		reduced	201189.4
LM	Ruc1	full	211811.1
		reduced	211771.8
LM	Rrc2	full	203993.7
		reduced	203956
LM	Ruc2	full	206858.7
		reduced	206794.4
LM	Rrc3	full	189787.8
		reduced	189748.8
LM	Ruc3	full	196491.1
		reduced	196457.6
LM	Rrc4	full	194499.2
		reduced	194467
LM	Ruc4	full	199481.5
		reduced	199445.4

Table 13 Continued

Analytical Method	Response Variable	Model	AIC
LM	Rrc5	full	184451.4
		reduced	184424.5
LM	Ruc5	full	175911.4
		reduced	175878.3
LM	Third	full	35375.39
		reduced	35088.32

The linear model with Third ridge-count as the response variable, with feature selected explanatory variables: Region, Longitude, frc.kmeans.2, frc.kmeans.4, and frc.kmeans.16, exhibited the greatest reduction in AIC (Table 13). The next lowest reduction in AIC value was for left ulnar count of digit 3, then for left ulnar count of digit 2, left ulnar count of digit 5, and right ulnar count of digit 2, in decreasing order. The remaining TRC, ARC, and finger ridge-count response variables exhibited a reduction in AIC through feature selection ranging between 21 and 39. Of note, the reduced models for the baseline linear models of all TRC, ARC, and finger ridge-counts as response variables indicated a marked improvement in the goodness-of-fit.

The best performance metric for the TRC, ARC, and finger ridge-counts response variables was the MAPE due to the presence of outliers that could unduly influence the modeling and the discrete, non-negative counts that were not truly continuous response variables (Table 4). The zeros in the response variables did not always result in an undefined MAPE, only for the ulnar ridge-counts for all digits on the left hand as well as right ulnar count of digit 1, right ulnar count of digit 3, right ulnar count of digit 5, and the Third ridge-count. The ulnar counts contain many more zeros than the radial counts, so the MAPE was more often undefined for the response variables of ulnar counts. When the MAPE was undefined (NaN), the second best performance metric for these response variables, the MAE, was considered.

No GLMs performed the best for any of the response variables of the TRC, ARC, and finger ridge-counts (Table 14). The mean regressor performed the best for three of the response variables: left ulnar count of digit 2, right ulnar count of digit 2, and right ulnar count of digit 4, and the MAPE was used as the best performance metric for these. The median regressor performed the best for five of the response variables: left radial count of digit 2, left ulnar count of digit 5, right radial count of digit 2, right ulnar count of digit 5, and the Third ridge count.

Table 14. Performance metrics of final GLM formulation as a result of LASSO regularization and lambda optimization as compared to baselines for TRC, ARC, and finger ridge-counts; ARC = absolute ridge-count; MAE = mean absolute error; MAPE = mean absolute percentage error; NA = Not available; NaN = Not a number; RMSE = root mean squared error; TRC = total ridge-count; best performance metric and model in bold

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline for GLM	mean TRC	NA	NA	39.898	49.731	0.285
Baseline for GLM	median TRC	NA	NA	39.616	49.928	0.273
Baseline LM	TRC	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	23.615	30.002	0.195
GLM	TRC	0.004	Landmass + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	135.656	144.380	27.299
Baseline for GLM	mean ARC	NA	NA	72.992	89.953	0.397
Baseline for GLM	median ARC	NA	NA	72.743	90.550	0.416
Baseline LM	ARC	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	33.715	42.126	0.229
GLM	ARC	0.007	Landmass + Sex + Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	180.373	201.364	34.173
Baseline for GLM	mean Lrc1	NA	NA	5.382	6.802	0.353
Baseline for GLM	median Lrc1	NA	NA	5.314	6.832	0.332

Table 14 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline LM	Lrc1	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	4.198	5.331	0.315
GLM	Lrc1	0.004	Landmass + Sex + Latitude + frc.kmeans.2 + frc.kmeans.10 + frc.kmeans.16	12.963	14.291	4.74
Baseline for GLM	mean Luc1	NA	NA	7.350	8.123	1.183
Baseline for GLM	median Luc1	NA	NA	6.270	10.262	NaN
Baseline LM	Luc1	NA	Ethnic Group + Sex + Latitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.16	4.946	6.299	3.369
GLM	Luc1	0.007	Ethnic Group + Landmass + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	6.573	8.747	2.365
Baseline for GLM	mean Lrc2	NA	NA	5.843	6.702	0.641
Baseline for GLM	median Lrc2	NA	NA	5.775	6.756	0.577
Baseline LM	Lrc2	NA	Ethnic Group + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	4.410	5.321	0.616
GLM	Lrc2	0.006	Landmass + Latitude + Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	7.708	9.208	2.801

Table 14 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline for GLM	mean Luc2	NA	NA	7.235	7.968	1.060
Baseline for GLM	median Luc2	NA	NA	6.830	9.342	3.415
Baseline LM	Luc2	NA	Region + Sex + Latitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	4.630	5.978	1.204
GLM	Luc2	0.008	Landmass + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	6.406	9.135	3.638
Baseline for GLM	mean Lrc3	NA	NA	5.135	6.354	0.422
Baseline for GLM	median Lrc3	NA	NA	5.038	6.401	0.388
Baseline LM	Lrc3	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	3.766	4.708	0.370
GLM	Lrc3	0.005	Sex + Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	10.186	11.541	4.049
Baseline for GLM	mean Luc3	NA	NA	5.770	7.171	1.496
Baseline for GLM	median Luc3	NA	NA	3.970	8.196	NaN
Baseline LM	Luc3	NA	Region + Sex + Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	3.638	5.002	1.460
GLM	Luc3	0.006	Landmass + Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10	4.006	7.233	8.441

Table 14 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline for GLM	mean Lrc4	NA	NA	4.950	6.398	0.312
Baseline for GLM	median Lrc4	NA	NA	4.892	6.484	0.288
Baseline LM	Lrc4	NA	Ethnic Group + Sex + Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	3.716	4.751	0.262
GLM	Lrc4	0.004	Landmass + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	13.376	14.608	4.809
Baseline for GLM	mean Luc4	NA	NA	6.891	7.627	1.164
Baseline for GLM	median Luc4	NA	NA	6.0276	9.721	NaN
Baseline LM	Luc4	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	3.950	5.220	2.608
GLM	Luc4	0.007	Landmass + Longitude + frc.kmeans.1 + frc.kmeans.4 + frc.kmeans.322	5.601	8.411	3.985
Baseline for GLM	mean Lrc5	NA	NA	4.149	5.170	0.319
Baseline for GLM	median Lrc5	NA	NA	4.103	5.256	0.293
Baseline LM	Lrc5	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	3.200	4.019	0.274

Table 14 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
GLM	Lrc5	0.003	Landmass + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	10.621	11.652	4.116
Baseline for GLM	mean Luc5	NA	NA	2.539	3.883	1.701
Baseline for GLM	median Luc5	NA	NA	1.503	4.164	NaN
Baseline LM	Luc5	NA	Region + Sex + Latitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	2.214	3.484	2.910
GLM	Luc5	0.004	Landmass + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	2.159	3.942	2.738
Baseline for GLM	mean Rrc1	NA	NA	5.243	6.728	0.303
Baseline for GLM	median Rrc1	NA	NA	5.157	6.758	0.286
Baseline LM	Rrc1	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	4.180	5.343	0.270
GLM	Rrc1	0.006	Landmass + Sex + Latitude + frc.kmeans.2 + frc.kmeans.16	14.743	15.958	5.148
Baseline for GLM	mean Ruc1	NA	NA	7.948	8.578	1.050
Baseline for GLM	median Ruc1	NA	NA	7.690	11.520	NaN
Baseline LM	Ruc1	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	5.051	6.255	1.221

Table 14 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
GLM	Ruc1	0.006	Region + Landmass + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	7.180	10.003	3.817
Baseline for GLM	mean Rrc2	NA	NA	5.806	6.664	0.640
Baseline for GLM	median Rrc2	NA	NA	5.757	6.727	0.576
Baseline LM	Rrc2	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	4.645	5.582	0.629
GLM	Rrc2	0.007	Landmass + Sex + Latitude + Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	7.824	9.488	3.564
Baseline for GLM	mean Ruc2	NA	NA	7.717	8.409	1.019
Baseline for GLM	median Ruc2	NA	NA	7.508	9.605	2.503
Baseline LM	Ruc2	NA	Region + Sex + Latitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	4.408	5.805	1.731
GLM	Ruc2	0.008	Landmass + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	6.980	9.820	3.773
Baseline for GLM	mean Rrc3	NA	NA	4.731	5.855	0.401
Baseline for GLM	median Rrc3	NA	NA	4.671	5.967	0.359

Table 14 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline LM	Rrc3	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	3.561	4.483	0.346
GLM	Rrc3	0.004	Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	9.734	10.997	3.918
Baseline for GLM	mean Ruc3	NA	NA	5.495	7.057	1.547
Baseline for GLM	median Ruc3	NA	NA	3.676	7.956	NaN
Baseline LM	Ruc3	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	3.570	4.957	1.275
GLM	Ruc3	0.006	Region + Sex + Longitude + frc.kmeans.2 + frc.kmeans.4	3.956	7.062	2.931
Baseline for GLM	mean Rrc4	NA	NA	4.813	6.191	0.306
Baseline for GLM	median Rrc4	NA	NA	4.732	6.308	0.278
Baseline LM	Rrc4	NA	Ethnic Group + Sex + Latitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	3.764	4.790	0.265
GLM	Rrc4	0.003	Landmass + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	13.202	14.391	4.763
Baseline for GLM	mean Ruc4	NA	NA	7.191	7.868	1.007

Table 14 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline for GLM	median Ruc4	NA	NA	7.112	8.525	1.778
Baseline LM	Ruc4	NA	Ethnic Group + Sex + Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	3.988	5.165	2.580
GLM	Ruc4	0.007	Region + Landmass + Sex + Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	6.660	9.208	3.702
Baseline for GLM	mean Rrc5	NA	NA	4.253	5.194	0.333
Baseline for GLM	median Rrc5	NA	NA	4.218	5.196	0.324
Baseline LM	Rrc5	NA	Ethnic Group + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	3.285	4.085	0.284
GLM	Rrc5	0.003	Region + Landmass + Sex + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16	10.392	11.472	4.059
Baseline for GLM	mean Ruc5	NA	NA	3.051	4.288	1.654
Baseline for GLM	median Ruc5	NA	NA	1.919	4.698	NaN
Baseline LM	Ruc5	NA	Ethnic Group + Sex + Latitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	2.495	3.631	4.630

Table 14 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
GLM	Ruc5	0.004	Region + Sex + Long + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.10 + frc.kmeans.16 + frc.kmeans.322	2.464	4.308	2.966
Baseline for GLM	mean Third	NA	NA	0.057	0.430	1.875
Baseline for GLM	median Third	NA	NA	0.027	0.431	NaN
Baseline LM	Third	NA	Region + Longitude + frc.kmeans.2 + frc.kmeans.4 + frc.kmeans.16	0.060	0.429	2.034
GLM	Third	0.001	Region + Landmass + frc.kmeans.2	3.929	4.055	1.009

Both the left and right ulnar count of digit 5 used the MAE as the best performance metric due to the undefined MAPE. The baseline linear model performed the best for fifteen response variables, and of those left ulnar count of digit 1 (MAE = 4.946), left ulnar count of digit 3 (MAE = 3.638), left ulnar count of digit 4 (MAE = 3.950), right ulnar count of digit 1 (MAE = 5.051), and right ulnar count of digit 3 (MAE = 3.570) used the MAE instead of the MAPE as the best performance metric (Table 14). Because the MAE and not RMSE or MAPE can be compared across models with different response variables given most other conditions were consistent, and that the MAE was not ideal for the TRC, ARC, and finger ridge-count response variables for the GLM, no comparisons across response variables were attempted. However, for the baseline linear models that did use the MAE due to undefined MAPE, the best performing response variable according to these MAE values was right ulnar count of digit 3 with explanatory variables: Ethnic Group, Sex, frc.kmeans.2, frc.kmeans.4, frc.kmeans.10, frc.kmeans.16, and frc.kmeans.322, also the most common feature selected explanatory variables for the linear models of TRC, ARC, and the finger ridge-counts. Because the baseline linear model clearly outperformed the other baselines and the GLM, the interpretation of the linear models in context of their assumptions, like the increase in prediction error due to heteroscedasticity, becomes more important.

4.1.5 Generalized linear modeling (GLM)

For the LASSO regularization of the GLM of TRC, ARC, and finger ridge-count response variables the final training set contained 32580 observations of 11 variables: Ethnic Group, Region, Landmass, Sex, Latitude, Longitude, frc.kmeans.2, frc.kmeans.4, frc.kmeans.10, frc.kmeans.16, and frc.kmeans.322, and for evaluating the final model on a separate test set to

calculate performance metrics, the final test set contained 13961 observations of the same 11 variables.

Calculating a goodness-of-fit for each model serves to indicate how well the explanatory variables in the model explain the observed variation in the response variable. However, comparing goodness-of-fit values for multiple models with different response variables is not appropriate. The goodness-of-fit for the GLM was the Kullback-Leibler R^2 value, for which higher values indicate a better model fit, and the full model was compared to the reduced model containing the same response variable and model formulation just with different explanatory variables to indicate how the feature selection process improved the fit of the model (Schneider, Hommel, and Blettner 2010). The desired higher Kullback-Leibler R^2 value from $1 - (model\ deviance \div null\ model\ deviance)$ reflects a higher null deviance in the denominator, and because a smaller deviance indicates a better model fit a higher null deviance suggests the null model fits the data poorly compared to the fitted or reduced model represented in the numerator.

All GLM of TRC, ARC, and finger ridge-counts exhibited a reduction in the Kullback-Leibler R^2 value of the full model compared to the feature-selected reduced model, suggesting the LASSO regularization did not improve the model fit enough to offset the benefit from just reducing the explanatory variables in the model. Therefore, a greater reduction in the Kullback-Leibler R^2 value between the full model and the reduced model indicates a greater decrease in model fit, and a smaller reduction in the Kullback-Leibler R^2 value indicates a smaller decrease in model fit or a maintenance of goodness-of-fit. The GLM with right ulnar count of digit 4 as the response variable, with feature selected explanatory variables: Region, Landmass, Sex, Longitude, frc.kmeans.2, frc.kmeans.4, frc.kmeans.10, frc.kmeans.16, and frc.kmeans.322,

exhibited the smallest reduction in Kullback-Leibler R^2 value (Table 15). The next smallest reductions in Kullback-Leibler R^2 value were for ARC, then right ulnar count of digit 2, right radial count of digit 3, right ulnar count of digit 5, right radial count of digit 2, right ulnar count of digit 1, left ulnar count of digit 2, left radial count of digit 1, and TRC. The next smallest reductions in Kullback-Leibler R^2 value were for right radial count of digit 4, left ulnar count of digit 3, right ulnar count of digit 3, left radial count of digit 3, left ulnar count of digit 5, left ulnar count of digit 4, right radial count of digit 5, left radial count of digit 4, right radial count of digit 1, left radial count of digit 2, and left radial count of digit 5. The GLM with Third ridge-count and left ulnar count of digit 1 exhibited the greatest reduction in Kullback-Leibler R^2 value, so a greater decrease in model fit after LASSO regularization. A small trend appeared where the ridge-counts from the right hand generally exhibited a smaller reduction in the Kullback-Leibler R^2 value so LASSO regularization hurt these models less. No trend of radial vs ulnar or by-digit was apparent. The reductions in Kullback-Leibler R^2 value ranged from 0.0001 for right ulnar count of digit 4 to 0.0964 for left ulnar count of digit 1. Of note, the reduced models for the GLM of all TRC, ARC, and finger ridge-counts as response variables indicated a small reduction in the goodness-of-fit.

To interpret the results of regression models, often the model architecture assumes equal within-group variance, and the betadisper function from the vegan package can approximately test this, because the beta dispersion represents the absolute value distance between groups from a multivariate centroid, and the anova function tests for significance in dispersion across groups (Oksanen et al. 2022). For the majority of response variables, significant differences in dispersion across the groups were present.

Table 15. Goodness-of-fit for full and reduced GLM of TRC, ARC, and finger ridge-counts. Via Kullback-Leibler R^2 value; ARC = absolute ridge-count; GLM = generalized linear model; TRC = total ridge-count

Analytical Method	Response Variable	Model	Kullback-Leibler R^2 value
GLM	TRC	full	0.4016
		reduced	0.4007
GLM	ARC	full	0.5624
		reduced	0.5623
GLM	Lrc1	full	0.2009
		reduced	0.2000
GLM	Luc1	full	0.2957
		reduced	0.1994
GLM	Lrc2	full	0.1958
		reduced	0.1934
GLM	Luc2	full	0.3090
		reduced	0.3084
GLM	Lrc3	full	0.2322
		reduced	0.2310
GLM	Luc3	full	0.5094
		reduced	0.5083
GLM	Lrc4	full	0.2603
		reduced	0.2588
GLM	Luc4	full	0.3659
		reduced	0.3645
GLM	Lrc5	full	0.2438
		reduced	0.2403
GLM	Luc5	full	0.2996
		reduced	0.2983
GLM	Rrc1	full	0.1992
		reduced	0.1971
GLM	Ruc1	full	0.3076
		reduced	0.3071
GLM	Rrc2	full	0.1615
		reduced	0.1611
GLM	Ruc2	full	0.3757
		reduced	0.3755
GLM	Rrc3	full	0.2234
		reduced	0.2231
GLM	Ruc3	full	0.5348
		reduced	0.5337
GLM	Rrc4	full	0.2328
		reduced	0.2317
GLM	Ruc4	full	0.3551
		reduced	0.3550

Table 15 Continued

Analytical Method	Response Variable	Model	Kullback-Leibler R^2 value
GLM	Rrc5	full	0.2471
		reduced	0.2456
GLM	Ruc5	full	0.3564
		reduced	0.3561
GLM	Third	full	0.1022
		reduced	0.0895

Fewer response variables exhibited no significant differences and thus no dispersion across groups: left ulnar count of digit 1 for Sex, left ulnar count of digit 2 for Sex, left ulnar count of digit 3 for Sex, left ulnar count of digit 5 for Ethnic Group, left ulnar count of digit 5 for Region, left ulnar count of digit 5 for Sex, left ulnar count of digit 5 for Landmass, right ulnar count of digit 1 for Region, right ulnar count of digit 1 for Landmass, right radial count of digit 2 for Sex, right ulnar count of digit 2 for Sex, right radial count of digit 3 for Sex, right ulnar count of digit 3 for Sex, right ulnar count of digit 5 for Sex, Third ridge-count for Ethnic Group, Third ridge-count for Region, Third ridge-count for Sex, and Third ridge-count for Landmass.

Demonstrated another way, the grouping Sex exhibited no significant dispersion for left ulnar count of digit 1, left ulnar count of digit 3, left ulnar count of digit 5, right radial count of digit 2, right ulnar count of digit 2, right radial count of digit 3, right ulnar count of digit 3, right ulnar count of digit 5, and Third ridge-count. The grouping Ethnic Group exhibited no significant dispersion for left ulnar count of digit 5 and Third ridge-count. The groupings Region and Landmass exhibited no significant dispersion for left ulnar count of digit 5, right ulnar count of digit 1, and Third ridge-count.

Using LASSO regularization for feature selection of the GLMs by fitting a negative binomial regression model and using k-fold cross-validation to test multiple lambda values, the best lambda values were calculated. The final model included the best lambda value and only the explanatory variables with non-zero coefficients, and the magnitude and sign of the non-zero coefficients indicated the importance and direction of influence of the explanatory variable. Figure 39 illustrated the explanatory variables with the greatest magnitude of importance for the GLM as indicated by high negative or high positive coefficients. For all response variables, the intercept expressed the highest importance with both negative and positive coefficient values,

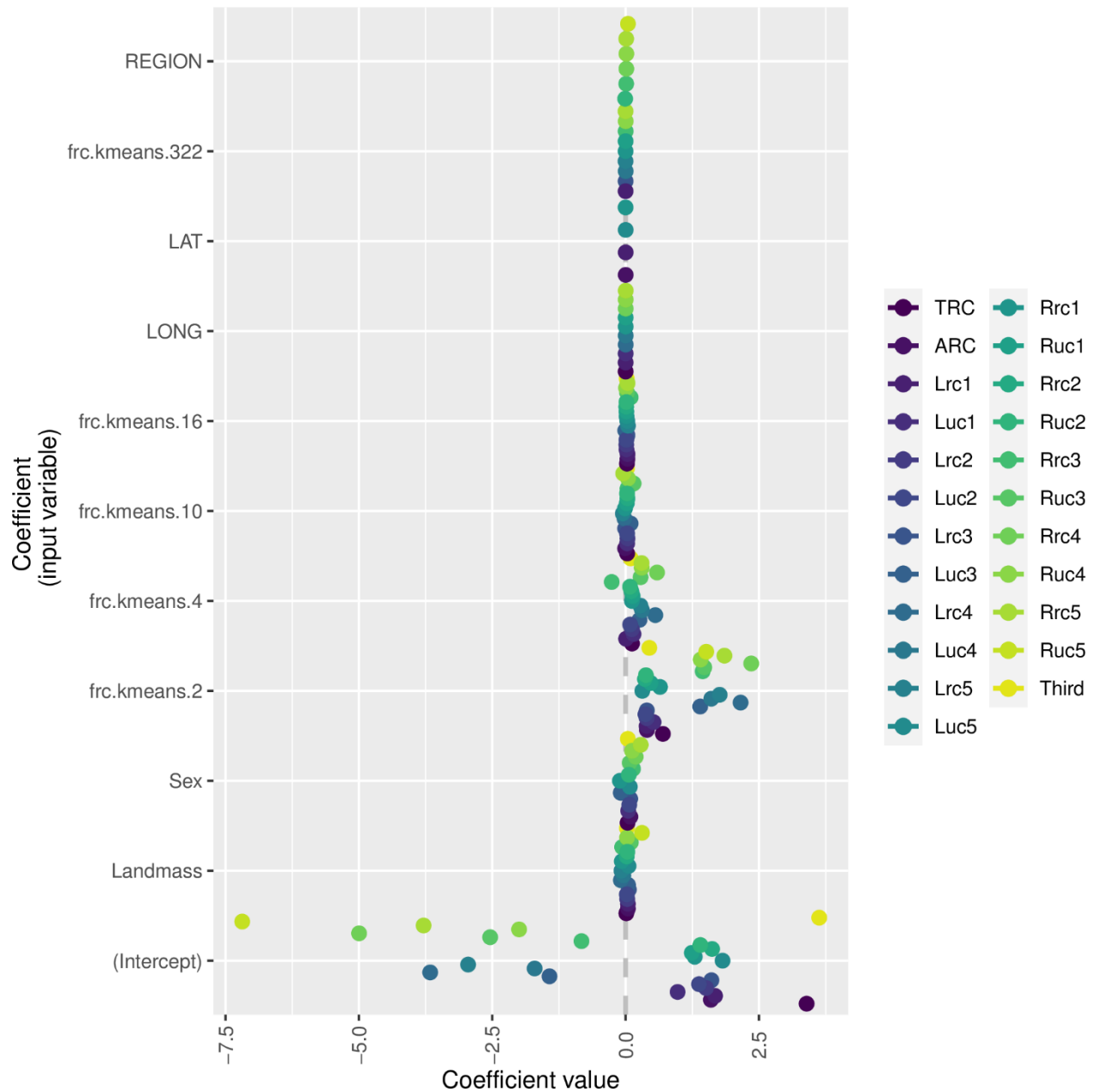


Figure 39. GLM coefficient value for each explanatory variable, colored by TRC, ARC, or finger ridge-counts response variable. (N = 13961); ARC = absolute ridge-count; TRC = total ridge-count

and explanatory variable `frc.kmeans.2` expressed the second highest importance but solely with positive coefficient values.

Because the intercept and `frc.kmeans.2` variable distended the coefficient value range, Figure 40 illustrated the same final model coefficients but with a more narrow range from -0.1 to 0.1. Within this more defined range, explanatory variables `Landmass`, `Sex`, and `frc.kmeans.4` expressed the next highest importance with mostly positive coefficient values. Explanatory variables `frc.kmeans.10` and `frc.kmeans.16` also expressed importance within this -0.1 to 0.1 range with mostly positive coefficients (Figure 40). Many GLMs excluded `Region` from their final models, so very few non-zero coefficients were even mapped, but for those on the right hand the coefficients were positive for `Region` (Table 14). Figure 41 illustrated the same final model coefficients but with a more narrow range from -0.002 to 0.002 to capture the details of explanatory variable coefficients of `frc.kmeans.322`, `Latitude`, and `Longitude`. Within this more defined range, explanatory variables `Latitude` and `Longitude` express similar and solely positive coefficient values. Explanatory variable `frc.kmeans.322` expresses both positive and negative coefficient values with no apparent trend toward left vs right, radial vs ulnar, or digit number.

Comparing the general makeup of the baseline linear models to the GLM, the linear models included more narrow population grouping variables like `Ethnic Group` vs `Region` or `Region` vs `Landmass`. Most models included many k-means cluster assignment variables suggesting those variables were not redundant and did indicate additional useful variation in the response dermatoglyphic features compared to the other explanatory variables (Table 14). The best performance metric for the TRC, ARC, and finger ridge-counts response variables was the MAPE due to the presence of outliers that could unduly influence the modeling and the discrete, non-negative counts that were not truly continuous response variables (Table 4).

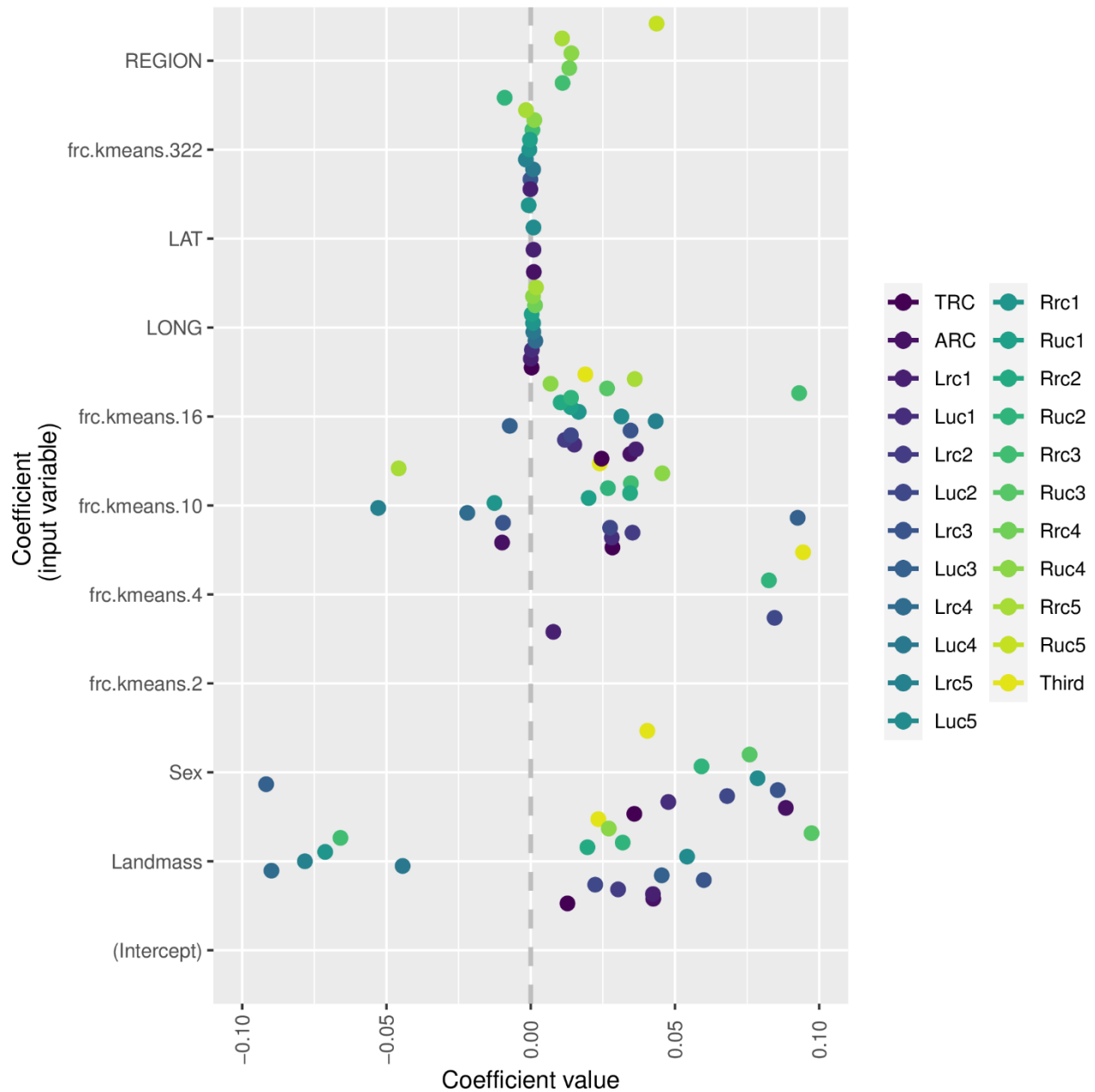


Figure 40. GLM coefficient value for each explanatory variable, colored by TRC, ARC, or finger ridge-counts response variable, zoomed to range -0.1 to 0.1. (N = 13961); ARC = absolute ridge-count; TRC = total ridge-count

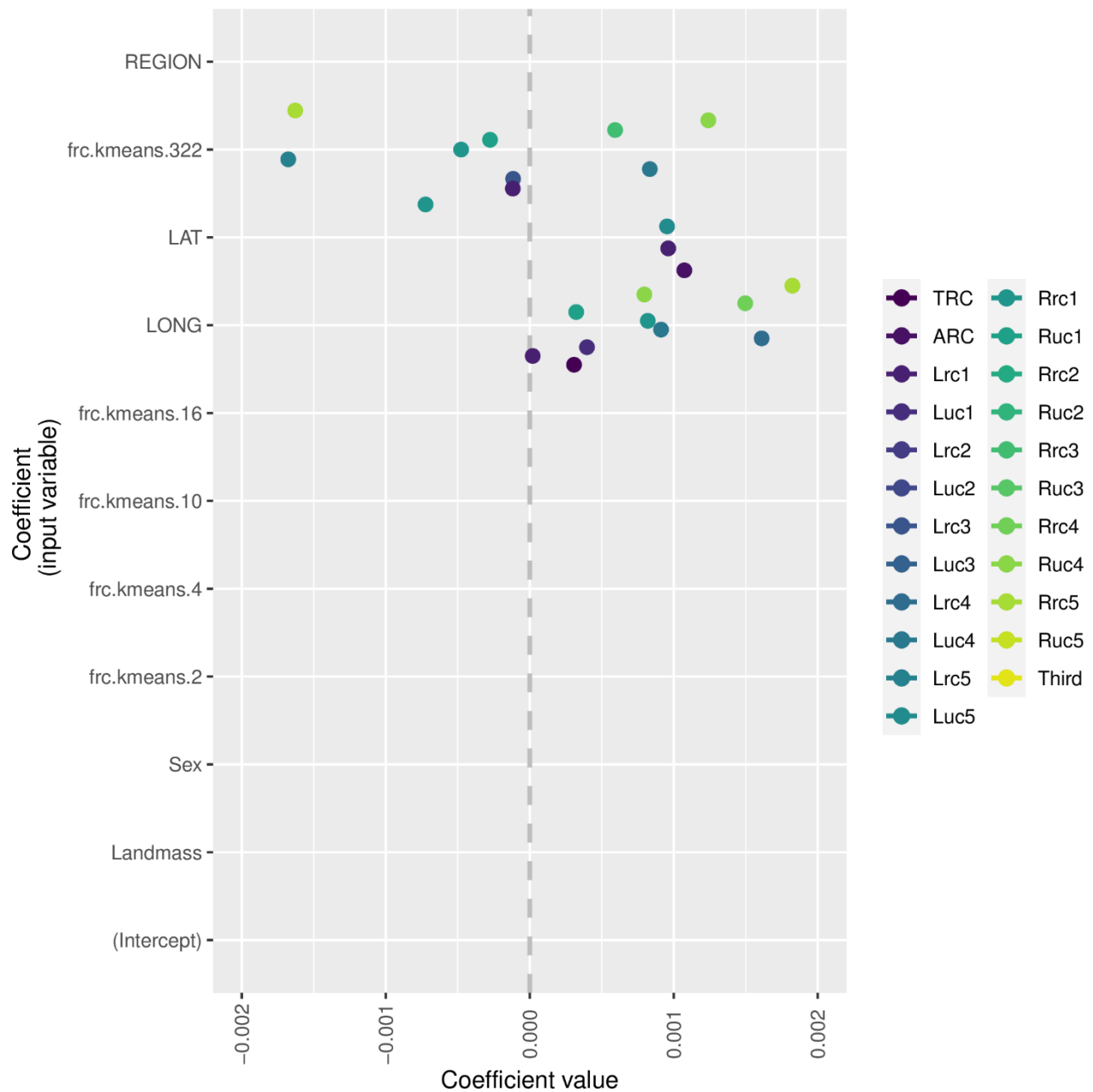


Figure 41. GLM coefficient value for each explanatory variable, colored by TRC, ARC, or finger ridge-counts response variable, zoomed to range -0.002 to 0.002. (N = 13961); ARC = absolute ridge-count; TRC = total ridge-count

When the MAPE was undefined (NaN), the second best performance metric for these response variables, the MAE, was considered. For performance, no single modeling approach exhibited the best performance metric scores. No GLMs performed the best for any of the response variables of the TRC, ARC, and finger ridge-counts (Table 14). The GLM of the Third ridge-count response variable exhibited the lowest MAPE 1.009 with the fitted explanatory variables: Region, Landmass, and frc.kmeans.2; and the next lowest MAPE (2.364) of GLM was for response variable left ulnar count of digit 1 including the fitted explanatory variables: Ethnic Group, Landmass, Sex, frc.kmeans.2, frc.kmeans.4, frc.kmeans.10, frc.kmeans.16, and frc.kmeans.322.

Overall, the baseline linear model performed best for the majority of response variables, the mean or median regressor performed best for left radial count of digit 2, left ulnar count of digit 2, left ulnar count of digit 5, right radial count of digit 2, right ulnar count of digit 2, right ulnar count of digit 4, right ulnar count of digit 5, and Third ridge-count, and the GLM performed best for no response variables of the TRC, ARC, and finger ridge-counts.

4.1.6 Baseline models for support vector regression

For the SVR of TRC and ARC response variables the 50% training set contained 23319 observations of 6 attributes: Sex, Ethnic Group, Region, Latitude, Longitude, and Landmass, for grid searching hyperparameter optimization the 25% validation set contained 11625 observations of the same 6 attributes, and for evaluating the final model a separate 25% test set containing 11604 observations of the same 6 attributes was used to calculate performance metrics.

To evaluate the assumptions of the mean regressor as an approximation of the central tendency of the response variables, the similarity to the median regressor was visualized, and for the TRC and ARC response variables the mean of each appeared similar to the median of each

(Figure 8, Figure 9). To evaluate whether the baseline linear model could be interpreted in terms of goodness-of-fit or performance, initial model checking was performed to see whether the assumptions of linear models were met or could be approximately met, since dermatoglyphic features like TRC do not exhibit a normal distribution but instead were negatively skewed and flattened for both females and males from a British population (Holt 1955).

Both baseline linear models for the SVR exhibited high multicollinearity (high > 10 , moderate 5-10, low < 5) in the explanatory variables as computed via the variance inflation factor, which assesses how much the linear correlation of an explanatory variable with other explanatory variables increases the variance of the coefficient estimate compared to theoretically no correlation (Table 16). The linear model with TRC as the response variable exhibited high collinearity for explanatory variables Ethnic Group and Longitude (Figure S65). The linear model with ARC as the response variable exhibited high collinearity (> 1000) for explanatory variables Ethnic Group and Longitude (Figure S63). For both of these linear models, the explanatory variable Sex was the only other feature selected variable, which exhibited low collinearity. The consequences of continuing to use a model known to exhibit high collinearity between explanatory variables are primarily relegated to interpretation of coefficients and resulting p values, which were not the focus of these models in this dissertation. Further, when collinearity is present, the predictions, precision of the predictions, and the goodness-of-fit statistics do not change (Neter 1996:289). Solutions to collinearity between explanatory variables are to remove the highly correlated explanatory variables, to perform LASSO regression that can handle collinearity, etc. (Frost 2017).

For both linear models with response variables TRC and ARC, the Anderson-Darling normality test p value was < 0.05 , so reject the null hypotheses the errors were from a normal

Table 16. Violation of standard regression assumptions of residuals for baseline linear models for SVR of TRC and ARC; ARC = absolute ridge-count; SVR = support vector regression; TRC = total ridge-count; p value < 0.05 in bold

SVR	Multicollinearity	Anderson-Darling Normality		Durbin-Watson Autocorrelation		Breusch-Pagan Heteroscedasticity
Response variable	Variance Inflation Factor > 10	A test statistic	p value < 0.001	d test statistic	p value	p value < 0.001
TRC	✓	83.038	✓	1.988	0.002	✓
ARC	✓	61.449	✓	1.989	0.008	✓

distribution (Table 16). For a visual assessment of normality, the normality of residuals plots exhibit an almost perfect one-to-one line toward the center comparing the standard normal distribution quantiles on the x-axis as the residuals from a perfect model to the sample quantiles on the y-axis as the real residuals from our models, but the relationship breaks down as seen from the dots falling away from the line particularly when the standard normal distribution quantiles on the x-axis dip below -2.5 and exceed 2.5 at the tails, and for the histograms of residuals the lower and upper tails were unequal (Figure S63, Figure S64, Figure S65, Figure S66). Thus, both linear models with response variables TRC and ARC exhibited non-normal distribution of errors, which could be attributed to autocorrelation. Other reasons for non-normal distribution of errors could be an important explanatory variable not included in the model or some polynomial or interaction term not included in the model. The consequences of continuing to use a model known to violate the normality of residuals assumption are primarily biased estimate of standard errors affecting interpretation of confidence intervals and the p values, which were not the focus of these models in this dissertation, and with large sample sizes where the number of observations per variable exceeds 10 the impact of the negative consequences diminishes (Schmidt and Finan 2018).

One potential source of the non-normally distributed errors from linear models with response variables TRC and ARC was evaluated with the Durbin-Watson test for autocorrelation of residuals, indicated by a p value < 0.05 (Table 16). Autocorrelation of residuals was detected by the Durbin-Watson test for both response variables. However, in practical use any Durbin-Watson test statistic value equal to 2 indicated no autocorrelation and between 1.5 and 2.5 could be approximated as no autocorrelation (Bedre 2019; SE n.d.). While non-normality of residuals does not pose a serious threat to the interpretation of the baseline linear models, the

consequences of continuing to use a model known to exhibit autocorrelation of residuals, even with large sample sizes, are great. Autocorrelation of residuals reduces the interpretability of the estimates of the coefficients but also their standard errors. As a result, the p values would be biased, and the standard error of the regression coefficients may underestimate the true standard deviation of the estimated regression coefficients, so the mean squared error performance metric may greatly underestimate the true variance of the errors. Similarly, the RMSE would be unreliable (Broersen 2005). Given no Durbin-Watson test statistics exceeded the bounds of the practical use of the Durbin-Watson test, neither of the two baseline linear models with response variables TRC and ARC were treated as autocorrelated (Bedre 2019; SE n.d.).

Other violations of linear regression assumptions that can seriously bias the interpretation of the model even with a large sample size are the presence of extreme observations that unduly influence the regression and also the heteroscedasticity that reduces the interpretation of the coefficients. For all linear models with response variables TRC and ARC, the results of the heteroscedasticity check exhibited a p value < 0.001 , so reject the null hypothesis the error variance was constant. Further, the homogeneity of variance plots for TRC and ARC as the response variables did not show a flat and horizontal reference line comparing the fitted values on the x-axis to the square root of the standardized residuals on the y-axis (Figure S63, Figure S65). Heteroscedasticity can occur when a large range separates the smallest and largest observed values, like with TRC and ARC response variables. The consequences of continuing to use a model known to exhibit heteroscedasticity rely on identifying the purity of the heteroscedasticity and so do the solutions. For impure heteroscedasticity where the model was not correctly specified by leaving an important variable out, and the generation of the coefficient values was not biased, so opportunities exist to solve the problem with additional explanatory

variables. For pure heteroscedasticity where the model is correctly specified and yet non-constant variance in the residuals persists, the issues with interpretation are the same as with autocorrelation, and the prediction error significantly changes over the range of prediction of the model, so the performance metrics, which rely on the prediction error likely represent an underestimate of the performance error, or a minimum. For the baseline linear models exhibiting significant heteroscedasticity, the performance metrics were interpreted as a minimum value, meaning if the comparison in performance between the baseline linear model that exhibits heteroscedasticity and the mean regressor were close in value, the mean regressor would likely be identified as the better performing baseline. Solutions to pure heteroscedasticity are the use of weighted least squares, obtaining standard errors corrected for heteroscedasticity, or retooling the model.

Calculating a goodness-of-fit for each model indicates how well the explanatory variables in the model explain the observed variation in the response variable. However, comparing goodness-of-fit values for multiple models with different response variables is not appropriate. So, only the AIC of the full model was compared to the reduced model containing the same response variable and model formulation to indicate how the feature selection process improved the fit of the model (Schneider, Hommel, and Blettner 2010). For the mean regressor, median regressor, and linear regression model as baseline comparisons for the SVR, only the linear model incorporated explanatory variables in the model construction, and so the goodness-of-fit was only assessed for the linear model as a baseline via the AIC value, for which lower values indicate a better model fit (Table 17).

The linear model with TRC as the response variable exhibited the greatest reduction in AIC (Table 17). The linear models with TRC and ARC as the response variables for SVR

Table 17. Goodness-of-fit for full and reduced baseline linear models of TRC, ARC, and finger ridge-counts, of same SVR data set;

AIC = Akaike information criterion; ARC = absolute ridge-count; SVR = support vector regression; TRC = total ridge-count

Analytical Method	Response Variable	Explanatory Variables	Model	AIC
LM	TRC	Sex + Ethnic Group + Longitude	full	246254.8
			reduced	246221.7
LM	ARC	Sex + Ethnic Group + Longitude	full	273853.6
			reduced	273822.7

exhibited a reduction in AIC through feature selection approximately ranging between 30 and 33, which was consistent with the reduction in AIC through feature selection for TRC, ARC, and finger ridge-count response variables for GLM (Table 13). Of note, the reduced models for the baseline linear models of both TRC and ARC as response variables for SVR indicated a marked improvement in the goodness-of-fit.

The best performance metric for the TRC and ARC response variables was the MAPE due to the presence of outliers that could unduly influence the modeling, the discrete, non-negative counts that were not truly continuous response variables, and the large values that can directly inflate and be reflected in the performance metric calculation, that make the other two performance metrics inappropriate in this context (Table 4). The zeros in the response variables did not result in an undefined MAPE.

The SVR performed the best for the response variable TRC using the MAPE (Table 18). Subsequently, the median regressor, linear model, and then mean regressor followed in MAPE performance. The linear model performed the best for the response variable ARC using the MAPE. Subsequently, the mean regressor, SVR, and then median regressor followed in MAPE performance. Although the TRC considers only the largest (usually radial) ridge-count for all digits and the ARC considers all ridge-counts for all digits, which is no great calculation difference, and both response variables exhibited the same feature selected explanatory variables, the order of performance of the models differed greatly. Because the MAE and not RMSE or MAPE can be compared across models with different response variables given most other conditions were consistent, and that the MAE was inappropriate for the TRC and ARC response variables for the SVR, no comparisons across response variables were attempted. Because the baseline linear model outperformed the other baselines and the SVR for response variable ARC,

Table 18. Sample sizes and performance metrics of final optimized SVR models for TRC and ARC; ARC = absolute ridge-count; MAE = mean absolute error; MAPE = mean absolute percentage error; RMSE = root mean squared error; TRC = total ridge-count; best performance metric and model in bold

Analytical Method	Response variable	Training data N	Validation data N	Test data N	MAE	RMSE	MAPE
Baseline for SVR	mean TRC	23319	11625	11604	39.334	49.090	0.280
Baseline for SVR	median TRC	23319	11625	11604	39.281	49.543	0.269
Baseline LM	TRC	23319	11625	11604	37.585	46.997	0.273
SVR	TRC	23319	11625	11604	37.643	47.570	0.262
Baseline for SVR	mean ARC	23319	11625	11604	71.989	88.004	0.387
Baseline for SVR	median ARC	23319	11625	11604	71.366	88.167	0.405
Baseline LM	ARC	23319	11625	11604	68.691	84.276	0.376
SVR	ARC	23319	11625	11604	68.272	84.423	0.391

the interpretation of the linear model in context of their assumptions, like the increase in prediction error due to heteroscedasticity, becomes more important.

4.1.7 Support vector regression (SVR)

The goodness-of-fit metric for the GLM, the Kullback-Leibler R^2 value, was not appropriate for SVR models, and traditional goodness-of-fit measures like the R^2 value or an information criterion were also not appropriate for SVR models, so no goodness-of-fit measure was evaluated for SVR models.

For the baseline support vector regression model, hyperparameter gamma was set to 1/number of column variables in the training data, which equated to 1/7 for the seven variables: response variable like TRC or ARC, and explanatory variables Sex, Ethnic Group, Region, Latitude, Longitude, and Landmass, and hyperparameter cost was set to 1. Figure S67 and Figure S68 illustrated the progression of grid searching for TRC and ARC and how the lower/upper boundaries of the optimal hyperparameters were identified as a vein of hyperparameters with lower RMSE values surrounded by hyperparameters with higher RMSE values. The coarse grid search for hyperparameters of the SVR model for TRC encompassed a gamma of 10^{-8} to 10^8 and a cost of 2^{-16} to 2^{16} , incremented by values of 1, and all SVR models started grid searching with this range of values except for ARC, for which the coarse grid search began with gamma values 10^{-5} to 10^5 and cost values of 2^{-10} to 2^{10} (Table S12). Even without an initially broader range for coarse grid searching, the limits of the optimal hyperparameters for the SVR ARC model were identified (Figure S68).

Because the performance metrics for the SVR models can be most directly compared when the models were generated from the same data sets, Table S12 illustrates the sample size for each model formulation, and because the TRC and ARC SVR models contained the same

underlying data, the final performance metrics can be directly compared between the two models. Table S12 reports the performance metrics: MAE, RMSE, and MAPE for each step of grid searching for TRC and ARC. Because the procession into a narrower grid search step relied on the RMSE, the RMSE and performance metrics overall decrease throughout grid searching, indicating a successful identification of ranges of optimal hyperparameter values. Neither of the final models exhibited higher RMSE values as a result of overfitting to the training data, so when the model was applied to the test data the model did generalize well (Table S12).

The best performance metric for the TRC and ARC response variables was the MAPE due to the presence of outliers that could unduly influence the modeling, the discrete, non-negative counts that were not truly continuous response variables, and the large values that can directly inflate and be reflected in the performance metric calculation, that make the other two performance metrics inappropriate in this context (Table 4). The zeros in the response variables did not result in an undefined MAPE. The SVR performed the best for the response variable TRC using the MAPE (Table 18). Overall, the ARC response variable exhibited performance metrics close to double those for the TRC response variable, which may represent the ARC inclusion of all 20 finger ridge-counts and the reduction of TRC to include only the largest ridge-count from the 10 digits.

4.2 Results for directional asymmetry, fluctuating asymmetry, and indices of asymmetry and diversity

The results for the Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response variables were separated by the method of analysis.

4.2.1 k-means clustering

For Directional Asymmetry, the optimal K using the gap statistic was 1 cluster, around 10 clusters for the within-cluster sum of squares approach, and 2 clusters for the average silhouette approach (Figure 42). For Fluctuating Asymmetry, the optimal K using the gap statistic was 1 cluster, around 8 clusters for the within-cluster sum of squares approach, and 2 clusters for the average silhouette approach (Figure 43). The single cluster suggested by the gap statistic did not lend any new information so was not included in any subsequent evaluations or analyses.

The internal performance of k-means clustering for Directional Asymmetry, evaluated by calculating the Dunn Index for the cluster assignments of different optimal K values, ranged from 0.016 to 0.032. Because the Dunn Index should be maximized, the optimal K value of 322 clusters from the number of Ethnic Groups with Dunn Index of 0.032 represents the most appropriate number of clusters of those evaluated (Table 19). But among the three approaches of optimal K selection, the optimal K value of 10 clusters from the within-cluster sum of squares approach with the Dunn Index of 0.018 represents the most appropriate number of clusters.

The internal performance of k-means clustering for Fluctuating Asymmetry, evaluated by calculating the Dunn Index for the cluster assignments of different optimal K values, ranged from 0.021 to 0.045. Because the Dunn Index should be maximized, the optimal K value of 322 clusters from the number of Ethnic Groups with Dunn Index of 0.045 represents the most appropriate number of clusters of those evaluated (Table 20). But among the three approaches of optimal K selection, the optimal K value of 8 clusters from the within-cluster sum of squares approach with the Dunn Index of 0.026 represents the most appropriate number of clusters.

The cluster assignments from k-means clustering of Directional Asymmetry variables and Fluctuating Asymmetry variables were included in subsequent GLM and SVR as additional

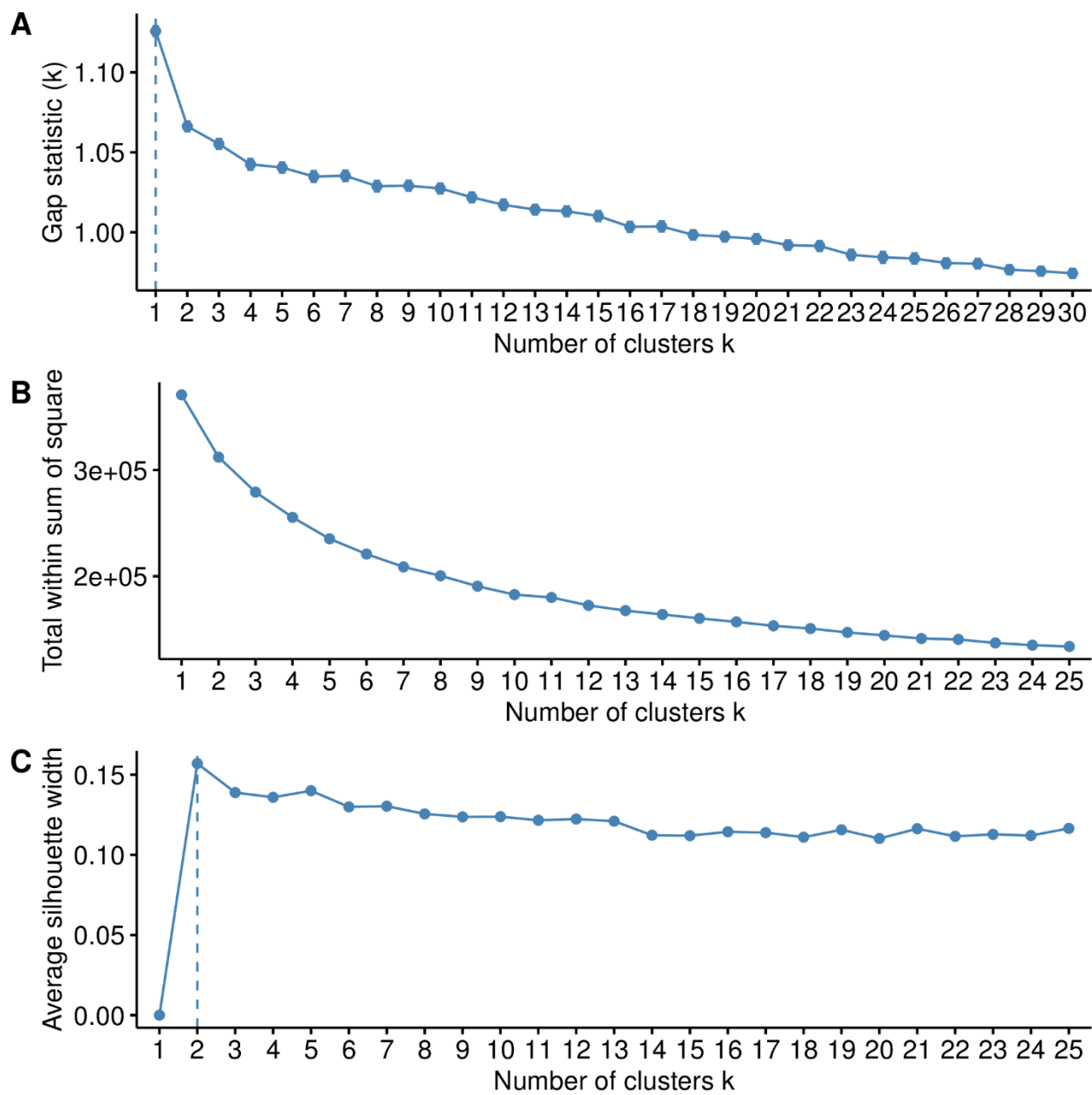


Figure 42. Directional Asymmetry optimal K for three approaches: (A) the gap statistic, (B) within-cluster sum of squares, and (C) average silhouette.

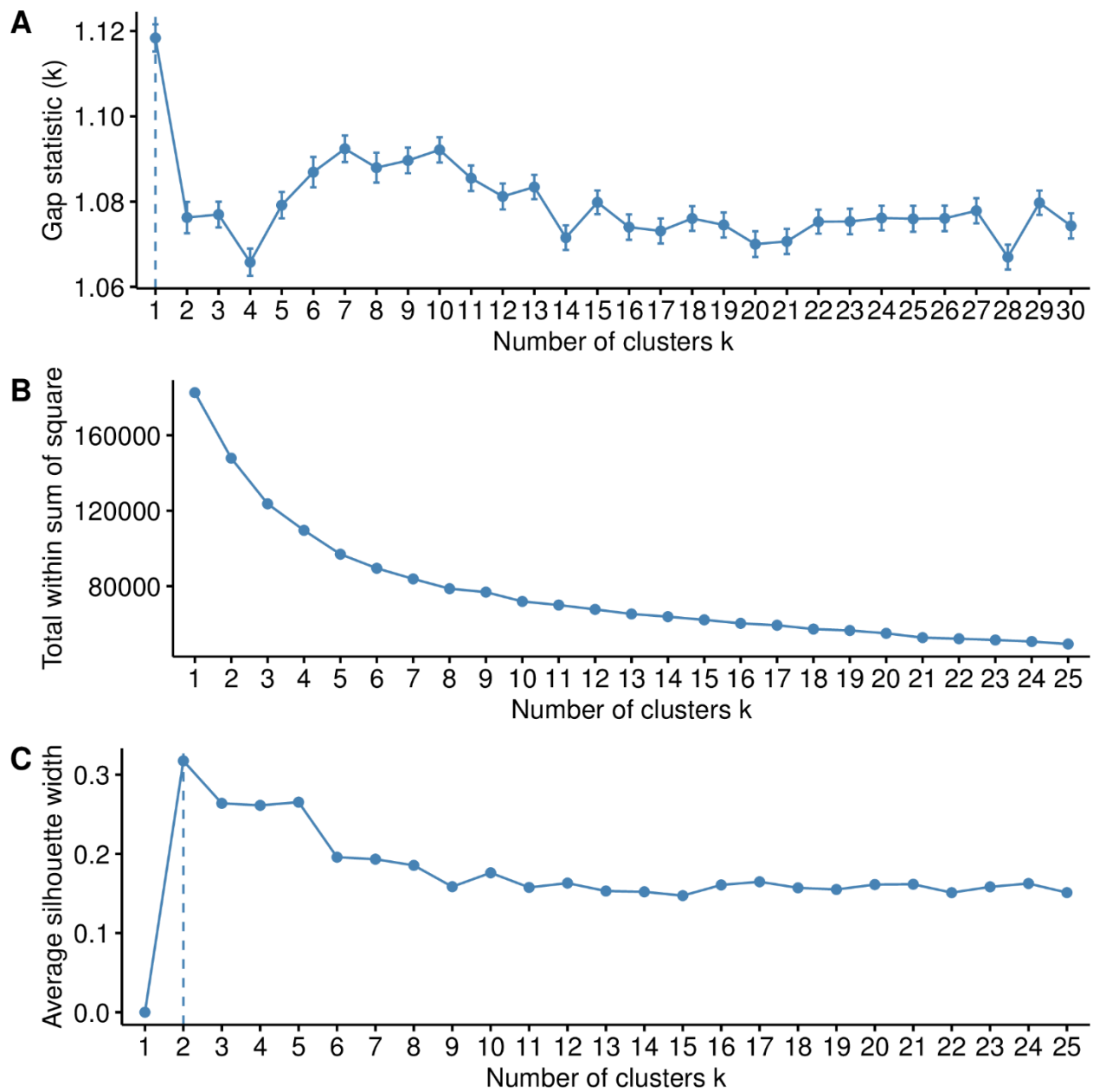


Figure 43. Fluctuating Asymmetry optimal K for three approaches: (A) the gap statistic, (B) within-cluster sum of squares, and (C) average silhouette.

Table 19. Dunn Index values for optimal K values for Directional Asymmetry. Dunn Index values in decreasing order; the Dunn Index should be maximized.

Number of Clusters	Approach for K selection	Dunn Index
322	Number of Ethnic Groups	0.032
10	Within sum of squares	0.018
4	Number of Landmasses	0.017
2	Average silhouette	0.016

Table 20. Dunn Index values for optimal K values for Fluctuating Asymmetry. Dunn Index values in decreasing order; the Dunn Index should be maximized.

Number of Clusters	Approach for K selection	Dunn Index
322	Number of Ethnic Groups	0.045
8	Within sum of squares	0.026
4	Number of Landmasses	0.023
2	Average silhouette	0.021

explanatory variables. To interpret the cluster assignments, stacked bar charts were created for Directional Asymmetry $K = 2, 4$, and 10 , Fluctuating Asymmetry $K = 2, 4$, and 8 , and compared to known groupings for Directional Asymmetry or Fluctuating Asymmetry using packages tidyverse, viridis, viridisLite, scales to set the viridis color palette for accessibility, cowplot for converting base R plots into ggplot-friendly objects, and ggpubr for multipanel plots (Wickham et al. 2019; Wilke 2020; Garnier et al. 2022; Wickham, Seidel, and RStudio 2022; Kassambara 2023; Wickham and RStudio 2023).

If the cluster assignments matched known groupings, then a known grouping would contain the predominately one cluster. For example, if $K = 2$ corresponded to Sex, then females would primarily contain one cluster, and males would primarily contain the second cluster (Figure 44, Figure 45). If $K = 4$ corresponded to Landmass, then each Landmass would primarily contain one cluster assignment each (Figure 46, Figure 47). The cluster assignments for $K = 10$ for Directional Asymmetry and $K = 8$ for Fluctuating Asymmetry did not readily match the number of levels in any known groupings, and the closest number of levels in a known grouping was 16, the number of Ethnic Groups. If $K = 10$ or 8 approximately corresponded to Ethnic Group, then each Ethnic Group would primarily contain one cluster assignment each (Figure 48, Figure 49). The cluster assignments for $K = 322$ were not compared to the number of Ethnic Groups in a stacked bar chart, because the chart would not be readable with 322 groups on the x-axis.

The stacked bar charts for Directional Asymmetry $K = 2, 4$, and 10 exhibited evenly distributed clusters for known groupings of Landmass (Figure 44A, Figure 46A, Figure 48A), Region (Figure 44B, Figure 46B, Figure 48B), and Sex (Figure 44C, Figure 46C, Figure 48C), which suggests $K = 2$ does not represent Sex, $K = 4$ does not represent Landmass, and $K = 10$

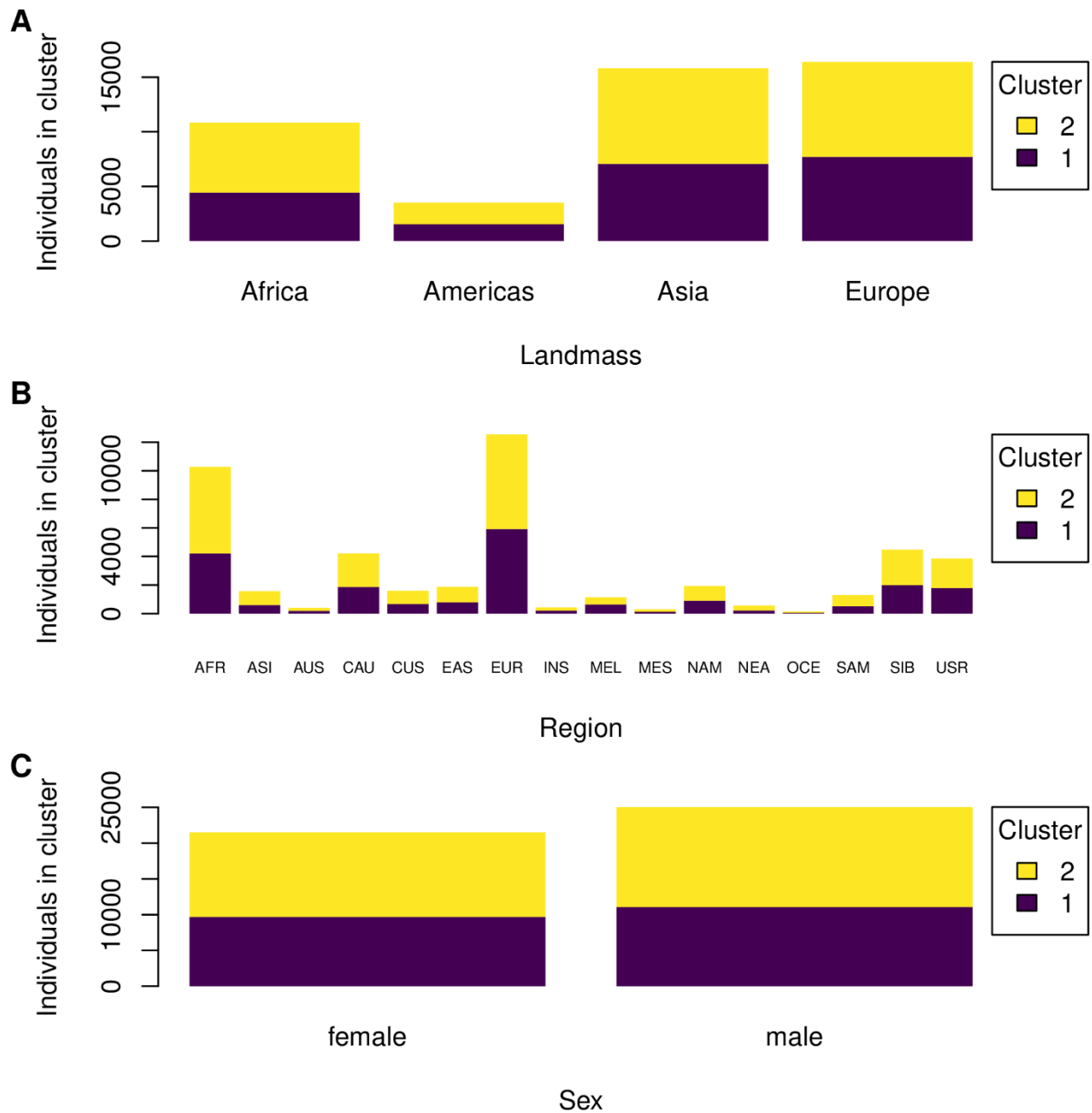


Figure 44. Stacked bar chart for $K = 2$ from the average silhouette method for Directional Asymmetry. Grouped (A) by-Landmass, (B) by-Region, and (C) by-Sex

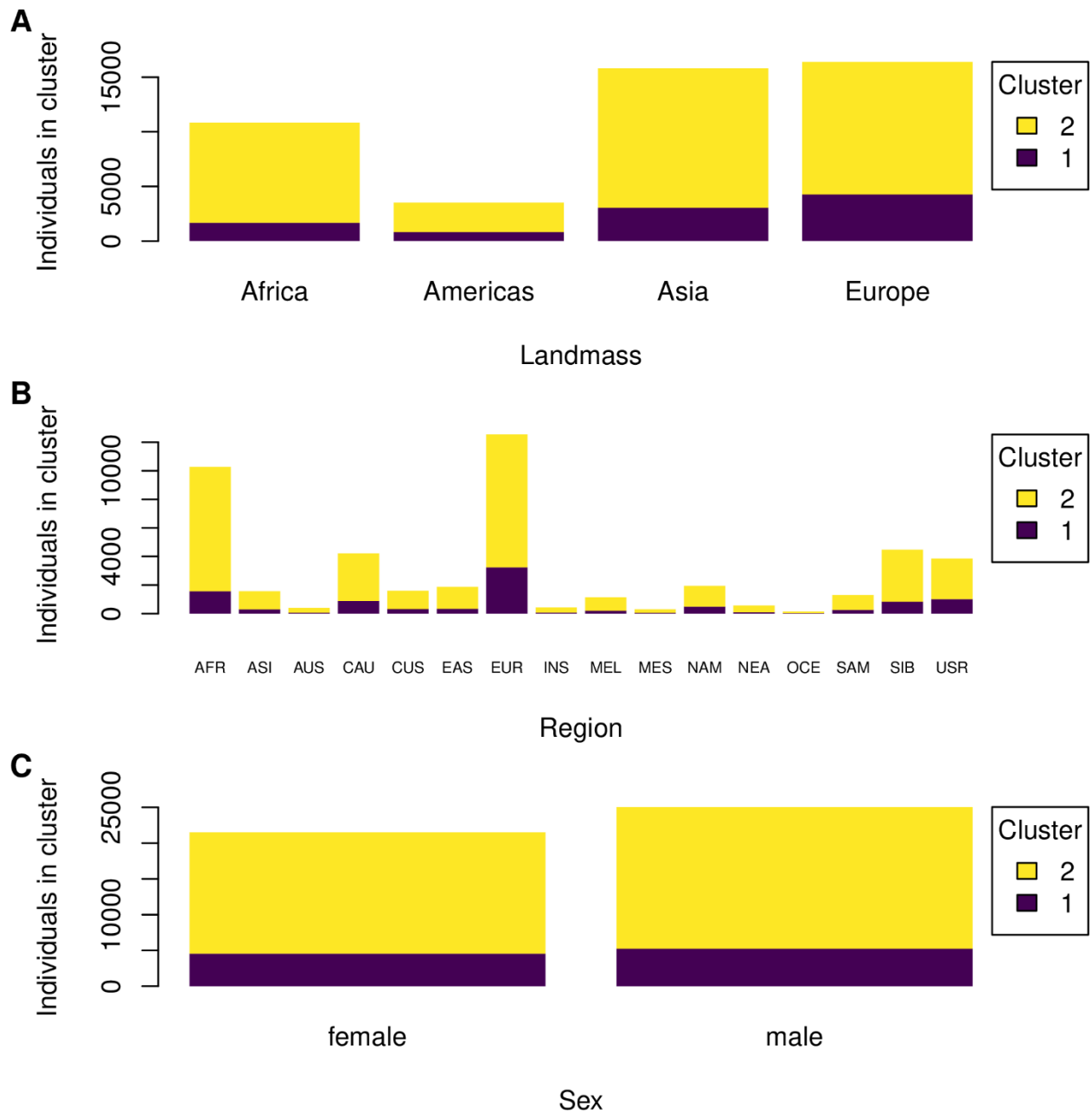


Figure 45. Stacked bar chart for $K = 2$ from the average silhouette method for Fluctuating Asymmetry. Grouped (A) by-Landmass, (B) by-Region, and (C) by-Sex

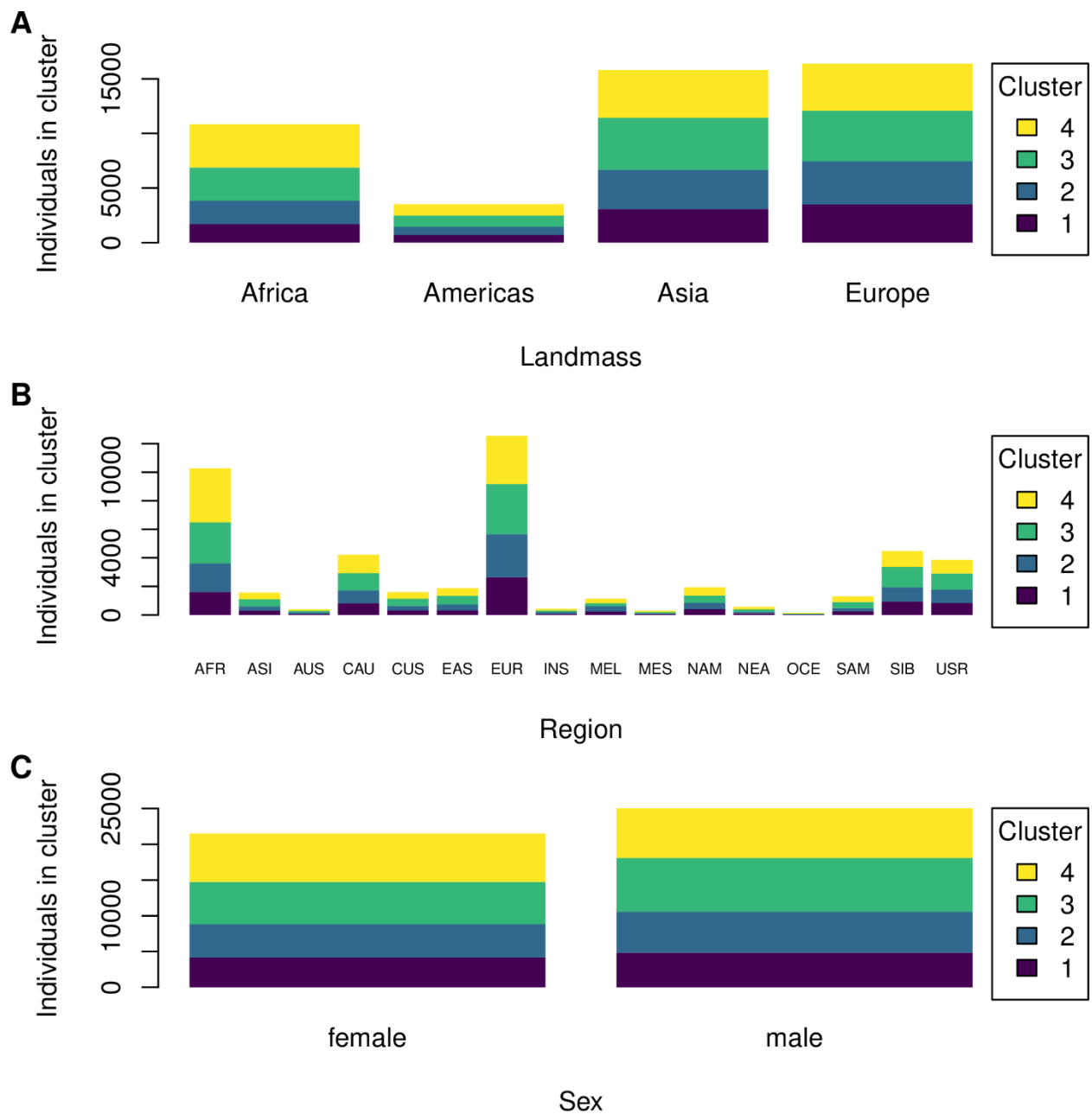


Figure 46. Stacked bar chart for $K = 4$ for Directional Asymmetry. Grouped (A) by-Landmass, (B) by-Region, and (C) by-Sex

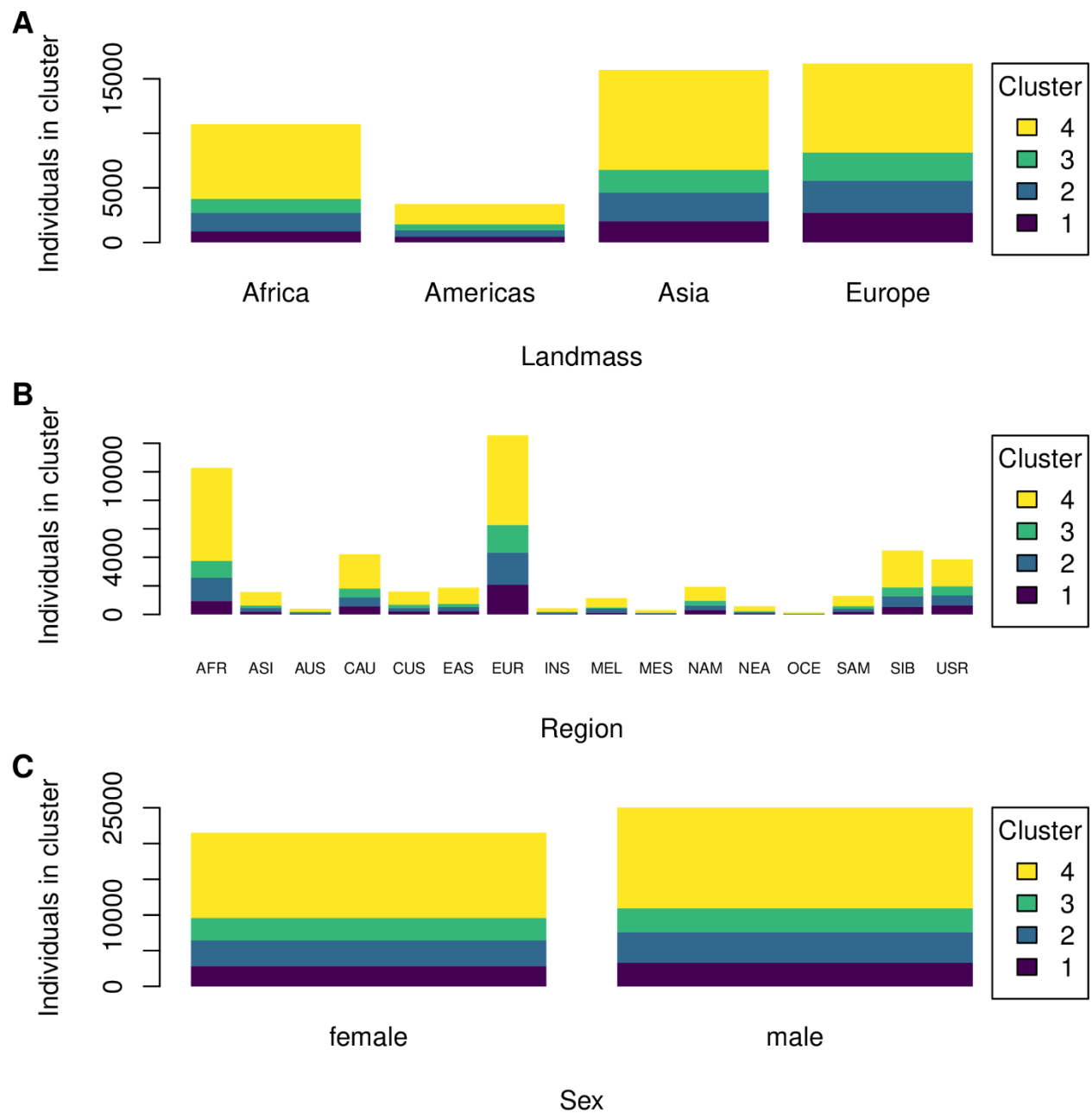


Figure 47. Stacked bar chart for $K = 4$ for Fluctuating Asymmetry. Grouped (A) by-Landmass, (B) by-Region, and (C) by-Sex

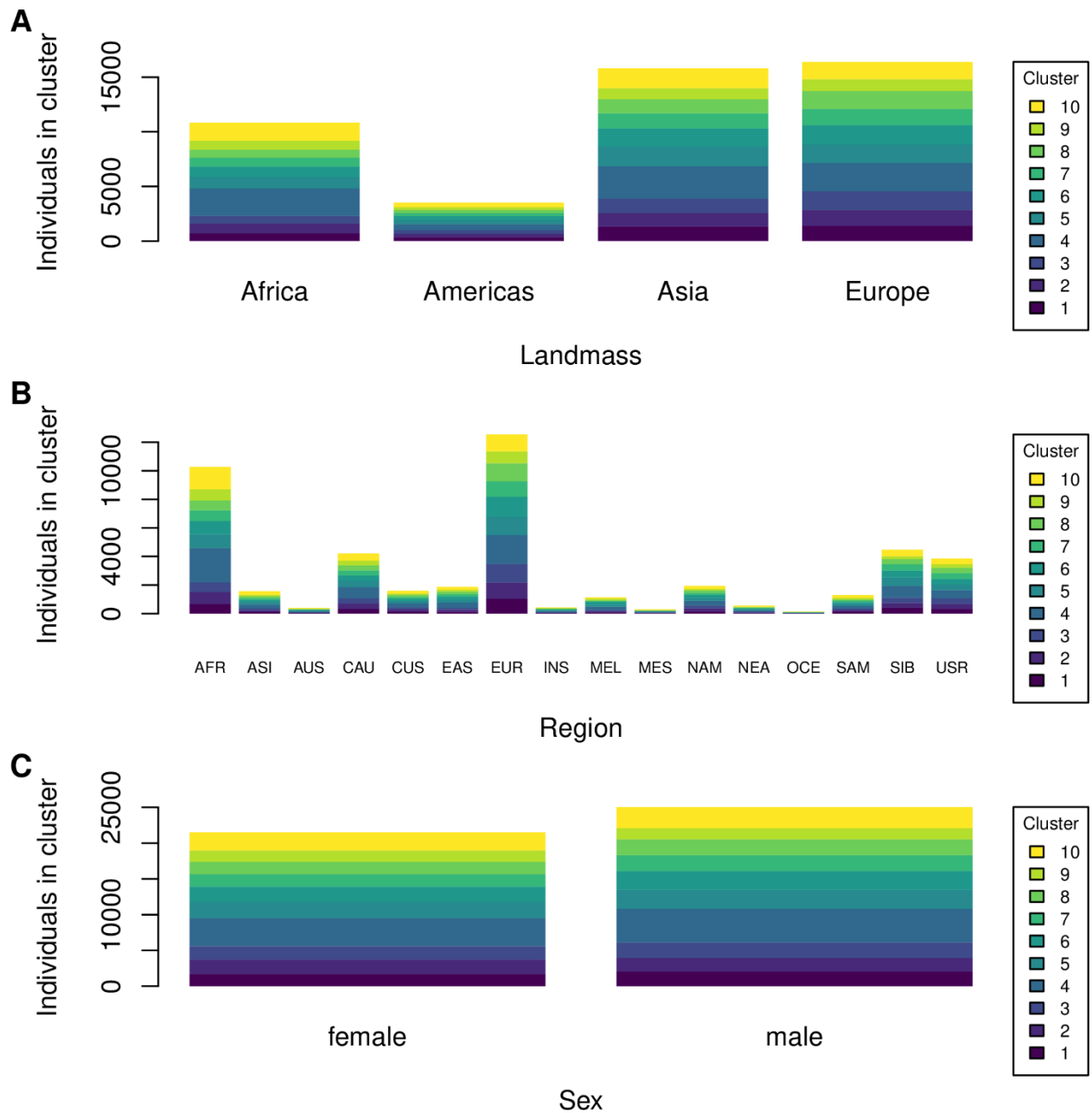


Figure 48. Stacked bar chart for $K = 10$ from the within-cluster sum of squares method for Directional Asymmetry. Grouped (A) by-Landmass, (B) by-Region, and (C) by-Sex

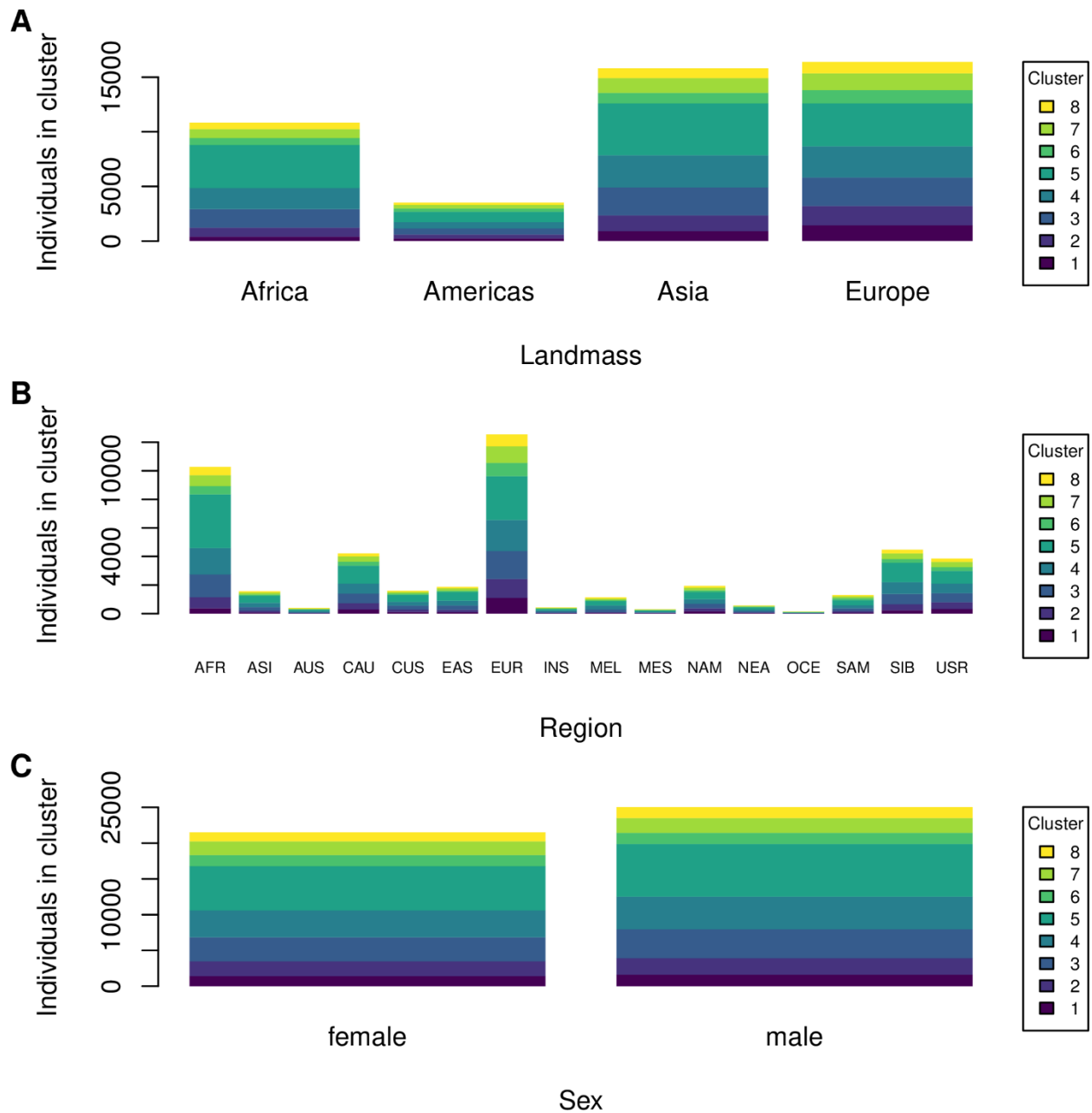


Figure 49. Stacked bar chart for $K = 8$ from the within-cluster sum of squares method for Fluctuating Asymmetry. Grouped (A) by-Landmass, (B) by-Region, and (C) by-Sex

does not represent Ethnic Group. The stacked bar charts for Fluctuating Asymmetry $K = 2, 4,$ and 8 exhibited clusters not primarily related to known groupings of Landmass (Figure 45A, Figure 47A, Figure 49A), Region (Figure 45B, Figure 47B, Figure 49B), and Sex (Figure 45C, Figure 47C, Figure 49C), which suggests $K = 2$ does not represent Sex, $K = 4$ does not represent Landmass, and $K = 8$ does not represent Ethnic Group. For Fluctuating Asymmetry, the clusters were not evenly distributed across the known groupings as was seen with Directional Asymmetry. Each stacked bar chart for Fluctuating Asymmetry $K = 2, 4,$ and 8 exhibited a single popular assignment, no matter the known grouping. For example, the stacked bar charts for Directional Asymmetry $K = 2$ (Figure 44) exhibited even distribution of the two clusters across all four Landmasses, all 16 Regions, and both Sexes, but the stacked bar charts for Fluctuating Asymmetry $K = 2$ (Figure 45) revealed cluster 2 dominated the assignment in all four Landmasses, all 16 Regions, and both Sexes. The argument that the domination of a single cluster for Fluctuating Asymmetry stacked bar charts demonstrated a bias of raw frequencies of cluster assignments rather than relative frequencies of cluster assignments does not hold, as evident for $K = 2$, which reflects both a raw frequency and inherently a relative frequency. The cluster assignments from the various k-means clustering of Directional Asymmetry variables and of Fluctuating Asymmetry variables with different optimal K represent explanatory variables presently unknown.

4.2.2 Spatial analysis via distance-based Moran's eigenvector maps (dbMEM)

The sample size and thus the truncation distance for the Directional Asymmetry and Fluctuating Asymmetry for each Landmass was the same, reflecting how the rows present to calculate Directional Asymmetry were the same as those to calculate Fluctuating Asymmetry (Table 21). The truncation distance for Landmass Asia was an order of magnitude greater than

Table 21. Truncation distances for Directional Asymmetry and Fluctuating Asymmetry within
Landmass; m = meters

Dermatoglyphic feature	Landmass	Sample Size	Truncation distance (m)
Directional Asymmetry	Africa	10809	3067568
Fluctuating Asymmetry	Africa	10809	3067568
Directional Asymmetry	Americas	3522	3051902
Fluctuating Asymmetry	Americas	3522	3051902
Directional Asymmetry	Asia	15816	10702414
Fluctuating Asymmetry	Asia	15816	10702414
Directional Asymmetry	Europe	16402	1240025
Fluctuating Asymmetry	Europe	16402	1240025

for every other Landmass, Landmasses Africa and Americas exhibited similar truncation distances, and Landmass Europe had the smallest truncation distance.

The dbMEM eigenvectors with positive spatial correlation were recorded in Table 22 as number of dbMEMs. Each Landmass shared the same number of dbMEMs for both Directional Asymmetry and Fluctuating Asymmetry. The Landmasses Americas and Asia also shared the same number of dbMEMs at six, but the adjusted R^2 value for Landmass Americas Directional Asymmetry was smaller by an order of magnitude compared to Landmass Americas Fluctuating Asymmetry and both Landmass Asia Directional Asymmetry and Fluctuating Asymmetry.

The adjusted R^2 value for Landmass Americas Fluctuating Asymmetry and Landmass Asia Directional Asymmetry and Fluctuating Asymmetry was on the same order of magnitude as Landmass Europe Directional Asymmetry and Fluctuating Asymmetry (10^{-4}). The adjusted R^2 value for Landmass Americas Directional Asymmetry was on the same order of magnitude as Landmass Africa Directional Asymmetry and Fluctuating Asymmetry (10^{-3}).

The number of significant dbMEMs was identified through forward selection of the detrended response variables and the full dbMEM variables using the adjusted R^2 value as the threshold for selection, and all Landmasses and dermatoglyphic features exhibited a single significant dbMEM, except Landmass Americas Fluctuating Asymmetry for which no dbMEMs were significant. After calculating the significant dbMEMs the new adjusted R^2 values were on the same order of magnitude as before, except for Landmass Africa Fluctuating Asymmetry whose adjusted R^2 value were now an order of magnitude lower even though the magnitude of reduction was consistent with other Landmasses (Table 22). The adjusted R^2 values represent the percent variance explained by the dbMEM analysis of the detrended response variable.

Table 22. Number of dbMEMs and the adjusted R^2 and the number of significant dbMEMs and the adjusted R^2 for Directional Asymmetry and Fluctuating Asymmetry within Landmass; dbMEMs = distance-based Moran's eigenvector maps; R^2_{adj} = adjusted R squared

Dermatoglyphic feature	Landmass	Number of dbMEMs	R^2_{adj}	Number of significant dbMEMs	R^2_{adj}	Significant dbMEMs
Directional Asymmetry	Africa	17	0.0024	1	0.0010	4
Fluctuating Asymmetry	Africa	17	0.0018	1	0.0008	4
Directional Asymmetry	Americas	6	0.0036	1	0.0023	6
Fluctuating Asymmetry	Americas	6	0.0006	NA	NA	NA
Directional Asymmetry	Asia	6	0.0010	1	0.0003	1
Fluctuating Asymmetry	Asia	6	0.0005	1	0.0002	4
Directional Asymmetry	Europe	22	0.0005	1	0.0002	14
Fluctuating Asymmetry	Europe	22	0.0005	1	0.0003	4

The single significant dbMEM was 4 for many Landmasses and dermatoglyphic features, except for Directional Asymmetry of Landmasses Americas, Asia, and Europe.

The cumulative variance explained by the significant canonical axes out of the total variance is found in Table 23, and the plots representing the spatially structured variation in the detrended response variables are found in Figure S69 through Figure S75. Because each Landmass and dermatoglyphic feature exhibited a single significant dbMEM variable and a single significant canonical axis, Figure S69 through Figure S75 allowed discrimination between broad and fine scales of spatial patterning, so no additional maps were necessary. The relation of the significant canonical axes to the environmental explanatory variables through linear modeling created normally-distributed residuals as indicated by the Anderson-Darling normality test with an $\alpha = 0.05$ threshold for all dermatoglyphic feature types and Landmasses, except Landmass Americas Fluctuating Asymmetry for which no significant dbMEMs or axes were present.

For both the single significant axis for Landmass Africa Directional Asymmetry and for Landmass Africa Fluctuating Asymmetry, Ethnic Group Egyptians and Sex male were not significant explanatory variables of the spatially structured variation in the detrended finger patterns, and Ethnic Group Egyptians-Arab exhibited NA results (Figure S69, Figure S70). For the single significant axis for Landmass Americas Directional Asymmetry, Ethnic Groups Mixtecs, Yukpa, and Longitude exhibited NA results (Figure S71). For the single significant axis for Landmass Asia Directional Asymmetry, Ethnic Group Easter-Island and Sex male were not significant explanatory variables of the spatially structured variation in detrended Directional Asymmetry, and Ethnic Groups Easter-Island-Polynesians, Pinai, Rembranga, Timorese, Ulchis, Vietnamese, and Yaghnobis exhibited NA results (Figure S72). The results for the single

Table 23. Variance explained by the significant canonical axes for Directional Asymmetry and Fluctuating Asymmetry. Where the R^2_{adj} is the variance explained by the dbMEM analysis of the detrended response variables, the proportion of variance in the significant axes includes only the significant axes with respect to the total explained variance. The variance explained by the significant axes is the product of the proportion of variance in significant axes and the R^2_{adj} ; R^2_{adj} = adjusted R squared

Dermatoglyphic feature	Landmass	Number of significant axes	Proportion of variance in significant axes	R^2_{adj}	Variance explained by significant axes
Directional Asymmetry	Africa	1	1	0.0010	0.0010
Fluctuating Asymmetry	Africa	1	1	0.0008	0.0008
Directional Asymmetry	Americas	1	1	0.0023	0.0023
Fluctuating Asymmetry	Americas	NA	NA	NA	NA
Directional Asymmetry	Asia	1	1	0.0003	0.0003
Fluctuating Asymmetry	Asia	1	1	0.0002	0.0002
Directional Asymmetry	Europe	1	1	0.0002	0.0002
Fluctuating Asymmetry	Europe	1	1	0.0003	0.0003

significant axis for Landmass Asia Fluctuating Asymmetry were the same as for Directional Asymmetry, except Ethnic Group Meskhetian-Turks was also not a significant explanatory variable of the spatially structured variation in detrended Fluctuating Asymmetry (Figure S73). For the single significant axis for Landmass Europe Directional Asymmetry, only Ethnic Group Åland-Mariehamm was not a significant explanatory variable of the spatially structured variation in the detrended Directional Asymmetry (Figure S74). The results for the single significant axis for Landmass Europe Fluctuating Asymmetry were the same as for Directional Asymmetry, except Ethnic Group Volga-Tatars was also not a significant explanatory variable of the spatially structured variation in detrended Fluctuating Asymmetry (Figure S75).

Because each Landmass and dermatoglyphic feature only exhibited a single significant dbMEM, no predefined separation of dbMEMs within each Landmass for broad, medium, and fine scale was possible. The map of Landmass Africa Directional Asymmetry (Figure S69) matches the distribution of points as the map of the Landmass Africa Fluctuating Asymmetry (Figure S70), except the sign indicated by color is swapped, reflecting the proportional relationship between the calculation of Directional Asymmetry and Fluctuating Asymmetry. The maps of Landmass Americas Directional Asymmetry, Landmass Asia Directional Asymmetry, and Landmass Asia Fluctuating Asymmetry appeared compressed due to outlying groups (Figure S71, Figure S72, Figure S73). The map of Landmass Europe Directional Asymmetry exhibits complex patterning suggesting more fine scale patterning, while the map of Landmass Europe Fluctuating Asymmetry exhibited clear patterning suggesting broad scale patterning indicated by the significant dbMEM (Figure S74, Figure S75). Overall, the spatial patterning of Landmass Africa (Figure S69, Figure S70), Landmass Americas Directional Asymmetry (Figure S71), Landmass Asia Directional Asymmetry (Figure S72), and Landmass Europe Fluctuating

Asymmetry (Figure S75) exhibit layering or sandwiching such that a sign indicated by color is surrounded on multiple sides by the opposite sign indicated by the opposing color.

The results of variation partitioning should only be interpreted if the pure environmental, pure linear trend, pure broad spatial scale, and pure fine spatial scale components were significant at $\alpha = 0.05$ threshold, and the four components were significant at $\alpha = 0.05$ threshold for the majority of Landmasses and dermatoglyphic features. However, Landmass Americas Fluctuating Asymmetry pure trend scale component was not significant at $\alpha = 0.05$ threshold, and Landmass Europe Fluctuating Asymmetry has no constrained component for the pure environmental scale component so no significance was tested. The majority of variation of the undetrended response variables was not captured by the fractions considered during variation partitioning, as evidenced by the high residual values (Figure 50). Note the total of the partitions including the residual value sometimes exceeds one or does not total to one, due to the variation partitioning protocol subtracting of R^2 values that were adjusted based on a different number of explanatory variables. Any sum of partitions was reached by summing the actual values in each partition rather than subtracting other partitions from the value one. The amount of spatially structured variation captured by the variation partitioning was 0.5% for Landmass Africa Directional Asymmetry, 0.3% for Landmass Africa Fluctuating Asymmetry, 1.1% for Landmass Americas Directional Asymmetry, 0.4% for Landmass Asia Directional Asymmetry, 0.2% for Landmass Asia Fluctuating Asymmetry, 0.2% for Landmass Europe Directional Asymmetry, and 0% for Landmass Europe Fluctuating Asymmetry.

Considering how little variation was captured by the variation partitioning, the variation appears to be close to evenly distributed in Landmass Africa Directional Asymmetry

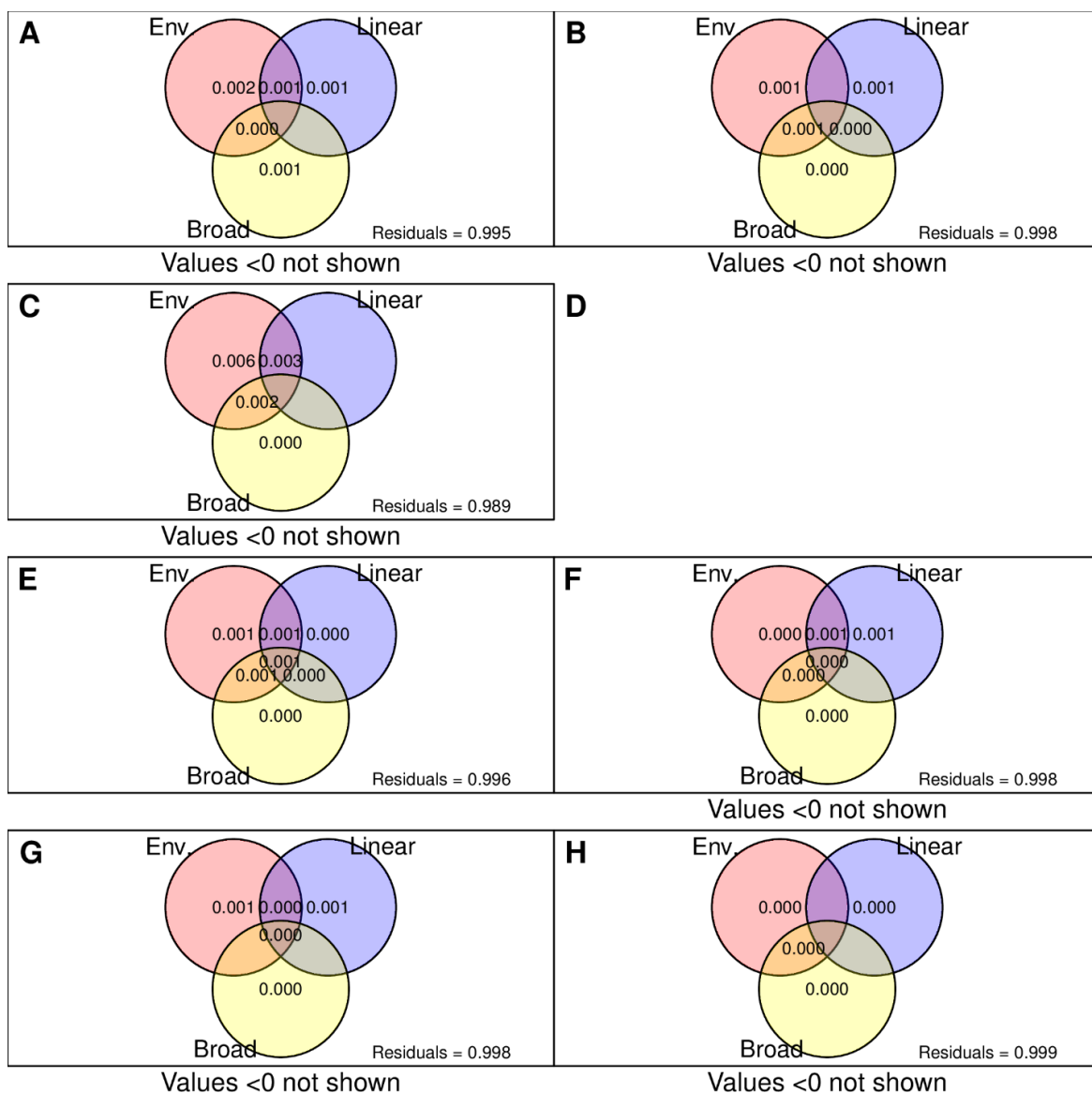


Figure 50. Variation partitioning of undetrended Directional Asymmetry and Fluctuating Asymmetry response variables into and environmental component, a linear trend, a broad scale, and a fine scale for (A) Landmass Africa Directional Asymmetry, (B) Landmass Africa Fluctuating Asymmetry, (C) Landmass Americas Directional Asymmetry, (D) nothing for Landmass Americas Fluctuating Asymmetry, (E) Landmass Asia Directional Asymmetry, (F) Landmass Asia Fluctuating Asymmetry, (G) Landmass Europe Directional Asymmetry, and (H) Landmass Europe Fluctuating Asymmetry; Env. = environmental

(Figure 50A), but no variation was purely broad scale in Landmass Africa Fluctuating Asymmetry (Figure 50B). The environmental component and associated was the only scale with any percentage for Landmass Americas Directional Asymmetry, indicating all the variation captured in the variation partitioning was not spatially structured (Figure 50C). Similarly, for Landmass Asia Directional Asymmetry the environmental component and associated was the only scale with any percentage (Figure 50E), and for Landmass Asia Fluctuating Asymmetry the linear trend and associated was the only scale with any percentage (Figure 50F). The Landmass Europe Directional Asymmetry variation partition was exclusively environmental and linear with no broad scale and no overlapping variation (Figure 50G). The Landmass Europe Fluctuating Asymmetry variation partition exhibited zero values for all components (Figure 50H). No fine scale patterning was possible, because each Landmass and dermatoglyphic feature only a single significant dbMEM variable, so the broad scale was solely represented by the single significant dbMEM variable. The lack of variation contained in the broad scale component across Landmasses and dermatoglyphic features may be due to the limitations of a single dbMEM variable to contain any breadth of variation. Thus, as would be expected based on the lack of breadth, the linear trend, which indicates patterning larger than the sampling area, and the environmental component, which indicates non-spatially structured variation, were more important for explaining the variation in Directional Asymmetry and Fluctuating Asymmetry.

4.2.3 Principal component analysis (PCA)

The loadings of the principal components illustrated the importance of Directional Asymmetry and Fluctuating Asymmetry, and by visualizing the loadings of the five principal components the variables that contributed the most to each component were apparent. Although the dimensionality reduction benefit of PCA was not particularly relevant for Directional

Asymmetry and Fluctuating Asymmetry variables with five variables each, the interpretability of each principal component was new information. To interpret the loadings that ranged from -1 to 1, a higher absolute value indicated the Directional Asymmetry variable or Fluctuating Asymmetry variable more heavily influenced the component than an absolute value close to zero.

For PC1, all Directional Asymmetry variables exhibited positive values indicating a positive correlation of Directional Asymmetry with PC1, and Directional Asymmetry of digit 2 was particularly positive (Figure 51A). In fact, Directional Asymmetry of digit 2 was positive for PC1, 2, 3, and 4, and zero for PC5. PC2 captured the negative correlation of Directional Asymmetry of digit 1 against the other digits but also exhibited positive values for Directional Asymmetry of digit 2 compared to negative values for Directional Asymmetry of digits 1, 3, 4, and 5 (Figure 51B). PC3 captured the negative correlation of Directional Asymmetry of digit 3 against the other digits but also exhibited positive values for Directional Asymmetry of digits 1 and 2 compared to negative values for Directional Asymmetry of digits 3, 4, and 5 (Figure 51C). PC4 exhibited a higher positive correlation of Directional Asymmetry of digit 4 against the other digits but also exhibited positive values for Directional Asymmetry of digits 2, 4, and 5 compared to negative values for Directional Asymmetry of digits 1 and 3 (Figure 51D). PC5 exhibited negative correlation of Directional Asymmetry of digit 5 against the other digits but also exhibited positive values for Directional Asymmetry of digits 1, 3, and 4 compared to negative values for Directional Asymmetry of digit 5 (Figure 51E).

The first principal component explained 27.8% of the variation, and the second principal component explained 23.5% of the variation in Directional Asymmetry. Visualizing the biplot of PC1 vs PC2 for only the centroids of different groupings illustrated little separation of the

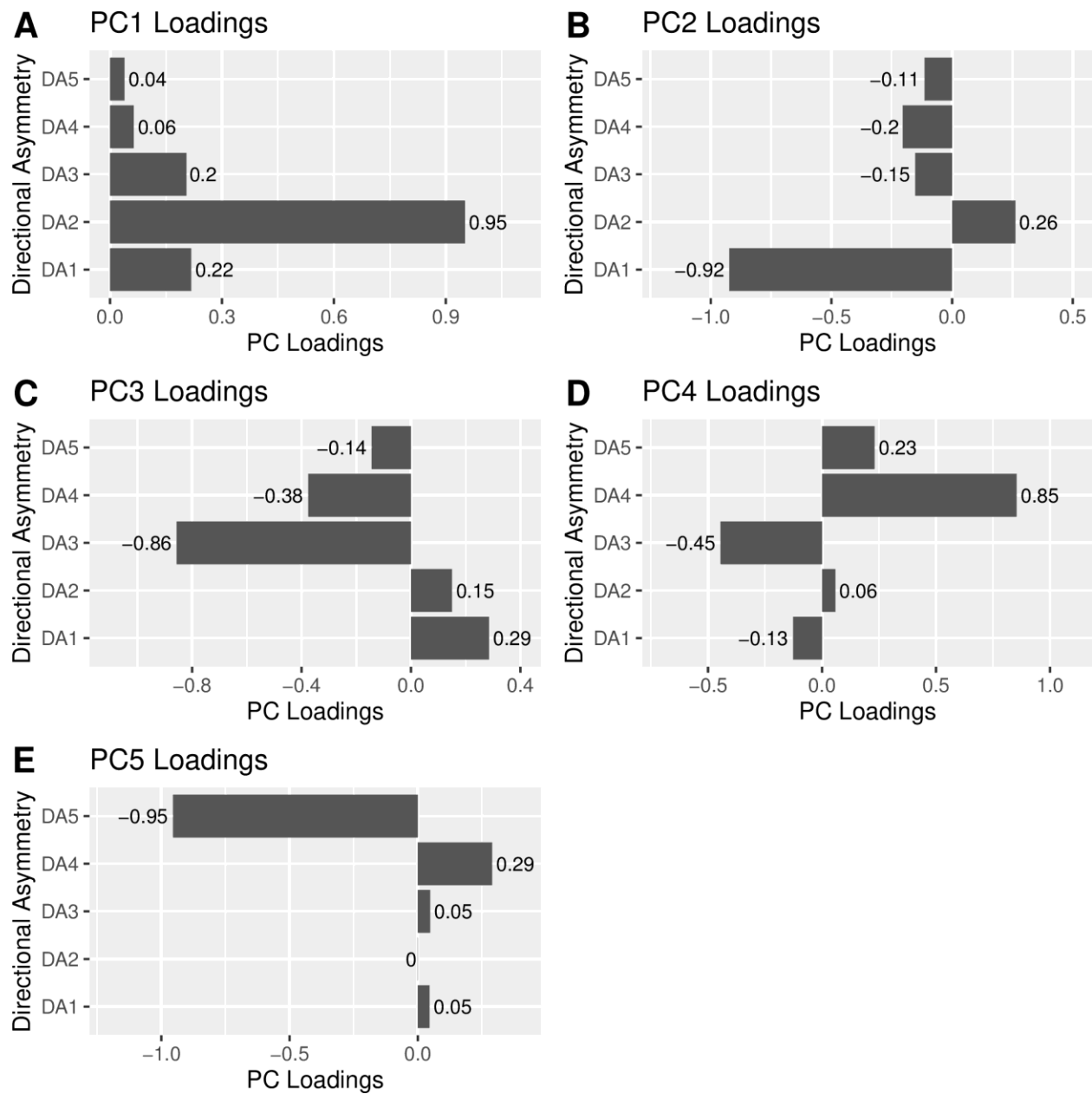


Figure 51. PC 1-5 loadings from PCA of covariance matrix of Directional Asymmetry. (N = 46549)

Landmasses, and only Regions Australia, Indonesia, and Melanesia, which are in Landmass Asia, were distant from other Regions across all other Landmasses (Figure 52). All Landmasses and both sexes appeared to exhibit similar standard errors (Figure 52A, Figure 52C).

For PC1 through PC4, all Fluctuating Asymmetry variables exhibited similar broad trends in correlation direction and magnitude as the Directional Asymmetry variables (Figure 53A, Figure 53B, Figure 53C, Figure 53D). However, PC5 exhibited positive correlation of Fluctuating Asymmetry of digit 5 against the other digits, where PC5 for Directional Asymmetry exhibited negative correlation of Directional Asymmetry of digit 5 (Figure 51E, Figure 53E). Some specific details of the Fluctuating Asymmetry loadings were different from Directional Asymmetry, as well. For example, PC2 exhibited a positive correlation for Fluctuating Asymmetry of digit 3, whereas PC2 exhibited a small negative correlation for Directional Asymmetry of digit 3. Further, PC4 exhibited a positive correlation for Fluctuating Asymmetry of digits 4 and 5 compared to negative values for Fluctuating Asymmetry of digits 1, 2, and 3, where the value for digit 2 was very close to zero; however, PC4 exhibited a positive value for Directional Asymmetry of digit 2. Directional Asymmetry of digit 2 was zero for PC5, while Fluctuating Asymmetry of digit 2 was closest to zero for PC4.

The first principal component explained 29.7% of the variation, and the second principal component explained 24.7% of the variation in Fluctuating Asymmetry. Visualizing the biplot of PC1 vs PC2 for only the centroids of different groupings illustrated separation of the Landmasses along PC1 but not along PC2, and the Regions appeared placed along PC1 similarly according to their Landmass placements (Figure 54). For Directional Asymmetry, however, the Regions did not appear placed similarly according to their Landmass placements (Figure 52). Both sexes appeared to exhibit similar placement along PC1 and PC2 and standard errors for Fluctuating

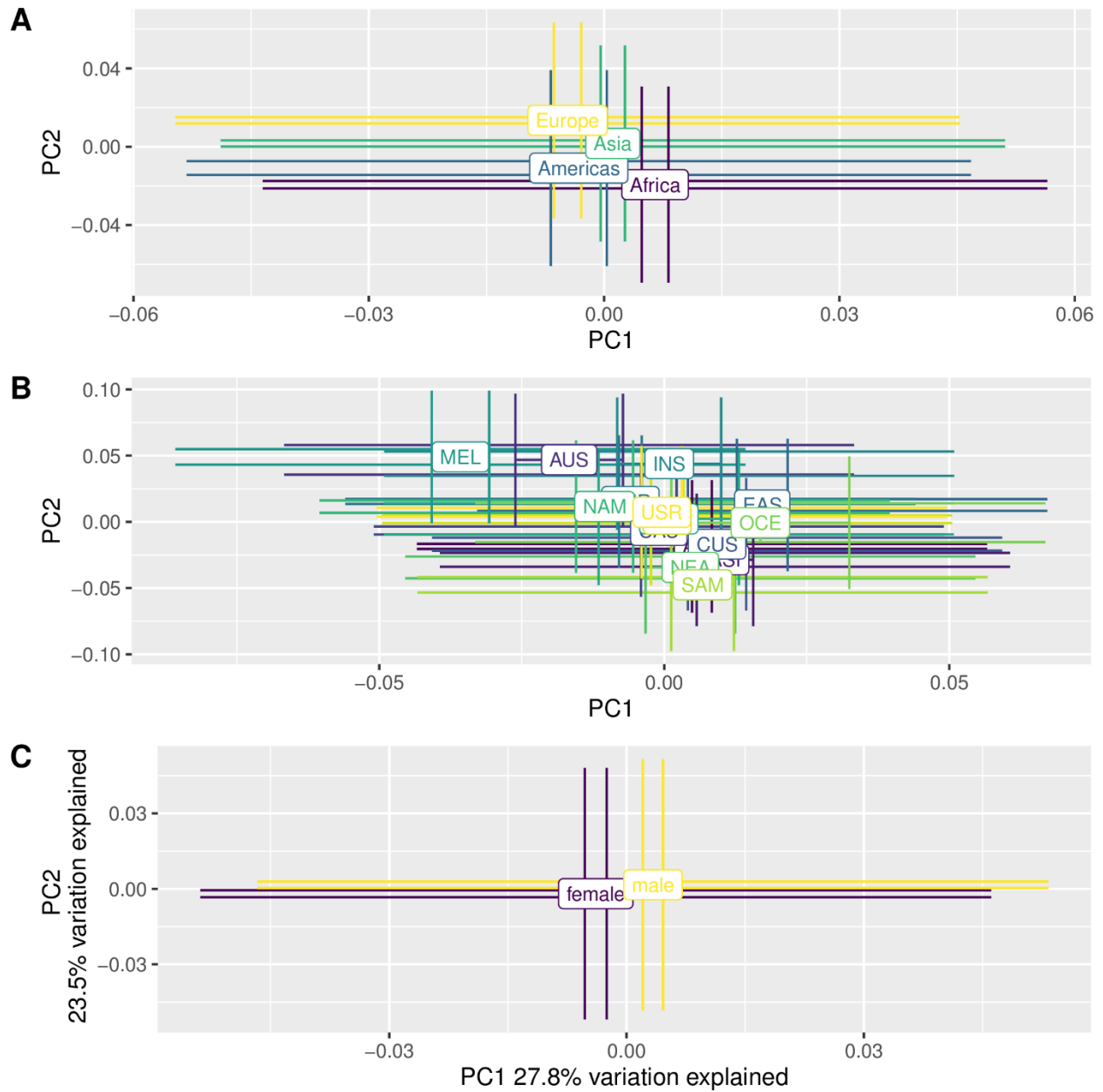


Figure 52. Principal component analysis of covariance matrix of Directional Asymmetry. Viewed by different grouping centroids: (A) by-Landmass, (B) by-Region, and (C) by-Sex (N = 46549). Error bars indicate the standard error.

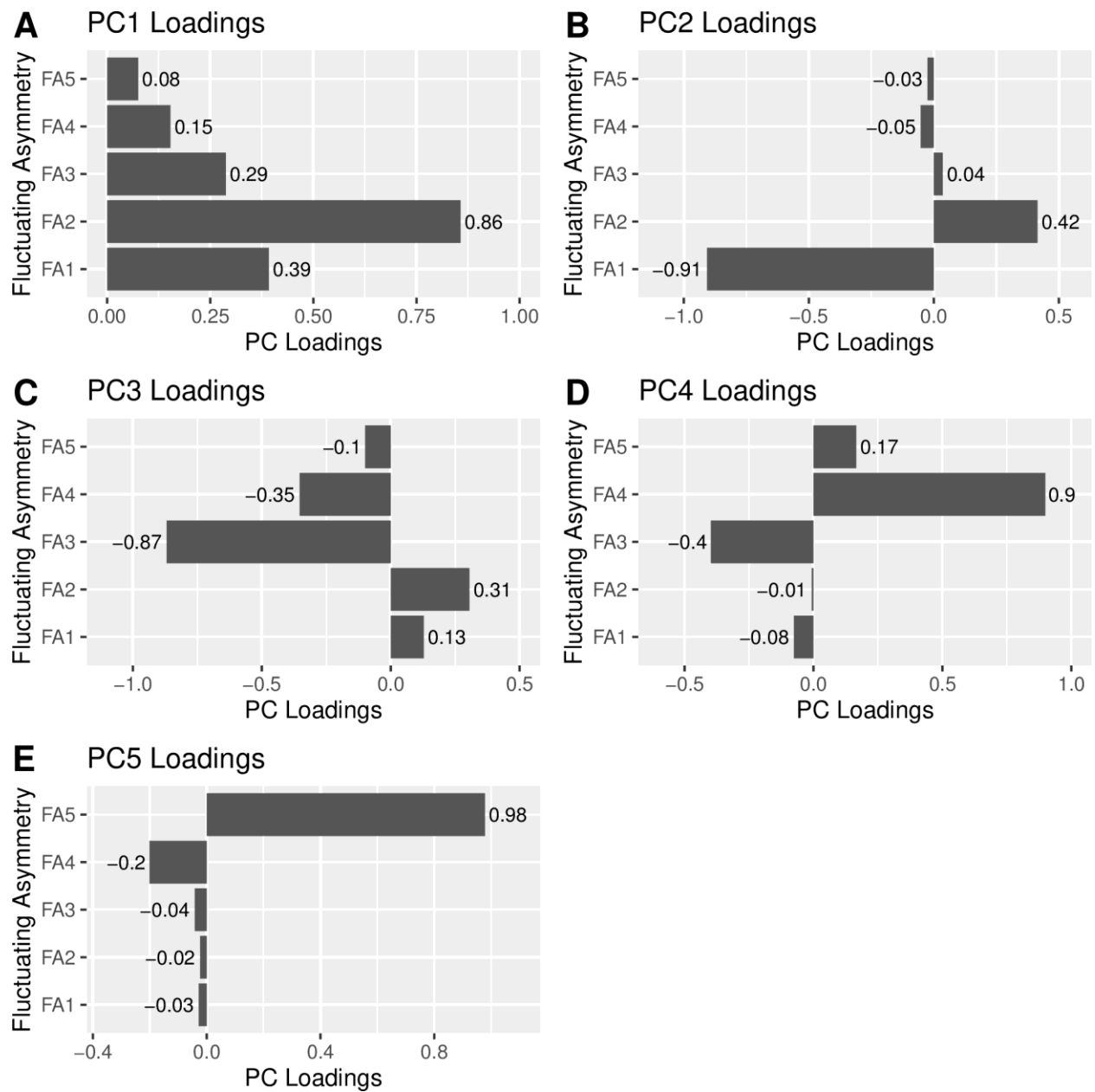


Figure 53. PC 1-5 loadings from PCA of covariance matrix of Fluctuating Asymmetry. (N = 46549)

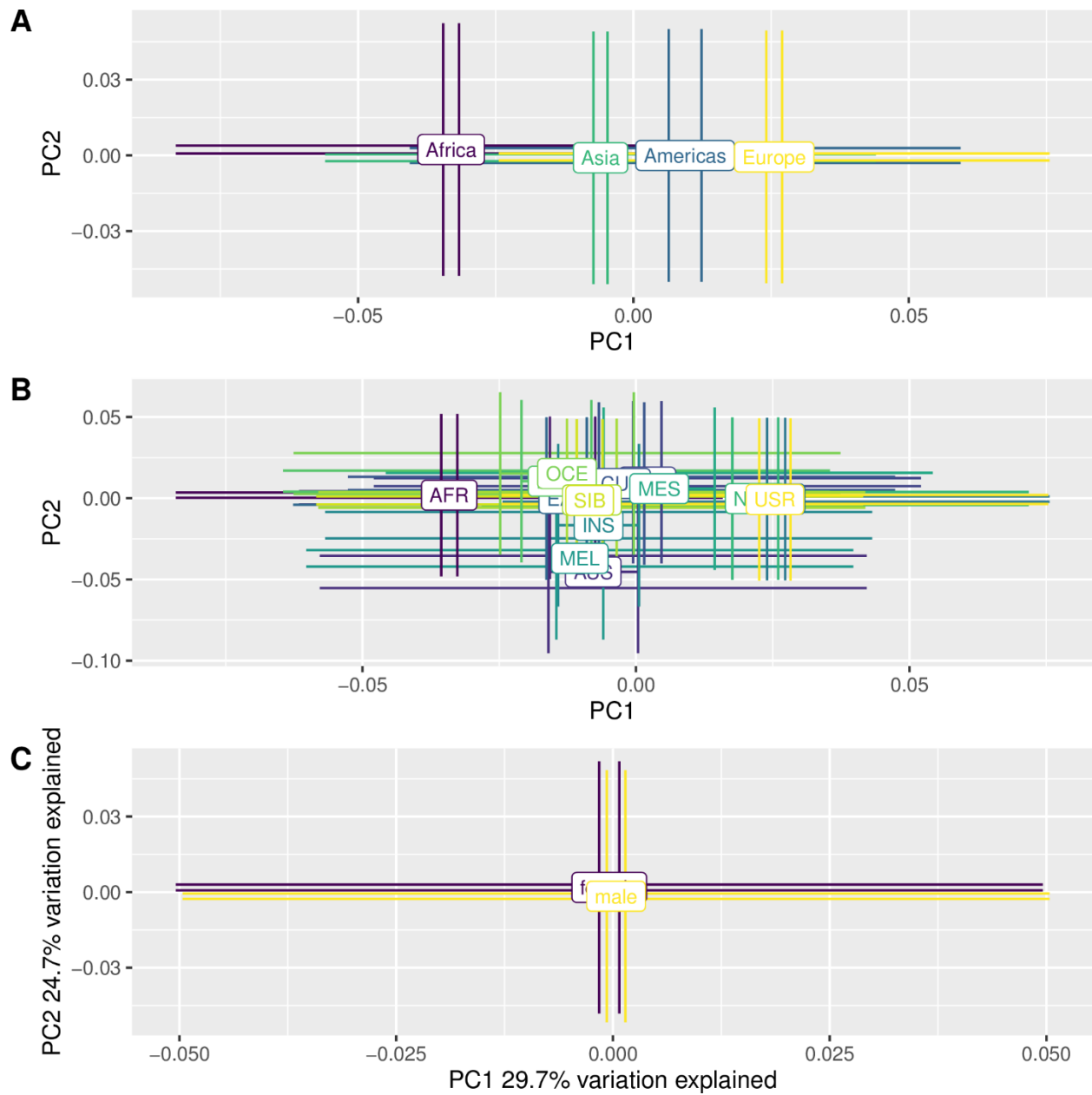


Figure 54. Principal component analysis of covariance matrix of Fluctuating Asymmetry. Viewed by different grouping centroids: (A) by-Landmass, (B) by-Region, and (C) by-Sex (N = 46549). Error bars indicate the standard error.

Asymmetry (Figure 54C).

When the full data set was split by-Sex, the loadings of the principal components illustrated the differences between females and males in Directional Asymmetry (Figure 55, Figure 56). For PC1, the Directional Asymmetry of females exhibited negative values indicating a negative correlation of Directional Asymmetry of females with PC1, and males exhibited all positive values of similar magnitude as the females (Figure 55A, Figure 56A). PC2 and PC5 exhibited the same trends of similar value for females and males, except the male Directional Asymmetry of digit 2 exhibited a zero value (Figure 55B, Figure 55E, Figure 56B, Figure 56E). PC3 and PC4 exhibited the same trends but opposite values of similar magnitude for females and males (Figure 55C, Figure 55D, Figure 56C, Figure 56D). Overall, the broad trends in correlation direction and magnitude of the Directional Asymmetry variables were present when the full data set was split by-Sex (Figure 51, Figure 55, Figure 56).

The biplot of Directional Asymmetry by-Landmass for females and males exhibited similar organization along PC1 and PC2 (Figure 57A, Figure 57C). For Directional Asymmetry by-Region for females and males, the organization along PC1 was similar but along PC2 the Regions Australia, Indonesia, and Melanesia were distant from the other Regions only for male Directional Asymmetry (Figure 57B, Figure 57D). In females, the first principal component explained 27.4% of the variation in the Directional Asymmetry, and the second principal component explained 23.2% of the variation in the Directional Asymmetry (Figure 57A, Figure 57B). In males, the first principal component explained 28.3% of the variation in the Directional Asymmetry, and the second principal component explained 23.8% of the variation in the Directional Asymmetry (Figure 57C, Figure 57D).

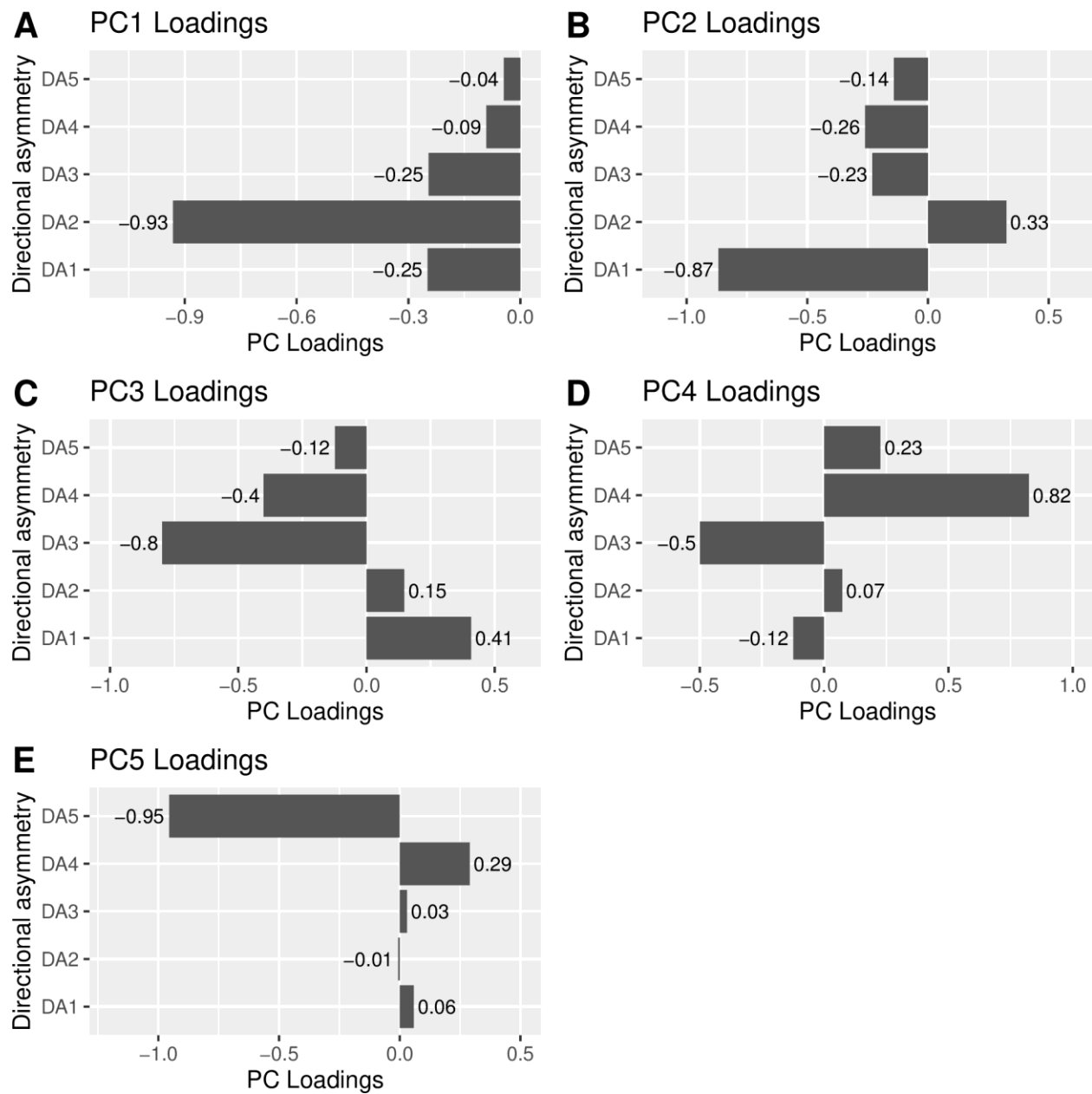


Figure 55. PC 1-5 loadings from PCA of covariance matrix of female Directional Asymmetry.
(N = 21501)

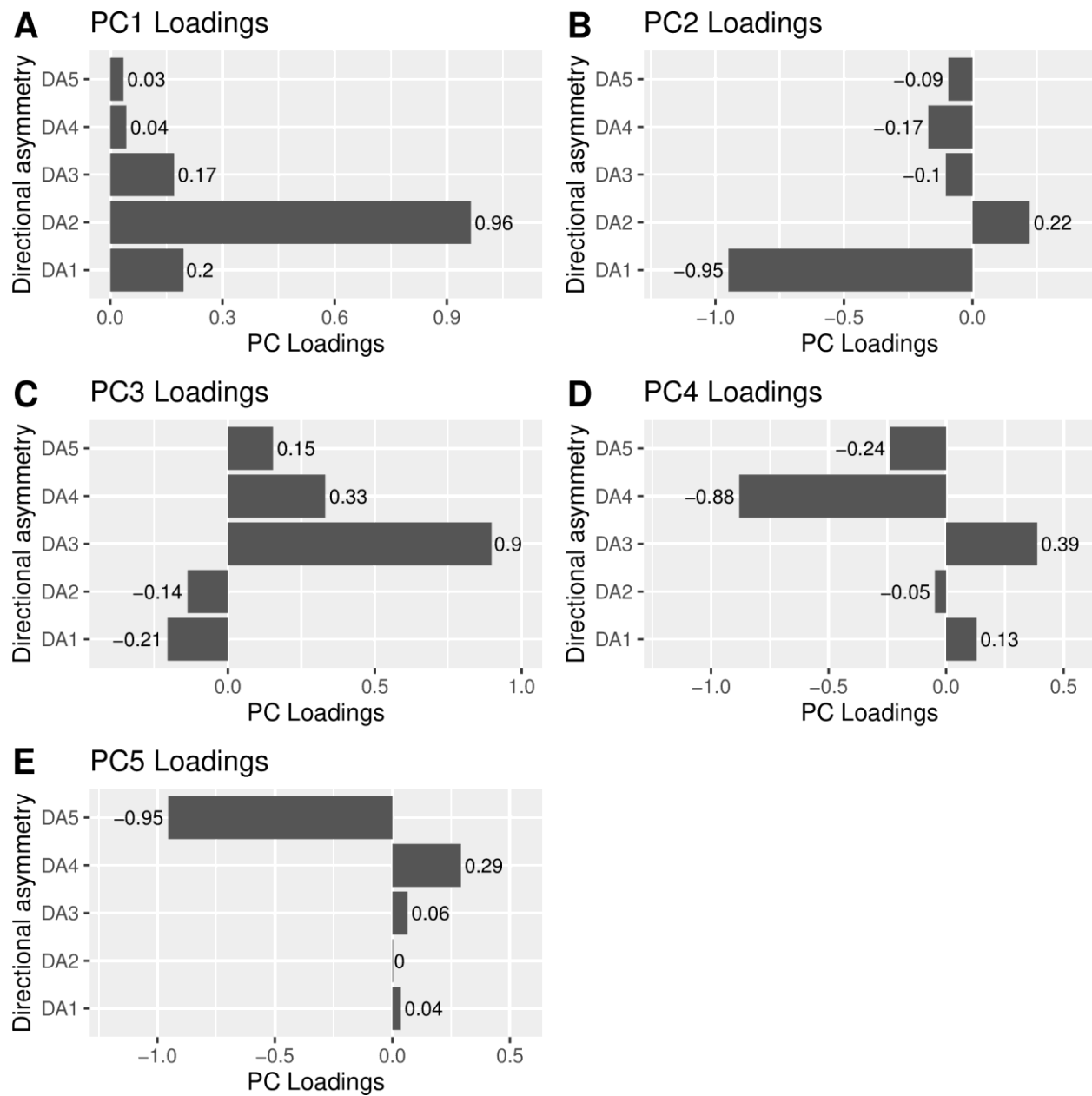


Figure 56. PC 1-5 loadings from PCA of covariance matrix of male Directional Asymmetry. (N = 25048)

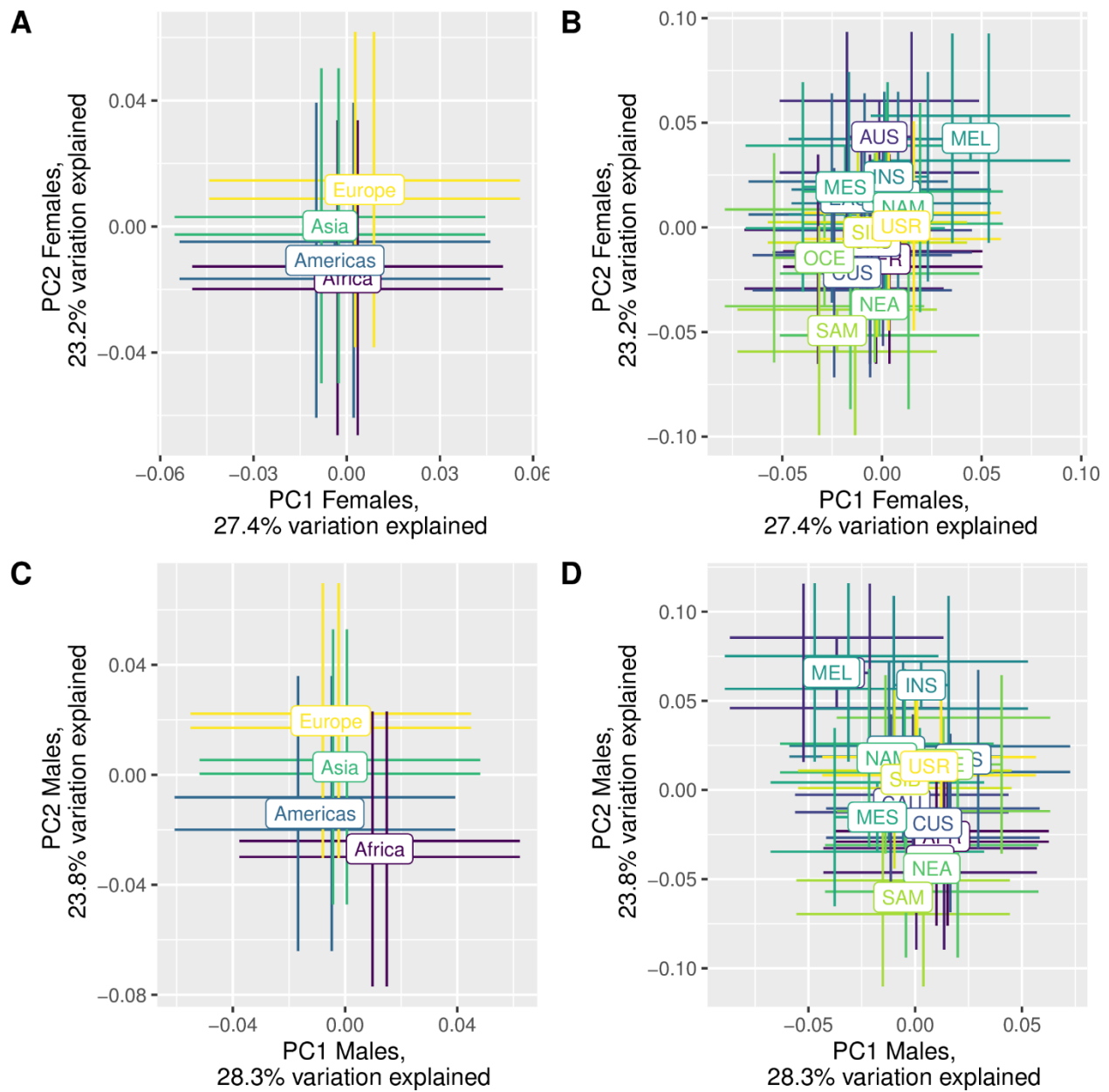


Figure 57. Principal component analysis of covariance matrix of female and male Directional Asymmetry. Viewed by different grouping centroids: (A) females by-Landmass, (B) females by-Region, (C) males by-Landmass, and (D) males by-Region (N = 21501 females, 25048 males).

Error bars indicate the standard error.

When the full data set was split by-Sex, the loadings of the principal components illustrated the differences between females and males of Fluctuating Asymmetry (Figure 58, Figure 59).

For PC1, PC3, and PC5, the Fluctuating Asymmetry of females exhibited the same trends of similar value as the males indicating a similar correlation of Fluctuating Asymmetry of females and males with the respective principal components. PC2 exhibited the same trends but opposite values of similar magnitude for females and males (Figure 58B, Figure 59B).

Specifically, PC2 for females exhibited positive values for the Fluctuating Asymmetry of digits 1, 4, and 5 compared to negative values for Fluctuating Asymmetry of digits 2 and 3. PC4 for females exhibited positive values for Fluctuating Asymmetry of digits 2, 4, and 5 compared to negative values for Fluctuating Asymmetry of digits 1 and 3, whereas PC4 for males exhibited positive values for Fluctuating Asymmetry of digits 4 and 5 only compared to negative values for Fluctuating Asymmetry of digits 1, 2, and 3.

The biplot of Fluctuating Asymmetry by-Landmass for females and males exhibited similar organization along PC1 and PC2, and the Landmasses appeared spread along PC1 but not PC2 (Figure 60A, Figure 60C). For Fluctuating Asymmetry by-Region for females and males, the organization along PC1 was similar but along PC2 the Regions appeared in opposite organization (Figure 60B, Figure 60D). In females, the first principal component explained 29.4% of the variation in the Fluctuating Asymmetry, and the second principal component explained 24.2% of the variation in the Fluctuating Asymmetry (Figure 60A, Figure 60B). In males, the first principal component explained 30% of the variation in the Fluctuating Asymmetry, and the second principal component explained 25.2% of the variation in the Fluctuating Asymmetry (Figure 60C, Figure 60D).

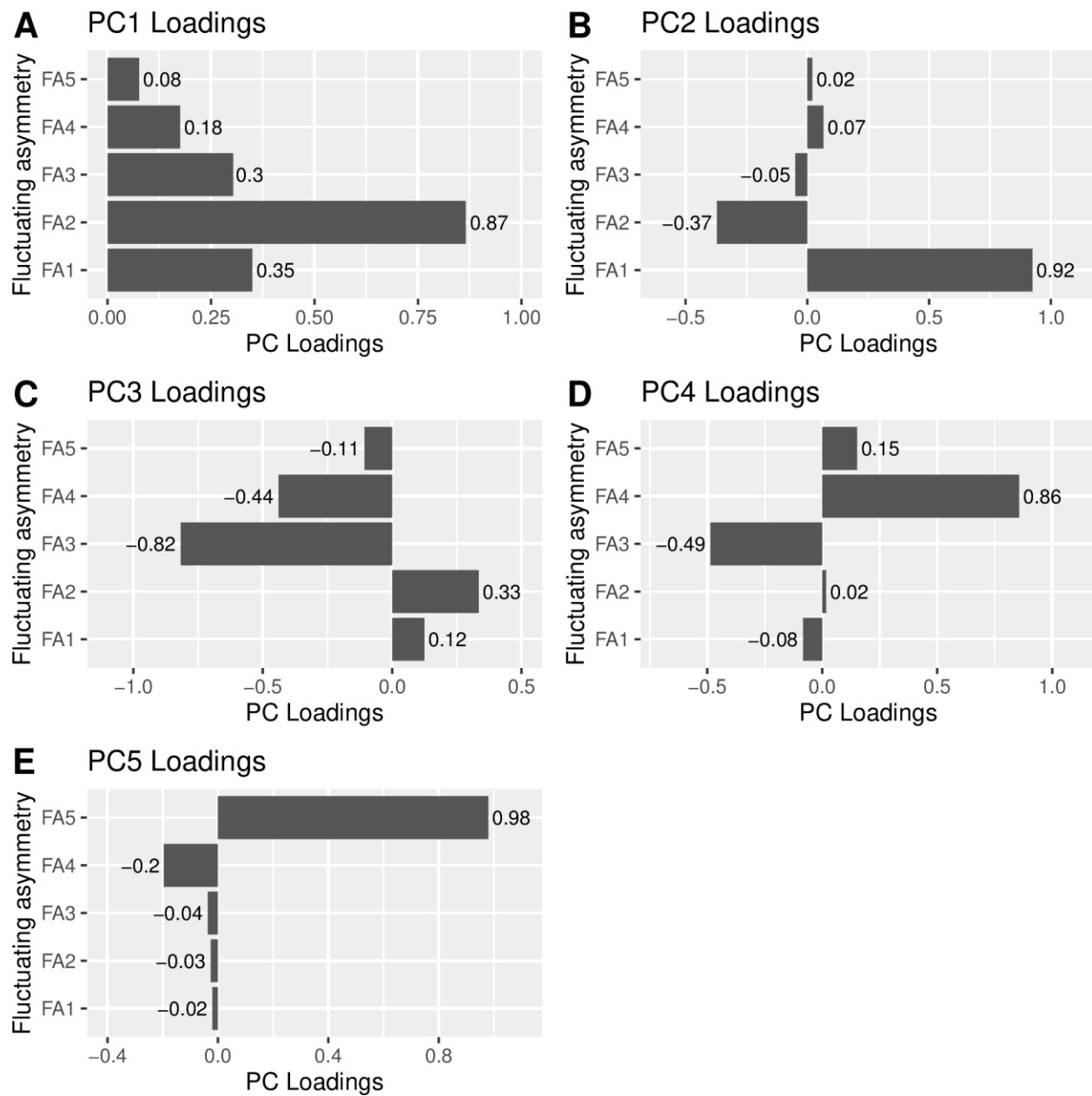


Figure 58. PC 1-5 loadings from PCA of covariance matrix of female Fluctuating Asymmetry.
(N = 21501)

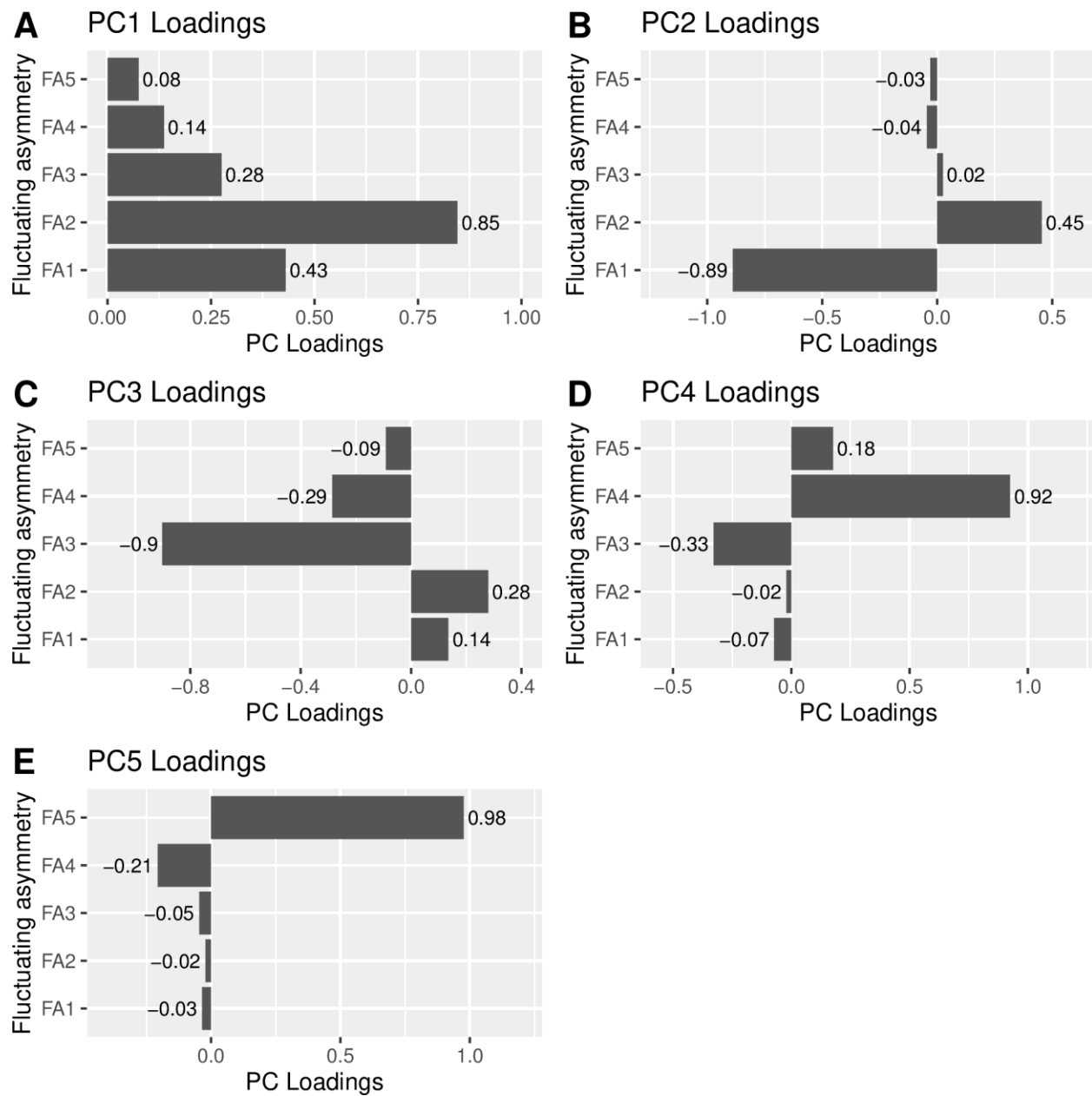


Figure 59. PC 1-5 loadings from PCA of covariance matrix of male Fluctuating Asymmetry. (N = 25048)

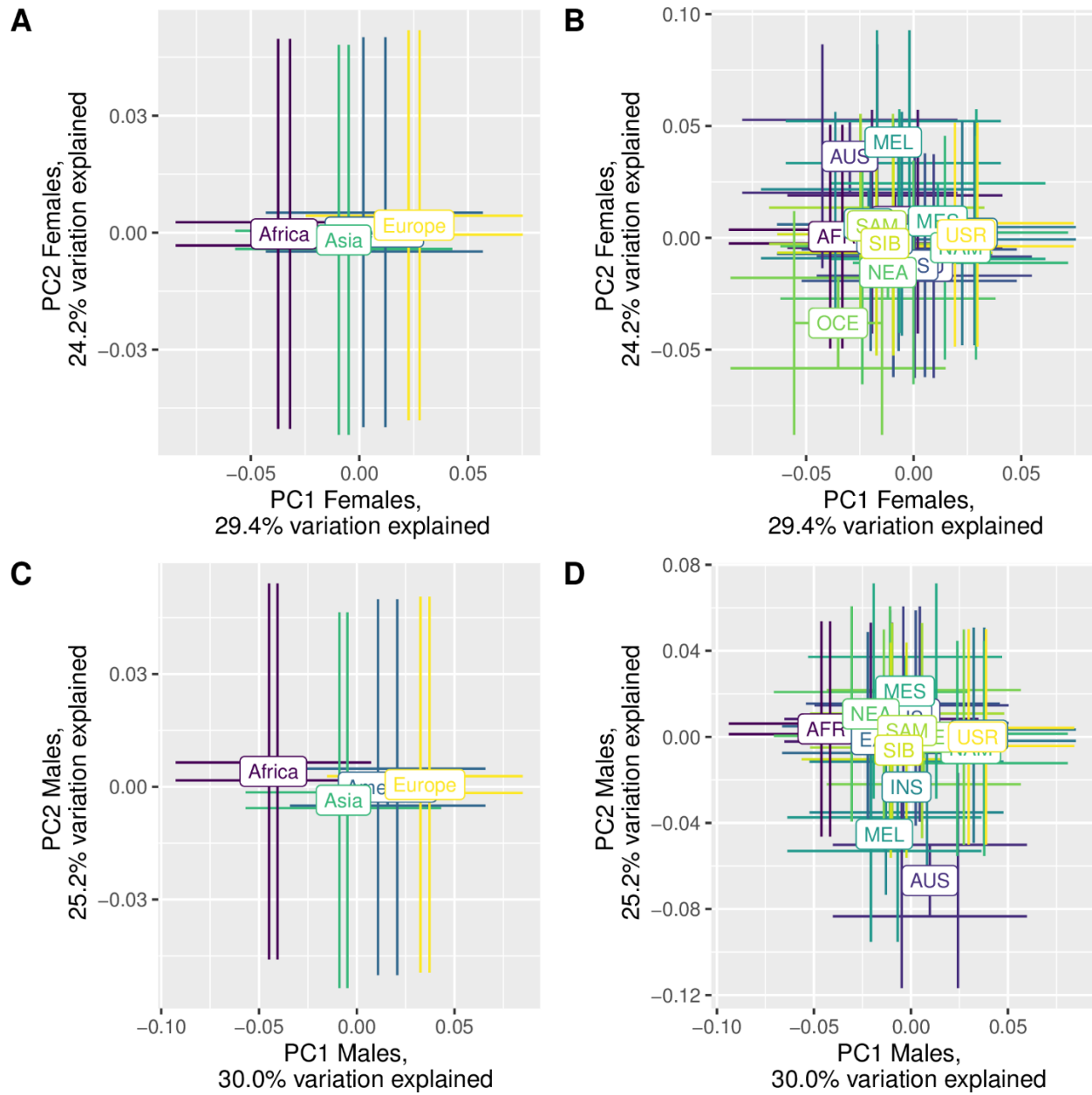


Figure 60. Principal component analysis of covariance matrix of female and male Fluctuating Asymmetry. Viewed by different grouping centroids: (A) females by-Landmass, (B) females by-Region, (C) males by-Landmass, and (D) males by-Region (N = 21501 females, 25048 males).

Error bars indicate the standard error.

When the full data set was split by-Landmass, the loadings of the principal components illustrated the importance of Directional Asymmetry in the each Landmass separately. For PC1, the Directional Asymmetry of Landmass Africa exhibited negative values indicating a negative correlation of Directional Asymmetry of Landmass Africa with PC1 (Figure 61A). Further of note, the largest value for PC1 for Directional Asymmetry of Landmass Africa was for digit 1, whereas the largest value for PC1 for Directional Asymmetry (Figure 51), female Directional Asymmetry (Figure 55), male Directional Asymmetry (Figure 56), Fluctuating Asymmetry (Figure 53), female Fluctuating Asymmetry (Figure 58), and male Fluctuating Asymmetry (Figure 59) was for digit 2. PC2 captured the negative correlation of digit one against the Directional Asymmetry of all other digits of Landmass Africa (Figure 61B). PC3 captured positive correlation for Directional Asymmetry of digits 1 and 2 versus negative correlation for Directional Asymmetry of digits 3, 4, and 5 of Landmass Africa (Figure 61C). PC4 exhibited a higher positive correlation for the Directional Asymmetry of digits 2, 4, and 5 and negative correlation for the Directional Asymmetry of digits 1 and 3 (Figure 61D). PC5 exhibited higher negative correlation for the Directional Asymmetry of digit 5 versus all other digits, and Directional Asymmetry of digit 2 was equal to zero (Figure 61E).

The first principal component explained 28.7% of the variation, and the second principal component explained 24.2% of the variation in Directional Asymmetry of Landmass Africa. Visualizing the biplot of PC1 versus PC2 for only the centroids of different groupings illustrated little separation of Region Africa from Region Northeast Africa and little separation of females from males for Directional Asymmetry (Figure 62). The centroids for Directional Asymmetry of females and males in Landmass Africa were separate, but the error bars representing the standard

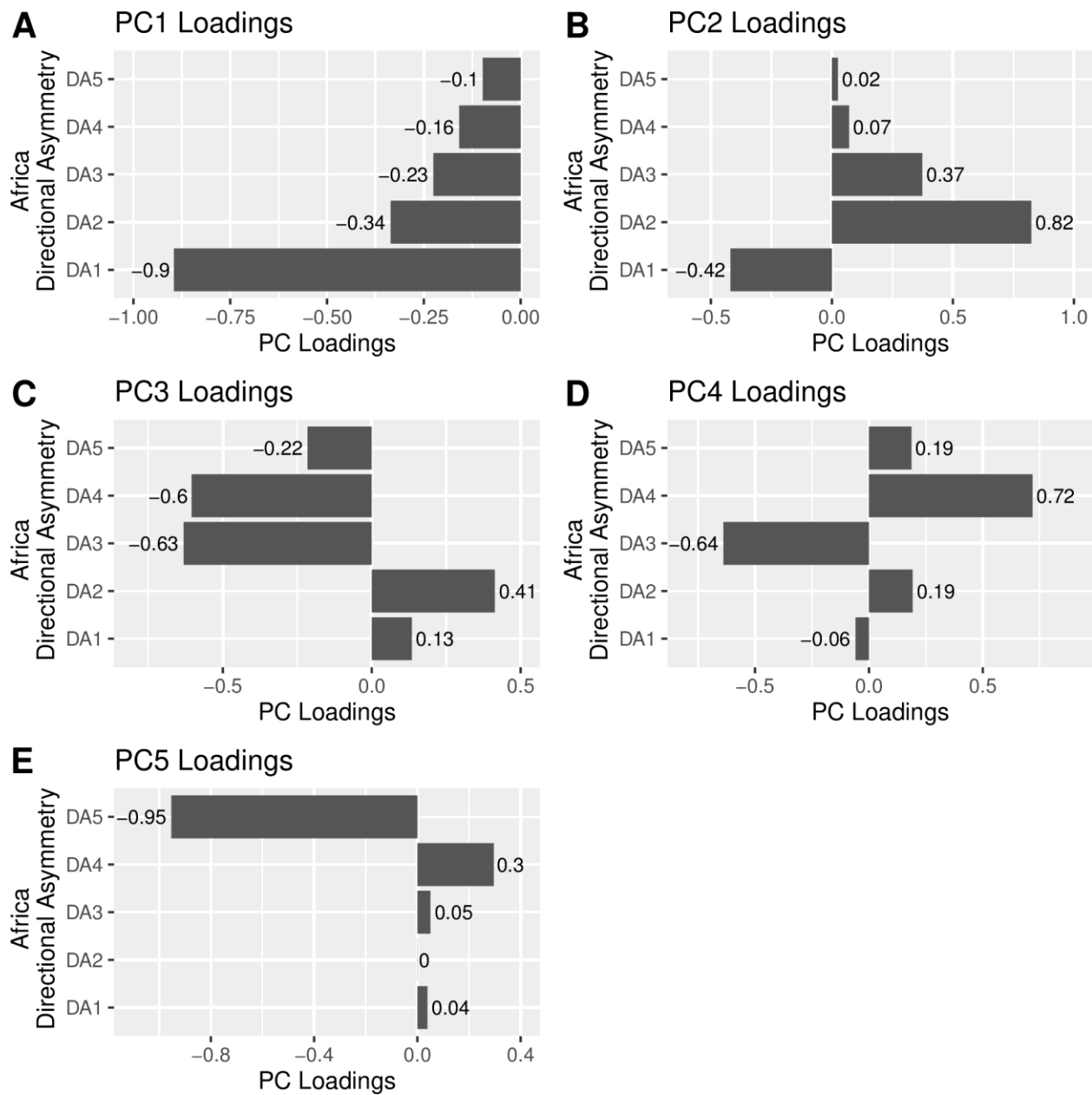


Figure 61. PC 1-5 loadings from PCA of covariance matrix of Landmass Africa Directional Asymmetry. (N = 10809)

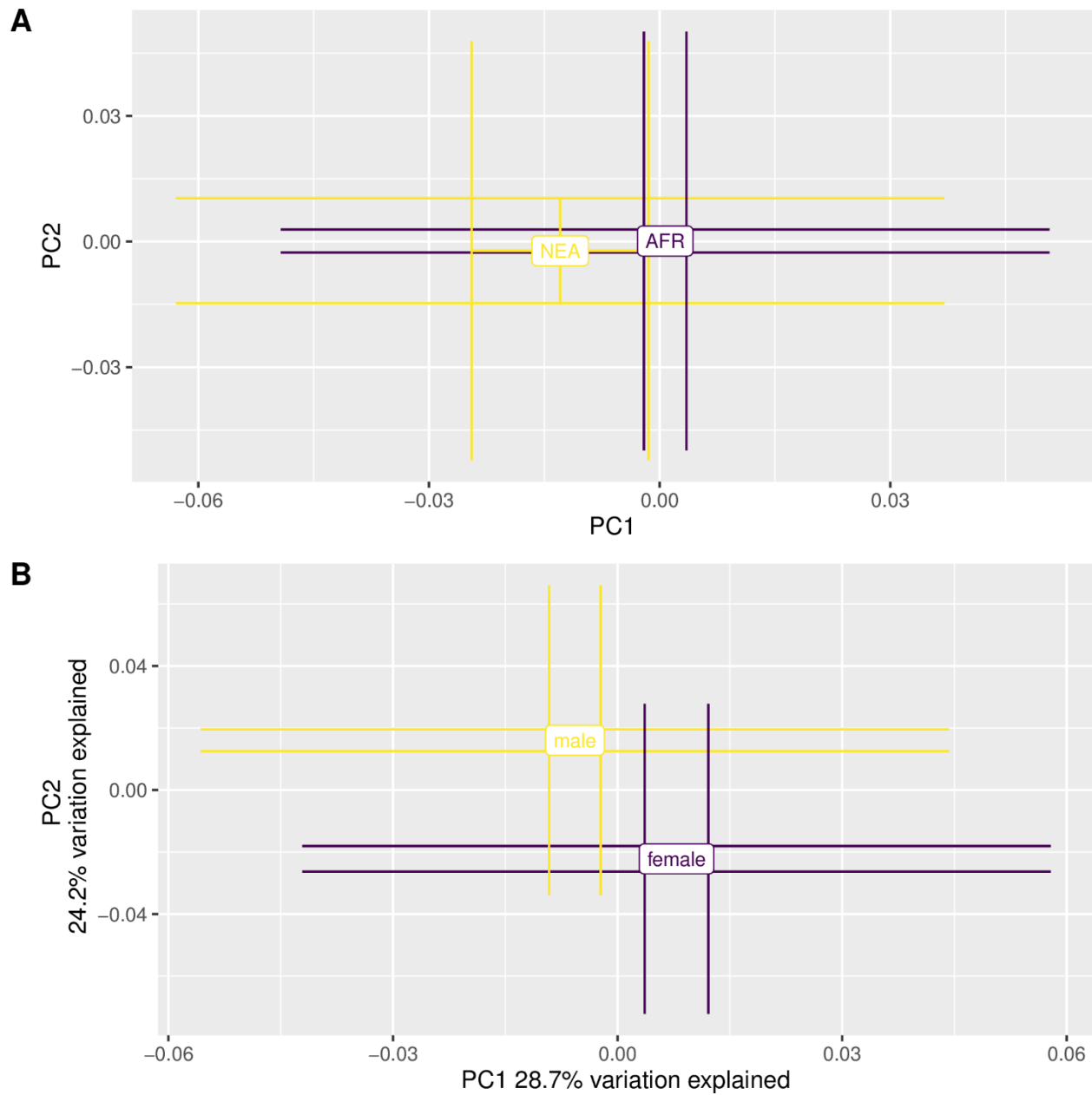


Figure 62. Principal component analysis of covariance matrix of Landmass Africa Directional Asymmetry. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 10809). Error bars indicate the standard error.

error still overlap on the PC1 axis and the PC2 axis (Figure 62B). Females and males of Landmass Africa appeared to exhibit a similar amount of variation.

Further splitting the Directional Asymmetry of Landmass Africa by-Sex, the two Regions exhibited opposing positions in the biplots for females and males. Region Northeast Africa appeared to contain more variation than Region Africa for Directional Asymmetry of both males and females. The first principal component for females exhibited 29.3% of the variation in Directional Asymmetry of Landmass Africa, and the second principal component for females exhibited 23.3% of the variation in Directional Asymmetry of Landmass Africa (Figure 63A). The first principal component for males exhibited 28.3% of the variation in Directional Asymmetry of Landmass Africa, and the second principal component for males exhibited 24.8% of the variation in Directional Asymmetry of Landmass Africa (Figure 63B).

When the full data set was split by-Landmass, the loadings of the principal components illustrated the importance of Fluctuating Asymmetry in the each Landmass separately. For PC1, the Fluctuating Asymmetry of Landmass Africa exhibited positive values indicating a positive correlation of Fluctuating Asymmetry of Landmass Africa with PC1 (Figure 64A). Similar to Directional Asymmetry, the largest value for PC1 for Fluctuating Asymmetry of Landmass Africa was for digit 1, whereas the largest value for PC1 for Directional Asymmetry (Figure 51), female Directional Asymmetry (Figure 55), male Directional Asymmetry (Figure 56), Fluctuating Asymmetry (Figure 53), female Fluctuating Asymmetry (Figure 58), and male Fluctuating Asymmetry (Figure 59) was for digit 2. PC2 through PC4 for Fluctuating Asymmetry of Landmass Africa exhibited the same patterns but opposite values than for Directional Asymmetry of Landmass Africa, and PC5 exhibited a high positive value for

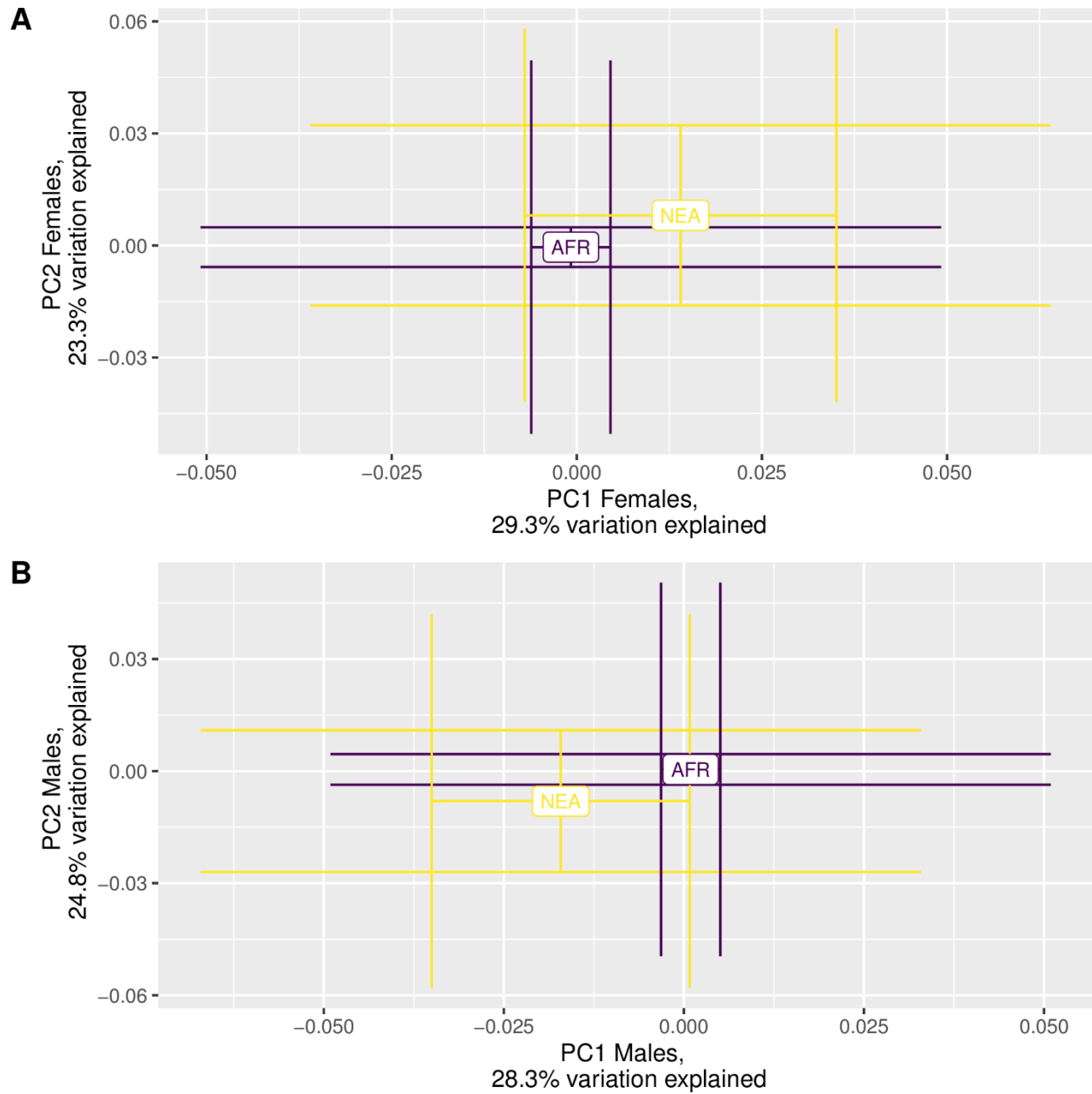


Figure 63. Principal component analysis of covariance matrix of female and male Landmass Africa Directional Asymmetry. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 4538 females, 6271 males). Error bars indicate the standard error.

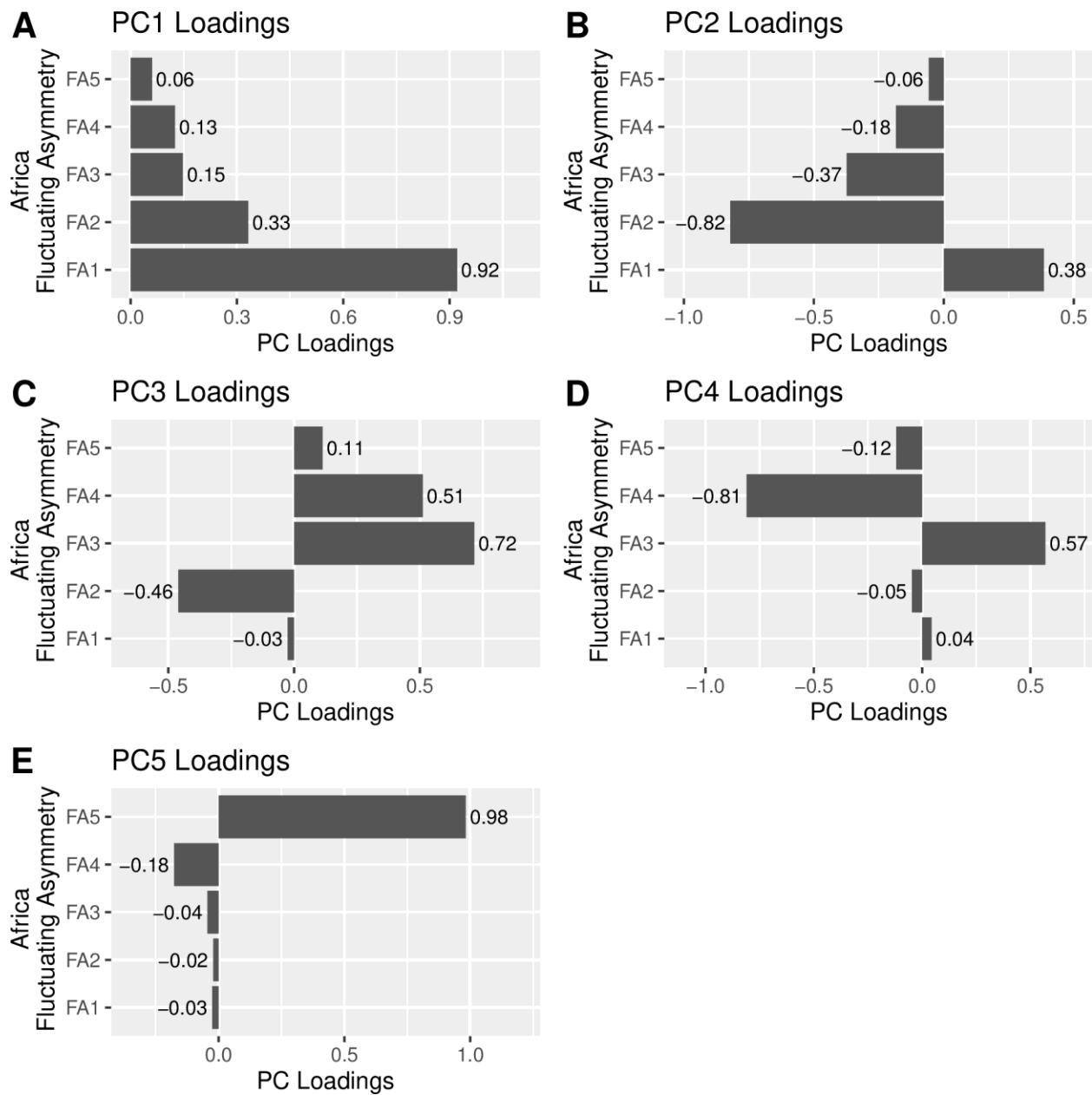


Figure 64. PC 1-5 loadings from PCA of covariance matrix of Landmass Africa Fluctuating Asymmetry. (N = 10809)

Fluctuating Asymmetry of digit 5 compared to negative values for all other digits (Figure 61, Figure 64).

The first principal component explained 30.7% of the variation, and the second principal component explained 25% of the variation in Fluctuating Asymmetry of Landmass Africa. Visualizing the biplot of PC1 versus PC2 for only the centroids of different groupings illustrated small distance of separation of Region Africa from Region Northeast Africa more on PC2 than PC1 and little separation of females from males for Fluctuating Asymmetry (Figure 65). The centroids for Fluctuating Asymmetry of females and males in Landmass Africa were separate, but the error bars representing the standard error still overlap on the PC1 axis and the PC2 axis (Figure 65B). Females and males of Landmass Africa appeared to exhibit a similar amount of variation.

Further splitting the Fluctuating Asymmetry of Landmass Africa by-Sex, the two Regions exhibited opposing positions in the biplots for females and males, and for females and males the two Regions exhibited opposite placement along PC1 and PC2 compared to Directional Asymmetry. Region Northeast Africa appeared to contain more variation than Region Africa for Fluctuating Asymmetry of both males and females. The first principal component for females exhibited 31% of the variation in Fluctuating Asymmetry of Landmass Africa, and the second principal component for females exhibited 24.7% of the variation in Fluctuating Asymmetry of Landmass Africa (Figure 66A). The first principal component for males exhibited 30.5% of the variation in Fluctuating Asymmetry of Landmass Africa, and the second principal component for males exhibited 25.3% of the variation in Fluctuating Asymmetry of Landmass Africa (Figure 66B).

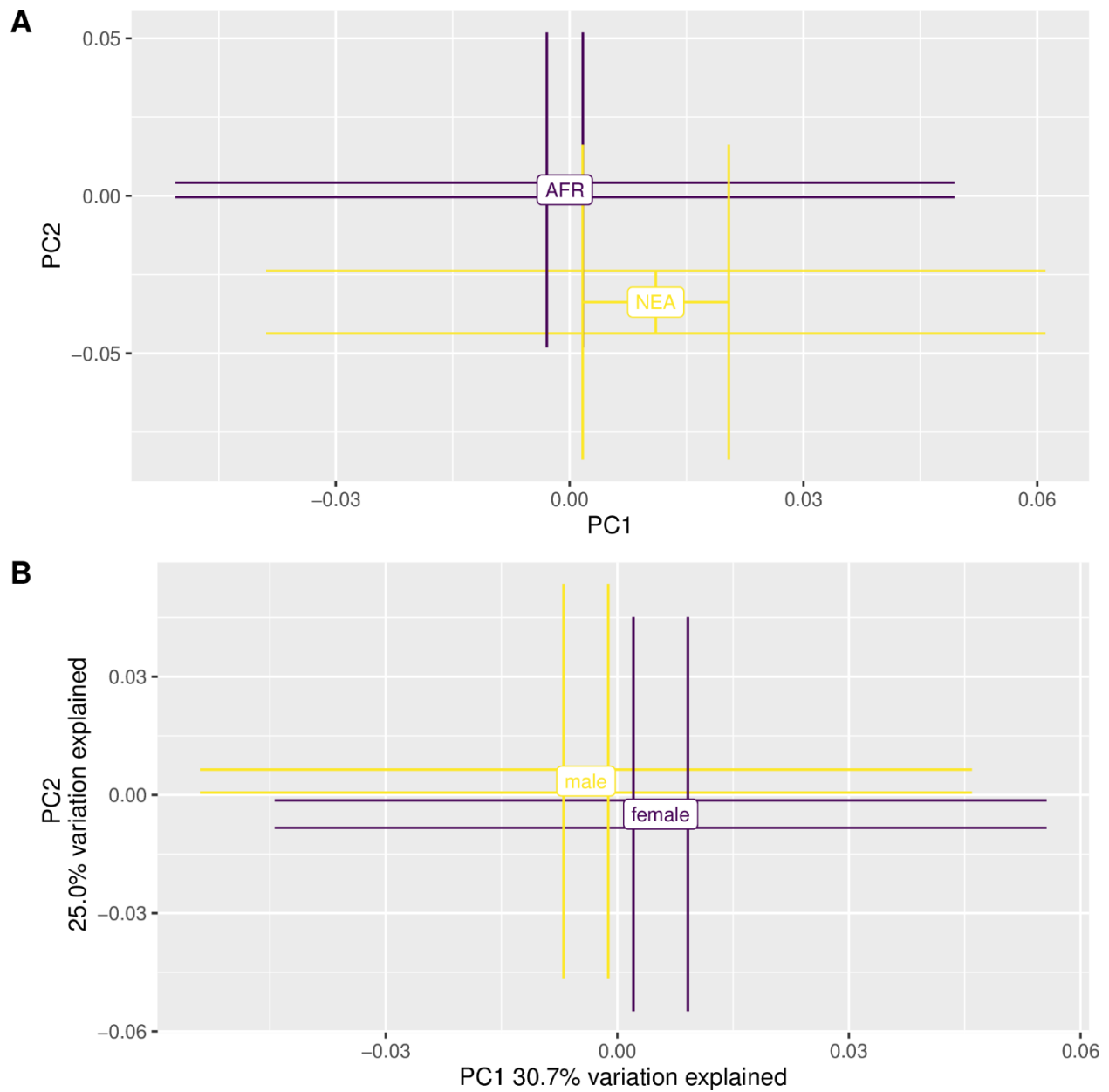


Figure 65. Principal component analysis of covariance matrix of Landmass Africa Fluctuating Asymmetry. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 10809). Error bars indicate the standard error.

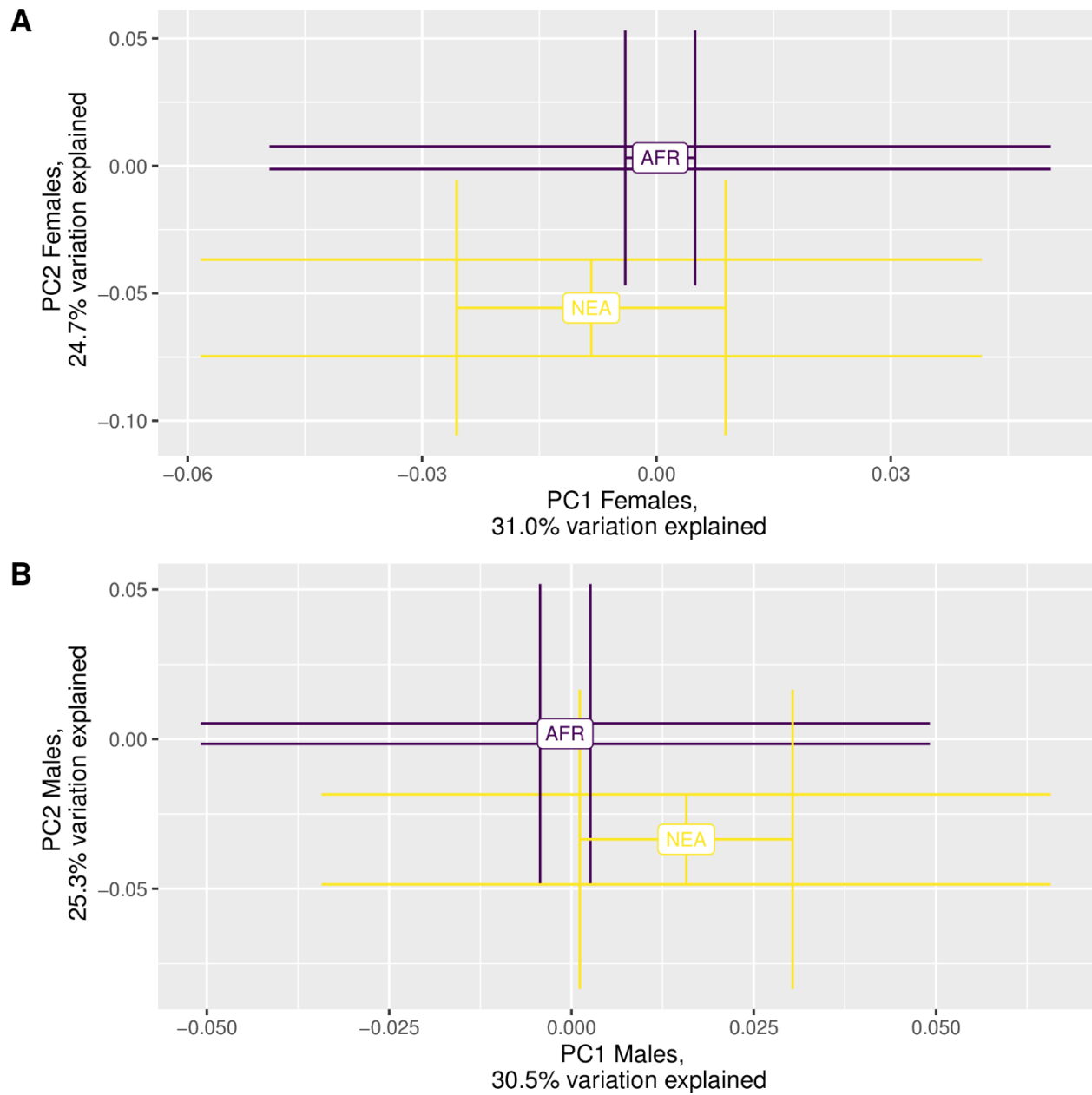


Figure 66. Principal component analysis of covariance matrix of female and male Landmass Africa Fluctuating Asymmetry. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 4537 females, 6269 males). Error bars indicate the standard error.

For PC1, the Directional Asymmetry of Landmass Americas exhibited negative values indicating a negative correlation of Directional Asymmetry of Landmass Americas with PC1 (Figure 67A). The largest value for PC1 for Directional Asymmetry of Landmass Americas was for digit 2, which was similar to Directional Asymmetry (Figure 51), female Directional Asymmetry (Figure 55), male Directional Asymmetry (Figure 56), Fluctuating Asymmetry (Figure 53), female Fluctuating Asymmetry (Figure 58), and male Fluctuating Asymmetry (Figure 59), but different from Directional Asymmetry of Landmass Africa whereas the largest value for PC1 was digit 1 (Figure 61). PC2 captured the largest negative correlation of digit 1 against all other Directional Asymmetry of Landmass Americas (Figure 67B). PC3 captured positive correlation for Directional Asymmetry of digits 1 and 2 versus negative correlation for Directional Asymmetry of digits 3, 4, and 5 of Landmass Americas (Figure 67C). PC4 exhibited a higher negative correlation for the Directional Asymmetry of digit 4 and positive correlation for the Directional Asymmetry of digit 3 (Figure 67D). PC5 exhibited higher positive correlation for the Directional Asymmetry of digit 5 versus all other digits (Figure 67E).

The first principal component explained 27.8% of the variation, and the second principal component explained 22.8% of the variation in Directional Asymmetry of Landmass Americas. Visualizing the biplot of PC1 versus PC2 for only the centroids of different groupings illustrated an increasingly positive value along PC1 from Region South America, to Mesoamerica, to North America. Region South America exhibited negative values along PC2, whereas Regions Mesoamerica and North America exhibited positive values along PC2 (Figure 68A). The centroids for Directional Asymmetry of females and males in Landmass Americas were separate, but the error bars representing the standard error still overlap on the PC1 axis and the PC2 axis

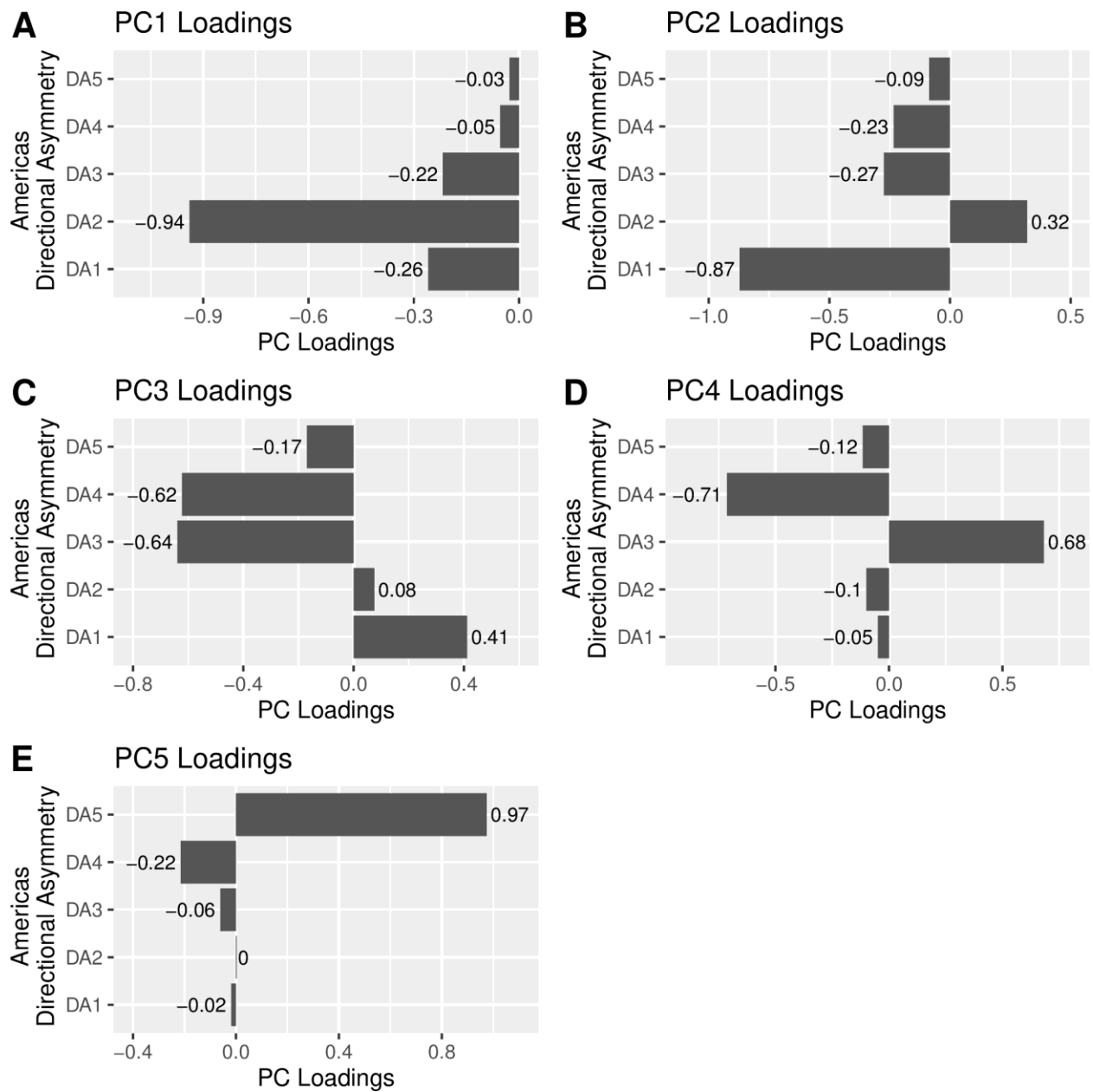


Figure 67. PC 1-5 loadings from PCA of covariance matrix of Landmass Americas Directional Asymmetry. (N = 3522)

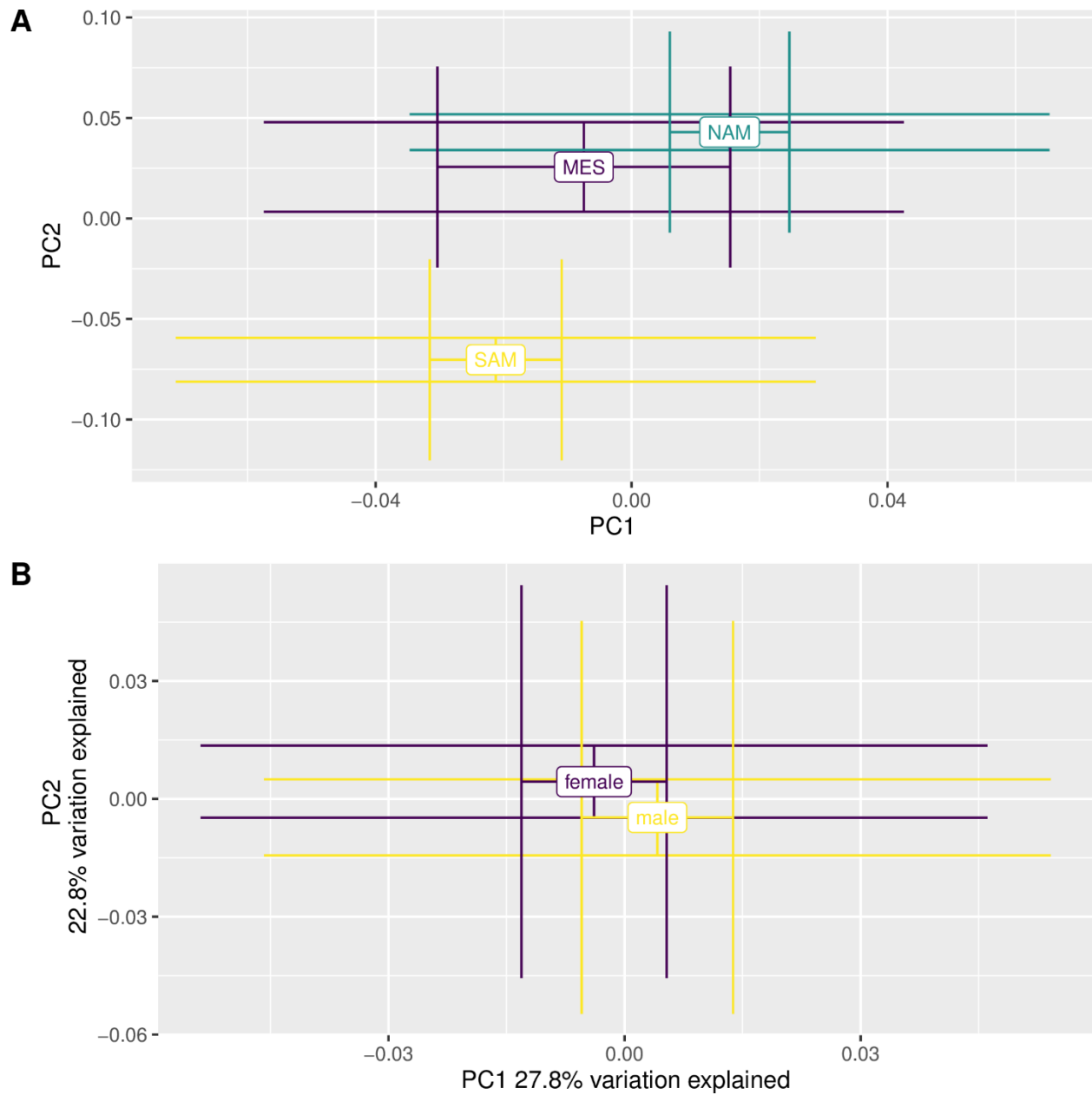


Figure 68. Principal component analysis of covariance matrix of Landmass Americas Directional Asymmetry. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 3522). Error bars indicate the standard error.

(Figure 68B). Females and males of Landmass Americas appeared to exhibit a similar amount of variation.

Further splitting the Directional Asymmetry of Landmass Americas by-Sex, females of Region South America exhibited high positive values along PC2 but males exhibited high negative values along PC2, and females of Region North America exhibited high negative values along PC2 but males exhibited high positive values along PC2 (Figure 69). The first principal component for females exhibited 27.5% of the variation in Directional Asymmetry of Landmass Americas, and the second principal component for females exhibited 22.1% of the variation in Directional Asymmetry of Landmass Americas (Figure 69A). The first principal component for males exhibited 28.2% of the variation in Directional Asymmetry of Landmass Americas, and the second principal component for males exhibited 23.3% of the variation in Directional Asymmetry of Landmass Americas (Figure 69B).

The loadings of Fluctuating Asymmetry of Landmass Americas exhibited the same magnitude and direction for loading of Directional Asymmetry of Landmass Americas for PC1 and 3. PC2 exhibited positive loadings for Fluctuating Asymmetry of digits 1, 4, and 5 compared to negative values for Fluctuating Asymmetry of digits 2 and 3 in Landmass Americas (Figure 70). Further, the PC4 exhibited positive loadings for Fluctuating Asymmetry of digits 2, 4, and 5 compared to negative values for Fluctuating Asymmetry of digits 1 and 3 in Landmass Americas. Finally, PC5 exhibited large positive loadings for Fluctuating Asymmetry of digit 5 compared to digits 2 and 3 loadings equal to zero and negative values for Fluctuating Asymmetry of digits 1 and 4 in Landmass Americas.

For PC1, PC3, and PC5, the Fluctuating Asymmetry of Landmass Americas exhibited similar trends and values as for Directional Asymmetry of Landmass Americas, and PC2 and

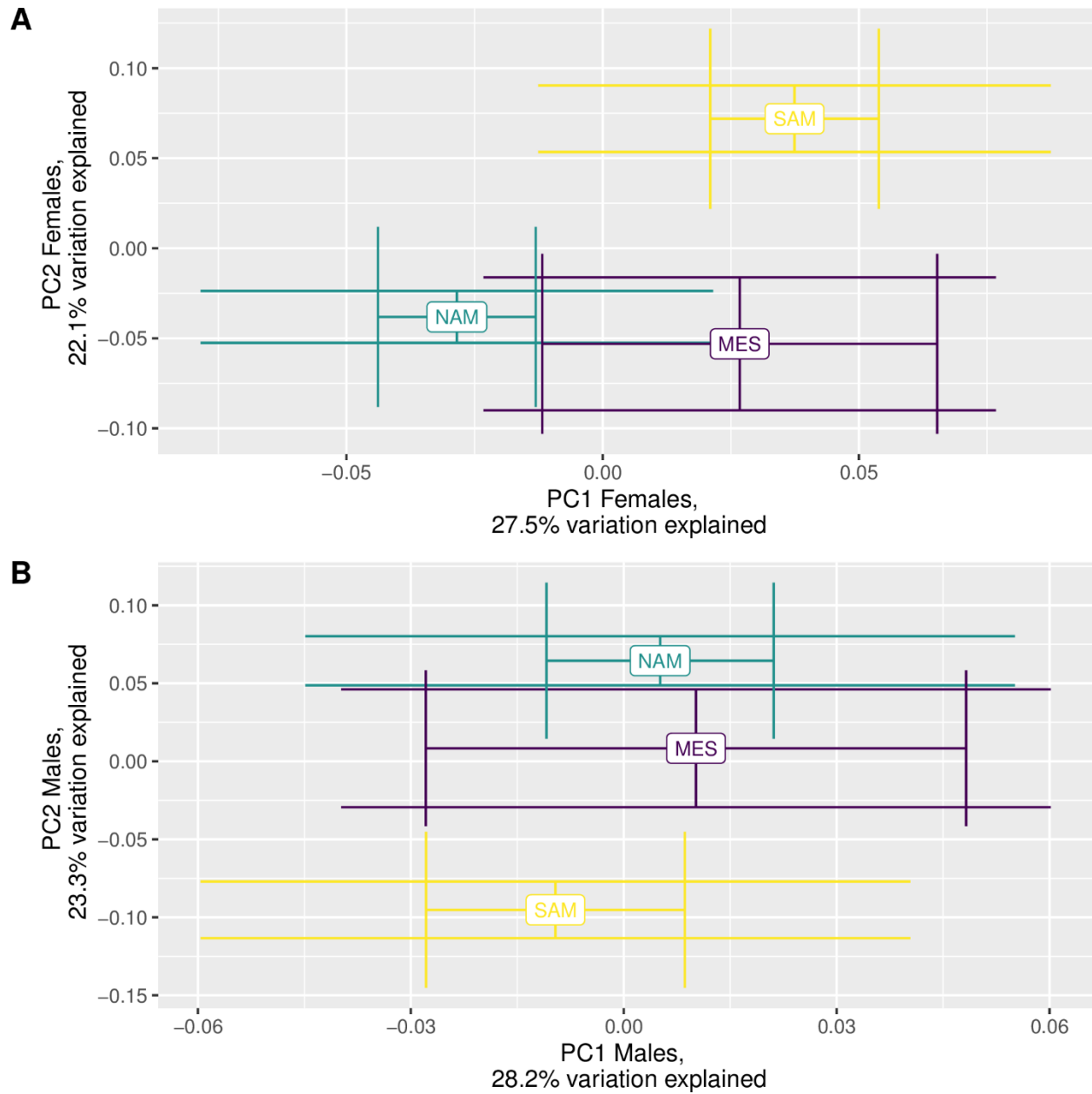


Figure 69. Principal component analysis of covariance matrix of female and male Landmass Americas Directional Asymmetry. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 1825 females, 1697 males). Error bars indicate the standard error.

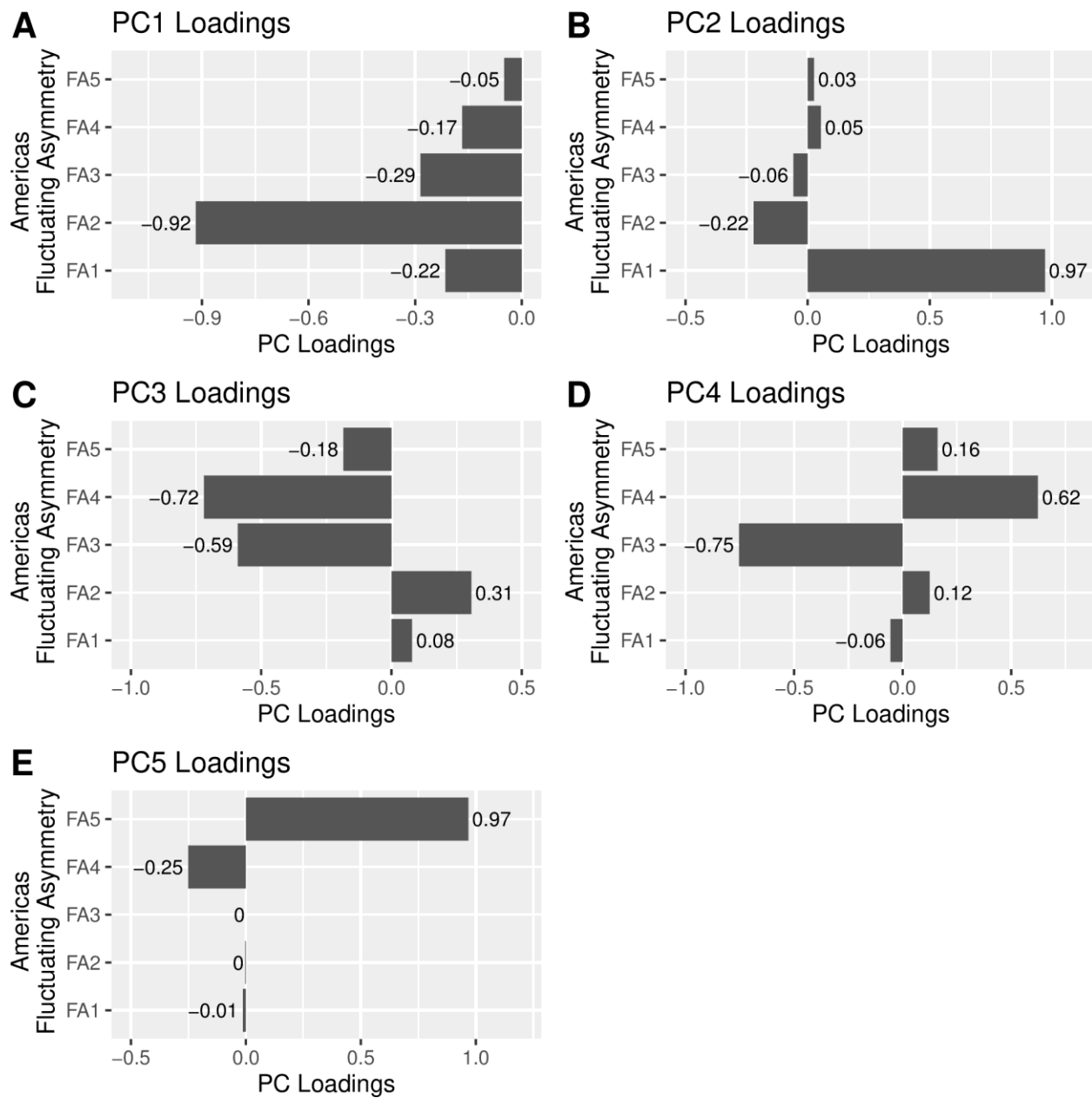


Figure 70. PC 1-5 loadings from PCA of covariance matrix of Landmass Americas Fluctuating Asymmetry. (N = 3522)

PC4 exhibited similar trends but opposite values as for Directional Asymmetry of Landmass Americas. PC2 captured the largest positive correlation of digit 1 against Fluctuating Asymmetry of all other digits in Landmass Americas (Figure 70B). PC4 exhibited a higher positive correlation for the Fluctuating Asymmetry of digit 4 and negative correlation for the Fluctuating Asymmetry of digit 3 (Figure 70D).

The first principal component explained 29.6% of the variation, and the second principal component explained 23.7% of the variation in Fluctuating Asymmetry of Landmass Americas. Visualizing the biplot of PC1 versus PC2 for only the centroids of different groupings illustrated the same pattern but opposite values as for Directional Asymmetry, now for Fluctuating Asymmetry an increasingly negative value along PC1 from Region South America, to Mesoamerica, to North America (Figure 71A). The centroids for Fluctuating Asymmetry of females and males in Landmass Americas were separate, but the error bars representing the standard error still overlap on the PC1 axis and the PC2 axis (Figure 71B). Females and males of Landmass Americas appeared to exhibit a similar amount of variation.

Further splitting the Fluctuating Asymmetry of Landmass Americas by-Sex, females of Region South America exhibited negative values along PC1 but males exhibited positive values along PC1, and females of Region North America exhibited positive values along PC1 but males exhibited negative values along PC1 (Figure 72). The first principal component for females exhibited 29.8% of the variation in Fluctuating Asymmetry of Landmass Americas, and the second principal component for females exhibited 23.5% of the variation in Fluctuating Asymmetry of Landmass Americas (Figure 72A). The first principal component for males exhibited 29.5% of the variation in Fluctuating Asymmetry of Landmass Americas, and the second principal component for males exhibited 24% of the variation in Fluctuating Asymmetry

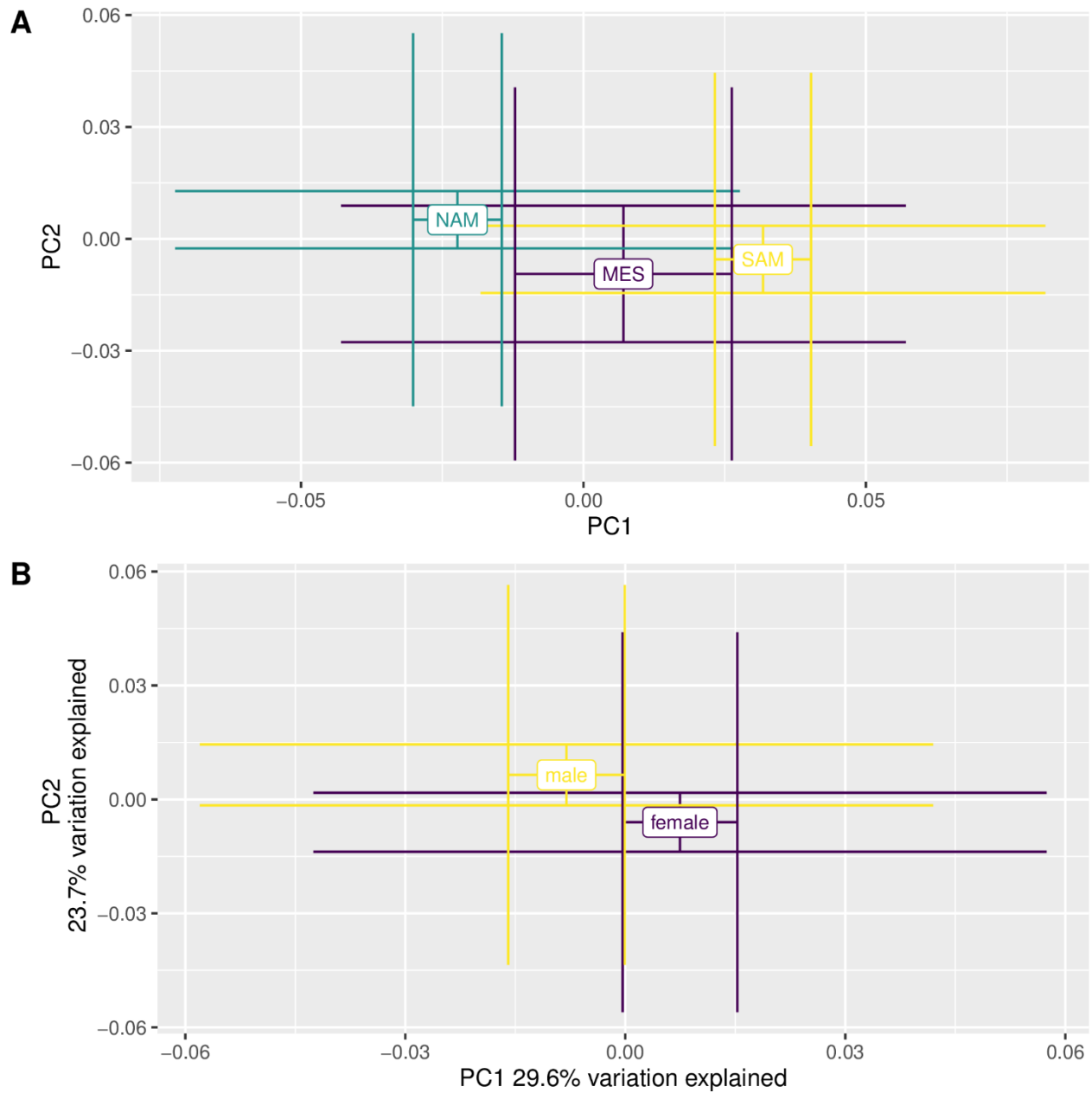


Figure 71. Principal component analysis of covariance matrix of Landmass Americas Fluctuating Asymmetry. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 3522). Error bars indicate the standard error.

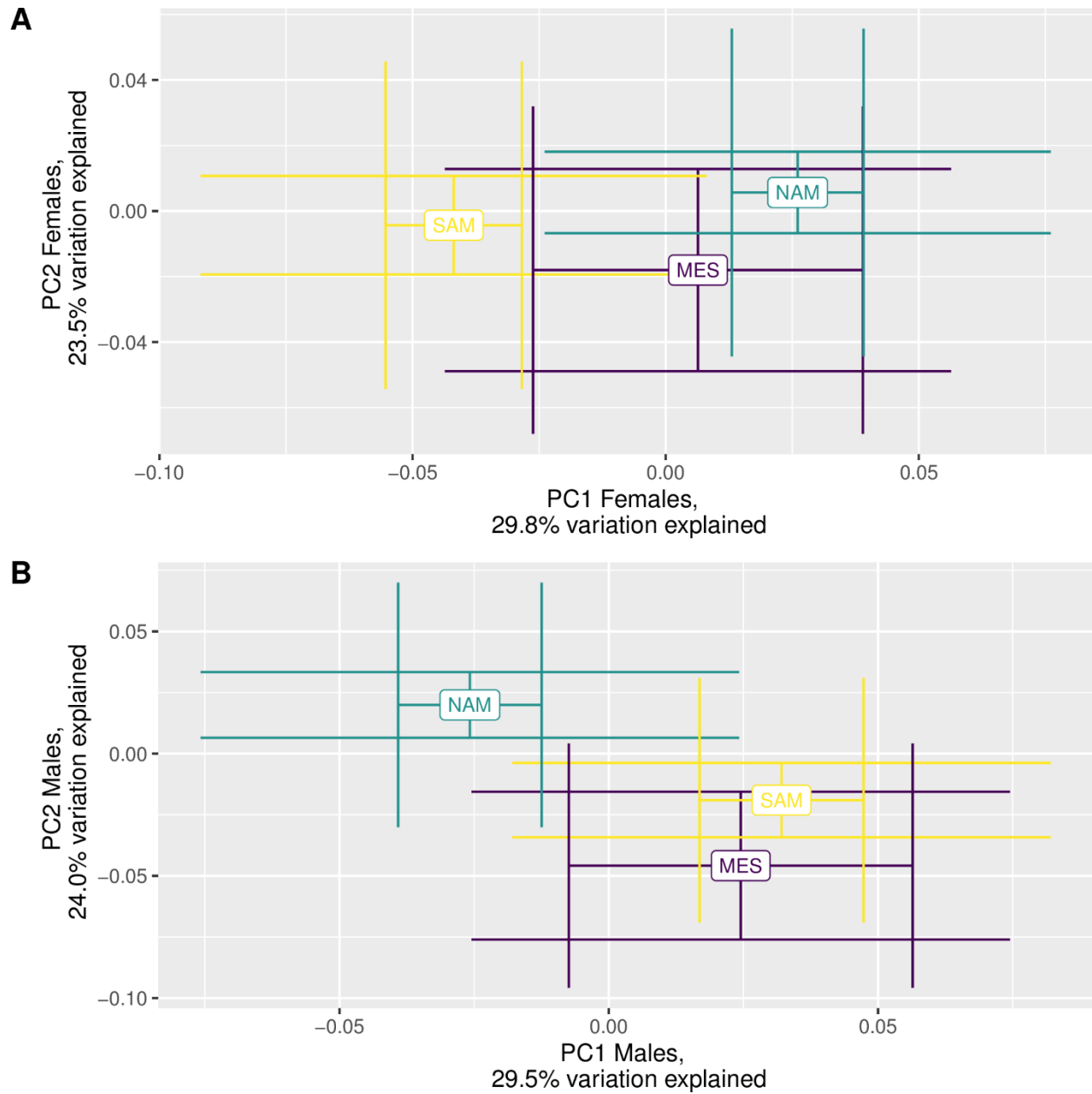


Figure 72. Principal component analysis of covariance matrix of female and male Landmass Americas Fluctuating Asymmetry. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 1825 females, 1697 males). Error bars indicate the standard error.

of Landmass Americas (Figure 72B).

For PC1, the Directional Asymmetry of Landmass Asia exhibited positive values indicating a positive correlation of Directional Asymmetry of Landmass Asia with PC1 (Figure 73A). The largest value for PC1 for Directional Asymmetry of Landmass Asia was for digit 2, which was similar to Directional Asymmetry (Figure 51), female Directional Asymmetry (Figure 55), male Directional Asymmetry (Figure 56), Landmass Americas Directional Asymmetry (Figure 67), Fluctuating Asymmetry (Figure 53), female Fluctuating Asymmetry (Figure 58), male Fluctuating Asymmetry (Figure 59), and Landmass Americas Fluctuating Asymmetry (Figure 70), but different from Landmass Africa Directional Asymmetry (Figure 61) and Landmass Africa Fluctuating Asymmetry (Figure 64) whereas the largest value for PC1 was digit 1. PC2 captured the largest negative correlation of digit 1 against the Directional Asymmetry of all other digits of Landmass Asia (Figure 73B). PC3 captured positive correlation for Directional Asymmetry of digits 1 and 2 versus negative correlation for Directional Asymmetry of digits 3, 4, and 5 of Landmass Asia (Figure 73C). PC4 exhibited a higher positive correlation for the Directional Asymmetry of digit 4 (Figure 73D). PC5 exhibited higher negative correlation for the Directional Asymmetry of digit 5 versus all other digits (Figure 73E).

The first principal component explained 27.6% of the variation, and the second principal component explained 23.9% of the variation in Directional Asymmetry of Landmass Asia. Visualizing the biplot of PC1 versus PC2 for only the centroids of different groupings illustrated a separation of Regions Australia, Indonesia, and Melanesia via more negative values along PC1 and more positive values along PC2 compared to the other six Regions in Landmass Asia (Figure 74A). The centroids for Directional Asymmetry of females and males in Landmass Asia were overlapping, and the error bars representing the standard error still overlap on the PC1 axis and

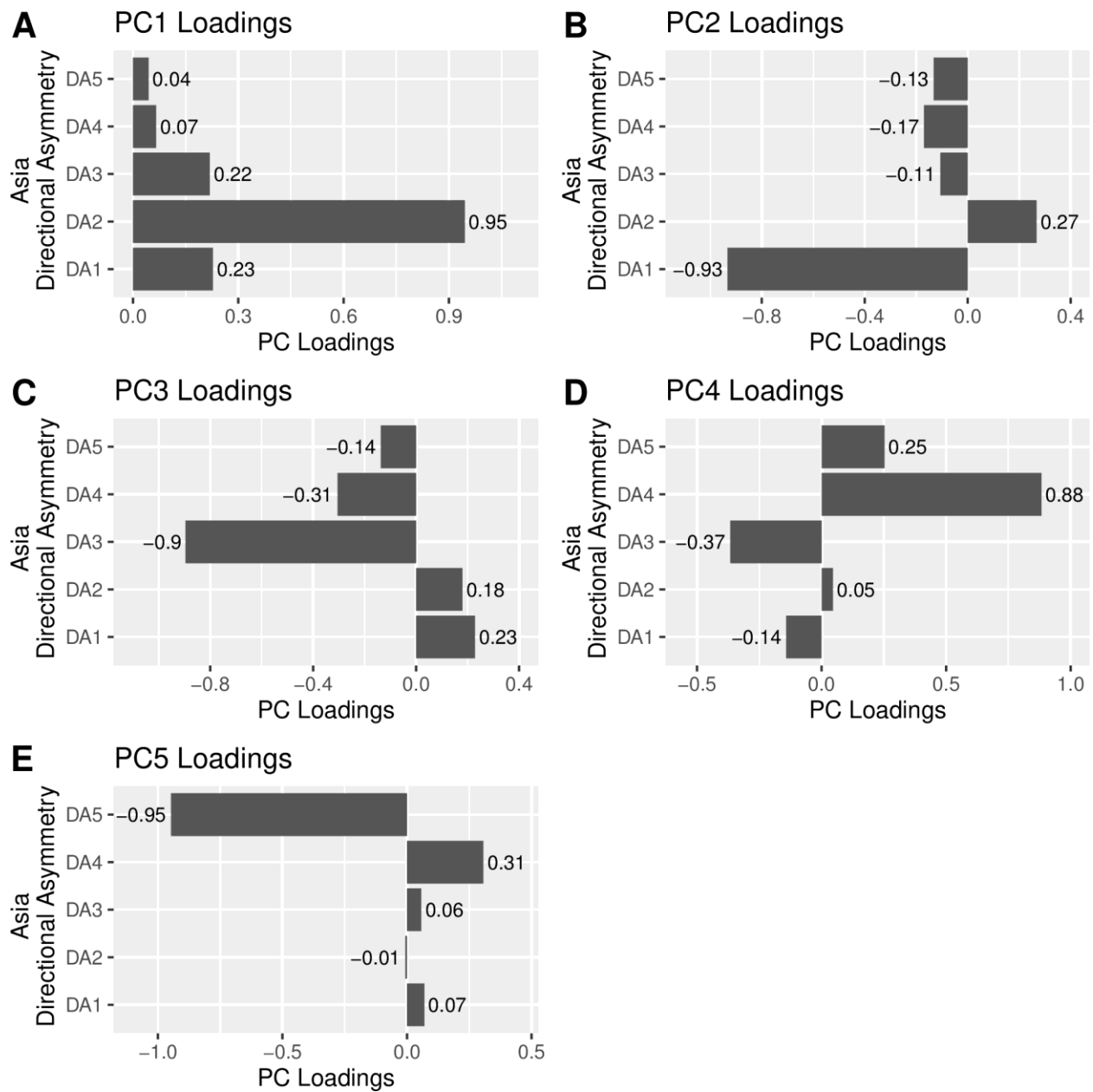


Figure 73. PC 1-5 loadings from PCA of covariance matrix of Landmass Asia Directional Asymmetry. (N = 15816)

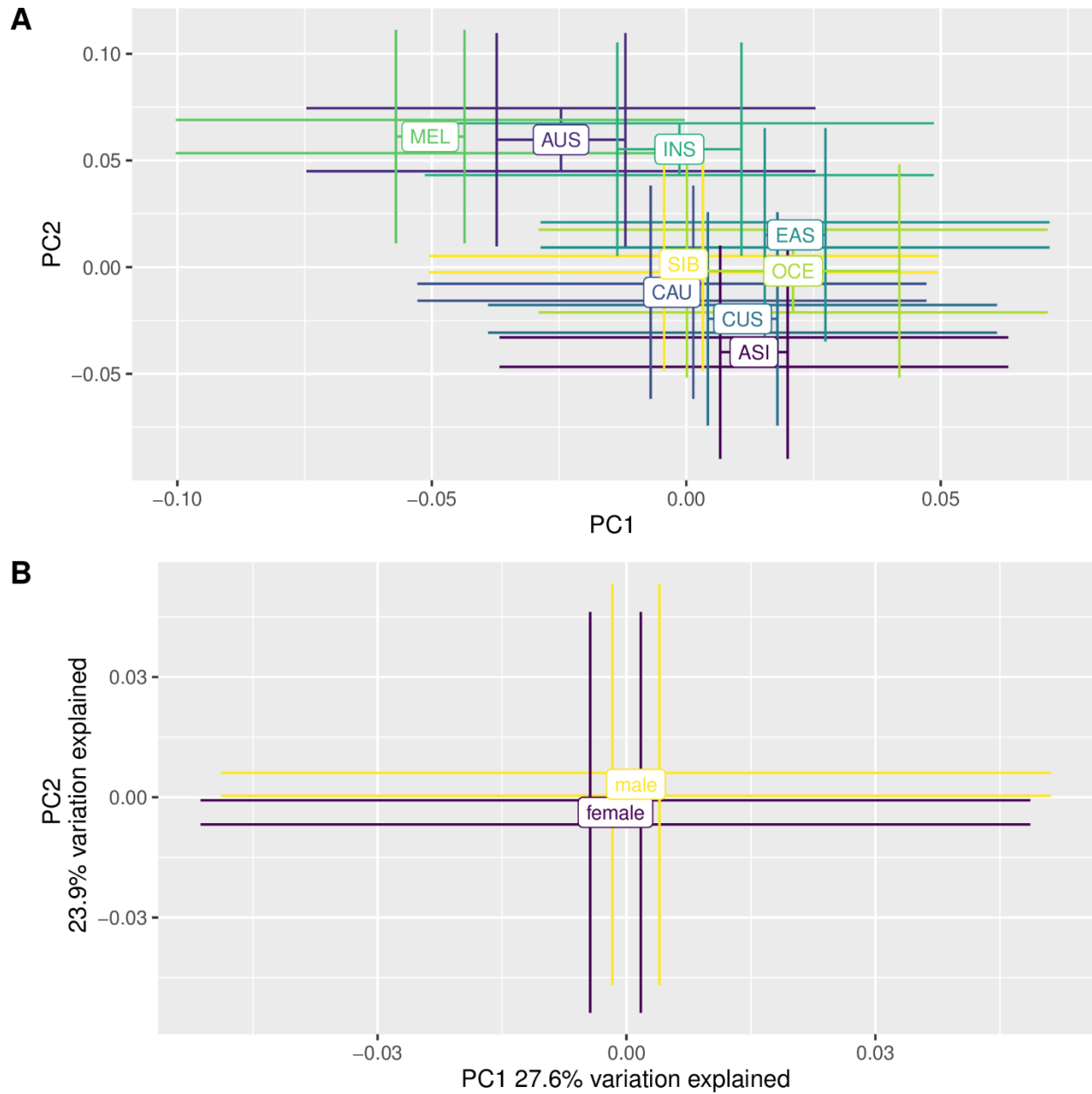


Figure 74. Principal component analysis of covariance matrix of Landmass Asia Directional Asymmetry. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 15816). Error bars indicate the standard error.

the PC2 axis (Figure 74B). Females and males of Landmass Asia appeared to exhibit a similar amount of variation.

Further splitting the Directional Asymmetry of Landmass Asia by-Sex, females differ from males primarily along PC2 (Figure 75). Regions Australia, Melanesia, and Indonesia can still be detected as separate from the other six Regions in Landmass Asia for both females and males separately as when the sexes were combined (Figure 74, Figure 75), although Regions Australia and Indonesia appeared closer to the other six Regions for females and not males. Females in Region Melanesia exhibited negative values along PC2 but males exhibited positive values along PC2, and females of Regions Asia and Central Soviet Union exhibited positive values along PC2 but males exhibited negative values along PC2 (Figure 75). Regions Siberia and the Caucasus in Landmass Asia remained around zero for PC1 and PC2 for both female and male Directional Asymmetry.

The first principal component for females exhibited 27.3% of the variation in Directional Asymmetry of Landmass Asia, and the second principal component for females exhibited 23.2% of the variation in Directional Asymmetry of Landmass Asia (Figure 75A). The first principal component for males exhibited 28.2% of the variation in Directional Asymmetry of Landmass Asia, and the second principal component for males exhibited 24.6% of the variation in Directional Asymmetry of Landmass Asia (Figure 75B).

For PC1, the Fluctuating Asymmetry of Landmass Asia exhibited positive values indicating a positive correlation of Fluctuating Asymmetry of Landmass Asia with PC1 (Figure 76A). The largest value for PC1 for Fluctuating Asymmetry of Landmass Asia was for digit 2, which was similar to Directional Asymmetry (Figure 52), female Directional Asymmetry (Figure 55), male Directional Asymmetry (Figure 56), Landmass Americas Directional Asymmetry

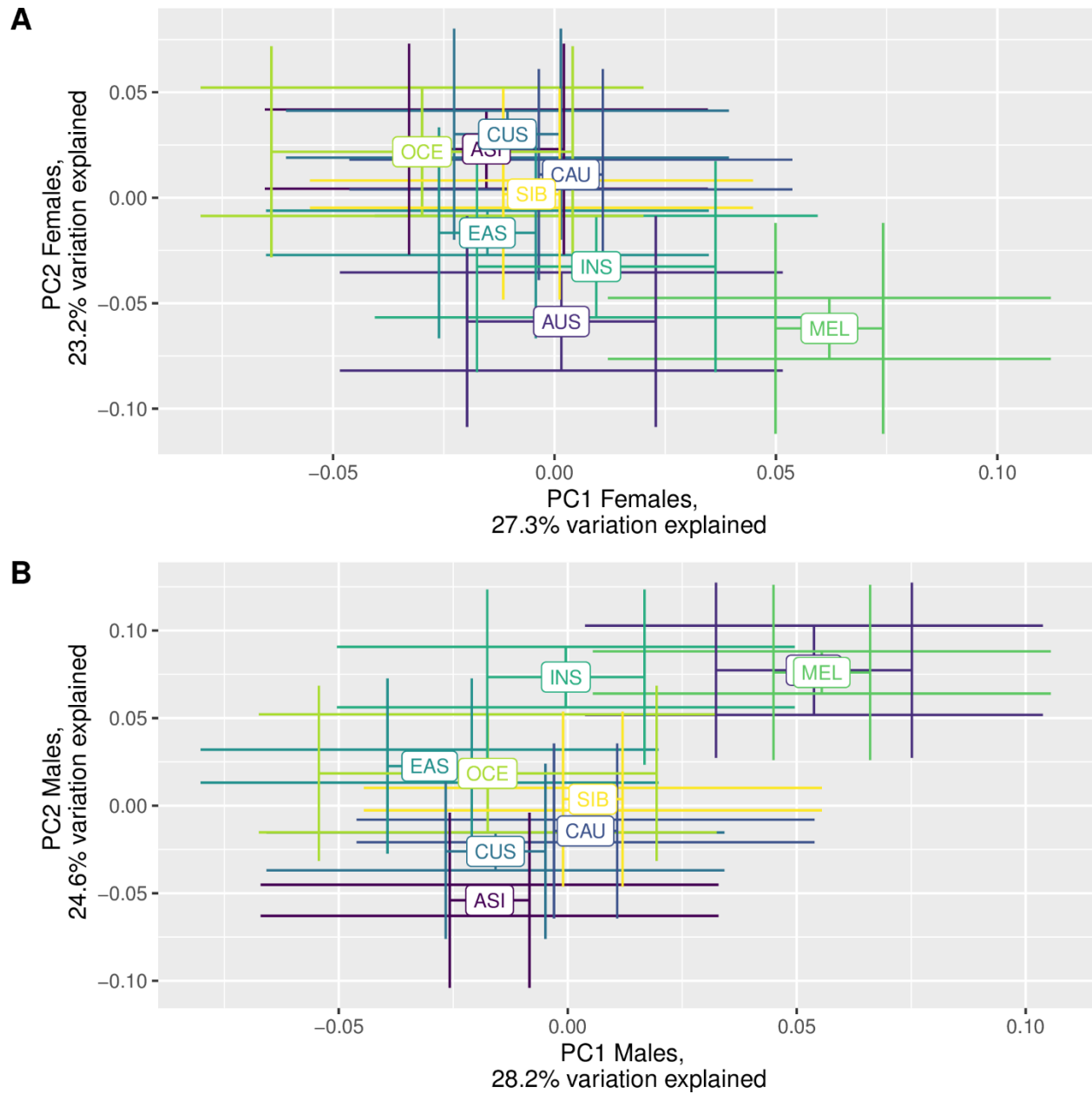


Figure 75. Principal component analysis of covariance matrix of female and male Landmass Asia Directional Asymmetry. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 7316 females, 8500 males). Error bars indicate the standard error.

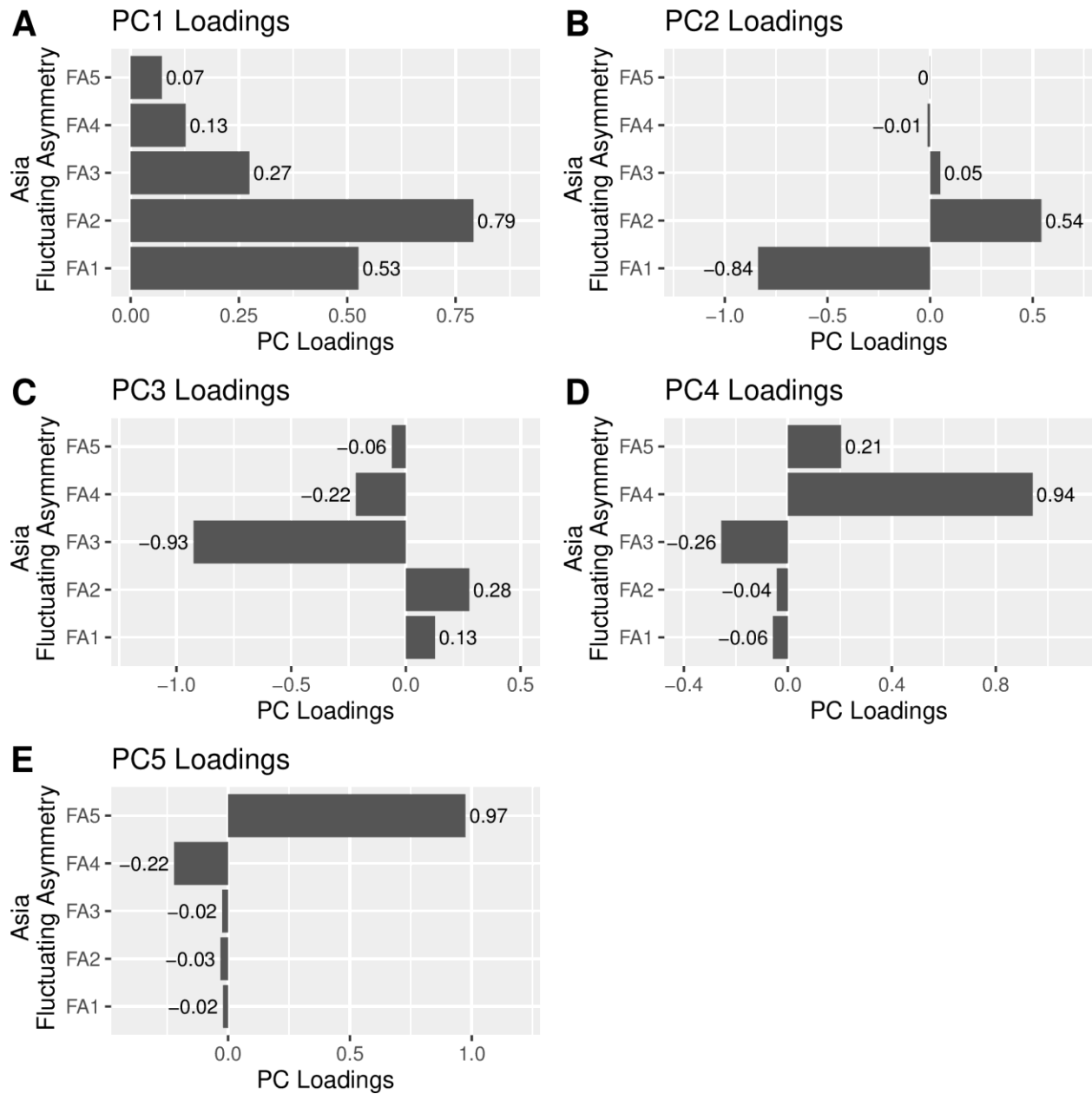


Figure 76. PC 1-5 loadings from PCA of covariance matrix of Landmass Asia Fluctuating Asymmetry. (N = 15816)

(Figure 67), Landmass Asia Directional Asymmetry (Figure 73), Fluctuating Asymmetry (Figure 53), female Fluctuating Asymmetry (Figure 58), male Fluctuating Asymmetry (Figure 59), and Landmass Americas Fluctuating Asymmetry (Figure 70), but different from Landmass Africa Directional Asymmetry (Figure 61) and Landmass Africa Fluctuating Asymmetry (Figure 64) whereas the largest value for PC1 was digit 1. The loadings of Fluctuating Asymmetry of Landmass Asia exhibited the same magnitude and direction for loading of Directional Asymmetry of Landmass Asia for PC1 and 3. PC2 exhibited positive loadings for Fluctuating Asymmetry of digits 2 and 3 compared to digit 5 loadings equal to zero and negative values for Fluctuating Asymmetry of digits 1 and 4, in Landmass Asia (Figure 76). Further, the PC4 exhibited positive loadings for Fluctuating Asymmetry of digits 4 and 5 compared to negative values for Fluctuating Asymmetry of digits 1, 2, and 3 in Landmass Asia. Finally, PC5 exhibited large positive loadings for Fluctuating Asymmetry of digit 5 compared to negative values for Fluctuating Asymmetry of all other digits in Landmass Asia(Figure 76E).

The first principal component explained 29.1% of the variation, and the second principal component explained 25.4% of the variation in Fluctuating Asymmetry of Landmass Asia. Visualizing the biplot of PC1 versus PC2 for only the centroids of different groupings illustrated a separation of Regions Australia, Indonesia, and Melanesia via more negative values along PC2 only compared to the other six Regions in Landmass Asia (Figure 77A). The centroids for Fluctuating Asymmetry of females and males in Landmass Asia were separate, but the error bars representing the standard error still overlap on the PC1 axis and the PC2 axis (Figure 77B). Females and males of Landmass Asia appeared to exhibit a similar amount of variation.

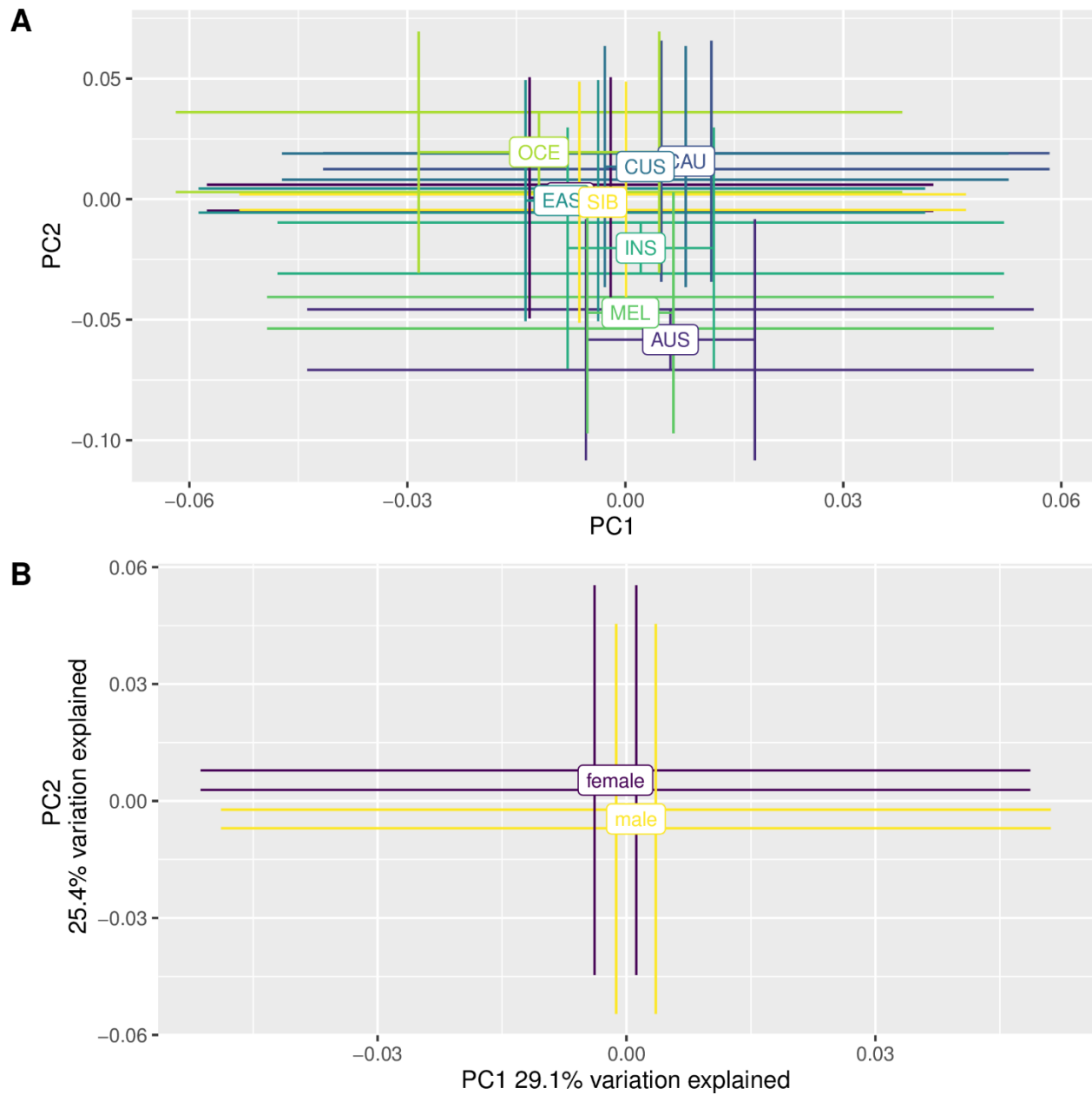


Figure 77. Principal component analysis of covariance matrix of Landmass Asia Fluctuating Asymmetry. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 15816). Error bars indicate the standard error.

Further splitting the Fluctuating Asymmetry of Landmass Asia by-Sex, females differ from males primarily along PC2 (Figure 78). Females in Regions Australia, Melanesia, and Indonesia can no longer be detected as separate from the other six Regions in Landmass Asia like was possible when the sexes were combined (Figure 77, Figure 78). Females in Regions Australia and Melanesia exhibited positive values along PC2 but males exhibited negative values along PC2, and females of Region Oceania exhibited negative values along PC2 but males exhibited values close to zero along PC2 (Figure 78). The remaining Regions in Landmass Asia remained around zero for PC1 and PC2 for both female and male Fluctuating Asymmetry.

The first principal component for females exhibited 28.2% of the variation in Fluctuating Asymmetry of Landmass Asia, and the second principal component for females exhibited 24.5% of the variation in Fluctuating Asymmetry of Landmass Asia (Figure 78A). The first principal component for males exhibited 30% of the variation in Fluctuating Asymmetry of Landmass Asia, and the second principal component for males exhibited 26.2% of the variation in Fluctuating Asymmetry of Landmass Asia (Figure 78B).

For PC1, the Directional Asymmetry of Landmass Europe exhibited positive values indicating a positive correlation of Directional Asymmetry of Landmass Europe with PC1 (Figure 79A). The largest value for PC1 for Directional Asymmetry of Landmass Europe was for digit 2, which was similar to Directional Asymmetry (Figure 51), female Directional Asymmetry (Figure 55), male Directional Asymmetry (Figure 56), Landmass Americas Directional Asymmetry (Figure 67), Landmass Asia Directional Asymmetry (Figure 73), Fluctuating Asymmetry (Figure 53), female Fluctuating Asymmetry (Figure 58), male Fluctuating Asymmetry (Figure 59), Landmass Americas Fluctuating Asymmetry (Figure 70), and Landmass Asia Fluctuating Asymmetry (Figure 76), but different from Landmass Africa Directional

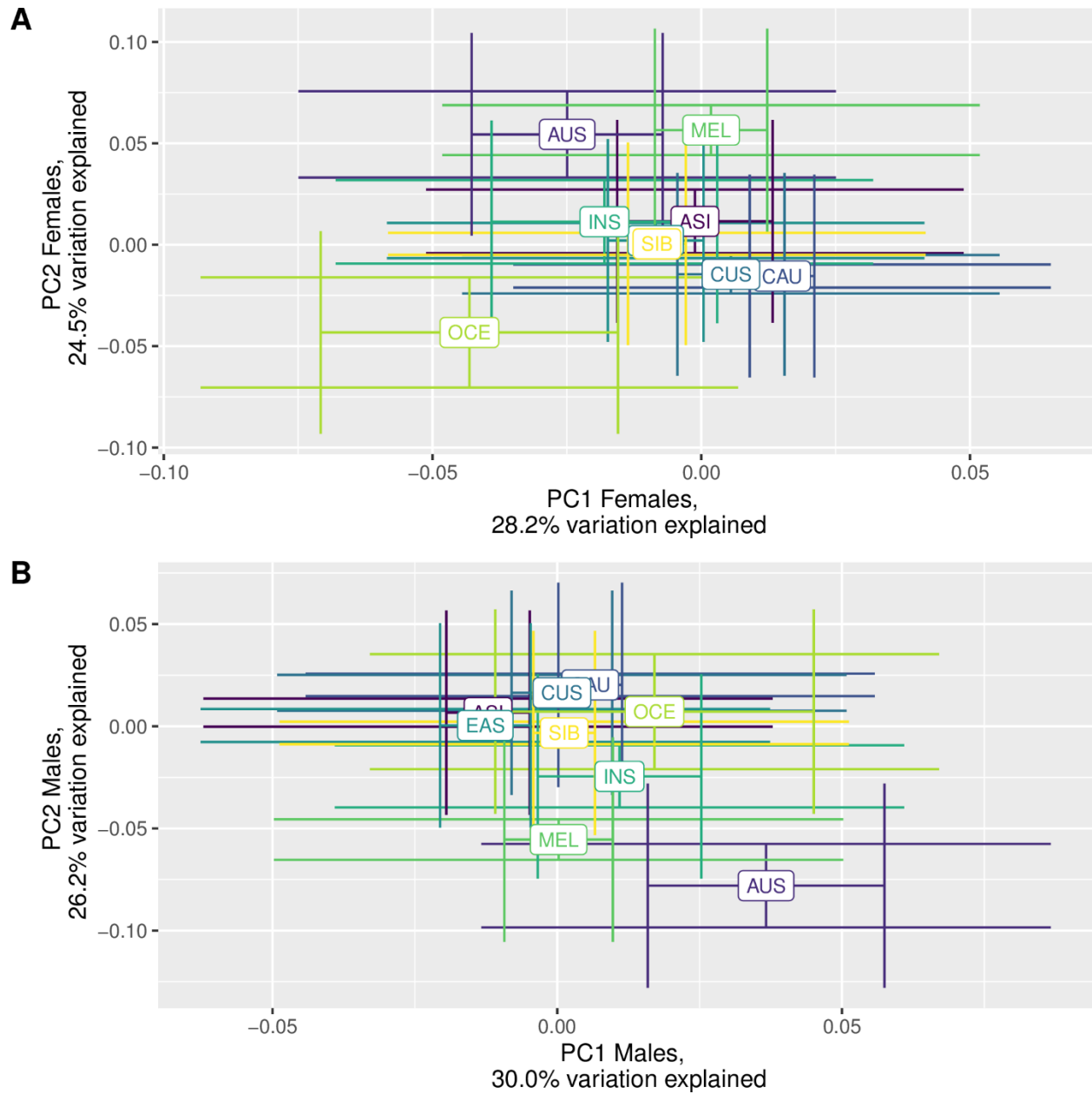


Figure 78. Principal component analysis of covariance matrix of female and male Landmass Asia Fluctuating Asymmetry. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 7316 females, 8500 males). Error bars indicate the standard error.

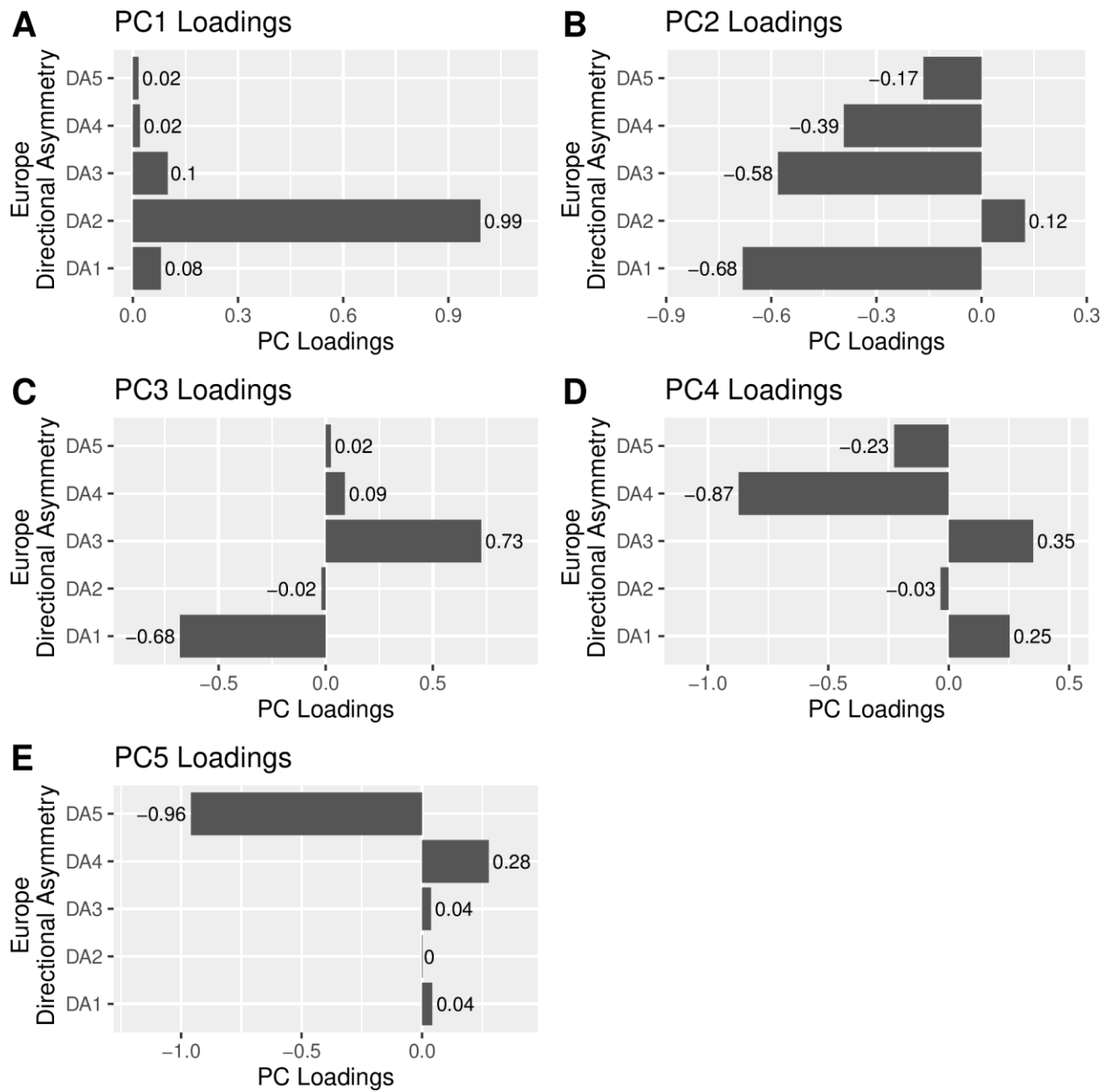


Figure 79. PC 1-5 loadings from PCA of covariance matrix of Landmass Europe Directional Asymmetry. (N = 16402)

Asymmetry (Figure 61) and Landmass Africa Fluctuating Asymmetry (Figure 64) whereas the largest value for PC1 was digit 1. PC2 captured the positive correlation of digit 2 against the negative correlation in Directional Asymmetry for all other digits of Landmass Europe (Figure 79B). PC3 captured negative correlation for Directional Asymmetry of digits 1 and 2 versus positive correlation for Directional Asymmetry of digits 3, 4, and 5 of Landmass Europe (Figure 79C). PC4 exhibited a higher negative correlation for the Directional Asymmetry of digit 4 (Figure 79D). PC5 exhibited higher negative correlation for the Directional Asymmetry of digit 5 versus all other digits (Figure 79E).

The first principal component explained 29.7% of the variation, and the second principal component explained 21.8% of the variation in Directional Asymmetry of Landmass Europe. Visualizing the biplot of PC1 versus PC2 for only the centroids of different groupings illustrated a slight separation of the centroids for the two Regions along PC2 (Figure 80A). The centroids for Directional Asymmetry of females and males in Landmass Europe were separate, but the error bars representing the standard error still overlap on the PC1 axis and the PC2 axis (Figure 80B). Females and males of Landmass Europe appeared to exhibit a similar amount of variation.

Further splitting the Directional Asymmetry of Landmass Europe by-Sex, females differ from males primarily along PC2 (Figure 81). The two Regions can still be detected as separate in Landmass Europe for both females and males separately as when the sexes were combined (Figure 80, Figure 81). The first principal component for females exhibited 28.9% of the variation in Directional Asymmetry of Landmass Europe, and the second principal component for females exhibited 22.5% of the variation in Directional Asymmetry of Landmass Europe (Figure 81A). The first principal component for males exhibited 30.5% of the variation in Directional Asymmetry of Landmass Europe, and the second principal component for males

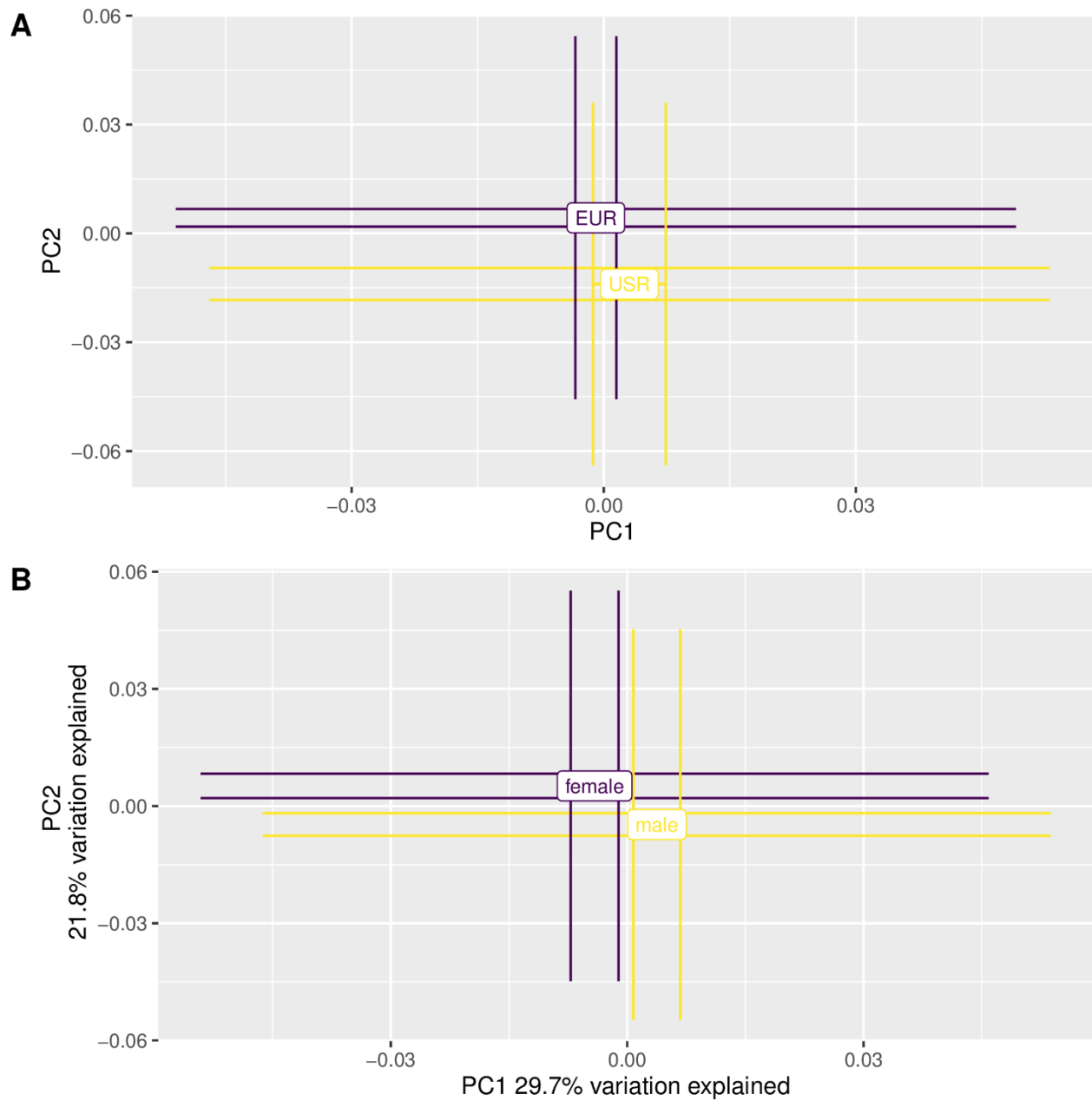


Figure 80. Principal component analysis of covariance matrix of Landmass Europe Directional Asymmetry. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 16402). Error bars indicate the standard error.

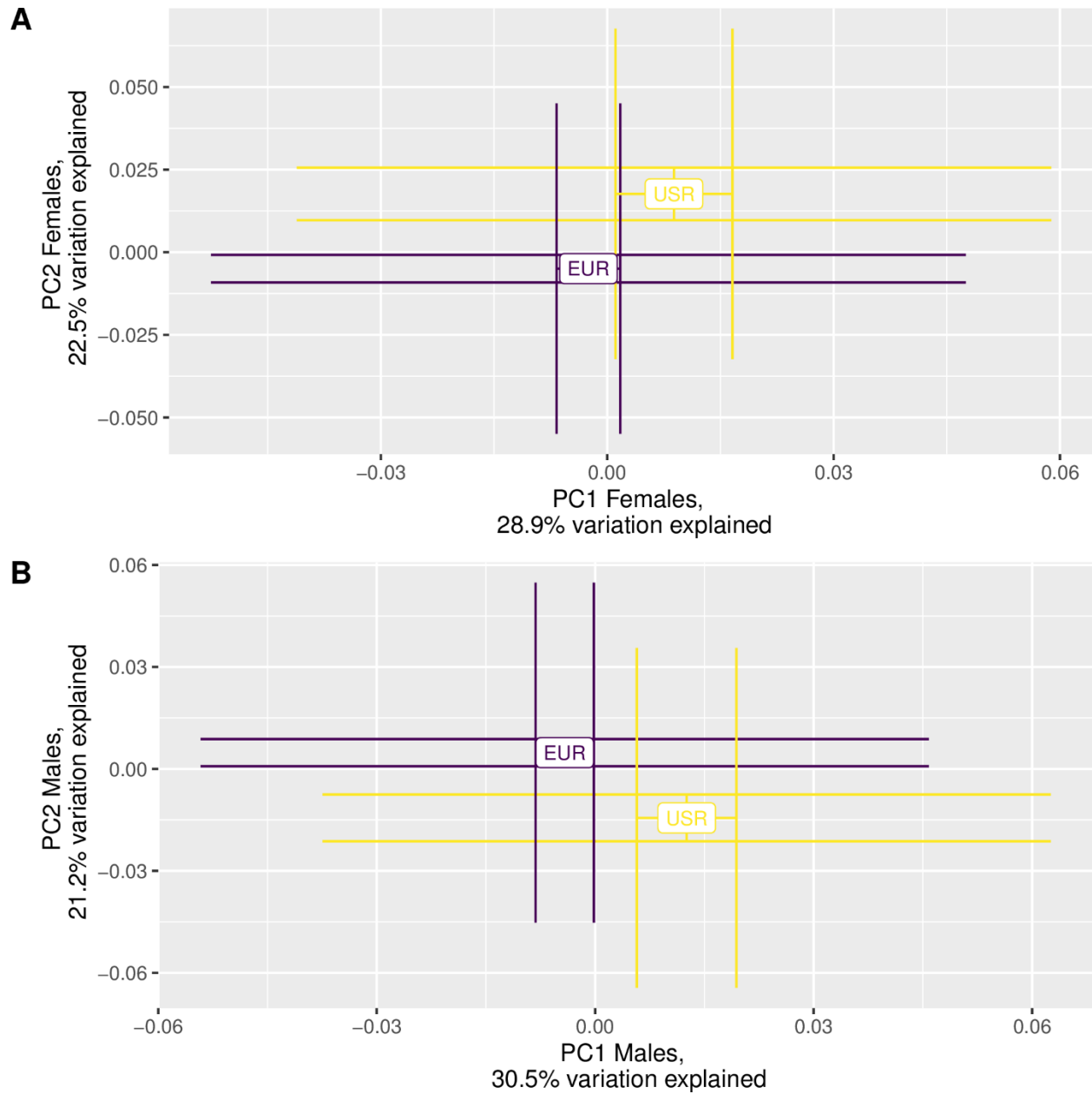


Figure 81. Principal component analysis of covariance matrix of female and male Landmass Europe Directional Asymmetry. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 7822 females, 8580 males). Error bars indicate the standard error.

exhibited 21.2% of the variation in Directional Asymmetry of Landmass Europe (Figure 81B).

For PC1, the Fluctuating Asymmetry of Landmass Europe exhibited negative values indicating a negative correlation of Fluctuating Asymmetry of Landmass Europe with PC1 (Figure 82A). The largest value for PC1 for Fluctuating Asymmetry of Landmass Europe was for digit 2, which was similar to Directional Asymmetry (Figure 51), female Directional Asymmetry (Figure 55), male Directional Asymmetry (Figure 56), Landmass Americas Directional Asymmetry (Figure 67), Landmass Asia Directional Asymmetry (Figure 73), Landmass Europe Directional Asymmetry (Figure 79), Fluctuating Asymmetry (Figure 54), female Fluctuating Asymmetry (Figure 58), male Fluctuating Asymmetry (Figure 59), Landmass Americas Fluctuating Asymmetry (Figure 70), and Landmass Asia Fluctuating Asymmetry (Figure 76), but different from Landmass Africa Directional Asymmetry (Figure 61) and Landmass Africa Fluctuating Asymmetry (Figure 64) whereas the largest value for PC1 was digit 1. The loadings of Fluctuating Asymmetry of Landmass Europe exhibited the opposite values but same magnitude and direction for loading of Directional Asymmetry of Landmass Europe for PC1, 2, and 3. PC2 captured the higher positive correlation of digit 1 against the lower values of correlation in Fluctuating Asymmetry for all other digits of Landmass Europe (Figure 82B). PC3 captured positive correlation for Fluctuating Asymmetry of digits 1 and 2 versus negative correlation for Fluctuating Asymmetry of digits 3, 4, and 5 of Landmass Europe (Figure 82C). Further, the PC4 exhibited positive loadings for Fluctuating Asymmetry of digits 4 and 5 compared to digit 2 loadings equal to zero and negative values for Fluctuating Asymmetry of digits 1 and 3 in Landmass Europe (Figure 82D). Finally, PC5 exhibited large positive loadings for Fluctuating Asymmetry of digit 5 compared to negative values for Fluctuating Asymmetry of all other digits in Landmass Europe (Figure 82E).

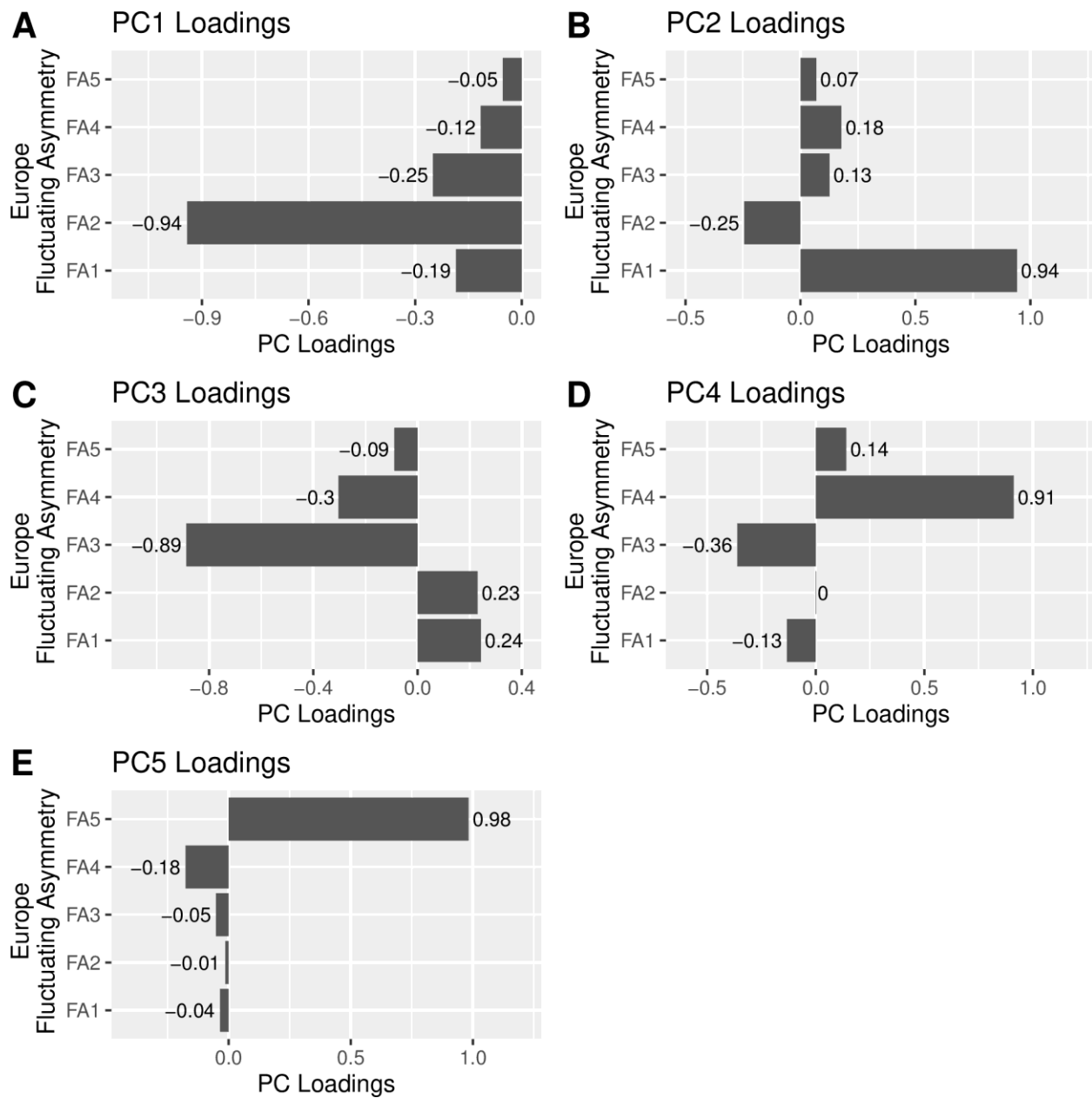


Figure 82. PC 1-5 loadings from PCA of covariance matrix of Landmass Europe Fluctuating Asymmetry. (N = 16402)

The first principal component explained 30.7% of the variation, and the second principal component explained 23.1% of the variation in Fluctuating Asymmetry of Landmass Europe. Visualizing the biplot of PC1 versus PC2 for only the centroids of different groupings illustrated the centroids for Fluctuating Asymmetry of females and males in Landmass Europe were overlapping, and the error bars representing the standard error overlapped on the PC1 axis and the PC2 axis (Figure 83A). The centroids for Fluctuating Asymmetry of females and males in Landmass Europe were separate, but the error bars representing the standard error still overlap on the PC1 axis and the PC2 axis (Figure 83B). Females and males of Landmass Europe appeared to exhibit a similar amount of variation.

Further splitting the Fluctuating Asymmetry of Landmass Europe by-Sex, females differ from males primarily along PC1, but the two Regions cannot be detected as separate (Figure 84). The first principal component for females exhibited 30.7% of the variation in Fluctuating Asymmetry of Landmass Europe, and the second principal component for females exhibited 22.8% of the variation in Fluctuating Asymmetry of Landmass Europe (Figure 84A). The first principal component for males exhibited 30.7% of the variation in Fluctuating Asymmetry of Landmass Europe, and the second principal component for males exhibited 23.4% of the variation in Fluctuating Asymmetry of Landmass Europe (Figure 84B).

4.2.4 Baseline models for generalized linear modeling

For the LASSO regularization of the GLM and thus baseline models of the GLM of Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit, the final training set contained 32585 observations, and for evaluating the final model on a separate test set to calculate performance metrics, the final test set contained 13964 observations. For GLM of Directional Asymmetry variables for each digit, the 10 explanatory variables

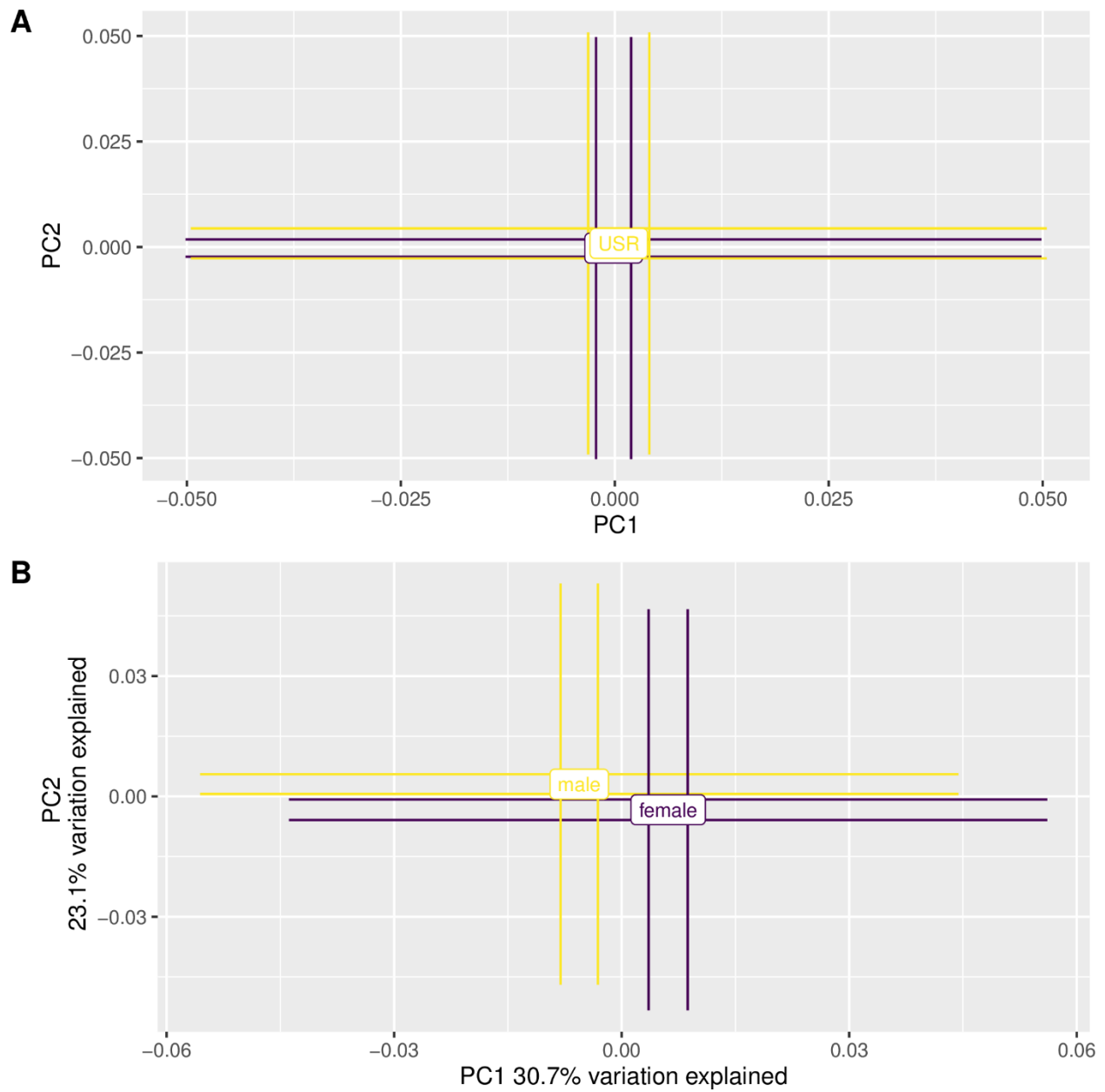


Figure 83. Principal component analysis of covariance matrix of Landmass Europe Fluctuating Asymmetry. Viewed by different grouping centroids: (A) by-Region and (B) by-Sex (N = 16402). Error bars indicate the standard error.

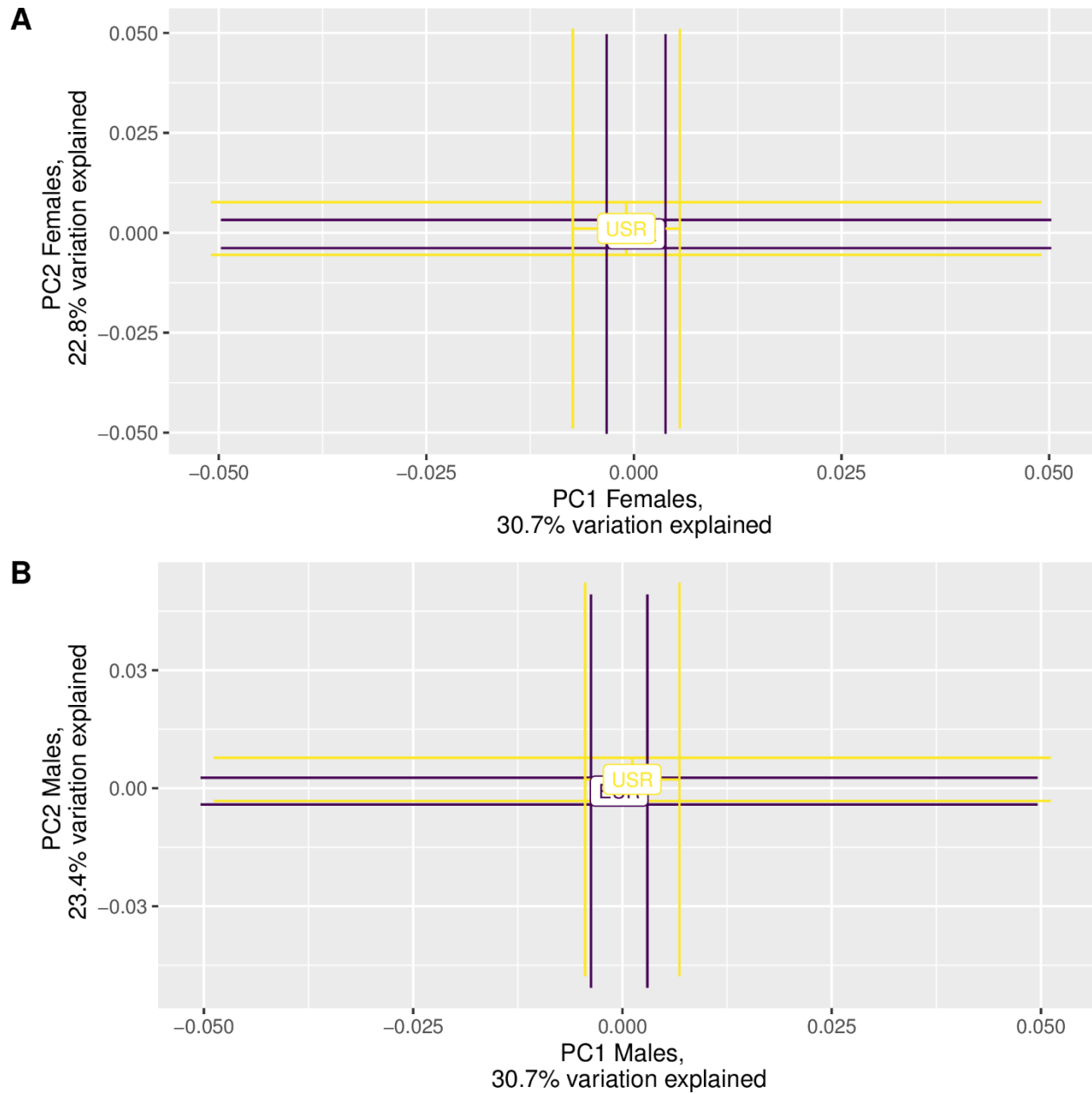


Figure 84. Principal component analysis of covariance matrix of female and male Landmass Europe Fluctuating Asymmetry. Viewed by different grouping centroids: (A) females by-Region, (B) males by-Region (N = 7822 females, 8580 males). Error bars indicate the standard error.

included were Ethnic Group, Region, Landmass, Sex, Latitude, Longitude, DA.kmeans.2, DA.kmeans.4, DA.kmeans.10, and DA.kmeans.322. For GLM of Fluctuating Asymmetry variables for each digit, the 10 explanatory variables included were Ethnic Group, Region, Landmass, Sex, Latitude, Longitude, FA.kmeans.2, FA.kmeans.4, FA.kmeans.8, and FA.kmeans.322. For LASSO regularization of the GLM and thus baseline models of the GLM of Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt, the final training set contained 32585 observations of the same 10 explanatory variables as for Directional Asymmetry variables for each digit, and the final test set contained 13964 observations of the same 10 variables.

To evaluate the assumptions of the mean regressor as an approximation of the central tendency of the response variables, the similarity to the median regressor was visualized, and for the Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response variables the mean of each appeared similar to the median of each (Figure 10, Figure 11, Figure 12). To evaluate whether the baseline linear model could even be interpreted in terms of goodness-of-fit or performance, initial model checking was performed to see whether the assumptions of linear models were met or could be approximately met.

Six baseline linear models for the GLM exhibited multicollinearity (high > 10 , moderate 5-10, low < 5) in the explanatory variables as computed via the variance inflation factor, which assesses how much the linear correlation of an explanatory variable with other explanatory variables increases the variance of the coefficient estimate compared to theoretically no correlation. For brevity, only the moderate or high explanatory variables were reported here, and the remaining explanatory variables not specified in that model were reported as low collinearity. The linear model with Directional Asymmetry of digit 3 as the response variable exhibited

moderate collinearity for explanatory variables Longitude and Region (Figure S80). The linear model with Directional Asymmetry of digit 4 as the response variable exhibited moderate collinearity for explanatory variables Latitude and Region (Figure S82). The linear model with Directional Asymmetry of digit 5 as the response variable exhibited high (> 30) collinearity for explanatory variable Region and moderate collinearity for explanatory variables Latitude and Longitude (Figure S84). The linear model with Fluctuating Asymmetry of digit 2 as the response variable exhibited moderate collinearity for explanatory variable Landmass (Figure S88). The linear model with Fluctuating Asymmetry of digit 5 as the response variable exhibited high (> 30) collinearity for explanatory variable Region and moderate collinearity for explanatory variables Latitude and Longitude (Figure S94). The linear model with Index.Asymmetry as the response variable exhibited high (> 30) collinearity for explanatory variable Region and moderate collinearity for explanatory variables Latitude and Longitude (Figure S96). The consequences of continuing to use a model known to exhibit high collinearity between explanatory variables are primarily relegated to interpretation of coefficients and resulting p values, which were not the focus of these models in this dissertation. Further, when collinearity is present, the predictions, precision of the predictions, and the goodness-of-fit statistics do not change (Neter 1996:289). Solutions to collinearity between explanatory variables are to remove the highly correlated explanatory variables, to perform LASSO regression that can handle collinearity, etc. (Frost 2017).

For all linear models with response variables of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity, the Anderson-Darling normality test p value was < 0.05 , so reject the null hypotheses the errors were from a normal distribution (Table 24). For a visual assessment of normality, the

Table 24. Violation of standard regression assumptions of residuals for baseline linear models for GLM of Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry of diversity; DA = directional asymmetry; FA = fluctuating asymmetry; GLM = generalized linear model; p value < 0.05 in bold

GLM	Multicollinearity	Anderson-Darling Normality		Durbin-Watson Autocorrelation		Breusch-Pagan Heteroscedasticity
		A test statistic	p value < 0.001	d test statistic	p value	
Response variable	Variance Inflation Factor > 10					p value < 0.001
DA1		92.266	✓	1.968	0.002	
DA2		114.79	✓	1.995	0.606	✓
DA3		121.67	✓	2.002	0.936	
DA4		70.777	✓	1.978	0.040	
DA5	✓	83.897	✓	1.975	0.012	✓
FA1		133.91	✓	1.985	0.116	✓
FA2		184.57	✓	2.006	0.572	✓
FA3		547.36	✓	1.991	0.356	✓
FA4		752.31	✓	1.979	0.048	✓
FA5	✓	410.92	✓	1.986	0.180	✓
Index.Asymmetry	✓	519.3	✓	1.975	0.028	✓
Index.Diversity		260.27	✓	1.995	0.034	✓
Index.Diversity.Holt		255.53	✓	1.989	0.006	✓

normality of residuals plots exhibit an almost perfect one-to-one line toward the center comparing the standard normal distribution quantiles on the x-axis as the residuals from a perfect model to the sample quantiles on the y-axis as the real residuals from our models, but the relationship breaks down as seen from the dots falling away from the line particularly when the standard normal distribution quantiles on the x-axis dip below -2.5 and exceed 2.5 at the tails, and for the histograms of residuals the lower and upper tails were unequal (Figure S76 through Figure S101). The normality of residual plots for Fluctuating Asymmetry variables digits 1 and 2 exhibit deviation of dots away from the line only when the standard normal distribution quantiles exceed 2.5 at the upper tail (Figure S88, Figure S90). Thus, all linear models with response variables of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity exhibited non-normal distribution of errors, which could be attributed to autocorrelation. Other reasons for non-normal distribution of errors could be an important explanatory variable not included in the model or some polynomial or interaction term not included in the model. The consequences of continuing to use a model known to violate the normality of residuals assumption are primarily biased estimate of standard errors affecting interpretation of confidence intervals and the p values, which were not the focus of these models in this dissertation, and with large sample sizes where the number of observations per variable exceeds 10 the impact of the negative consequences diminishes (Schmidt and Finan 2018).

One potential source of the non-normally distributed errors from linear models with response variables of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity was evaluated with the Durbin-Watson test for autocorrelation of residuals, indicated by a p value < 0.05 (Table 24).

Autocorrelation of residuals was detected by the Durbin-Watson test for response variables: Directional Asymmetry of digit 1, Directional Asymmetry of digit 4, Directional Asymmetry of digit 5, Fluctuating Asymmetry of digit 4, Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt (Table 24). However, in practical use any Durbin-Watson test statistic value equal to 2 indicated no autocorrelation and between 1.5 and 2.5 could be approximated as no autocorrelation (Bedre 2019; SE n.d.). The highest Durbin-Watson test statistic was 2.006181 for response variable Fluctuating Asymmetry of digit 2, and the p value was 0.572 indicating the null hypothesis of no autocorrelation of residuals was supported. The lowest Durbin-Watson test statistic with a p value > 0.05 (actual p value = 0.116) indicating no autocorrelation of residuals was 1.984592 for response variable Fluctuating Asymmetry of digit 1. This test statistic was less than 2. The lowest Durbin-Watson test statistic with a p value < 0.05 (actual p value = 0.002) indicating autocorrelation of residuals was 1.968382 for response variable Directional Asymmetry of digit 1. This test statistic where the p value indicated autocorrelation was similar to the previously stated test statistic where the p value indicated no autocorrelation. While non-normality of residuals does not pose a serious threat to the interpretation of the baseline linear models, the consequences of continuing to use a model known to exhibit autocorrelation of residuals, even with large sample sizes, are great. Autocorrelation of residuals reduces the interpretability of the estimates of the coefficients but also their standard errors. As a result, the p values would be biased, and the standard error of the regression coefficients may underestimate the true standard deviation of the estimated regression coefficients, so the mean squared error performance metric may greatly underestimate the true variance of the errors. Similarly, the RMSE would be unreliable (Broersen 2005). Given no Durbin-Watson test statistics exceeded the bounds of the practical use of the Durbin-Watson test, no baseline linear models with

response variables of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity were treated as autocorrelated (Bedre 2019; SE n.d.).

Other violations of linear regression assumptions that can seriously bias the interpretation of the model even with a large sample size are the presence of extreme observations that unduly influence the regression and also the heteroscedasticity that reduces the interpretation of the coefficients. Except for response variables Directional Asymmetry of digit 1, Directional Asymmetry of digit 3, and Directional Asymmetry of digit 4, all other linear models with response variables of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity, the results of the heteroscedasticity check exhibited a p value < 0.001 , so reject the null hypothesis the error variance was constant (Table 24). Further, the homogeneity of variance plots for Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity as the response variables did not show a flat and horizontal reference line comparing the fitted values on the x-axis to the square root of the standardized residuals on the y-axis (Figure S76, Figure S78, Figure S80, Figure S82, Figure S84, Figure S86, Figure S88, Figure S90, Figure S92, Figure S94, Figure S96, Figure S98, Figure S100). The consequences of continuing to use a model known to exhibit heteroscedasticity rely on identifying the purity of the heteroscedasticity and so do the solutions. For impure heteroscedasticity where the model was not correctly specified by leaving an important variable out, and the generation of the coefficient values was not biased, so opportunities exist to solve the problem with additional explanatory variables. For pure heteroscedasticity where the model is correctly specified and yet non-constant variance in the residuals persists, the issues with interpretation are the same as with

autocorrelation, and the prediction error significantly changes over the range of prediction of the model, so the performance metrics, which rely on the prediction error likely represent an underestimate of the performance error, or a minimum. For the baseline linear models exhibiting significant heteroscedasticity, the performance metrics were interpreted as a minimum value, meaning if the comparison in performance between the baseline linear model that exhibits heteroscedasticity and the mean regressor were close in value, the mean regressor would likely be identified as the better performing baseline. Solutions to pure heteroscedasticity are the use of weighted least squares, obtaining standard errors corrected for heteroscedasticity, or retooling the model (Bucevska 2011).

Calculating a goodness-of-fit for each model serves to indicate how well the explanatory variables in the model explain the observed variation in the response variable. However, comparing goodness-of-fit values for multiple models with different response variables is not appropriate. So, only the AIC of the full model was compared to the reduced model containing the same response variable and model formulation to indicate how the feature selection process improved the fit of the model (Schneider, Hommel, and Blettner 2010). For the mean regressor, median regressor, and linear regression model as baseline comparisons for the GLM, only the linear model incorporated explanatory variables in the model construction, and so the goodness-of-fit was only assessed for the linear model as a baseline via the AIC value, for which lower values indicate a better model fit (Table 25).

The linear model with Fluctuating Asymmetry of digit 2 as the response variable, with feature selected explanatory variables: Landmass, Sex, Latitude, Longitude, FA.kmeans.2, FA.kmeans.4, FA.kmeans.8, and FA.kmeans.322, exhibited the greatest reduction in AIC (Table 25). The next lowest reductions in AIC value were for the remaining Fluctuating Asymmetry

Table 25. Goodness-of-fit for full and reduced baseline linear models of Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry and diversity, of same GLM data set; AIC = Akaike information criterion; DA = directional asymmetry; FA = fluctuating asymmetry; GLM = generalized linear model; LM = linear model

Analytical Method	Response Variable	Model	AIC
LM	DA1	full	177287.1
		reduced	177139.9
LM	DA2	full	177270.9
		reduced	177013.2
LM	DA3	full	174271.3
		reduced	174006.3
LM	DA4	full	168681.8
		reduced	168542.2
LM	DA5	full	166371.2
		reduced	166309.5
LM	FA1	full	152188.8
		reduced	151921.6
LM	FA2	full	133473.5
		reduced	133107.3
LM	FA3	full	153487.8
		reduced	153166.2
LM	FA4	full	150954.8
		reduced	150682.8
LM	FA5	full	137658.4
		reduced	137377.3
LM	Index.Asymmetry	full	176634
		reduced	176508.7
LM	Index.Diversity	full	166036.7
		reduced	165999.8
LM	Index.Diversity.Holt	full	120656.1
		reduced	120619.8

variables of the other digits. Then, the next lowest reductions in AIC value were for Directional Asymmetry of digit 3, Directional Asymmetry of digit 2, Directional Asymmetry of digit 1, Directional Asymmetry of digit 4, and Index.Asymmetry. Only the remaining response variables for Directional Asymmetry of digit 5, Index.Diversity, and Index.Diversity.Holt exhibited a reduction under 100 AIC value. The majority of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response variables exhibited a reduction in AIC through feature selection ranging between 125 and 366. Of note, the reduced models for the baseline linear models of all Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity as response variables indicated a marked improvement in the goodness-of-fit.

The best performance metric for the Directional Asymmetry variables for each digit response variables was the MAE due to the presence of outliers that could unduly influence the modeling, the zeros in the response variables that did not always results in an undefined MAPE, and the arbitrary zeros since directional asymmetry represents the signed difference between the left and right sides (Table 5). The best performance metric for the Fluctuating Asymmetry variables for each digit and the indices of asymmetry and diversity was either the MAE or MAPE due to the presence of outliers and the zeros. When the MAPE was undefined (NaN), the second best performance metric for these response variables, the MAE, was considered.

The GLM performed the best for only the response variables Index.Asymmetry (Table 26). The mean regressor performed the best for only the response variable Fluctuating Asymmetry of digit 1 and only considering the lowest MAPE, because the lowest MAE for that response variable was the linear model. The median regressor performed the best for three

Table 26. Performance metrics of final GLM formulation as a result of LASSO regularization and lambda optimization as compared to baselines for Directional Asymmetry, Fluctuating Asymmetry, and indices of asymmetry and diversity; MAE = mean absolute error; MAPE = mean absolute percentage error; NA = Not available; NaN = Not a number; RMSE = root mean squared error; best performance metric and model in bold

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline for GLM	mean DA1	NA	NA	3.214	4.234	1.686
Baseline for GLM	median DA1	NA	NA	3.211	4.235	1.606
Baseline LM	DA1	NA	Region + Sex + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10 + DA.kmeans.322	2.717	3.609	10.311
GLM	DA1	0.006	Landmass + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10 + DA.kmeans.322	2.719	3.616	9.627
Baseline for GLM	mean DA2	NA	NA	3.332	4.594	7.108
Baseline for GLM	median DA2	NA	NA	3.312	4.626	NaN
Baseline LM	DA2	NA	Landmass + Sex + Latitude + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10	2.692	3.571	2.371
GLM	DA2	0.012	DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10	2.694	3.576	1.605
Baseline for GLM	mean DA3	NA	NA	2.906	3.948	9.324
Baseline for GLM	median DA3	NA	NA	2.886	3.958	NaN

Table 26 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline LM	DA3	NA	Region + Sex + Longitude + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10	2.606	3.498	8.939
GLM	DA3	0.003	DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10	2.609	3.505	7.169
Baseline for GLM	mean DA4	NA	NA	2.634	3.611	2959.766
Baseline for GLM	median DA4	NA	NA	2.634	3.611	NaN
Baseline LM	DA4	NA	Region + Sex + Latitude + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10	2.4076	3.170	9.211
GLM	DA4	0.003	Region + Sex + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10	2.4079	3.172	8.841
Baseline for GLM	mean DA5	NA	NA	2.353	3.131	9.764
Baseline for GLM	median DA5	NA	NA	2.325	3.139	NaN
Baseline LM	DA5	NA	Region + Sex + Latitude + Longitude + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10	2.326	3.090	25.160
GLM	DA5	0.001	Region + Latitude + DA.kmeans.2 + DA.kmeans.4	2.326	3.092	43.147
Baseline for GLM	mean FA1	NA	NA	2.347	3.081	0.672
Baseline for GLM	median FA1	NA	NA	2.246	3.118	0.749

Table 26 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline LM	FA1	NA	Region + Sex + FA.kmeans.2 + FA.kmeans.4 + FA.kmeans.8 + FA.kmeans.322	1.891	2.453	0.722
GLM	FA1	0.003	FA.kmeans.2 + FA.kmeans.4 + FA.kmeans.8 + FA.kmeans.322	2.613	3.680	2.308
Baseline for GLM	mean FA2	NA	NA	2.426	3.229	0.730
Baseline for GLM	median FA2	NA	NA	2.302	3.485	1.151
Baseline LM	FA2	NA	Landmass + Sex + Latitude + Longitude + FA.kmeans.2 + FA.kmeans.4 + FA.kmeans.8 + FA.kmeans.322	1.447	1.846	0.627
GLM	FA2	0.006	FA.kmeans.2 + FA.kmeans.4 + FA.kmeans.8	2.532	3.617	2.505
Baseline for GLM	mean FA3	NA	NA	2.011	2.709	0.695
Baseline for GLM	median FA3	NA	NA	1.913	2.851	0.957
Baseline LM	FA3	NA	Landmass + Sex + Latitude + FA.kmeans.2 + FA.kmeans.8 + FA.kmeans.322	1.937	2.543	0.689
GLM	FA3	0.005	Landmass + Latitude + FA.kmeans.2 + FA.kmeans.4 + FA.kmeans.8 + FA.kmeans.322	2.172	3.220	2.147
Baseline for GLM	mean FA4	NA	NA	1.816	2.470	0.682
Baseline for GLM	median FA4	NA	NA	1.692	2.550	0.846

Table 26 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline LM	FA4	NA	Region + Sex + FA.kmeans.2 + FA.kmeans.4 + FA.kmeans.8 + FA.kmeans.322	1.777	2.398	0.665
GLM	FA4	0.003	Landmass + FA.kmeans.2 + FA.kmeans.4 + FA.kmeans.8 + FA.kmeans.322	1.962	2.953	2.061
Baseline for GLM	mean FA5	NA	NA	1.594	2.109	0.678
Baseline for GLM	median FA5	NA	NA	1.498	2.134	0.749
Baseline LM	FA5	NA	Region + Sex + Latitude + Longitude + FA.kmeans.2 + FA.kmeans.4 + FA.kmeans.8 + FA.kmeans.322	1.518	1.991	0.659
GLM	FA5	0.003	Landmass + FA.kmeans.2 + FA.kmeans.4 + FA.kmeans.8 + FA.kmeans.322	1.769	2.536	2.234
Baseline for GLM	mean Index.Asymmetry	NA	NA	2.989	3.851	0.363
Baseline for GLM	median Index.Asymmetry	NA	NA	2.924	3.911	0.391
Baseline LM	Index.Asymmetry	NA	Region + Sex + Latitude + Longitude + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10 + DA.kmeans.322	2.777	3.628	0.343
GLM	Index.Asymmetry	0.005	Landmass + Latitude + Longitude + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10 + DA.kmeans.322	2.756	3.569	0.341

Table 26 Continued

Analytical Method	Response Variable	Best lambda value	Explanatory Variables	MAE	RMSE	MAPE
Baseline for GLM	mean	NA	NA	2.504	3.142	0.365
Baseline for GLM	Index.Diversity	NA	NA	2.466	3.179	0.389
Baseline LM	median	NA	NA	2.466	3.179	0.389
Baseline LM	Index.Diversity	NA	Ethnic Group + Sex + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10 + DA.kmeans.322	2.428	3.062	0.354
GLM	Index.Diversity	0.001	Landmass + Sex + Latitude + Longitude + DA.kmeans.4	2.463	3.088	0.358
Baseline for GLM	mean	NA	NA	1.274	1.584	0.319
Baseline for GLM	Index.Diversity. Holt	NA	NA	1.253	1.602	0.337
Baseline LM	median	NA	NA	1.253	1.602	0.337
Baseline LM	Index.Diversity. Holt	NA	Ethnic Group + Sex + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10 + DA.kmeans.322	1.218	1.526	0.306
GLM	Index.Diversity. Holt	0.001	Region + Landmass + Sex + Latitude + Longitude + DA.kmeans.2 + DA.kmeans.4 + DA.kmeans.10 + DA.kmeans.322	1.235	1.537	0.310

response variables: Fluctuating Asymmetry of digit 3, Fluctuating Asymmetry of digit 4, and Fluctuating Asymmetry of digit 5 only considering the lowest MAE, because the lowest MAPE for those response variables was the linear model. The baseline linear model performed the best for the response variable for Directional Asymmetry of digit 1 only considering the lowest MAE. The baseline linear model performed best for three response variables: Directional Asymmetry of digit 2, Directional Asymmetry of digit 3, and Directional Asymmetry of digit 4, only considering the lowest MAE, because the MAPE was undefined. The baseline linear model performed best for three response variables: Fluctuating Asymmetry of digit 2, Index.Diversity, and Index.Diversity.Holt, considering both the lowest MAE and MAPE.

Because the MAE and not RMSE or MAPE can be compared across models with different response variables given most other conditions were consistent, the baseline linear models where the MAE was used exhibited the best performance for Index.Diversity.Holt with explanatory variables: Ethnic Group, Sex, DA.kmeans.2, DA.kmeans.4, DA.kmeans.10, and DA.kmeans.322. Because the baseline linear model clearly outperformed the other baselines and the GLM, the interpretation of the linear models in context of their assumptions, like the increase in prediction error due to heteroscedasticity, becomes more important.

4.2.5 Generalized linear modeling (GLM)

For the LASSO regularization of the GLM of Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit, the final training set contained 32585 observations, and for evaluating the final model on a separate test set to calculate performance metrics, the final test set contained 13964 observations. For GLM of Directional Asymmetry variables for each digit, the 10 explanatory variables included were Ethnic Group, Region, Landmass, Sex, Latitude, Longitude, DA.kmeans.2, DA.kmeans.4, DA.kmeans.10, and

DA.kmeans.322. For GLM of Fluctuating Asymmetry variables for each digit, the 10 explanatory variables included were Ethnic Group, Region, Landmass, Sex, Latitude, Longitude, FA.kmeans.2, FA.kmeans.4, FA.kmeans.8, and FA.kmeans.322. For LASSO regularization of the GLM of Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt, the final training set contained 32585 observations of the same 10 explanatory variables as for Directional Asymmetry variables for each digit, and the final test set contained 13964 observations of the same 10 variables.

Calculating a goodness-of-fit for each model serves to indicate how well the explanatory variables in the model explain the observed variation in the response variable. However, comparing goodness-of-fit values for multiple models with different response variables is not appropriate. The goodness-of-fit for the GLM was the Kullback-Leibler R^2 value, for which higher values indicate a better model fit, and the full model was compared to the reduced model containing the same response variable and model formulation just with different explanatory variables to indicate how the feature selection process improved the fit of the model (Schneider, Hommel, and Blettner 2010). The desired higher Kullback-Leibler R^2 value from $1 - (model\ deviance \div null\ model\ deviance)$ reflects a higher null deviance in the denominator, and because a smaller deviance indicates a better model fit a higher null deviance suggests the null model fits the data poorly compared to the fitted or reduced model represented in the numerator.

All GLM of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response variables exhibited a reduction in the Kullback-Leibler R^2 value of the full model compared to the feature-selected reduced model, suggesting the LASSO regularization did not improve the model fit enough to

offset the benefit from just reducing the explanatory variables in the model (Table 27, Table 28, Table 29). Therefore, a greater reduction in the Kullback-Leibler R^2 value between the full model and the reduced model indicated a greater decrease in model fit, and a smaller reduction in the Kullback-Leibler R^2 value indicated a smaller decrease in model fit or a maintenance of goodness-of-fit. The GLM with Directional Asymmetry of digit 4, with feature selected explanatory variables: Region, Sex, DA.kmeans.2, DA.kmeans.4, and DA.kmeans.10, exhibited the smallest reduction in Kullback-Leibler R^2 value (Table 27). The next smallest reductions in Kullback-Leibler R^2 value were for Directional Asymmetry of digit 1, Directional Asymmetry of digit 2, and Directional Asymmetry of digit 5. The GLM with Directional Asymmetry of digit 3 exhibited the greatest reduction in Kullback-Leibler R^2 value, so a greater decrease in model fit after LASSO regularization. The reductions in Kullback-Leibler R^2 value ranged from 0.0007 for Directional Asymmetry of digit 4 to 0.0021 for Directional Asymmetry of digit 3. Of note, the reduced models for the GLM of Directional Asymmetry variables for each digit as response variables indicated a small reduction in the goodness-of-fit.

To interpret the results of regression models, often the model architecture assumes equal within-group variance, and the `betadisper` function from the `vegan` package can approximately test this, because the beta dispersion represents the absolute value distance between groups from a multivariate centroid, and the `anova` function tests for significance in dispersion across groups (Oksanen et al. 2022). For the majority of response variables, significant differences in dispersion across the groups were present.

Fewer response variables exhibited no significant differences and thus no dispersion across groups: Directional Asymmetry of digit 1 for Sex, Directional Asymmetry of digit 2 for Sex, Fluctuating Asymmetry of digit 1 for Ethnic Group, Fluctuating Asymmetry of digit 1 for

Table 27. Goodness-of-fit for full and reduced GLM of Directional Asymmetry variables for each digit. Via Kullback-Leibler R^2 value; DA = directional asymmetry; GLM = Generalized linear model

Analytical Method	Response Variable	Model	Kullback-Leibler R^2 value
GLM	DA1	full	0.2690
		reduced	0.2682
GLM	DA2	full	0.3816
		reduced	0.3805
GLM	DA3	full	0.2179
		reduced	0.2157
GLM	DA4	full	0.2335
		reduced	0.2328
GLM	DA5	full	0.0267
		reduced	0.0253

Table 28. Goodness-of-fit for full and reduced GLM of Fluctuating Asymmetry variables for each digit. Via Kullback-Leibler R^2 value; FA = fluctuating asymmetry; GLM = Generalized linear model

Analytical Method	Response Variable	Model	Kullback-Leibler R^2 value
GLM	FA1	full	0.3039
		reduced	0.3028
GLM	FA2	full	0.5263
		reduced	0.5255
GLM	FA3	full	0.1134
		reduced	0.1129
GLM	FA4	full	0.0639
		reduced	0.0623
GLM	FA5	full	0.0959
		reduced	0.0947

Table 29. Goodness-of-fit for full and reduced GLM of Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt. Via Kullback-Leibler R^2 value; GLM = Generalized linear model

Analytical Method	Response Variable	Model	Kullback-Leibler R^2 value
GLM	Index.Asymmetry	full	0.1415
		reduced	0.1413
GLM	Index.Diversity	full	0.0345
		reduced	0.0329
GLM	Index.Diversity.Holt	full	0.0560
		reduced	0.0558

Region, Fluctuating Asymmetry of digit 1 for Sex, Fluctuating Asymmetry of digit 1 for Landmass, Fluctuating Asymmetry of digit 2 for Sex, Fluctuating Asymmetry of digit 3 for Sex, Index.Asymmetry for Sex, Index.Diversity for Sex, and Index.Diversity.Holt for Sex.

Demonstrated another way, the grouping Sex exhibited no significant dispersion for Directional Asymmetry of digit 1, Directional Asymmetry of digit 2, Fluctuating Asymmetry of digit 1, Fluctuating Asymmetry of digit 2, Fluctuating Asymmetry of digit 3, Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt. The grouping Ethnic Group, Region, and Landmass exhibited no significant dispersion for Fluctuating Asymmetry of digit 1 response variable.

Using LASSO regularization for feature selection of the GLMs by fitting a Gaussian regression model for Directional Asymmetry variables for each digit and the indices of asymmetry and diversity and a negative binomial regression model for Fluctuating Asymmetry variables for each digit and using k-fold cross-validation to test multiple lambda values, the best lambda values were calculated. The final model included the best lambda value and only the explanatory variables with non-zero coefficients, and the magnitude and sign of the non-zero coefficients indicated the importance and direction of influence of the explanatory variable. Figure 85 illustrates the explanatory variables with the greatest magnitude of importance for the GLM of Directional Asymmetry variables for each digit as indicated by high negative or high positive coefficients. For all response variables, the intercept expressed the highest importance with four negative and one positive coefficient values, and explanatory variable DA.kmeans.2 expressed the second highest importance but solely with positive coefficient values. Many GLMs excluded Sex from their final models for the GLM of Directional Asymmetry variables for each digit, so the only non-zero coefficient was positive for Directional Asymmetry of digit 4 (Table 26).

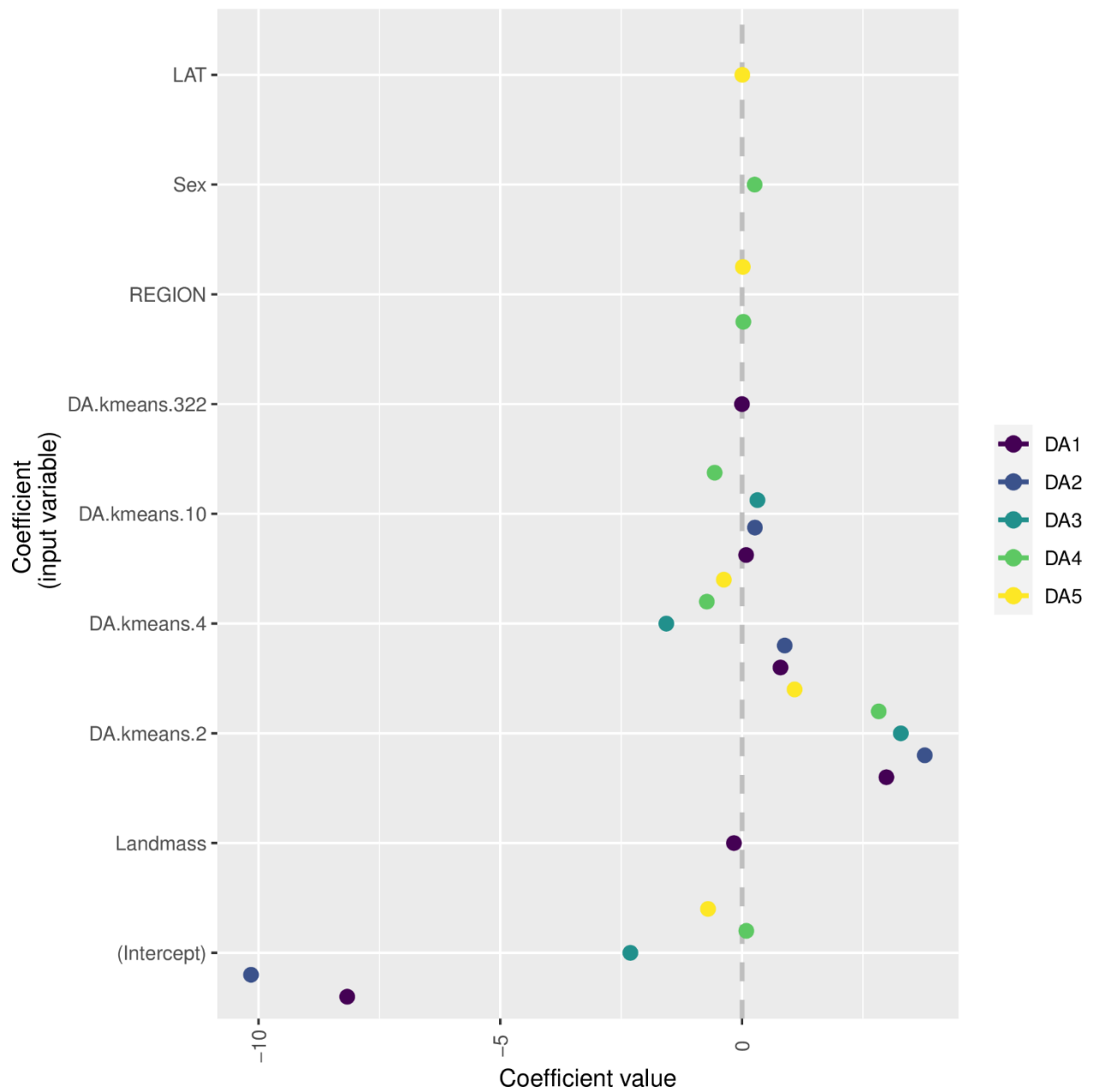


Figure 85. GLM coefficient value for each explanatory variable, colored by Directional Asymmetry response variable. (N = 13964)

Because the intercept and DA.kmeans.2 variable distended the coefficient value range, Figure 86 illustrates the same final model coefficients for the GLM of Directional Asymmetry variables for each digit but with a more narrow range from -2 to 2. Within this more defined range, explanatory variables DA.kmeans.4 and DA.kmeans.10 expressed the next highest importance with both negative and positive coefficient values. Many GLMs excluded Landmass from their final models for the GLM of Directional Asymmetry variables for each digit, so the only non-zero coefficient was negative for Directional Asymmetry of digit 1 (Table 26).

Because the remaining explanatory variables still appeared constricted around zero due to the range of the other variables, Figure 87 illustrates the same final model coefficients for the GLM of Directional Asymmetry variables for each digit but with a more narrow range from -0.1 to 0.1. Within this most defined range, explanatory variable Region exhibited positive coefficients for Directional Asymmetry of digit 4 and Directional Asymmetry of digit 5. The single non-zero coefficient of Latitude was positive, and the single non-zero coefficient of DA.kmeans.322 was slightly negative. Of note, no final models for the GLM of Directional Asymmetry variables for each digit included Ethnic Group or Longitude with a non-zero coefficient (Table 26).

The best performance metric for the Directional Asymmetry variables for each digit response variables was the MAE due to the presence of outliers that could unduly influence the modeling, the zeros in the response variables that did not always results in an undefined MAPE, and the arbitrary zeros since directional asymmetry represents the signed difference between the left and right sides (Table 5). No GLMs performed the best for any of the Directional Asymmetry variables for each digit response variables (Table 26). The baseline median regressor exhibited the lowest MAE for Directional Asymmetry of digit 5 response

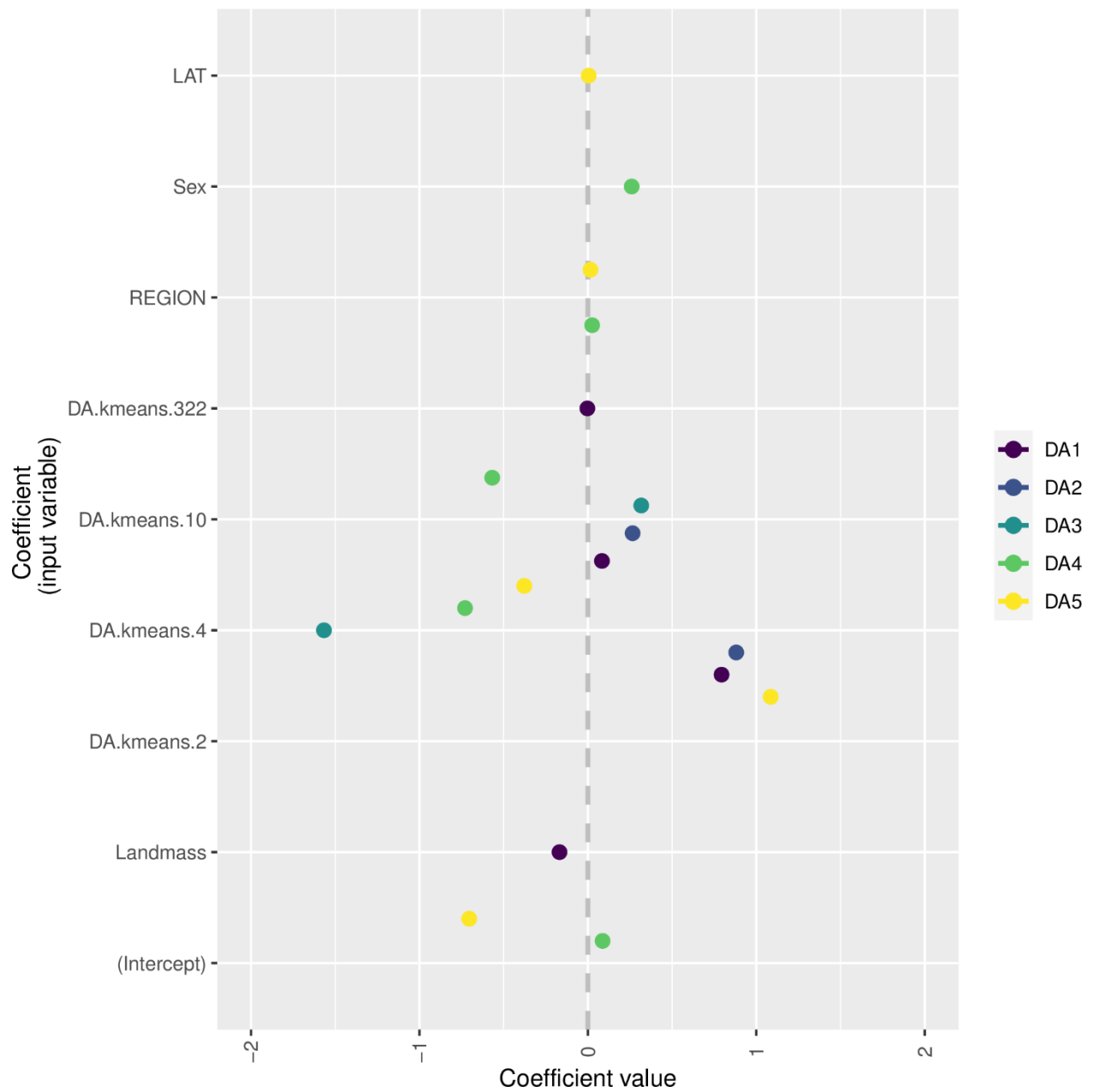


Figure 86. GLM coefficient value for each explanatory variable, colored by Directional Asymmetry response variable, zoomed to range -2 to 2. (N = 13964)

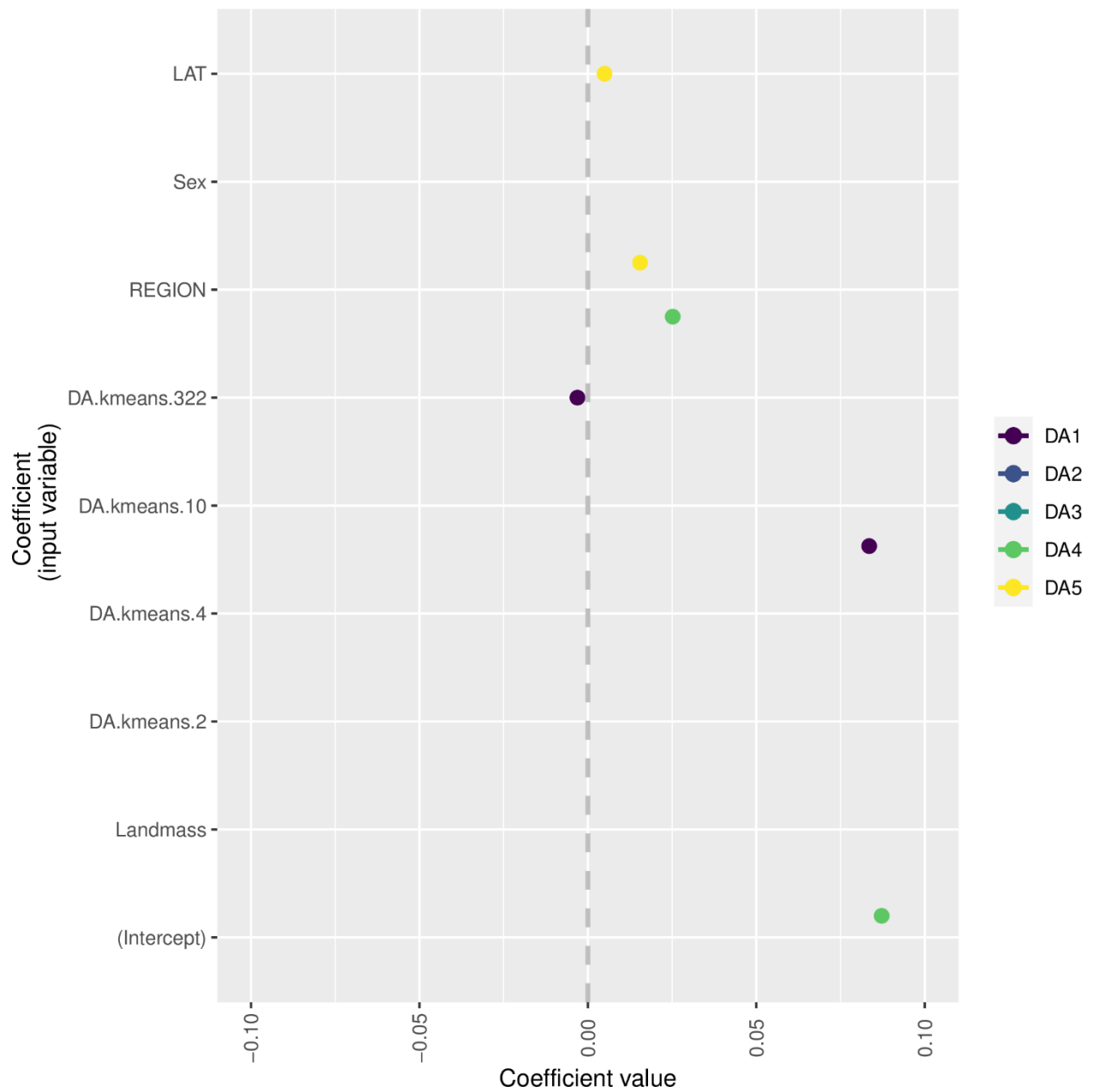


Figure 87. GLM coefficient value for each explanatory variable, colored by Directional Asymmetry response variable, zoomed to range -0.1 to 0.1. (N = 13964)

variable, while the baseline linear model exhibited the lowest MAE for the remaining Directional Asymmetry variables of the other digits. Across the baseline linear models for Directional Asymmetry variables for each digit, the lowest MAE was 2.408 for Directional Asymmetry of digit 4 response variable with fitted explanatory variables: Region, Sex, Latitude, DA.kmeans.2, DA.kmeans.4, and DA.kmeans.10. Although the GLMs did not perform the best, just considering Directional Asymmetry variables for each digit, the best performing final GLM with a Gaussian family with the lowest MAE was 2.326 for Directional Asymmetry of digit 5 response variable with fitted explanatory variables: Region, Latitude, DA.kmeans.2, and DA.kmeans.4. The GLM for Directional Asymmetry variables for digits 1 through 4 performed better than the baseline mean or medium regressors but not better than the baseline linear models.

The GLM with Fluctuating Asymmetry of digit 3, with feature selected explanatory variables: Landmass, Latitude, FA.kmeans.2, FA.kmeans.4, FA.kmeans.8, and FA.kmeans.322, exhibited the smallest reduction in Kullback-Leibler R^2 value between the full model and the reduced or fitted model (Table 28). The next smallest reductions in Kullback-Leibler R^2 value were for Fluctuating Asymmetry of digit 2, Fluctuating Asymmetry of digit 5, and Fluctuating Asymmetry of digit 1. The GLM with Fluctuating Asymmetry of digit 4 exhibited the greatest reduction in Kullback-Leibler R^2 value, so a greater decrease in model fit after LASSO regularization. The reductions in Kullback-Leibler R^2 value ranged from 0.0006 for Fluctuating Asymmetry of digit 3 to 0.0016 for Fluctuating Asymmetry of digit 4. Of note, the reduced models for the GLM of Fluctuating Asymmetry variables for each digit as response variables indicated a small reduction in the goodness-of-fit.

Figure 88 illustrates the explanatory variables with the greatest magnitude of importance for the GLM of Fluctuating asymmetry variables for each digit as indicated by high negative or high positive coefficients. For all response variables, the intercept expressed the highest importance with all positive coefficient values, and explanatory variable FA.kmeans.2 expressed the second highest importance but with mostly negative coefficient values and a positive coefficient value for Fluctuating Asymmetry of digit 1. Explanatory variable FA.kmeans.8 exhibited both negative and positive coefficient values.

Because the intercept and FA.kmeans.2 variable distended the coefficient value range, Figure 89 illustrates the same final model coefficients for the GLM of Fluctuating Asymmetry variables for each digit but with a more narrow range from -0.1 to 0.1. Within this more defined range, explanatory variable FA.kmeans.4 expressed the next highest importance with all negative coefficient values. Explanatory variable Landmass also expressed importance within this -0.1 to 0.1 range with positive coefficients for Fluctuating Asymmetry of digit 3, Fluctuating Asymmetry of digit 4, and Fluctuating Asymmetry of digit 5 (Figure 89).

Because the remaining explanatory variables still appeared constricted around zero due to the range of the other variables, Figure 90 illustrates the same final model coefficients for the GLM of Fluctuating Asymmetry variables for each digit but with a more narrow range from -0.002 to 0.002. Within this most defined range, explanatory variable FA.kmeans.322 exhibited all negative coefficients. The single non-zero coefficient of Latitude was positive for Fluctuating Asymmetry of digit 3. Of note, no final models for the GLM of Fluctuating Asymmetry variables for each digit included Ethnic Group, Region, Sex, and Longitude with a non-zero coefficient (Table 26).

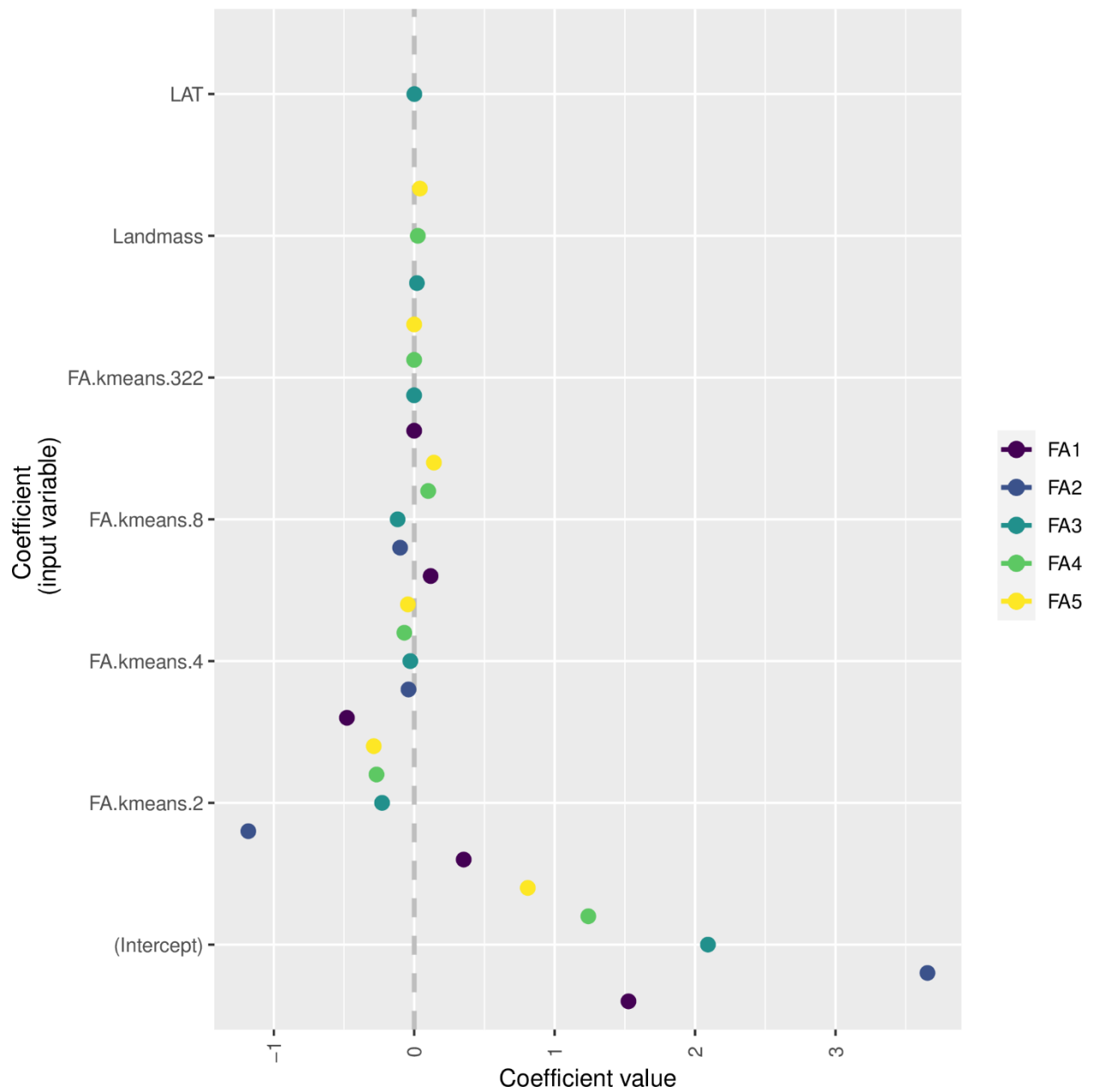


Figure 88. GLM coefficient value for each explanatory variable, colored by Fluctuating Asymmetry response variable. (N = 13964)

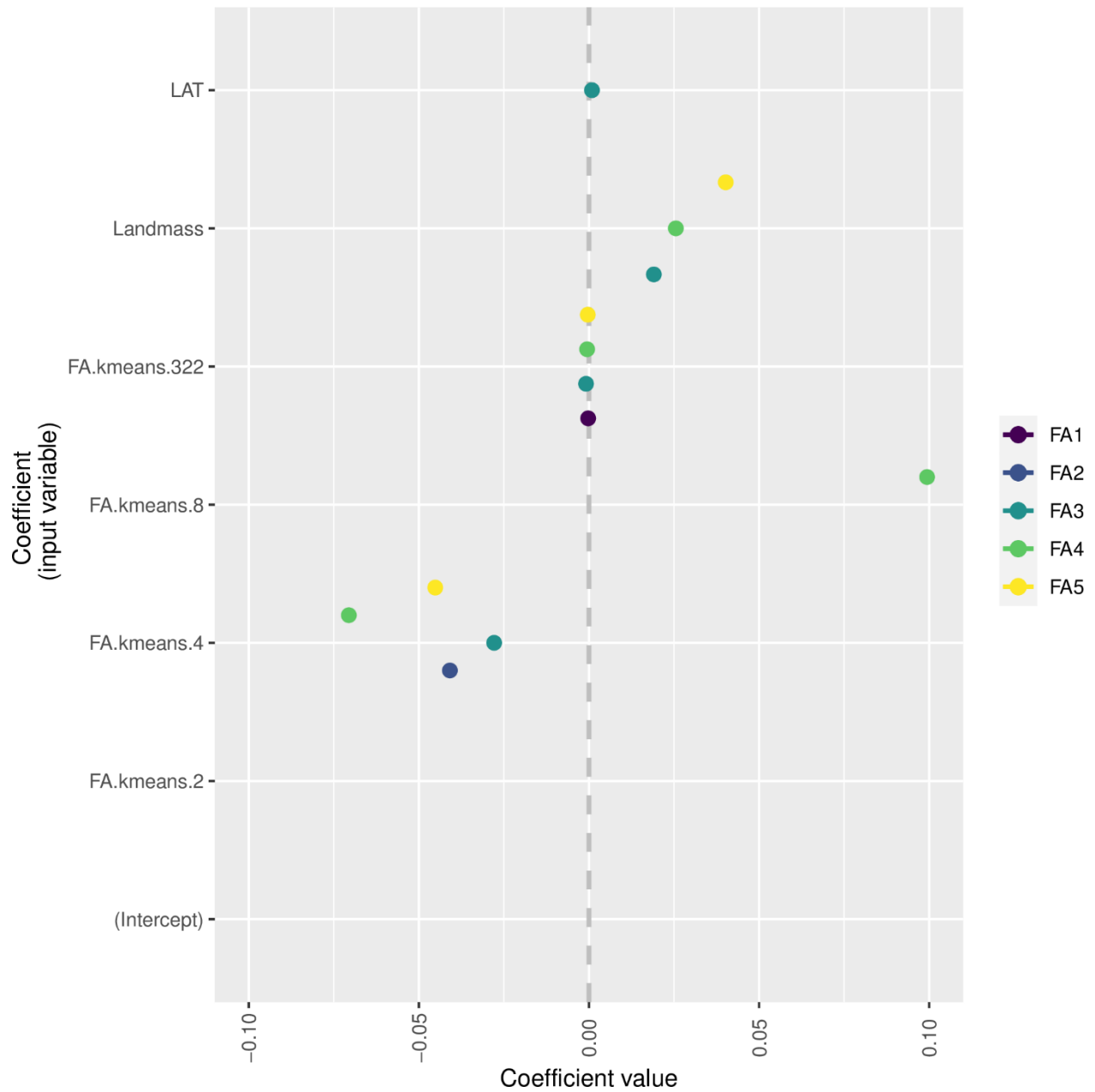


Figure 89. GLM coefficient value for each explanatory variable, colored by Fluctuating Asymmetry response variable, zoomed to range -0.1 to 0.1. (N = 13964)

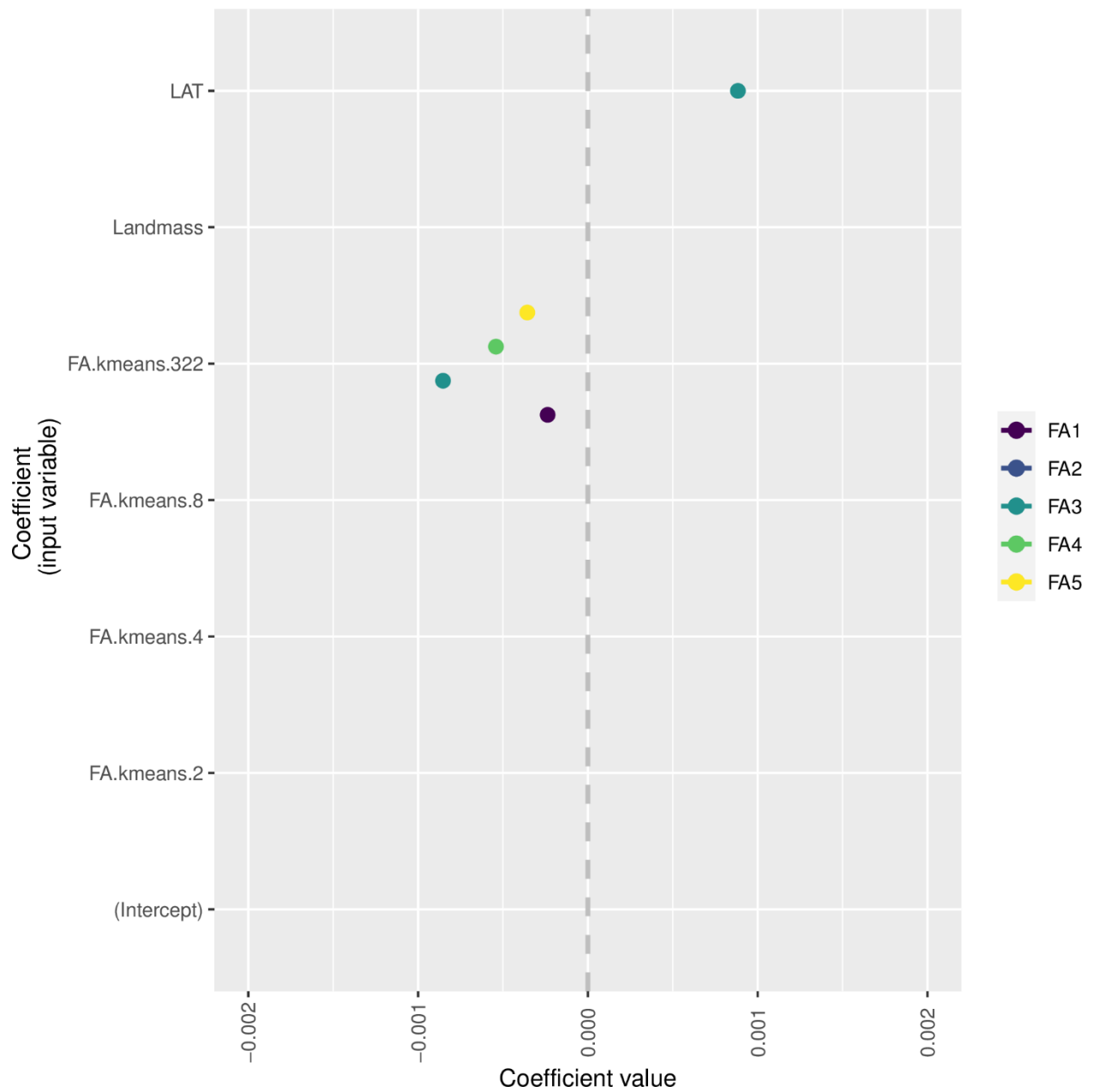


Figure 90. GLM coefficient value for each explanatory variable, colored by Fluctuating Asymmetry response variable, zoomed to range -0.002 to 0.002. (N = 13964)

The best performance metric for the Fluctuating Asymmetry variables for each digit and was either the MAE or MAPE due to the presence of outliers and the zeros (Table 5). When the MAPE was undefined (NaN), the second best performance metric for these response variables, the MAE, was considered. No GLMs performed the best for any of the Fluctuating Asymmetry variables for each digit response variables (Table 26). Fluctuating Asymmetry of digit 1 response variable exhibited the lowest MAE for the baseline linear model and the lowest MAPE for the mean regressor. Fluctuating Asymmetry of digit 2 response variable exhibited the lowest MAE and MAPE for the baseline linear model. Fluctuating Asymmetry of digits 3, 4, and 5 response variables exhibited the lowest MAE for the median regressor and the lowest MAPE for the baseline linear model. Across the baseline linear models for Fluctuating Asymmetry variables for each digit, the lowest MAE was 1.44743 and the lowest MAPE was 0.627 for Fluctuating Asymmetry of digit 2 response variable with fitted explanatory variables: Landmass, Sex, Latitude, Longitude, FA.kmeans.2, FA.kmeans.4, FA.kmeans.8, and FA.kmeans.322.

Although the GLMs did not perform the best, just considering Fluctuating Asymmetry variables for each digit, the best performing final GLM with a negative binomial family with the lowest MAE was 1.769 for Fluctuating Asymmetry of digit 5 response variable with fitted explanatory variables: Landmass, FA.kmeans.2, FA.kmeans.4, FA.kmeans.8, and FA.kmeans.322. The best performing final GLM with a negative binomial family with the lowest MAPE was 2.061 for Fluctuating Asymmetry of digit 4 response variable with the same fitted explanatory variables. No GLMs for Fluctuating Asymmetry variables for each digit performed better than the baseline models.

The GLM with Index.Diversity.Holt, with feature selected explanatory variables: Region, Landmass, Sex, Latitude, Longitude, DA.kmeans.2, DA.kmeans.4, DA.kmeans.10, and

DA.kmeans.322, exhibited the smallest reduction in Kullback-Leibler R^2 value between the full model and the reduced or fitted model (Table 29). The next smallest reduction in Kullback-Leibler R^2 value was for Index.Asymmetry. The GLM with Index.Diversity exhibited the greatest reduction in Kullback-Leibler R^2 value, so a greater decrease in model fit after LASSO regularization. The reductions in Kullback-Leibler R^2 value ranged from 0.0002 for Index.Diversity.Holt to 0.0016 for Index.Diversity. Of note, the reduced models for the GLM of Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt as response variables indicated a small reduction in the goodness-of-fit.

Figure 91 illustrates the explanatory variables with the greatest magnitude of importance for the GLM of Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt as indicated by high negative or high positive coefficients. For all response variables, the intercept expressed the highest importance with all positive coefficient values, and explanatory variable DA.kmeans.2 expressed the second highest importance but with solely negative coefficient values for Index.Asymmetry and Index.Diversity.Holt.

Because the intercept distended the coefficient value range, Figure 92 illustrates the same final model coefficients for the GLM of fluctuating asymmetry variables for each digit but with a more narrow range from -0.4 to 0.4. Within this more defined range, explanatory variable Landmass expressed the next highest importance with all positive coefficient values. DA.kmeans.4 exhibited negative coefficient values for Index.Diversity and Index.Diversity.Holt. Explanatory variable DA.kmeans.10 expressed importance within this -0.4 to 0.4 range with positive coefficients for Index.Asymmetry and Index.Diversity.Holt, and Sex exhibited positive coefficients for Index.Diversity and Index.Diversity.Holt (Figure 92).

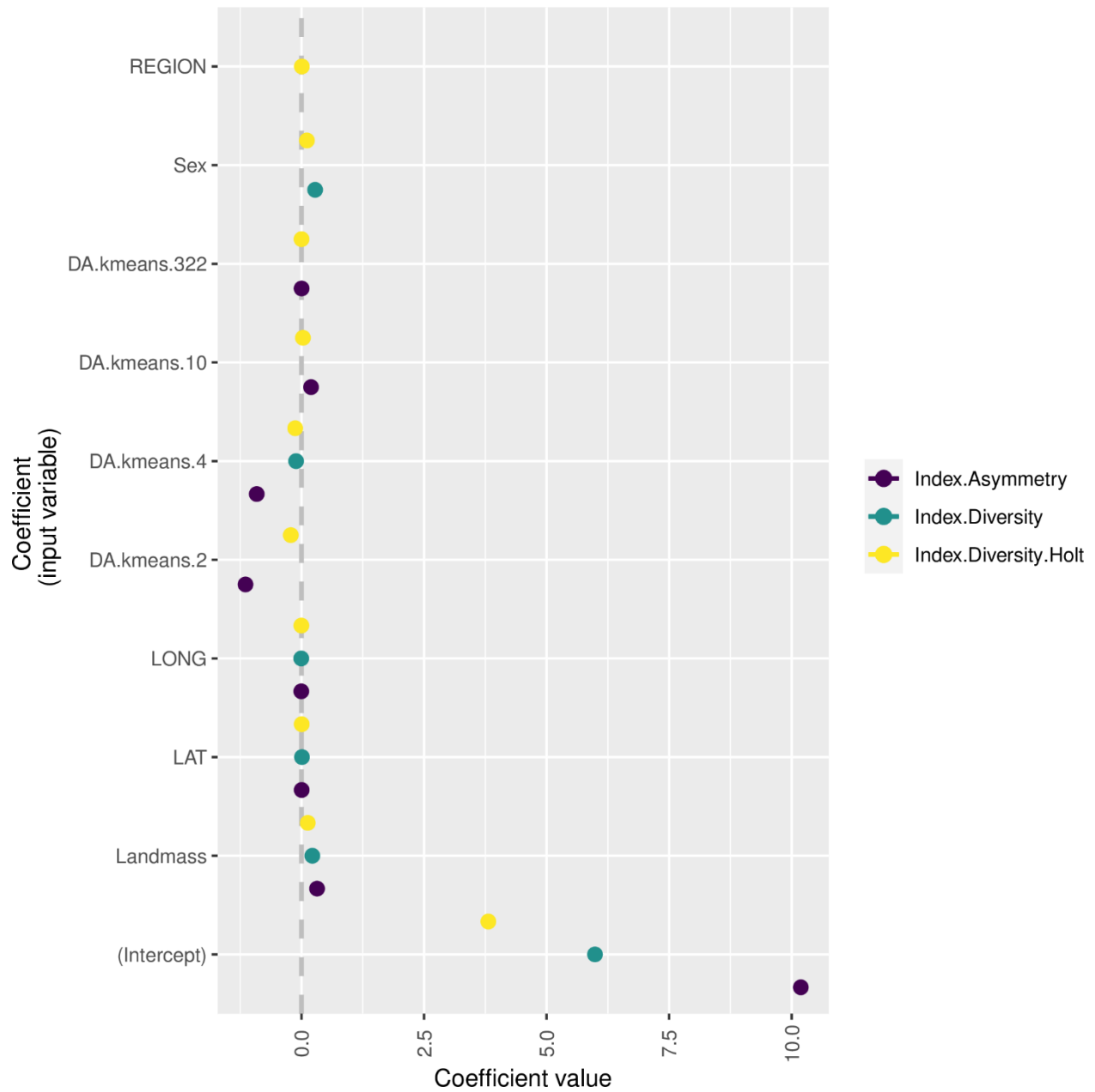


Figure 91. GLM coefficient value for each explanatory variable, colored by Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt response variable. (N = 13964)

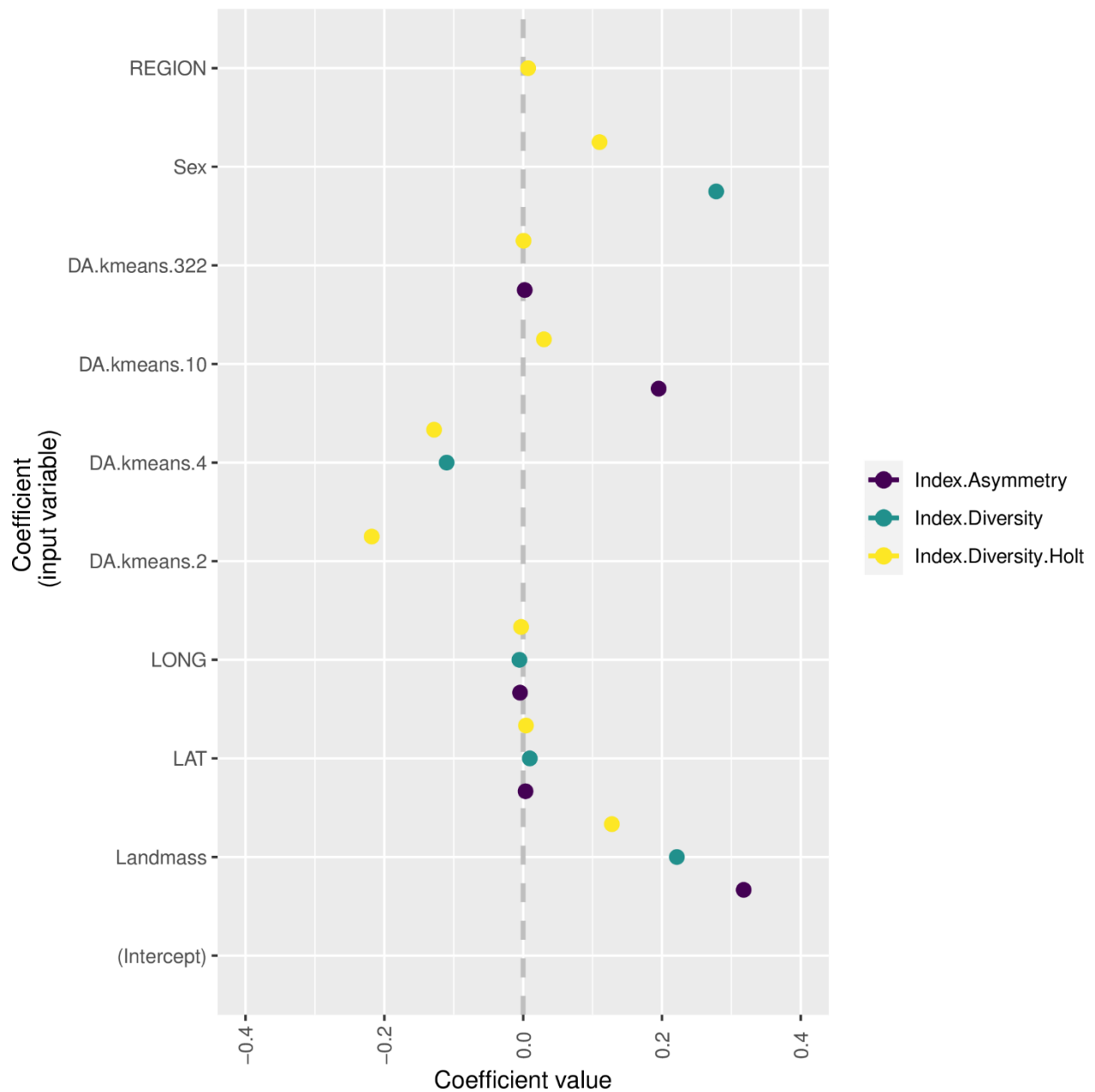


Figure 92. GLM coefficient value for each explanatory variable, colored by Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt response variable, zoomed to range -0.4 to 0.4. (N = 13964)

Because the remaining explanatory variables still appeared constricted around zero due to the range of the other variables, Figure 93 illustrates the same final model coefficients for the GLM of Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt but with a more narrow range from -0.04 to 0.04. Within this most defined range, explanatory variable Latitude exhibited all positive coefficients, and Longitude exhibited all negative coefficients. The single non-zero coefficient of Region was positive for Index.Diversity.Holt. Even with the more narrow range explanatory variable DA.frc.kmeans.322 appeared barely away from zero with both positive coefficients for Index.Asymmetry and Index.Diversity.Holt. Of note, no final models for the GLM of Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt included Ethnic Group with a non-zero coefficient (Table 26).

The best performance metric for the Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt was either the MAE or MAPE due to the presence of outliers and the zeros (Table 5). The best performing final GLM with a Gaussian family with both the lowest MAE (2.756) and the lowest MAPE (0.341) was Index.Asymmetry response variable with fitted explanatory variables: Landmass, Latitude, Longitude, DA.kmeans.2, DA.kmeans.4, DA.kmeans.10, and DA.kmeans.322 (Table 26). Both the Index.Diversity and Index.Diversity.Holt response variables exhibited the lowest MAE and MAPE values for the baseline linear model. Across the baseline linear models, the lowest MAE was 1.218 and the lowest MAPE was 0.306 for Index.Diversity.Holt response variable with fitted explanatory variables: Ethnic Group, Sex, DA.kmeans.2, DA.kmeans.4, DA.kmeans.10, and DA.kmeans.322. For GLM with Index.Diversity and Index.Diversity.Holt as the response variables, the baseline linear model performed better, but the GLM still performed better than the mean and median regressors. Otherwise, GLM Index.Asymmetry performed better than the three baseline models.

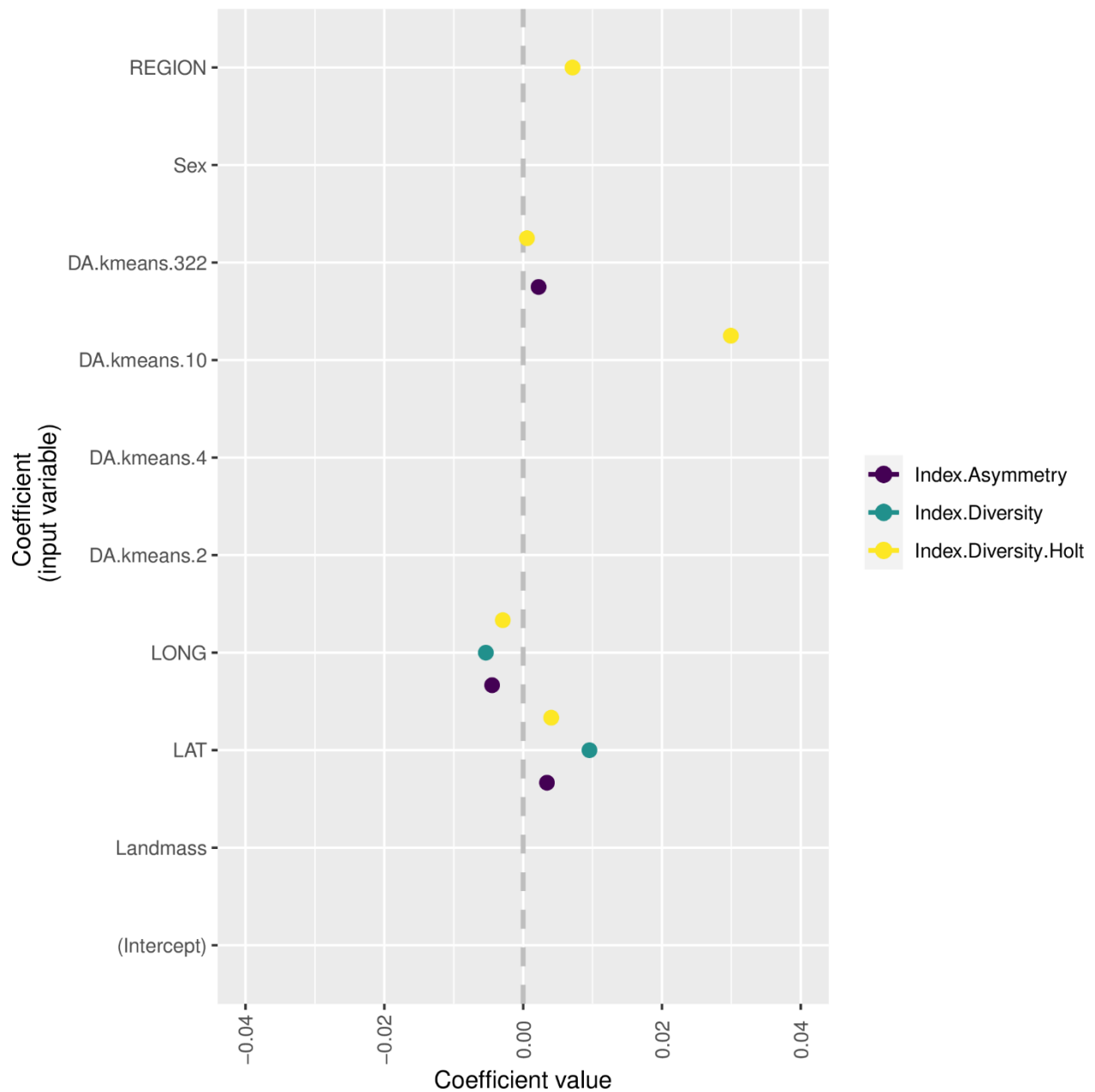


Figure 93. GLM coefficient value for each explanatory variable, colored by Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt response variable, zoomed to range -0.04 to 0.04. (N = 13964)

Comparing the general makeup of the baseline linear models to the GLM, most models included many k-means cluster assignment variables suggesting those variables were not redundant and did indicate additional useful variation in the response dermatoglyphic features than the other explanatory variables (Table 26). Overall, the baseline linear model performed best for the majority of response variables, the mean or median regressor performed best for Directional Asymmetry of digit 5, Fluctuating Asymmetry of digit 1 (MAPE only), and for Fluctuating Asymmetry of digits 3, 4, and 5 (MAE only). The GLM performed best for only the Index.Asymmetry response variable.

4.2.6 Baseline models for support vector regression

For the SVR of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response variables the 50% training set was used and contained 6 attributes: Sex, Ethnic Group, Region, Latitude, Longitude, and Landmass; for grid searching hyperparameter optimization the 25% validation set was used with same 6 attributes, and for evaluating the final model a separate 25% test set containing the same 6 attributes was used to calculate performance metrics.

To evaluate the assumptions of the mean regressor as an approximation of the central tendency of the response variables, the similarity to the median regressor was visualized, and for Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response variables the mean of each appeared similar to the median of each (Figure 10, Figure 11, Figure 12). To evaluate whether the baseline linear model could be interpreted in terms of goodness-of-fit or performance, initial model checking was performed to see whether the assumptions of linear models were met or could be approximately met.

Six baseline linear models for the SVR exhibited high multicollinearity (high > 10 , moderate 5-10, low < 5) in the explanatory variables as computed via the variance inflation factor, which assesses how much the linear correlation of an explanatory variable with other explanatory variables increases the variance of the coefficient estimate compared to theoretically no correlation (Table 30). The linear model with Directional Asymmetry of digit 1 as the response variable exhibited high collinearity (> 30) for explanatory variable Region and moderate collinearity for explanatory variables Latitude and Longitude (Figure S102). The linear model with Directional Asymmetry of digit 4 as the response variable exhibited moderate collinearity for explanatory variables Latitude and Region (Figure S108). For both of these linear models, the explanatory variable Sex was the only other feature selected variable, which exhibited low collinearity. The linear model with Directional Asymmetry of digit 5 as the response variable exhibited moderate collinearity for the only two feature selected explanatory variables in the model: Latitude and Region (Figure S110). The linear model with Fluctuating Asymmetry of digit 2 as the response variable exhibited moderate collinearity for explanatory variable Landmass (Figure S114). The linear model with Fluctuating Asymmetry of digit 5 as the response variable exhibited high collinearity (> 30) for explanatory variable Region and moderate collinearity for explanatory variables Latitude and Longitude (Figure S120). The linear model with Index.Asymmetry as the response variable exhibited moderate collinearity for explanatory variable Landmass (Figure S122). The consequences of continuing to use a model known to exhibit high collinearity between explanatory variables are primarily relegated to interpretation of coefficients and resulting p values, which were not the focus of these models in this dissertation. Further, when collinearity is present, the predictions, precision of the predictions, and the goodness-of-fit statistics do not change (Neter 1996:289). Solutions to

Table 30. Violation of standard regression assumptions of residuals for baseline linear models for SVR of Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry of diversity; DA = directional asymmetry; FA = fluctuating asymmetry; SVR = support vector regression; p value < 0.05 in bold

SVR	Multicollinearity	Anderson-Darling Normality		Durbin-Watson Autocorrelation		Breusch-Pagan Heteroscedasticity
		A test statistic	p value < 0.001	d test statistic	p value	p value < 0.001
Response variable	Variance Inflation Factor > 10					
DA1	✓	82.232	✓	1.986	0.228	✓
DA2		231.43	✓	2.014	0.306	✓
DA3		111.42	✓	2.011	0.484	✓
DA4		128.2	✓	2.004	0.914	✓
DA5		76.641	✓	1.983	0.148	✓
FA1		633.44	✓	1.987	0.300	✓
FA2		866.17	✓	1.980	0.104	✓
FA3		735.43	✓	2.006	0.664	✓
FA4		827.24	✓	1.998	0.786	✓
FA5	✓	675.21	✓	1.991	0.424	✓
Index.Asymmetry		312.66	✓	1.988	0.368	✓
Index.Diversity		192.61	✓	2.000	0.042	✓
Index.Diversity.Holt		184.08	✓	2.002	0.072	✓

collinearity between explanatory variables are to remove the highly correlated explanatory variables, to perform LASSO regression that can handle collinearity, etc. (Frost 2017).

For linear models with response variables Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity, the Anderson-Darling normality test p value was < 0.05 , so reject the null hypotheses the errors were from a normal distribution (Table 30). For a visual assessment of normality, the normality of residuals plots exhibit an almost perfect one-to-one line toward the center comparing the standard normal distribution quantiles on the x-axis as the residuals from a perfect model to the sample quantiles on the y-axis as the real residuals from our models, but the relationship breaks down as seen from the dots falling away from the line particularly when the standard normal distribution quantiles on the x-axis dip below -2.5 and exceed 2.5 at the tails, and for the histograms of residuals the lower and upper tails were unequal (Figure S102 through Figure S127). The normality of residual plots for Fluctuating Asymmetry variables digits 1 through 5 exhibit further deviation of dots away from the line when the standard normal distribution quantiles on the x-axis dip below -1.25 and exceed 1.25 at the tails (Figure S112, Figure S114, Figure S116, Figure S118, Figure S120). Thus, linear models with response variables Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity exhibited non-normal distribution of errors, which could be attributed to autocorrelation. Other reasons for non-normal distribution of errors could be an important explanatory variable not included in the model or some polynomial or interaction term not included in the model. The consequences of continuing to use a model known to violate the normality of residuals assumption are primarily biased estimate of standard errors affecting interpretation of confidence intervals and the p values, which were not the focus of these models

in this dissertation, and with large sample sizes where the number of observations per variable exceeds 10 the impact of the negative consequences diminishes (Schmidt and Finan 2018).

One potential source of the non-normally distributed errors from linear models with response variables Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity was evaluated with the Durbin-Watson test for autocorrelation of residuals, indicated by a p value < 0.05 (Table 30).

Autocorrelation of residuals was detected by the Durbin-Watson test only for response variable Index.Diversity. However, in practical use any Durbin-Watson test statistic value equal to 2 indicated no autocorrelation and between 1.5 and 2.5 could be approximated as no autocorrelation (Bedre 2019; SE n.d.). While non-normality of residuals does not pose a serious threat to the interpretation of the baseline linear models, the consequences of continuing to use a model known to exhibit autocorrelation of residuals, even with large sample sizes, are great. Autocorrelation of residuals reduces the interpretability of the estimates of the coefficients but also their standard errors. As a result, the p values would be biased, and the standard error of the regression coefficients may underestimate the true standard deviation of the estimated regression coefficients, so the mean squared error performance metric may greatly underestimate the true variance of the errors. Similarly, the RMSE would be unreliable (Broersen 2005). Given no Durbin-Watson test statistics exceeded the bounds of the practical use of the Durbin-Watson test, neither of the two baseline linear models with response variables Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity were treated as autocorrelated (Bedre 2019; SE n.d.).

Other violations of linear regression assumptions that can seriously bias the interpretation of the model even with a large sample size are the presence of extreme observations that unduly

influence the regression and also the heteroscedasticity that reduces the interpretation of the coefficients. For all linear models with response variables Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity, the results of the heteroscedasticity check exhibited a p value < 0.001 , so reject the null hypothesis the error variance was constant. Further, only the homogeneity of variance plots for Directional Asymmetry of digit two, Index.Diversity, and Index.Diversity.Holt response variables did not show a flat and horizontal reference line comparing the fitted values on the x-axis to the square root of the standardized residuals on the y-axis (Figure S104, Figure S124, Figure S126). The consequences of continuing to use a model known to exhibit heteroscedasticity rely on identifying the purity of the heteroscedasticity and so do the solutions. For impure heteroscedasticity where the model was not correctly specified by leaving an important variable out, and the generation of the coefficient values was not biased, so opportunities exist to solve the problem with additional explanatory variables. For pure heteroscedasticity where the model is correctly specified and yet non-constant variance in the residuals persists, the issues with interpretation are the same as with autocorrelation, and the prediction error significantly changes over the range of prediction of the model, so the performance metrics, which rely on the prediction error likely represent an underestimate of the performance error, or a minimum. For the baseline linear models exhibiting significant heteroscedasticity, the performance metrics were interpreted as a minimum value, meaning if the comparison in performance between the baseline linear model that exhibits heteroscedasticity and the mean regressor were close in value, the mean regressor would likely be identified as the better performing baseline. Solutions to pure heteroscedasticity are the use of weighted least squares, obtaining standard errors corrected for heteroscedasticity, or retooling the model.

Calculating a goodness-of-fit for each model indicates how well the explanatory variables in the model explain the observed variation in the response variable. However, comparing goodness-of-fit values for multiple models with different response variables is not appropriate. So, only the AIC of the full model was compared to the reduced model containing the same response variable and model formulation to indicate how the feature selection process improved the fit of the model (Schneider, Hommel, and Blettner 2010). For the mean regressor, median regressor, and linear regression model as baseline comparisons for the SVR, only the linear model incorporated explanatory variables in the model construction, and so the goodness-of-fit was only assessed for the linear model as a baseline via the AIC value, for which lower values indicate a better model fit (Table 31).

The linear model with Fluctuating Asymmetry of digit 2 as the response variable, with feature selected explanatory variables: Sex, Latitude, Longitude, and Landmass, exhibited the greatest reduction in AIC value (Table 31). The next greatest reduction in AIC value was for Directional Asymmetry of digit 2, and then Directional and Fluctuating Asymmetry of digit 3 exhibited the same and next greatest reduction in AIC value. The next greatest reductions in AIC values were for Fluctuating Asymmetry of digit 5, Fluctuating Asymmetry of digit 4, and Fluctuating Asymmetry of digit 1. Further, the response variables Directional Asymmetry of digit 4, Directional Asymmetry of digit 5, Index.Asymmetry, and Directional Asymmetry of digit 1 exhibited a reduction less than 200 AIC value and greater than 100 AIC value. Only the remaining response variables Index.Diversity and Index.Diversity.Holt exhibited a reduction under 100 AIC value. The majority of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response

Table 31. Goodness-of-fit for full and reduced baseline linear models of Directional Asymmetry and Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity, of same SVR data set; AIC = Akaike information criterion; DA = directional asymmetry; FA = fluctuating asymmetry; LM = linear model; SVR = support vector regression

Analytical Method	Response Variable	Explanatory Variables	Model	AIC
LM	DA1	Sex + Region + Latitude + Longitude	full	135521.3
			reduced	135399.3
LM	DA2	Sex	full	139758.8
			reduced	139489.9
LM	DA3	Sex + Region	full	132017.7
			reduced	131757.6
LM	DA4	Sex + Region + Latitude	full	128069.1
			reduced	127896.1
LM	DA5	Region + Latitude	full	121266.3
			reduced	121110.2
LM	FA1	Sex + Region	full	120892
			reduced	120670.1
LM	FA2	Sex + Latitude + Longitude + Landmass	full	123068.9
			reduced	122788.3
LM	FA3	Latitude + Landmass	full	113920.2
			reduced	113660.1
LM	FA4	Sex + Region	full	110640.2
			reduced	110416.1
LM	FA5	Sex + Region + Latitude + Longitude	full	102717.4
			reduced	102460.6
LM	Index.Asymmetry	Latitude + Longitude + Landmass	full	129359.3
			reduced	129219.2
LM	Index.Diversity	Sex + Ethnic Group	full	118923.3
			reduced	118896.9
LM	Index.Diversity.Holt	Sex + Ethnic Group	full	86877.88
			reduced	86855.35

variables for SVR exhibited a reduction in AIC through feature selection ranging between 122 and 281, which was similar with the reductions in AIC value through feature selection for the same response variables for GLM even though the specifics of response variables with the greatest reduction in AIC value were not the same across methodology (Table 25). Of note, the reduced models for the baseline linear models of all Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity as response variables indicated a marked improvement in the goodness-of-fit.

The best performance metric for the Directional Asymmetry variables for each digit response variables was the MAE due to the presence of outliers that could unduly influence the modeling, the zeros in the response variables that did not always results in an undefined MAPE, and the arbitrary zeros since directional asymmetry represents the signed difference between the left and right sides (Table 5). The best performance metric for the Fluctuating Asymmetry variables for each digit and the indices of asymmetry and diversity was either the MAE or MAPE due to the presence of outliers and the zeros. When the MAPE was undefined (NaN), the second best performance metric for these response variables, the MAE, was considered.

The mean regressor performed the best for only the response variable Fluctuating Asymmetry of digit 3 and only considering the lowest MAPE, because the lowest MAE for that response variable was the median regressor (Table 32). The median regressor performed the best for response variables Directional Asymmetry of digits 2 through 5 using the MAE, because the MAPE for the median regressor was NaN. The median regressor performed the best for response variable Directional Asymmetry of digit 1 for the MAPE, but the lowest MAE for that response variable was the SVR model. The median regressor performed the best for response variables Fluctuating Asymmetry of digits 1, 2, 4, and 5 for the MAE, and the lowest MAPE for those

response variables was the linear model. The linear model performed the best for the response variables Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt considering the MAPE, while the best performing model for these response variables considering the MAE was the SVR model. Overall, the SVR model performed the best considering the MAE for response variables Directional Asymmetry of digit 1, Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt. No response variables exhibited best performance when considering both the MAE and MAPE.

Because the MAE and not RMSE or MAPE can be compared across models with different response variables given most other conditions were consistent, the sample sizes and explanatory variables were comparable between Directional Asymmetry of each digit vs Fluctuating Asymmetry of each digit, and also between the three response variables Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt (Table 32). For example, for Directional Asymmetry vs Fluctuating Asymmetry of digit 1, the MAE of mean regressor, median regressor, linear model, and SVR model was lower for Fluctuating Asymmetry of digit 1. In fact, for all digits comparing the Directional Asymmetry to the Fluctuating Asymmetry, the MAE of all models was lower for the Fluctuating Asymmetry variable of that digit. Comparing the Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt the MAE of all models was lower for Index.Diversity.Holt. The poor performance of the baseline linear models of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity could be due to the violation of the linear model assumptions and likely increase in prediction error, but because the magnitude of an increase in prediction error is unknown, the level of certainty of the best performance metrics is also unknown.

Table 32. Sample sizes and performance metrics of final optimized SVR models for Directional Asymmetry and Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity; MAE = mean absolute error; MAPE = mean absolute percentage error; NA = Not available; NaN = Not a number; RMSE = root mean squared error; best performance metric and model in bold

Analytical Method	Response variable	Training data N	Validation data N	Test data N	MAE	RMSE	MAPE
Baseline for SVR	mean DA1	23626	11806	11782	3.212	4.233	1.667
Baseline for SVR	median DA1	23626	11806	11782	3.211	4.235	1.605
Baseline LM	DA1	23626	11806	11782	3.204	4.212	1.810
SVR	DA1	23626	11806	11782	3.183	4.209	2.143
Baseline for SVR	mean DA2	23618	11787	11762	3.371	4.692	6.821
Baseline for SVR	median DA2	23618	11787	11762	3.347	4.720	NaN
Baseline LM	DA2	23618	11787	11762	3.371	4.692	6.877
SVR	DA2	23618	11787	11762	3.352	4.702	32.971
Baseline for SVR	mean DA3	23665	11821	11791	2.934	4.005	9.857
Baseline for SVR	median DA3	23665	11821	11791	2.911	4.013	NaN
Baseline LM	DA3	23665	11821	11791	2.919	3.994	46.018
SVR	DA3	23665	11821	11791	2.924	3.997	31.396
Baseline for SVR	mean DA4	23631	11804	11784	2.705	3.724	99.887
Baseline for SVR	median DA4	23631	11804	11784	2.701	3.724	NaN
Baseline LM	DA4	23631	11804	11784	2.711	3.711	25.504
SVR	DA4	23631	11804	11784	2.706	3.711	70.291

Table 32 Continued

Analytical Method	Response variable	Training data N	Validation data N	Test data N	MAE	RMSE	MAPE
Baseline for SVR	mean DA5	23650	11818	11777	2.386	3.157	8.802
Baseline for SVR	median DA5	23650	11818	11777	2.352	3.162	NaN
Baseline LM	DA5	23650	11818	11777	2.379	3.152	116.303
SVR	DA5	23650	11818	11777	2.378	3.157	86.351
Baseline for SVR	mean FA1	23626	11806	11782	2.371	3.101	0.677
Baseline for SVR	median FA1	23626	11806	11782	2.259	3.131	0.753
Baseline LM	FA1	23626	11806	11782	2.359	3.098	0.675
SVR	FA1	23626	11806	11782	2.261	3.143	0.781
Baseline for SVR	mean FA2	23618	11787	11762	2.467	3.328	0.743
Baseline for SVR	median FA2	23618	11787	11762	2.332	3.590	1.166
Baseline LM	FA2	23618	11787	11762	2.451	3.306	0.736
SVR	FA2	23618	11787	11762	2.336	3.419	0.933
Baseline for SVR	mean FA3	23665	11821	11791	2.049	2.762	0.711
Baseline for SVR	median FA3	23665	11821	11791	1.955	2.908	0.977
Baseline LM	FA3	23665	11821	11791	2.052	2.755	0.712
SVR	FA3	23665	11821	11791	2.041	2.761	0.726
Baseline for SVR	mean FA4	23631	11804	11784	1.835	2.563	0.698
Baseline for SVR	median FA4	23631	11804	11784	1.732	2.657	0.866
Baseline LM	FA4	23631	11804	11784	1.827	2.557	0.695

Table 32 Continued

Analytical Method	Response variable	Training data N	Validation data N	Test data N	MAE	RMSE	MAPE
SVR	FA4	23631	11804	11784	1.749	2.632	0.833
Baseline for SVR	mean FA5	23650	11818	11777	1.591	2.113	0.681
Baseline for SVR	median FA5	23650	11818	11777	1.502	2.142	0.751
Baseline LM	FA5	23650	11818	11777	1.582	2.109	0.677
SVR	FA5	23650	11818	11777	1.620	2.132	0.678
Baseline for SVR	mean Index.Asymmetry	23319	11625	11604	3.062	3.968	0.373
Baseline for SVR	median Index.Asymmetry	23319	11625	11604	3.010	4.055	0.406
Baseline LM	Index.Asymmetry	23319	11625	11604	3.036	3.927	0.370
SVR	Index.Asymmetry	23319	11625	11604	2.996	3.959	0.384
Baseline for SVR	mean Index.Diversity	23319	11625	11604	2.528	3.176	0.369
Baseline for SVR	median Index.Diversity	23319	11625	11604	2.502	3.225	0.395
Baseline LM	Index.Diversity	23319	11625	11604	2.477	3.121	0.362
SVR	Index.Diversity	23319	11625	11604	2.451	3.156	0.384
Baseline for SVR	mean Index.Diversity.Holt	23319	11625	11604	1.285	1.603	0.323
Baseline for SVR	median Index.Diversity.Holt	23319	11625	11604	1.273	1.629	0.343
Baseline LM	Index.Diversity.Holt	23319	11625	11604	1.253	1.570	0.315
SVR	Index.Diversity.Holt	23319	11625	11604	1.245	1.582	0.327

4.2.7 Support vector regression (SVR)

The grid searching figures illustrated the progression of grid searching how the lower/upper boundaries of the optimal hyperparameters were identified as a vein of hyperparameters with lower RMSE values surrounded by hyperparameters with higher RMSE values (Figure S128 through Figure S140). The coarse grid search for hyperparameters of the SVR model for most response variables encompassed a gamma of 10^{-8} to 10^8 and a cost of 2^{-16} to 2^{16} , incremented by values of one, except for Directional Asymmetry of digit 2, Directional Asymmetry of digit 3, and Directional Asymmetry of digit 4 SVR models, for which the coarse grid search began with gamma values of 10^{-6} to 10^6 and cost values of 2^{-14} to 2^{14} (Figure S130, Figure S132, Figure S134, Table S13). Even without an initially broader range for coarse grid searching, the limits of the optimal hyperparameters for a SVR Directional Asymmetry of digit 2, Directional Asymmetry of digit 3, and Directional Asymmetry of digit 4 models were identified (Figure S130, Figure S132, Figure S134).

In models where the lower/upper boundaries of the optimal hyperparameters were not identified as a vein of hyperparameters with lower RMSE values surrounded by hyperparameters with higher RMSE values, an additional grid searching step was performed for Directional Asymmetry of digit 3, Fluctuating Asymmetry of digits 3, 4, and 5, and Index.Asymmetry SVR models (Figure S132, Figure S133, Figure S135, Figure S137, Figure S138). Specifically, Directional Asymmetry of digit 3 had two medium grid searches, and the first was cost of 2^6 to 2^{16} , gamma values of 10^{-8} to 10^{-4} , by 0.5 increments. Fluctuating Asymmetry of digit 3 had two medium grid searches, and the first was cost of 2^{15} to 2^{20} , gamma values of 10^{-15} to 10^{-5} , by 1 increments. Fluctuating Asymmetry of digit 4 had two medium grid searches, and the first was cost of 2^{-25} to 2^{-10} , gamma values of 10^{-20} to 10^{-11} , by 1 increments. Fluctuating Asymmetry of

digit 5 had two medium grid searches, and the first was cost of 2^{13} to 2^{25} , gamma values of 10^{-20} to 10^{-7} , by 1 increments. Index.Asymmetry had two medium grid searches, and the first was cost of 2^{10} to 2^{25} , gamma values of 10^{-7} to 10^{-4} , by 0.5 increments. The second medium grid search for these five response variables is found in Table S13.

Table S13 also reports the performance metrics: MAE, RMSE, and MAPE for each step of grid searching for the response variables. Because the procession into a narrower grid search step relied on the RMSE, the RMSE and performance metrics overall decrease throughout grid searching, indicating a successful identification of ranges of optimal hyperparameter values. The final models for all except Directional Asymmetry of digit 1 and Fluctuating Asymmetry of digit 1 response variables exhibited higher RMSE values as a result of overfitting to the training data, so when the model was applied to the test data the model did not generalize well (Table S13).

Because the performance metrics for the SVR models can be most directly compared when the models were generated from the same data sets, Table 32 illustrates the sample size for each model formulation, and because the SVR models for Directional Asymmetry of digit 1 and Fluctuating Asymmetry of digit 1 response variables, for example, contained the same underlying data, the final performance metrics can be directly compared between the two models. The best performance metric for the Directional Asymmetry variables for each digit response variables was the MAE due to the presence of outliers that could unduly influence the modeling, the zeros in the response variables that did not always results in an undefined MAPE, and the arbitrary zeros since directional asymmetry represents the signed difference between the left and right sides (Table 5). The best performance metric for the Fluctuating Asymmetry variables for each digit and the indices of asymmetry and diversity was either the MAE or

MAPE due to the presence of outliers and the zeros. When the MAPE was undefined (NaN), the second best performance metric for these response variables, the MAE, was considered.

The median regressor performed the best for response variable Directional Asymmetry of digit 1 for the MAPE, but the lowest MAE for that response variable was the SVR model. The linear model performed the best for the response variables Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt considering the MAPE, while the best performing model for these response variables considering the MAE was the SVR model. Overall, the SVR model performed the best considering the MAE for response variables Directional Asymmetry of digit 1, Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt.

4.3 Discussion for finger ridge-counts and summaries of ridge-counts

Anthropologists who struggle in their dermatoglyphics research with what variables and units of measurements are most appropriate for population studies did not have access to a worldwide dermatoglyphics data set, and the results of the different methodological approaches for analyzing finger ridge-counts and summaries of ridge-counts as response variables corroborate previous findings as well as offer new information to understand patterns of worldwide human variation (Meier 1980:171). First, different dermatoglyphic features were considered as response variables using perspectives of data comparison like univariate vs multivariate and summary vs raw. Second, different explanatory variables were considered using perspectives of sample size and composition as well as levels of population study and comparison. Finally, different methods of descriptive modeling and analysis were considered using perspectives of unsupervised vs supervised and distribution-dependent vs distribution-free (Figure 1).

4.3.1 Response variables

The PCA loadings revealed the differential importance of the multivariate finger ridge-count response variables. For all finger ridge-counts the loadings of PC1 were all of the same direction, supporting previous research of PC1 as ARC, the primary influencer of the component (Jantz and Hawkinson 1980). Here, the greatest amount of variation in finger ridge-counts can be referenced by univariate, summary variable ARC. For PC2 the loadings separated digit 1 from all other digits, and for PC3 the loadings separated radial versus ulnar sides of the digits. In Jantz and Hawkinson (1980), the loadings for PC2 represented the radial-ulnar contrast. Further absent in Jantz and Hawkinson (1980) but present in this dissertation for PC4, the loadings separated the left and right ulnar count of digit 2 versus the equally negative left ulnar count of digit 4, and for PC5 the loadings separated the radial counts of digits 4 and 5 compared to right radial count of digit 2, suggesting the involvement of the 2D:4D ratio with finger ridge-counts (Kalichman, Batsevich, and Kobylansky 2019; Jantz 2021; Manning and Fink 2023). The results of PCA of finger ridge-counts via the loadings suggest finger pattern intensity encompassing both size and type drive the majority of variation in finger ridge-counts, whereas digit 1 variation and perhaps development, radial versus ulnar contrasts, and 2D:4D trends represent important subordinate drivers of variation in finger ridge-counts. The multivariate, raw finger ridge-counts were required to attain these results. The trends, but not the exact magnitude and direction, in loadings were consistent even after finger ridge-counts were separated by-Landmass and then by-Sex, further suggesting a common, perhaps developmental trajectory driving variation in finger ridge-counts rather than population relationships dominating (Jantz 1987; Pigliucci 2010).

The interpretability of the performance of the different regression models with different univariate response variables for raw finger ridge-counts and summary TRC or ARC was dictated by the adherence to model assumptions. For the baseline linear models, the five

assumptions of multicollinearity, normality of the residuals, no autocorrelation, no outliers, and no heteroscedasticity were all repeatedly violated (Table 12, Table 16). The consequences of collinearity were diminished through subsequent LASSO regression in GLM, and the collinearity did not affect interpretation of the baseline linear model performance metrics. The impact of the violation of normality of the residuals in baseline linear models diminished in this dissertation due to the large sample sizes in the worldwide dermatoglyphics data set, where the number of observations per variable exceeded 10 in most cases, and transformation of the response variables was purported as a solution, which was directly addressed and not supported with summary variables TRC and ARC and transformation variables like Directional Asymmetry variables of each digit, Fluctuating variables of each digit, and indices of asymmetry and diversity (Graham and Özener 2016; Schmidt and Finan 2018). For practical purposes, autocorrelation of residuals was considered absent in all baseline linear models (Bedre 2019; SE n.d.). The effect of outliers on the regression models was not directly addressed in this dissertation, so the impact of the outliers on performance metrics of baseline models was unknown. Finally, all baseline linear models contained statistically significant heteroscedasticity, which results in an underestimate of prediction error and thus the performance metric values that rely on the prediction error. The type of heteroscedasticity, impure or pure, was unknown and not explored in this dissertation, so no solutions to the heteroscedasticity were applied (Bucevska 2011). Thus, the performance metrics of the baseline linear models should be interpreted as minimum values. The baseline linear model for GLM performed best for 15 response variables compared to the GLM, and no GLM performed best for TRC, ARC, or finger ridge-count response variable models. The differences in the appropriate performance metric MAPE between the best performing baseline linear models for GLM and the GLM were large enough to discount

the influence of heteroscedasticity on the model selection for TRC, ARC, and finger ridge-count response variables. Because the MAE was not ideal for the TRC, ARC, and finger ridge-count response variables for the GLM, no comparisons across response variables were attempted. However, for the baseline linear models that did use the MAE due to undefined MAPE, the best performing response variable according to these MAE values was right ulnar count of digit 3.

Further complicating the interpretation of the GLM was assessment of goodness-of-fit of the LASSO regularized models compared to the full models as well as the beta dispersion to test for equal within-group variance. All GLM of TRC, ARC, and finger ridge-count response variables exhibited small declines in goodness-of-fit using the Kullback-Leibler R^2 value suggesting the LASSO regularization did not improve the model fit enough to offset the benefit of gaining simplicity from just reducing the explanatory variables in the model (Rissanen 1978; Rissanen 1983). The finger ridge-count variables on the right hand exhibited a smaller decline in goodness-of-fit so the LASSO regularization hurt these models least. The beta dispersion was evaluated on different groupings, here the explanatory variables. Sex exhibited no dispersion for the most finger ridge-count response variables, and Ethnic Group, Region, and Landmass exhibited no dispersion for only a few finger ridge-count response variables. A grouping exhibiting dispersion may unduly influence the performance of the model given its difference in within-group variance. Thus, any groupings that exhibited no dispersion for a particular response variable can be readily interpreted in the GLM. Given that most models contained groupings exhibiting dispersion, the poor performance of the GLM and the comparably high performance metrics compared to linked baseline models for GLM could be explained.

The differences in the appropriate performance metric MAPE and the best performing baseline linear models for SVR and the SVR of TRC and ARC were absolutely small values (~

0.02) but relatively large, and for the SVR with TRC as the response variable the SVR model actually exhibited superior performance compared to the baseline linear model, so here heteroscedasticity was not an issue for model selection. The SVR with ARC as the response variable exhibited larger relative differences in the appropriate performance metric MAPE so the influence of heteroscedasticity on model selection here was likely small. Because SVR of finger ridge-count response variables were not included in this dissertation, the utility of comparing SVR to GLM that did include finger ridge-count response variables and that outperformed summary TRC and ARC response variables is low. Thus, just comparing summary TRC and ARC response variables between GLM and SVR and their baseline models, for GLM the baseline linear model for GLM performed best considering the MAPE for TRC and ARC response variables, with the GLM MAPE dramatically higher than the baseline models. For SVR, not only were the MAPE values of the SVR models for TRC and ARC response variables comparable to the baseline models for SVR, the SVR actually exhibited the best performance considering the MAPE for TRC response variable.

4.3.2 Explanatory variables

The results of the unsupervised k-means clustering of finger ridge-counts effectively generated new explanatory variables that were directly linked to the existing individuals in the data set and were confirmed not to be redundant with the existing explanatory variables like Sex, Landmass, or Ethnic Group (Figure 17, Figure 18, Figure 19). The k-means cluster assignments as explanatory variables were also confirmed not to be redundant with each other, as evidenced by their inclusion in final regression models after feature selection and regularization (Table 14). In the worldwide dermatoglyphics data set, the explanatory variables were already collected, and adding new information to this data set would be difficult to locate the same individuals or

families, to approximate the same sample groups, and to match the effort of 100+ researchers over thirty years. The results of k-means clustering generated novel explanatory variables, thereby avoiding the need to physically gather more information in the field. Of note, the optimal K values were calculated using a 10% subset shown to be comparable to the overall dermatoglyphics data set to ease the computational burden, and the success of the k-means clustering results as explanatory variables in subsequent regression models confirmed the flexibility of the sample size and composition in the worldwide dermatoglyphics data set. This means subsequent analyses that could benefit from reduction of sample size still maintaining a degree of sample overlap with the overall data set for computational reasons can attain meaningful results. A purported reason for the non-normal distribution of errors in the baseline linear models was important explanatory variables were not included in the model, even though the k-means clustering assignments were included as explanatory variables in the baseline linear models. Thus, the k-means cluster assignments for the finger ridge-counts did not complete the model formulation or completely solve the problem of explaining the remaining variation in finger ridge-counts the existing explanatory variables did not. Further, the biological meaning of the k-means cluster assignments remains unknown, and because k-means clustering imposes a structure to accommodate multivariate finger ridge-counts, the interpretability of these explanatory variables may be limited in the same way PCA is limited (Leguebe, Vrydagh, and Ducros 1981).

The spatial analysis of finger patterns and finger ridge-counts via distance-based Moran's eigenvector maps (dbMEM) addressed the appropriate level of population study and comparison by constraining the analyses to within the Landmasses and also by explicitly testing broad, medium, and fine scale spatial patterns. For all Landmasses and for both qualitative finger

patterns and quantitative finger ridge-counts, Sex was not a significant explanatory variable of the spatially structured variation in the detrended response variables. In the Americas, the truncation distance for the finger ridge-counts was 2182080 m greater than the truncation distance for the finger patterns, likely due to the different sample size and composition for each dermatoglyphic feature, after rows containing NA were removed (Table 9). Further, the small sample sizes overall in the Americas compared to other Landmasses likely limited the range of variation present for this Landmass, thereby affecting the truncation distance.

For Landmasses Africa, Asia, and Europe, the finger patterns exhibited a higher number of dbMEMs than finger ridge-counts, suggesting general qualitative pattern codes were better explained by spatial variables compared to the variation in quantitative finger ridge-counts. However, when considering the significance of the dbMEM variables, Landmasses Asia and Europe did not differ much between finger patterns and finger ridge-counts, while Landmass Africa exhibited six significant dbMEMs for finger ridge-counts compared to two significant dbMEMs for finger patterns. The smallest reduction, by one dbMEM, in the number of dbMEMs to the number of significant dbMEMs was observed in the Americas for both finger patterns and finger ridge-counts, suggesting the variation in the response variables was characterized well by the explanatory spatial variables (Table 10). Both Africa and Europe exhibited a high number of dbMEMs but a low number of significant dbMEMs, suggesting the variation in the response variables was not well characterized by the explanatory spatial variables or that the scale was not appropriate for these Landmasses. Perhaps larger linear trends were dominating the variation in the response variables, or perhaps even smaller, more regional trends were ignored as noise given the selection of the truncation distances for these Landmasses (Table 9) (Dennis and Sunderland 1979).

The highest portion of variation captured by the variation partitioning at all was purely environmental, which can indicate the presence of a non-spatially structured component of the environmental variables, non-spatially structured components of environmental variables that are not measured in the analysis, or some historical causal factor without spatial structuring in the analysis (Legendre and Legendre 2012a:874). During variation partitioning, the higher portions of spatially structured environmental variation were found in Landmass Americas in finger ridge-counts as 1.8%, 1.8%, and 0.8% for a total of 4.4% and can be considered induced spatial dependence, when spatial structures in the environment produce similar spatial structures in the response data (Figure 20D). No causal relationship between the environment and the response variables for individuals can be suggested, though, because any correlations between spatial variables and environmental variables can either be due to direct induced spatial dependence by a direct influence of the environment on individuals, to spatial variation induced by a historical causal factor through some underlying process that is spatially structured and influencing the response variables and the environmental variables but is not measured, or to spatial autocorrelation between the environmental variables and the response variables for individuals (Legendre and Legendre 2012a:879).

If the arbitrary decisions assigning significant dbMEM variables to broad or fine scales were not appropriate, the variation in the undetrended response variables should still be found in either the broad or fine scales. But the majority of variation in the undetrended response variables was not found in the broad or fine scales, so even if the arbitrary assignments were incorrect, no loss of interpretation in broad or fine scale trends occurred. For variation partitioning overall, the majority of variation in the undetrended response variables for all Landmasses was associated with pure environment, the linear trends, and the residuals. If

variation in the undetrended response variables was due to spatial patterning larger than the sampling area, the portion of variation indicated by the linear trend should be higher, which is not the case. If variation in the undetrended response variables was due to environmental pressures that were not spatially structured, then the portion of variation indicated by the purely environmental component should be higher, and while it is the component with the highest portion of variation captured in the undetrended response variables the amount of variation overall that is captured remains low. Subsequent analyses of finger patterns and finger ridge-counts within these Landmasses would not benefit from first removing the spatial variation. Because the residuals were high, minimally above 90% for all undetrended response variables for all Landmasses, the source of the variation in the undetrended response variables was either not measured, not spatially structured at the scale of the analysis, some historical causal event not measured or spatially structured at the scale of the analysis, or some random variation and noise (Legendre and Legendre 2012a:879).

The PCA can address the explanatory variables driving variation in the dermatoglyphic response variables with the PCA biplots, where the individuals were colored according to their explanatory variable expression. In the biplot of PC1 vs PC2 for all finger ridge-counts, Landmass Americas appeared to contain more variation than the other three Landmasses, and Landmass Asia appeared separate from the other Landmasses. When zoomed into Region the three Regions Australia, Melanesia, and Oceania, which are all in Landmass Asia, were separate from the remaining Regions and likely driving the separation of Landmass Asia. Further, females and males in finger ridge-counts appeared entirely separate in the PC1 vs PC2 biplot, supporting the treatment of Sex as a subpopulation, even though females and males exhibited a similar amount of variation (Binkley 1985; Crawford and Duggirala 1992; Nagy and Pap 2004). Further,

even when separated by Landmass in PCA, females and males still exhibited separation for all Landmasses, suggesting the sex differences do not widely vary between Landmasses. From PCA, the migratory tendencies of females versus males were not apparent (Mbaka et al. 2016). Within Landmass, the PCA biplot of PC1 vs PC2 for Americas illustrated a larger range of standard errors for Region Mesoamerica compared to Regions North America and South America, and this suggests interpretation of Mesoamerica variation would benefit from further division of Mesoamerica into smaller subgroupings.

Within Landmass Asia, outlying Regions Australia, Melanesia, and Oceania highlight the limitation of creating for research purposes such a large Landmass Asia that incorporates many disparate groups in terms of dermatoglyphic variation as well as population history. The trade-off of collapsing multiple disparate groups into one large Landmass was these groups were represented in the worldwide dermatoglyphics data set in smaller numbers, so assigning them to their own Landmass would produce wildly imbalanced sample sizes between Landmasses. After removing the outlying Regions Australia, Melanesia, and Oceania, the PCA biplot of PC1 vs PC2 for Landmass Asia showed similarity between Region Siberia and Central Soviet Union, and Region Indonesia exhibited a greater range in standard error than other Regions. Because Landmass Europe encompassed a smaller area as well as groups with some shared population history, the differences between groups and even sexes were fewer than in other Landmasses (Dennis and Sunderland 1979).

The baseline linear regression model, GLM, and SVR model of finger ridge-counts addressed the level of population study and comparison through collinearity of explanatory variables and the performance of the different models with different explanatory variables. For the baseline linear models of the GLM data set and for the baseline linear models of the SVR

data set, explanatory variable Sex did not exhibit collinearity, further supporting the modeling of females and males as subpopulations within a larger grouping (Binkley 1985). The explanatory variables that exhibited the highest collinearity in the final, fitted baseline linear models, as evidenced by high variance inflation factor, were some combination of Ethnic Group or Region and Latitude or Longitude, and did not survive LASSO regularization as part of GLM (Table 14). For SVR, even though the explanatory variables Ethnic Group and Longitude exhibited high collinearity for TRC and ARC response variables, no feature selection step was utilized so no collinear explanatory variables were removed for modeling. The collinearity suggested any one of these explanatory variables contains useful information but multiple were redundant and not required for successful model formulation. In this dissertation, the interactions between the explanatory variables were not modeled so no conclusion as to which of the collinear variables was most informative was reached. In terms of performance, the baseline linear model with the lowest MAE, considered due to undefined MAPE, was for right ulnar count of digit 3 response variable with some of the most common final explanatory variables Ethnic Group, Sex, frc.kmeans.2, frc.kmeans.4, frc.kmeans.10, frc.kmeans.16, and frc.kmeans.322.

The negative binomial GLM for raw finger ridge-counts and summary TRC and ARC response variables did not apparently violate any model assumptions, so the interpretability of the non-zero coefficients was possible (Binkley 1985). With coefficients of greater magnitude and positive direction, explanatory variables frc.kmeans.2, Landmass, Sex, and frc.kmeans.4 exhibited importance in the GLM. Overall, explanatory variable Region survived LASSO regularization in four GLM models, and the k-means clustering assignments were included in most fitted GLM models. The two GLM models with the lowest MAPE exhibited similar explanatory variables: Landmass and frc.kmeans.2, and additionally different explanatory

variables: Region, Ethnic Group, frc.kmeans.4, frc.kmeans.10, frc.kmeans.16, and frc.kmeans.322.

4.3.3 Methods

Many of the statistical nuances of the methods that impacted performance have already been discussed in *4.3 Discussion for finger ridge-counts and summaries of ridge-counts* so will not be repeated. Here, the multivariate response variables were only analyzed using unsupervised methods, so the benefit of univariate versus multivariate response variables cannot be evaluated across the current methods. Specifically, because multivariate response variables were relegated to unsupervised methods, the benefits of the unsupervised methods to generate explanatory variables or to interpret the explanatory variables driving variation in multivariate finger ridge-counts could not be isolated from the benefits of the increased range of variation in the multivariate finger ridge-counts compared to previous research considering only summary variables like TRC or ARC. The unsupervised methods of k-means clustering and PCA of finger ridge-counts can accommodate multivariate finger ridge-counts to include a way of reducing the large amount of variation and covariation in finger ridge-counts compared to univariate models only containing TRC or ARC that mask this variation and covariation between the sides of digits, the digits, and the hands.

The methods of k-means clustering and PCA may not be hampered by their statistical distribution dependence, and provide a structure to interpret the finger ridge-counts but also provide a structure to use new methods that may rely on a different distribution or may abandon the dependence. Further, the methods of baseline linear regression and GLM heavily depend on the assumptions of the underlying statistical distribution so any violation of the assumptions may preclude interpretation of the performance of the models but offers a theory-bound justification

for subsequent modeling changes. The method of SVR does not depend on any underlying statistical distribution so is flexible enough to model data like finger ridge-counts that do not adhere well to any assumptions or requirements of a specific statistical distribution. The flexibility of SVR was limited through inclusion of only univariate or single output variables that could already mask variation and covariation.

Given dermatoglyphic data tend to not follow a normal distribution, the mean and median regressors as baselines were included with the expectation the high number of zeros caused the skew from normality in dermatoglyphic data and with the expectation the mean represents the central tendency for data closer to a normal distribution. Since ulnar ridge-counts exhibit more zero ridge-counts, the median regressor was expected to perform better than the mean regressor for these response variables. The mean regressor of the GLM performed best considering the MAPE for left ulnar count of digit 2, left ulnar count of digit 4, and right ulnar count of digit 4 response variables, which was unexpected given these were all ulnar ridge-counts. The median regressor of the GLM performed best considering the MAPE and MAE when MAPE undefined for the left radial count of digit 2, left ulnar count of digit 5, right radial count of digit 2, right ulnar count of digit 5, and the Third ridge-count response variables. Again, this was unexpected, because the skew in dermatoglyphic features was thought to stem from the zero inflation.

Finally, no multiple output models were included in this dissertation, but GLM and SVR both contain multiple output extensions that can accommodate the correlation structure between finger ridge-counts so all finger ridge-counts can be included as response variables in a single model.

4.4 Discussion for directional asymmetry, fluctuating asymmetry, and indices of asymmetry and diversity

4.4.1 Response variables

The PCA loadings revealed the differential importance of the multivariate Directional Asymmetry variables and Fluctuating Asymmetry variables. For both Directional Asymmetry variables and Fluctuating Asymmetry variables, the major trends of the loadings were the same: PC1 was dominated by positive values for digit 2, PC2 was dominated by negative values for digit 1, PC3 was dominated by negative values for digit 3, PC4 was dominated by positive values for digit 4, but PC5 differed with high negative values for digit 5 for Directional Asymmetry and high positive values for digit 5 for Fluctuating Asymmetry. The details of the magnitude and direction of the more minor digits for each PC were different for Directional Asymmetry and for Fluctuating Asymmetry. Here, the greatest amount of variation in Directional Asymmetry and Fluctuating Asymmetry can be referenced by digits 1 and 2. The trends including magnitude and direction in PC loadings of Directional Asymmetry overall were consistent for females, after Directional Asymmetry variables were separated by-Sex, and males exhibited the opposite direction for PC1, 3, and 4 but the same magnitude and direction as females for PC2 and 5. The trends in PC loadings of Fluctuating Asymmetry females were consistent with Directional Asymmetry females for PC1, 3, and 4.

The details of the magnitude and direction of the PC loadings of Directional Asymmetry variables and Fluctuating Asymmetry variables changed after separating the variables by-Landmass. The PC1 of Directional Asymmetry variables and Fluctuating Asymmetry variables within Africa was dominated by digit 1, whereas in the other Landmasses PC1 was still dominated by digit 2 as in the overall data sets. Further, for all Landmasses of Directional Asymmetry and Fluctuating Asymmetry, PC5 was dominated by digit 5. The details of PC2, 3,

and 4 were all different between Landmasses for Directional Asymmetry variables and Fluctuating Asymmetry variables. Because the trends in PC loadings depended on how the Directional Asymmetry variables and Fluctuating Asymmetry variables were subset, the Directional Asymmetry variables and Fluctuating Asymmetry variables do not represent some common developmental trajectory (Sengupta and Karmakar 2006; Dipierri et al. 2018). Sengupta and Karmakar (2006) purported an environmental component to asymmetry, which corroborated the differences in PC loadings across Landmasses but did not explain the differences in PC loadings across females and males. Even though the PC loadings of female Directional Asymmetry and Fluctuating Asymmetry variables were more similar to the pooled data sets containing both sexes, no rule for sex differences in asymmetry was clear (Jantz, Fohl, and Zahler 1979; Jantz and Brehme 1993).

The interpretability of the performance of the different regression models with different univariate response variables for Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity was dictated by the adherence to model assumptions. For the baseline linear models, the five assumptions of multicollinearity, normality of the residuals, no autocorrelation, no outliers, and no heteroscedasticity were all repeatedly violated (Table 24, Table 30). The consequences of collinearity were not readily diminished through subsequent LASSO regression in GLM, because many final, fitted GLM models contained the explanatory variables deemed collinear in the baseline linear models. The collinearity did not affect interpretation of the baseline linear model performance metrics. The impact of the violation of normality of the residuals in baseline linear models diminished in this dissertation due to the large sample sizes, where the number of observations per variable exceeded 10 in most cases (Graham and Özener 2016; Schmidt and

Finan 2018). For practical purposes, autocorrelation of residuals was considered absent in all baseline linear models (Bedre 2019; SE n.d.). The effect of outliers on the regression models was not directly addressed in this dissertation, so the impact of the outliers on performance metrics of baseline models was unknown. Finally, not all baseline linear models contained statistically significant heteroscedasticity. Baseline linear models for GLM with Directional Asymmetry of digits 1, 3, and 4 exhibited homoscedasticity, so the prediction errors and the performance metric values for these models were valid in this context. The remaining response variables exhibited heteroscedasticity and thus untrustworthy performance metric values that should be interpreted as minimum values. The identification of the best performing model became difficult in situations where the baseline linear model performance metric exhibited a slightly lower value (i.e. by 10^{-3}) compared to the associated GLM or SVR model performance, because the difference between the calculated and true performance metric for the baseline linear model was unknown. The type of heteroscedasticity, impure or pure, was unknown and not explored in this dissertation, so no solutions to the heteroscedasticity were applied (Bucevska 2011). The GLM actually performed best for Index.Asymmetry response variable. The baseline linear model for GLM performed best for Directional Asymmetry of digits 1, 2, 3, and 4 response variables considering the MAE only. The baseline linear model and the GLM for these response variables exhibited very similar MAE values, and luckily the baseline linear model of Directional Asymmetry of digits 1, 3, and 4 did not exhibit heteroscedasticity, so only the model selection between baseline linear model and GLM using MAE was questionable for Directional Asymmetry of digit 2 response variable due to heteroscedasticity. The best performance of the baseline linear model for Fluctuating Asymmetry of digit 2, Index.Diversity, and Index.Diversity.Holt response variables considered both the MAE and MAPE, and the

differences between the performance metrics for the baseline linear model and the GLM were not as small. But the effect of heteroscedasticity on the model selection between these models was still unclear. For the baseline linear models that did use the MAE due to undefined MAPE, the best performing response variable according to these MAE values was Index.Diversity.Holt.

Further complicating the interpretation of the GLM was assessment of goodness-of-fit of the LASSO regularized models compared to the full models as well as the beta dispersion to test for equal within-group variance. All GLM of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response variables exhibited small declines in goodness-of-fit using the Kullback-Leibler R^2 value suggesting the LASSO regularization did not improve the model fit enough to offset the benefit of gaining simplicity from just reducing the explanatory variables in the model (Rissanen 1978; Rissanen 1983). The Directional Asymmetry of digit 4, the Fluctuating Asymmetry of digit 3, and the Index.Diversity.Holt all exhibited the smallest decline in goodness-of-fit per their data set, so the LASSO regularization hurt these models least. The beta dispersion was evaluated on different groupings, here the explanatory variables. Sex exhibited no dispersion for the majority of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and all indices of asymmetry and diversity response variables. Ethnic Group, Region, and Landmass exhibited no dispersion for only Fluctuating Asymmetry of digit 1 response variable. A grouping exhibiting dispersion may unduly influence the performance of the model given its difference in within-group variance. Thus, any groupings that exhibited no dispersion for a particular response variable can be readily interpreted in the GLM. Of note, even though a grouping may exhibit no dispersion that does not guarantee the inclusion of the grouping in the final fitted model. For example, Fluctuating Asymmetry of digit 1 exhibited no significant beta

dispersion for explanatory variables Sex, Ethnic Group, Region, and Landmass, and none of these variables survived LASSO feature selection so the final GLM model contained only FA.kmeans.2, FA.kmeans.4, FA.kmeans.8, and FA.kmeans.322 explanatory variables.

For the baseline linear models for SVR, all models exhibited heteroscedasticity. For the indices of asymmetry and diversity response variables, the best performing models using both MAE and MAPE differed based on which performance metrics was used. The baseline linear model exhibited the lowest MAPE, while the SVR model exhibited the lowest MAE, and given the properties of the response variables no performance metric was clearly more appropriate for these data. Given the effect of heteroscedasticity to increase uncertainty around the baseline linear model performance metrics, model selection on Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt response variables for SVR resulted in a draw. The differences in the appropriate performance metrics MAE or MAPE and the best performing baseline linear models for SVR and the SVR of Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit were relatively large, so here heteroscedasticity was not an issue for model selection. For SVR, the MAE and MAPE did not correspond with the best performance for any Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, or indices of asymmetry and diversity response variables. In terms of performance, overall the baseline linear models for SVR performed poorly. For all baseline models and SVR models the MAE of Fluctuating Asymmetry variables for each digit was lower than the MAE of Directional Asymmetry variables for each digit (Dipierri et al. 2018). Considering just the MAE, the SVR models of Directional Asymmetry of digit 1, Index.Asymmetry, Index.Diversity, and Index.Diversity.Holt performed better than any of their baseline models.

4.4.2 Explanatory variables

The results of the unsupervised k-means clustering of multivariate Directional Asymmetry variables and Fluctuating Asymmetry variables effectively generated new explanatory variables that were directly linked to the existing individuals in the data set and were confirmed not to be redundant with the existing explanatory variables like Sex, Landmass, or Ethnic Group (Figure 44 through Figure 49). The k-means cluster assignments as explanatory variables were also confirmed not to be redundant with each other, as evidenced by their inclusion in final regression models after feature selection and regularization (Table 26). Of note, the optimal K values were calculated using a 10% subset shown to be comparable to the overall dermatoglyphics data set to ease the computational burden, and the success of the k-means clustering results as explanatory variables in subsequent regression models confirmed the flexibility of the sample size and composition in the worldwide dermatoglyphics data set. This means subsequent analyses that could benefit from reduction of sample size still maintaining a degree of sample overlap with the overall data set for computational reasons can attain meaningful results. A purported reason for the non-normal distribution of errors in the baseline linear models was important explanatory variables were not included in the model, even though the k-means clustering assignments were included as explanatory variables in the baseline linear models. Thus, the k-means cluster assignments for the Directional Asymmetry variables and Fluctuating Asymmetry variables did not complete the model formulation or completely solve the problem of explaining the remaining variation in Directional Asymmetry of Fluctuating Asymmetry that the existing explanatory variables did not. Further, the biological meaning of the k-means cluster assignments remains unknown, and because k-means clustering imposes a structure to accommodate multivariate Directional Asymmetry variables and Fluctuating

Asymmetry variables, the interpretability of these explanatory variables may be limited in the same way PCA is limited (Leguebe, Vrydagh, and Ducros 1981).

The spatial analysis of multivariate Directional Asymmetry variables and Fluctuating Asymmetry variables via distance-based Moran's eigenvector maps (dbMEM) addressed the appropriate level of population study and comparison by constraining the analyses to within the Landmasses but not explicit testing for broad, medium, and fine scale spatial patterns, because both the Directional Asymmetry variables and Fluctuating Asymmetry variables for all Landmasses exhibited only a single significant dbMEM. Only for Landmasses Africa and Asia for both Directional Asymmetry variables and Fluctuating Asymmetry variables, Sex was not a significant explanatory variable of the spatially structured variation in the detrended response variables. The truncation distances were the same for Directional Asymmetry variables and Fluctuating Asymmetry variables for all Landmasses. For Landmass Americas the Fluctuating Asymmetry variables exhibited no significant dbMEMs, perhaps suggesting the exclusion of spatial structuring of variation in fluctuating asymmetry.

Spatial analysis for Directional Asymmetry variables and Fluctuating Asymmetry variables captured little spatially structured variation, seen via the single significant dbMEM variables and the majority, albeit low, percentage of variation in environmental and linear trends during variation partitioning (Figure 20, Figure 50). This suggests the Directional Asymmetry variables and Fluctuating Asymmetry variables, which are transformations of finger ridge-counts, are devoid of any spatial patterning, and perhaps the variation in Directional Asymmetry variables or Fluctuating Asymmetry variables was not well characterized by the explanatory spatial variables or the scale was not appropriate for the Landmasses.

For variation partitioning overall, the majority of variation in the undetrended response variables for all Landmasses was associated with pure environment, the linear trends, and the residuals. If variation in the undetrended response variables was due to spatial patterning larger than the sampling area, the portion of variation indicated by the linear trend should be higher, which was not the case. If variation in the undetrended response variables was due to environmental pressures that were not spatially structured, then the portion of variation indicated by the purely environmental component should be higher, and while it was the component with the highest portion of variation captured in the undetrended response variables the amount of variation overall that was captured remains low. Because the residuals were high, minimally above 98% for all undetrended response variables across Landmasses, the source of the variation in the undetrended response variables was either not measured, not spatially structured at the scale of the analysis, some historical causal event not measured or spatially structured at the scale of the analysis, or some random variation and noise (Legendre and Legendre 2012a:879).

The PCA can address the explanatory variables driving variation in the dermatoglyphic response variables with the PCA biplots, where the individuals were colored according to their explanatory variable expression. In the biplot of PC1 vs PC2 for all Directional Asymmetry variables, Landmasses exhibited little separation, and Regions Australia, Indonesia, and Melanesia, which are all in Landmass Asia, were separate from the remaining Regions. In the biplot of PC1 vs PC2 for all Fluctuating Asymmetry variables, even though they all overlapped the Landmasses were differentiated along the PC1 axis but not the PC2 axis. Landmass Africa and Region Africa (but not Region Northeast Africa) appeared slightly separate from the other groups in that level. Females and males in Directional Asymmetry variables and Fluctuating Asymmetry variables appeared overlapping and when separated by Landmass in PCA, females

and males still exhibited considerable overlap for all Landmasses, advocating pooling of females and males in analyses including Directional Asymmetry variables and Fluctuating Asymmetry variables. For subsequent biplots considering subsets by-Landmass and by-Sex, very little separation between groups was present for both Directional Asymmetry variables and Fluctuating Asymmetry variables.

The baseline linear regression model, GLM, and SVR model of Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response variables addressed the level of population study and comparison through collinearity of explanatory variables and the performance of the different models with different explanatory variables. For the baseline linear models of the GLM data set and for the baseline linear models of the SVR data set, explanatory variable Sex did not exhibit collinearity. The explanatory variables that exhibited the highest collinearity in the final, fitted baseline linear models, as evidenced by high variance inflation factor, were some combination of Landmass or Region and Latitude or Longitude and some of these explanatory variables did survive LASSO regularization as part of GLM or in some cases Region was substituted for Landmass (Table 26). For SVR, even though the explanatory variable Region exhibited high collinearity for Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response variables, no feature selection step was utilized so no collinear explanatory variables were removed for modeling. In this dissertation, the interactions between the explanatory variables were not modeled so no conclusion as to which of the collinear variables was most informative was reached. In terms of performance, across all the baseline models and SVR models for the indices of asymmetry and

diversity, the MAE was lowest for Index.Diversity.Holt with explanatory variables Sex and Ethnic Group.

For Gaussian models of GLM for Directional Asymmetry variables for each digit, the coefficient with greatest magnitude and positive direction other than the intercept was explanatory variable DA.kmeans.2. For negative binomial models of GLM for Fluctuating Asymmetry variables for each digit, the coefficients with greater magnitude and both positive and negative direction were FA.kmeans.2 and FA.kmeans.8 explanatory variables. No final GLM models for Fluctuating Asymmetry variables for each digit included Ethnic Group, Region, Sex, or Longitude. For Gaussian models of GLM for indices of asymmetry and diversity, the coefficients with greater magnitude were solely negative DA.kmeans.2 and solely positive Landmass explanatory variables. No final GLM models for indices of asymmetry and diversity included Ethnic Group. Overall the respective k-means clustering assignments were included in most fitted GLM for Directional Asymmetry variables for each digit and for Fluctuating Asymmetry variables for each digit, and the explanatory variables Landmass, Latitude, Longitude, and the k-means clustering assignments were included in the fitted GLM for indices of asymmetry and diversity. The MAE for GLM models of Directional Asymmetry variables for each digit consistently decreased from digit 1 to 5, and the GLM model with the lowest MAE for Directional Asymmetry variables for each digit exhibited explanatory variables: Region, Latitude, DA.kmeans.2, and DA.kmeans.4. The two GLM models with the lowest MAE and then the lowest MAPE for Fluctuating Asymmetry variables for each digit exhibited the same explanatory variables: Landmass, FA.kmeans.2, FA.kmeans.4, FA.kmeans.8, and FA.kmeans.322.

4.4.3 Methods

Many of the statistical nuances of the methods that impacted performance have already been discussed in 4.4 *Discussion for directional asymmetry, fluctuating asymmetry, and indices of asymmetry and diversity* so will not be repeated. Here, the multivariate response variables were only analyzed using unsupervised methods, so the benefit of univariate versus multivariate response variables cannot be evaluated across the current methods. The unsupervised methods of k-means clustering and PCA can accommodate multivariate Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit to include a way of reducing the large amount of variation and covariation in Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit compared to univariate models only containing indices of asymmetry and diversity that mask this variation and covariation between the sides of digits, the digits, and the hands.

The methods of k-means clustering and PCA may not be hampered by their statistical distribution dependence and provide a structure to interpret the Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit but also provide a structure to use new methods that may rely on a different distribution or may abandon the dependence. The methods of baseline linear regression and GLM depend heavily on the assumptions of the underlying statistical distribution so any violation of the assumptions may preclude interpretation of the performance of the models but offers a theory-bound justification for subsequent modeling changes. The method of SVR does not depend on any underlying statistical distribution so is flexible enough to model data like Directional Asymmetry variables for each digit, Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity response variables that do not adhere well to any assumptions or requirements of a specific statistical distribution.

The flexibility of SVR was limited through inclusion of only univariate or single output variables that could already mask variation.

Given that raw finger ridge-count response variables and summary variables like TRC and ARC tend to not follow a normal distribution, the transformation of the raw finger ridge-counts into Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit could also transform the distribution of the data closer to a normal distribution. Thus, the differences in the performance just between the mean and median regressor were not expected to be high. The mean regressor of the SVR performed best for Fluctuating Asymmetry of digit 3 considering the MAPE, and the median regressor of the SVR performed best for Fluctuating Asymmetry of digit 3 considering the MAE. Further, the median regressor performed best for Directional Asymmetry of digits 2 through 5 for the MAE, given an undefined MAPE. The better performance of the median regressor suits Directional Asymmetry variables for each digit, because these counts exhibit both positive and negative values, which may skew the mean as the central tendency.

Finally, no multiple output models were included in this dissertation, but GLM and SVR both contain multiple output extensions that can accommodate the correlation structure between Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit so all Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit can be included as response variables in a single model.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

The goals of this dissertation were to identify the patterns of human dermatoglyphic variation in worldwide populations by specifically addressing the existing hurdles in interpretation and analysis of dermatoglyphic features as synthesized directly from Meier (1980:165) (Figure 1).

5.1 Conclusions of research

Addressing Factor I of the number and kinds of dermatoglyphic traits to use, the univariate finger ridge-counts performed better in regression models simply because each finger ridge-count contains a more narrow range of variance than summary variables like TRC or ARC. For GLM of summary variables TRC and ARC, the GLM models did not outperform their baseline linear models, whereas the SVR of TRC and ARC exhibited superior performance for TRC compared to its baseline linear model. This difference in performance of summary response variables TRC and ARC reflect the misfit of the imposed negative binomial statistical distribution for GLM, because SVR does not depends on any specific distribution. Also, the declines in goodness-of-fit using the Kullback-Leibler R^2 value for the GLM could also indicate the model formulation of the GLM was not optimal. The indices of asymmetry and diversity response variables consistently performed better in regression models compared to the Directional Asymmetry variables for each digit and the Fluctuating Asymmetry variables for each digit, and this was likely not due to any achievements in picking an optimal statistical distribution for GLM, because the indices of asymmetry and diversity performed best in SVR models, as well. Further supporting the concept that the superior performance in indices of asymmetry and diversity related more to the structure of the data variables rather than to model

formulation, Index.Diversity and Index.Diversity.Holt performed best in the baseline linear regression models for GLM. The indices of asymmetry and diversity all performed similarly well for the baseline linear models for SVR, and here no consensus in superior performance was possible due to the uncertainty in the prediction error due to heteroscedasticity. Overall, the univariate finger ridge-counts performed better than the summary TRC and ARC response variables, likely due to modeling constraints. The summary indices of asymmetry and diversity performed better than univariate Directional Asymmetry variables for each digit or Fluctuating Asymmetry variables for each digit, likely due to the structure of the data variables.

Addressing Factor III of the optimal sample size and composition as well as Factor V of the optimal level of population study, Latitude and Longitude explanatory variables were found to be redundant with either Ethnic Group and Region for finger ridge-counts or Landmass and Region for Directional Asymmetry and Fluctuating Asymmetry. Consistently, the k-means clustering assignments as novel explanatory variables in the baseline linear models and GLM models were not redundant and captured the most variation in raw finger ridge-counts, summary TRC and ARC, Directional Asymmetry variables, and Fluctuating Asymmetry variables, compared to the existing explanatory variables available in the original data set. Given the success of the explanatory variables of k-means clustering assignments, which were based on preset optimal K values explored using only a 10% subset of the worldwide dermatoglyphics data set, the sample size and composition of the worldwide dermatoglyphics data set can be proportionally subset and still offer valuable information concerning the drivers of patterns of variation. Additionally in terms of sample size and composition, the spatial analyses that cut the worldwide data set into smaller data sets within Landmasses, revealed subsequent analyses within Landmasses of qualitative finger patterns, quantitative finger ridge-counts, multivariate

Directional Asymmetry variables, and multivariate Fluctuating Asymmetry variables would not benefit from first removing the spatial variation before additional modeling, because the variation in these dermatoglyphic features was not spatially structured to any high degree. Spatial analysis of Fluctuating Asymmetry variables in Landmass Americas did not reveal a single significant dbMEM. Explanatory variable Sex never exhibited collinearity, and for analyses of finger ridge-counts Sex does capture enough variation such that females and males should be considered separately as subpopulations. However, for analyses of directional asymmetry and fluctuating asymmetry pooling females and males to bolster sample sizes is recommended, because Sex does not capture enough variation in these dermatoglyphic variables. In terms of the level of population study, when Directional Asymmetry variables and Fluctuating Asymmetry variables were split by Landmass, the details of the PCA changed suggesting finer scales of study may not contain structured variation and may instead be overwhelmed by stochastic signals difficult to interpret. The details of the PCA of finger ridge-counts split by-Landmass did not change relative to the overall data set, suggesting finer scales of study may contain structured variation and should be explored to what extent. Overall, the k-means clustering assignments consistently captured the most variation in the relevant dermatoglyphic features, but the remaining explanatory variables differentially captured variation depending on the specific dermatoglyphic feature.

Addressing Factor IV of the best methods of testing congruence and overall the best methods for analyzing dermatoglyphic data, the violations of the assumption of homoscedasticity that increased uncertainty in the performance metrics for the baseline linear models did not cloud the interpretation of the relative performance of univariate finger ridge-count or summary TRC and ARC response variables. The baseline linear models for GLM of Directional Asymmetry

variables of digits 1, 3, and 4 did not exhibit heteroscedasticity. Further, the interpretations of baseline linear models for SVR of the Directional Asymmetry variables for each digit and the Fluctuating Asymmetry variables for each digit were not affected by the heteroscedasticity. The heteroscedasticity present in baseline linear model for GLM of Directional Asymmetry of digit 2 was the only instance where the interpretation of model performance was clouded. The increasing complexity in model formulation and increased burden of computation for GLM and SVR was only merited if they exhibited superior performance compared to their associated baseline models. For finger ridge-counts, TRC, and ARC response variables, none of the best performing models included the negative binomial GLM models, suggesting the model formulation was not optimal. For TRC and ARC response variables, the SVR models performed better than the baseline models so perhaps another GLM model formulation is not required for these dermatoglyphic features, and SVR can be considered a successful model framework. For Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit analyzed via GLM, the differences in performance metric value for the best model compared to the others was relatively large, supporting the decisions in model selection, and generally the baseline models of some kind performed better. Oddly, the Directional Asymmetry variables for each digit GLM models assumed a Gaussian distribution, and even these Gaussian GLM performed worse than standard linear regression. Perhaps the feature selection approaches drove the differences in model performance between baseline linear models and Gaussian GLM models for Directional Asymmetry variables for each digit. For Directional Asymmetry variables for each digit and Fluctuating Asymmetry variables for each digit analyzed via SVR, the baseline models and the SVR all exhibited very similar performance metrics, and often the performance metric for the median regressor was only slightly ($\sim 10^{-3}$) better than the SVR. For indices of

asymmetry and diversity, the SVR exhibited superior performance compared to the baseline models. Direct comparison of the unsupervised methods to the supervised methods does not make sense, and each excelled in valuable ways. Unsupervised methods generated new information driving the variation in dermatoglyphic features, and supervised methods leveraged the contribution of existing information to patterns of variation in dermatoglyphic features.

From the results of this research, dermatoglyphic features can be selected based on their strengths when comparing to other biological data sets. Because dermatoglyphic morphology persists after fixation during intrauterine development, the role of the developmental environment in shaping patterns of variation in dermatoglyphic features cannot be dissociated from the variation itself. For example, the trends in PC loadings for finger ridge-count response variables reveal the importance of finger pattern intensity, radioulnar gradients, and 2D:4D contrasts that have strong links to hormone concentration and sensitivity during development (Dressler and Voracek 2011; Jantz 2021; Manning and Fink 2023; Polcerová et al. 2023).

This dissertation on the nature and significance of worldwide human dermatoglyphic variation provides a framework for continued resurgence in anthropological dermatoglyphics to discover the evolutionary forces acting on and maintaining the variation in discriminatory and persistent dermatoglyphic features that ostensibly serve no functional purpose.

5.2 Limitations of study

In this dissertation, the interactions between the explanatory variables were not modeled so no conclusion as to which of the collinear variables was most informative was reached. The effect of outliers on the regression models was not directly addressed in this dissertation, so the impact of the outliers on performance metrics of baseline models was unknown. The type of heteroscedasticity, impure or pure, was unknown and not explored in this dissertation, so no

solutions to the heteroscedasticity were applied (Bucevska 2011). Even after the goodness-of-fit test indicated the model was appropriate for the data, standard limitations of regression models prohibit interpretation or predictions for outside the range of samples in the model. If the modeled data truly represent a sample of the population, interpretation or predictions may be appropriate outside the range of samples in the model but should always stay within the scope of the population (Hilbe 2012c:76). The comparison of GLM versus SVR for raw finger ridge-counts was limited, because the SVR of finger ridge-counts was not included in this dissertation. Also, the k-means clustering assignments were not included in any of the SVR models, which also hurt the comparability between GLM and SVR models.

5.3 Recommendation for future research

Specifically, future research should directly address the limitations of this study, which is possible merely by reconfiguring the models and does not require additional data beyond the worldwide dermatoglyphics data set. The power in the results of the analyses of the worldwide dermatoglyphics data set lies in the generation of hypotheses for subsequent research to investigate the promising candidates in detail using targeted strategies (Huttenhower, Finn, and McHardy 2023). The k-means clustering assignments served as informative, non-redundant explanatory variables, so subsequent research should explore the biological meaning of the clusters, like perhaps finger patterns may be structuring the k-means cluster assignments for finger ridge-counts. Testing hypotheses of known variables as the possible cluster assignments can increase the interpretability of the k-means cluster assignments as explanatory variables in subsequent modeling.

The disappointing performance of the negative binomial GLM models of finger ridge-count, TRC, and ARC response variables may be attributed to the model formulation's inability

to handle zero inflation. Zero-inflated negative binomial GLM models can be attempted to better account for the zero inflation. While negative binomial models can handle overdispersion in count data, zero-inflated negative binomial regression contains a binary model for the excess zeros and a negative binomial model for the remaining non-zero counts and can thus handle both excess zeros and overdispersion in count data. To further combat the dependence on Gaussian or other ill-fitting distributions due to the limitations of the GLM formulation itself, Bayesian multilevel modeling can handle zero-inflated models as well as provide a flexible and powerful modeling framework for complex data sets.

Future research should explore multiple output methods that can accommodate all finger ridge-counts as response variables and can account for the dependence between the variables. Just by incorporating a truly multiple output model framework, these methods have surpassed performance of single output SVR and artificial neural networking (Chen et al. 2021). Neural networking can be applied to the dermatoglyphic features in incremental increases in complexity. In the same way the k-means clustering assignments captured the variation in the dermatoglyphic response variables, neural networking can generate new information about the important structures within dermatoglyphic features. This would be appropriate, because the supervised methods with existing explanatory variables did not predict the variation in the response variables well and in some cases the constraints of the distribution-dependent modeling framework hurt the performance of the models (Adlung et al. 2021:4).

Further, convolutional neural networks that work well with image data can accept the digital scans of the entire handprint for individuals in the worldwide dermatoglyphics data set to train models that can better handle categorical data and can learn the connections or relationships between dermatoglyphic variables. Accepting the digital scans of the entire handprints removes

the requirement for a researcher to process the physical handprints to assess finger patterns and count finger ridges. Although assessing finger patterns and counting finger ridges does not require any equipment, the training and reliability of the researcher is so vital to generating data sets. Attempting convolutional neural networking on digital scans of handprints also removes the downside of multiple, different researchers collecting raw data and then analyzing them thereby generating interobserver error, so new data could be added to the worldwide dermatoglyphics data set in this way with limited risk in reliability.

REFERENCES

- Abraham, Joshua, Christophe Champod, Chris Lennard, and Claude Roux
2013 Modern Statistical Models for Forensic Fingerprint Examinations: A Critical Review. *Forensic Science International* 232(1): 131–150.
- Adlung, Lorenz, Yotam Cohen, Uria Mor, and Eran Elinav
2021 Machine Learning in Clinical Decision Making. *Med* 2(6). Elsevier: 642–665.
- Alpaslan-Roodenberg, Songül, David Anthony, Hiba Babiker, et al.
2021 Ethics of DNA Research on Human Remains: Five Globally Applicable Guidelines. *Nature* 599(7883). Nature Publishing Group: 41–46.
- Alpaydin, Ethem
2014 Dimensionality Reduction. In *Introduction to Machine Learning*. 3rd edition Pp. 115–160. Cambridge, Massachusetts: Massachusetts Institute of Technology.
- Arnold, Jeffrey B.
2021 Ggthemes: Extra Themes, Scales and Geoms for “Ggplot2.” <https://CRAN.R-project.org/package=ggthemes>.
- Bedre, Renesh
2019 The Durbin-Watson Test in R: How to Analyze and Interpret Your Results. *Data Science Blog*. <https://www.reneshbedre.com/blog/durbin-watson-test.html>, accessed July 7, 2023.
- Binkley, Kenneth Melvin
1985 A Search for Systemic Variability in Inter-Finger Ridge Count Variation among European Groups. PhD Dissertation, University of Tennessee. https://trace.tennessee.edu/utk_graddiss/4030.
- Bivand, Roger
2022 R Packages for Analyzing Spatial Data: A Comparative Case Study with Areal Data. *Geographical Analysis* 54(3): 488–518.
- Bivand, Roger, Micah Altman, Luc Anselin, et al.
2023 Spdep: Spatial Dependence: Weighting Schemes, Statistics. <https://cran.r-project.org/web/packages/spdep/>, accessed June 22, 2023.
- Blanchet, F. Guillaume, Pierre Legendre, and Daniel Borcard
2008 Forward Selection of Explanatory Variables. *Ecology* 89(9): 2623–2632.
- Bonifay, Wes
2022 Increasing Generalizability via the Principle of Minimum Description Length. *Behavioral and Brain Sciences* 45. Cambridge University Press: e5.

- Borcard, Daniel, François Gillet, and Pierre Legendre
 2018 Spatial Analysis of Ecological Data. *In* Numerical Ecology with R. Daniel Borcard, François Gillet, and Pierre Legendre, eds. Pp. 299–367. Use R! Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-71404-2_7, accessed November 11, 2020.
- Borcard, Daniel, and Pierre Legendre
 2002 All-Scale Spatial Analysis of Ecological Data by Means of Principal Coordinates of Neighbour Matrices. *Ecological Modelling* 153: 51–68.
- Boroffice, Robert A.
 1978 Digital Dermatoglyphic Patterns in a Sample of the Nigerian Population. *American Journal of Physical Anthropology* 49: 167–170.
- Bougeard, Stéphanie, and Stéphane Dray
 2018 Supervised Multiblock Analysis in R with the Ade4 Package. *Journal of Statistical Software* 86(1): 1–17.
- Breusch, TS, and AR Pagan
 1979 A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* 47: 1287–1294.
- Brind'Amour, Anik, Stéphanie Mahévas, and Pierre Legendre
 2018 Application of Moran Eigenvector Maps (MEM) to Irregular Sampling Designs. *Spatial Statistics* 26: 56–68.
- Brock, Guy, Vasyl Pihur, Susmita Datta, and Somnath Datta
 2008 clValid: An R Package for Cluster Validation. *Journal of Statistical Software* 25(4): 1–22.
- Brodhage, G, and GG Wendt
 1951 Eine Notwendige Ergänzung Der Quantitativen Erbbiologischen Auswertung von Fingerleisten. *Zeitschrift Für Menschliche Vererbungs- Und Konstitutionslehre*: 212–220.
- Broersen, P.M.T.
 2005 The Uncertainty of Measured Autocorrelation Functions. *In* Proceedings of the 2005 IEEE International Workshop on Advanced Methods for Uncertainty Estimation in Measurement, 2005. Pp. 90–95. <https://ieeexplore.ieee.org/document/1594612>, accessed September 25, 2023.
- Brownlee, Jason
 2014a Why You Should Be Spot-Checking Algorithms on Your Machine Learning Problems. *MachineLearningMastery.Com*. <https://machinelearningmastery.com/why-you-should-be-spot-checking-algorithms-on-your-machine-learning-problems/>, accessed June 25, 2023.
 2014b How To Get Baseline Results And Why They Matter. *MachineLearningMastery.Com*. <https://machinelearningmastery.com/how-to-get-baseline-results-and-why-they-matter/>, accessed June 25, 2023.

Bucevska, Vesna

2011 Heteroscedasticity. *In* International Encyclopedia of Statistical Science. Miodrag Lovric, ed. Pp. 630–633. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-04898-2_628, accessed September 25, 2023.

Cameron, A. Colin, and Frank A. G. Windmeijer

1997 An R-Squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models. *Journal of Econometrics* 77(2): 329–342.

Chakravathy, Ganesh P., Arun Shirali, K. Nithyananda Chowta, et al.

2018 “Handy” Tool for Hypertension Prediction: Dermatoglyphics. *Indian Heart Journal* 70: 5116–6119.

Chao, Anne, Chun-Huo Chiu, and Lou Jost

2014 Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures through Hill Numbers. *Annual Review of Ecology Evolution and Systematics* 45: 297–324.

Chapelle, Olivier, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee

2002 Choosing Multiple Parameters for Support Vector Machines. *Machine Learning* 46(1): 131–159.

Chattopadhyay, Prasanta Kumar, and Dharmalingam Ganeson

1984 Inheritance of Dermal Ridges: Relation between Finger and Palmar a-b Ridge Counts. *Japanese Journal of Human Genetics* 29(1): 59–61.

Chen, Sixing, and Jukka-Pekka Onnela

2019 A Bootstrap Method for Goodness of Fit and Model Selection with a Single Observed Network. *Scientific Reports* 9(1). Nature Publishing Group: 16674.

Chen, Xiyi, Muzheng Xiao, Dawei Kang, et al.

2021 Prediction of Geometric Characteristics of Melt Track Based on Direct Laser Deposition Using M-SVR Algorithm. *Materials* 14(23). Multidisciplinary Digital Publishing Institute: 7221.

Chen, Ying-Ju, Fadel M. Megahed, L. Allison Jones-Farmer, and Steven E. Rigdon

2023 Basemodels: Baseline Models for Classification and Regression. <https://cran.r-project.org/web/packages/basemodels/>, accessed June 29, 2023.

Cortes, Corinna, and Vladimir Vapnik

1995 Support-Vector Networks. *Machine Learning* 20(3): 273–297.

Cox, DR, and EJ Snell

1989 Analysis of Binary Data. 2nd edition. London: Chapman and Hall.

Crawford, Michael H., and R. Duggirala

1992 Digital Dermatoglyphic Patterns of Eskimo and Amerindian Populations: Relationships between Geographic, Dermatoglyphic, Genetic, and Linguistic Distances. *Human Biology* 64(5): 683–704.

Crawley, Michael J.

2013a Statistical Modelling. *In* *The R Book*. 2nd edition Pp. 388–448. Imperial College London at Silwood Park, UK: John Wiley & Sons, Ltd.

2013b Count Data. *In* *The R Book*. 2nd edition Pp. 579–598. Imperial College London at Silwood Park, UK: John Wiley & Sons, Ltd.

Cummins, Harold, and Charles Midlo

1961 *Finger Prints, Palms and Soles: And Introduction to Dermatoglyphics*. New York: Dover Publications.

Daly, E. Susanne, Stephen Ousley, and Richard L. Jantz

2010 Effects of Distance from Africa on Within-Population Phenotypic Diversity in Modern Human Dermatoglyphics Finger Ridge Counts. *In* P. 90. Minneapolis, MO: American Journal of Physical Anthropology.

Dennis, R.L.H., and E. Sunderland

1979 Dermatoglyphic Variation in the Human Populations of the North Pennine Dales, North England. *American Journal of Physical Anthropology* 50: 309–324.

Dimensionality Reduction

N.d. Machine Learning. Stanford University. <https://www.coursera.org/learn/machine-learning>.

Dipierri, J.E., E. Gutierrez-Redomero, N. Rivalderia, M.C. Alonso-Rodriguez, and E. Alfaro

2018 The Asymmetry of Dermatoglyphic Finger Ridge Counts and the Geographic Altitude of the Jujenean Population in Northwest Argentina. *HOMO - Journal of Comparative Human Biology* 69: 364–376.

Dray, Stéphane, David Bauman, Guillaume Blanchet, et al.

2022 *Adespatial: Multivariate Multiscale Spatial Analysis*. <https://CRAN.R-project.org/package=adespatial>.

Dray, Stéphane, Anne-Béatrice Dufour, and Jean Thioulouse, et al.

2023 *Ade4: Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences*. <https://cran.r-project.org/web/packages/ade4/>, accessed June 22, 2023.

Dressler, Stefan G., and Martin Voracek

2011 No Association between Two Candidate Markers of Prenatal Sex Hormones: Digit Ratios (2D:4D and Other) and Finger-Ridge Counts. *Developmental Psychobiology* 53(1): 69–78.

Dunn, JC

1974 Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics* 4(1). Taylor & Francis: 95–104.

Espinoza, Eha, Ully Saputri, Faizal Hafiz Fadilah, and Dodi Devianto

2021 Modeling the Count Data of Public Health Service Visits with Overdispersion Problem by Using Negative Binomial Regression. *Journal of Physics: Conference Series* 1940(1). IOP Publishing: 012021.

Federal Bureau of Investigation

1990 The Science of Fingerprints: Classification and Uses. Washington D.C.: US Department of Justice.

Forgy, EW

1965 Cluster Analysis of Multivariate Data: Efficiency vs Interpretability of Classifications. *Biometrics* 21: 768–769.

Fournier, Nichole A., and Ann H. Ross

2016 Sex, Ancestral, and Pattern Type Variation of Fingerprint Minutiae: A Forensic Perspective on Anthropological Dermatoglyphics. *American Journal of Physical Anthropology* 160: 625–632.

Friedman, Jerome, Trevor Hastie, Rob Tibshirani, et al.

2023 Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. <https://cran.r-project.org/web/packages/glmnet/>, accessed June 22, 2023.

Frost, Jim

2017 Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. *Statistics By Jim*. <http://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/>, accessed September 24, 2023.

Garnier, Simon, Ross, et al.

2021 Viridis - Colorblind-Friendly Color Maps for R. <https://sjmgarnier.github.io/viridis/>.

2022 Viridis - Colorblind-Friendly Color Maps for R. <https://sjmgarnier.github.io/viridis/>.

Glover, James D., Zoe R. Sudderick, Barbara Bo-Ju Shih, et al.

2023 The Developmental Basis of Fingerprint Pattern Formation and Variation. *Cell* 0(0). Elsevier. [https://www.cell.com/cell/abstract/S0092-8674\(23\)00045-4](https://www.cell.com/cell/abstract/S0092-8674(23)00045-4), accessed February 10, 2023.

Glowacz, Adam

2021 Ventilation Diagnosis of Angle Grinder Using Thermal Imaging. *Sensors* 21(8). Multidisciplinary Digital Publishing Institute: 2853.

Graham, John H., and Bariş Özener

2016 Fluctuating Asymmetry of Human Populations: A Review. *Symmetry* 8(154): 1–36.

- Gross, Juergen, and Uwe Ligges
2015 *Nortest: Tests for Normality*. <https://CRAN.R-project.org/package=nortest>.
- Grünwald, Peter D
2007 *The Minimum Description Length Principle*. MIT Press.
<https://ieeexplore.ieee.org/book/6267274>, accessed June 26, 2023.
- Hamerly, Greg, and Jonathan Drake
2015 Accelerating Lloyd's Algorithm for k-Means Clustering. *In* *Partitional Clustering Algorithms*. M. Emre Celebi, ed. Pp. 41–78. Cham: Springer International Publishing.
http://link.springer.com/10.1007/978-3-319-09259-1_2, accessed August 4, 2021.
- Hansen, Mark H., and Bin Yu
2001 Model Selection and the Principle of Minimum Description Length. *Journal of the American Statistical Association* 96(454). [American Statistical Association, Taylor & Francis, Ltd.]: 746–774.
- Harich, N., E. Esteban, A. Chafik, and P. Moral
2002 Dermatoglyphic Characterization of Berbers from Morocco: Qualitative and Quantitative Digital and Palm Data. *Annals of Human Biology* 29(4): 442–456.
- Hawkinson, Cleone H.
1981 Dermatoglyphic Variability in Armenian Fingertips and Palms: A Directional Asymmetry Approach to Identify Developmental Stress. Master's Thesis, University of Tennessee. https://trace.tennessee.edu/utk_gradthes/4226.
- Hebbali, Aravind
2020 *Olsrr: Tools for Building OLS Regression Models*. <https://cran.r-project.org/web/packages/olsrr/>, accessed June 29, 2023.
- Hijmans, Robert J.
2022 *Geosphere: Spherical Trigonometry*. <https://CRAN.R-project.org/package=geosphere>.
- Hilbe, Joseph M.
2012a Introduction. *In* *Negative Binomical Regression*. 2nd edition Pp. 1–14. Cambridge, United Kingdom: Cambridge University Press.
2012b Overdispersion. *In* *Negative Binomical Regression*. 2nd edition Pp. 141–184. Cambridge, United Kingdom: Cambridge University Press.
2012c Assessment of Count Models. *In* *Negative Binomical Regression*. 2nd edition Pp. 61–76. Cambridge, United Kingdom: Cambridge University Press.
- Holt, Sarah B.
1955 Genetics of Dermal Ridges: Frequency Distributions of Total Finger Ridge-Count. *Annals of Human Genetics* 20(2): 159–170.

- 1958 Genetics of Dermal Ridgees: The Relation between Total Ridge-Count and the Variability of Counts from Finger to Finger. *Annals of Human Genetics* 22(4): 323–339.
- 1960 Genetics of Dermal Ridges: Familial Correlations for $(S/\sqrt{10})$, a Measurement of the Diversity of Ridge-Counts from Finger to Finger. *Annals of Human Genetics* 24(3): 253–269.
- 1968 The Genetics of Dermal Ridges. Springfield, Illinois: Charles C Thomas.

Huttenhower, Curtis, Robert D. Finn, and Alice Carolyn McHardy

- 2023 Challenges and Opportunities in Sharing Microbiome Data and Analyses. *Nature Microbiology*. Nature Publishing Group: 1–11.

Jantz, R. L., A. W. Eriksson, and H. Brehme

- 1993 Population Relationships of Lapps as Reflected by Quantitative Dermatoglyphics. *Human Biology* 65(5): 711–730.

Jantz, Richard L.

- 1974a Multivariate Analysis of Dermatoglyphic Variation in Man. *Yearbook of Physical Anthropology* 18: 121–139.
- 1974b Finger Ridge-Counts and Inter-Finger Variability in Negroes and Whites. *Human Biology* 46(4): 663–675.
- 1975 Population Variation in Asymmetry and Diversity from Finger to Finger to Digital Ridge-Counts. *American Journal of Physical Anthropology* 42: 215–224.
- 1977a Sex and Race Differences in Finger Ridge-Count Correlations. *American Journal of Physical Anthropology* 46: 171–176.
- 1977b The Genetic Significance of the Relationship between Total Finger Ridge Count and Its Variability. *Homo* 28: 1–11.
- 1987 Anthropological Dermatoglyphic Research. *Annual Review of Anthropology* 16: 161–177.
- 2021 Finger Ridge-Counts Correlate with the Second to Fourth Digit Ratio (2d:4d). *American Journal of Human Biology: The Official Journal of the Human Biology Council*: e23625.

Jantz, Richard L., and H Brehme

- 1993 Directional and Fluctuating Asymmetry in the Palmar Interdigital Ridge-Counts. *Anthropologischer Anzeiger* 51(1): 59–67.

Jantz, Richard L., H Brehme, and AW Eriksson

- 1992 Dermatoglyphic Variation among Finno-Ugric Speaking Populations: Methodological Alternatives. *American Journal of Physical Anthropology* 89: 1–10.

Jantz, Richard L., and V.P. Chopra

- 1983 A Comparison of Dermatoglyphic Methodologies in Population Studies. *American Journal of Physical Anthropology* 60: 61–67.

Jantz, Richard L., Fred K. Fohl, and James W. Zahler

- 1979 Finger Ridge-Counts and Handedness. *Human Biology* 51(1): 91–99.

Jantz, Richard L., and Cleone H. Hawkinson

1980 Components of Racial Variation in Finger Ridge-Counts. *American Journal of Physical Anthropology* 52: 139–144.

Jantz, Richard L., and Kenneth R. Parham

1978 Racial Differences in Dermal Ridge Breadth. *Human Biology* 50(1): 33–40.

Jost, Lou

2006 Entropy and Diversity. *Oikos* 113(2): 363–375.

2007 Partitioning Diversity into Independent Alpha and Beta Components. *Ecology* 88(10): 2427–2439.

Kalichman, Leonid, Valery Batsevich, and Eugene Kobylansky

2019 Heritability Estimation of 2D:4D Finger Ratio in a Chuvashian Population-Based Samples. *American Journal of Human Biology* 31: 1–5.

Karatzoglou, Alexandros, Alex Smola, Kurt Hornik, and Achim Zeileis

2004 Kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software* 11(9): 1–20.

Kassambara, Alboukadel

2017 Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. Edition 1. Multivariate Analysis, 1. Erscheinungsort nicht ermittelbar: STHDA.

2023 Ggpubr: “ggplot2” Based Publication Ready Plots. <https://cran.r-project.org/web/packages/ggpubr/>, accessed June 22, 2023.

Kassambara, Alboukadel, and Fabian Mundt

2020 Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. <https://CRAN.R-project.org/package=factoextra>.

Kobyliansky, Eugene, and G. Livshits

1986 Anthropometric Multivariate Structure and Dermatoglyphic Peculiarities in Biochemically and Morphologically Different Heterozygous Groups. *American Journal of Physical Anthropology* 70: 251–263.

Kondo, Shigeru

2017 An Updated Kernel-Based Turing Model for Studying the Mechanisms of Biological Pattern Formation. *Journal of Theoretical Biology* 414: 120–127.

Kowal, Emma, Laura S. Weyrich, Juan Manuel Argüelles, et al.

2023 Community Partnerships Are Fundamental to Ethical Ancient DNA Research. *Human Genetics and Genomics Advances* 4(2): 100161.

Kruschke, John K.

2015 Overview of the Generalized Linear Model. In *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd edition Pp. 419–448. Academic Press.

Krzanowski, W.J.

2000 Motivation and Fundamental Concepts. *In Principles of Multivariate Analysis: A User's Perspective*. Revised Edition. Pp. 3–32. Oxford University Press.

Kuhn, Max

2022 Caret: Classification and Regression Training. <https://CRAN.R-project.org/package=caret>.

Kvalseth, Tarald O.

1983 Note on the R² Measure of Goodness of Fit for Nonlinear Models. *Bulletin of the Psychonomic Society* 21(1): 79–80.

1985 Cautionary Note about R². *The American Statistician* 39(4). [American Statistical Association, Taylor & Francis, Ltd.]: 279–285.

Lander, Jared P.

2022 Coefplot: Plots Coefficients from Fitted Models. <https://cran.r-project.org/web/packages/coefplot/>, accessed June 22, 2023.

Larose, Daniel T., and Chantal D. Larose

2014 Model Evaluation Techniques. *In Discovering Knowledge in Data: An Introduction to Data Mining*. Somerset, UNITED STATES: John Wiley & Sons, Incorporated.

<http://ebookcentral.proquest.com/lib/utk/detail.action?docID=1699137>, accessed June 15, 2023.

Legendre, Pierre

1993 Spatial Autocorrelation: Trouble or New Paradigm? *Ecology* 74(6). Ecological Society of America: 1659–1673.

Legendre, Pierre, and Louis Legendre

2012a Multiscale Analysis: Spatial Eigenfunctions. *In Numerical Ecology*. 3rd edition Pp. 859–906. Oxford, Netherlands, The: Elsevier.

<http://ebookcentral.proquest.com/lib/utk/detail.action?docID=982554>, accessed July 6, 2021.

2012b Complex Ecological Datasets. *In Numerical Ecology*. 3rd edition Pp. 1–58. Oxford, Netherlands, The: Elsevier.

<http://ebookcentral.proquest.com/lib/utk/detail.action?docID=982554>, accessed July 6, 2021.

2012c Spatial Analysis. *In Numerical Ecology*. 3rd edition Pp. 785–858. Oxford, Netherlands, The: Elsevier. <http://ebookcentral.proquest.com/lib/utk/detail.action?docID=982554>, accessed July 6, 2021.

Leguebe, A.

1982 Biometric Methods for the Analysis of Dermatoglyphic Data. *Collegium Antropologicum* 6(2): 121–137.

Leguebe, A., S. Vrydagh, and J. Ducros

1981 Finger Ridge-Count Correlations: Race, Sex and Hand Differences. *Journal of Human Evolution* 10: 453–466.

- Leisch, Friedrich, and Evgenia Dimitriadou
2021 *MLbench: Machine Learning Benchmark Problems*.
- Li, Jinxi, James D. Glover, Haiguo Zhang, et al.
2022 *Limb Development Genes Underlie Variation in Human Fingerprint Patterns*. *Cell* 185(1): 95-112.e18.
- Lloyd, S.
1982 *Least Squares Quantization in PCM*. *IEEE Transactions on Information Theory* 28(2): 129–137.
- Loesch, Danuta Z., and Maureen Lafranchi
1990 *Relationships of Epidermal Ridge Patterns with Body Measurements and Their Possible Evolutionary Significance*. *American Journal of Physical Anthropology* 82: 183–189.
- Loesch, D.Z., and N.G. Martin
1984 *Finger Ridge Patterns and Tactile Sensitivity*. *Annals of Human Biology* 11(2). Taylor & Francis: 113–124.
- Lüdecke, Daniel, Dominique Makowski, Mattan S. Ben-Shachar, et al.
2023 *Performance: Assessment of Regression Models Performance*. <https://cran.r-project.org/web/packages/performance/>, accessed June 26, 2023.
- Lüdecke, Daniel, Dominique Makowski, Mattan S. Ben-Shachar, Indrajeet Patil, and Brenton M. Wiernik
2022 *Easystats: Framework for Easy Statistical Modeling, Visualization, and Reporting*. <https://cran.r-project.org/web/packages/easystats/>, accessed June 29, 2023.
- Lüdecke, Daniel, Dominique Makowski, Indrajeet Patil, et al.
2023 *See: Model Visualisation Toolbox for “easystats” and “Ggplot2.”* <https://cran.r-project.org/web/packages/see/>, accessed June 29, 2023.
- Maceo, Alice V.
2011 *Anatomy and Physiology of Adult Friction Ridge Skin*. In *The Fingerprint Sourcebook*. Alan McRoberts, ed. Pp. 3-3: 3-26. National Institute of Justice. <https://nij.ojp.gov/library/publications/fingerprint-sourcebook>.
- Machado, João Felipe, Paula Roquetti Fernandes, Ricardo Wagner Roquetti, and José Fernandes Filho
2010 *Digital Dermatoglyphic Heritability Differences as Evidenced by a Female Twin Study*. *Twin Research and Human Genetics* 13(5): 482–489.
- MacQueen, J.
1967 *Some Methods for Classification and Analysis of Multivariate Observations*. *Proceedings*

of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. University of California Press: 281–297.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik
2022 Cluster: Cluster Analysis Basics and Extensions. <https://CRAN.R-project.org/package=cluster>.

Manning, J. T., and B. Fink
2023 Digit Ratio (2D:4D) and Its Relationship to Foetal and Maternal Sex Steroids: A Mini-Review. *Early Human Development* 183: 105799.

Mao, Wentao, Xiaoxia Mu, Yanbin Zheng, and Guirong Yan
2014 Leave-One-out Cross-Validation-Based Model Selection for Multi-Input Multi-Output Support Vector Machine. *Neural Computing and Applications* 24(2): 441–451.

Marion, Zachary H., James A. Fordyce, and Benjamin M. Fitzpatrick
2015 Extending the Concept of Diversity Partitioning to Characterize Phenotypic Complexity. *The American Naturalist* 186(3): 348–361.

Massicotte, Philippe, Andy South, and Koen Hufkens
2023 Rnaturalearth: World Map Data from Natural Earth. <https://cran.r-project.org/web/packages/rnaturalearth/index.html>, accessed September 21, 2023.

Mbaka, Godwin, Adedayo Ejiwunmi, Olusegun Alabi, and Timothy Olatayo
2016 Digital Dermatoglyphic Variation and Migratory Pattern of Ethnic Liberians. *Egyptian Journal of Forensic Sciences* 6: 416–421.

McElreath, Richard
2020 Chapter 6: The Haunted DAG and the Causal Terror. *In* *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. 2nd edition Pp. 161–190. CRC Press. <https://www.routledge.com/Statistical-Rethinking-A-Bayesian-Course-with-Examples-in-R-and-STAN/McElreath/p/book/9780367139919>, accessed June 26, 2023.

Meier, Robert J.
1980 Anthropological Dermatoglyphics, a Review. *Yearbook of Physical Anthropology* 23: 147–178.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch
2022 E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. <https://CRAN.R-project.org/package=e1071>.

Miller, James R.
1973 Dermatoglyphics. *The Journal of Investigative Dermatology* 60(6): 435–442.

Miller, Maxine J.

1985 Turner Syndrome Dermatoglyphics: A Principal Components Analysis. Master's Thesis, University of Tennessee. https://trace.tennessee.edu/utk_gradthes/4195.

Morissette, Laurence, and Sylvain Chartier

2013 The K-Means Clustering Technique: General Considerations and Implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology* 9(1): 15–24.

Mouloudi, Saeed, Hadi Rahmanpanah, Soheil Gohari, et al.

2021 What Can Artificial Intelligence and Machine Learning Tell Us? A Review of Applications to Equine Biomechanical Research. *Journal of the Mechanical Behavior of Biomedical Materials* 123: 104728.

Murrell, Paul, and Zhijian Wen

2020 gridGraphics: Redraw Base Graphics Using “grid” Graphics. <https://cran.r-project.org/web/packages/gridGraphics/>, accessed June 22, 2023.

Nagelkerke, Nico J. D.

1991 A Note on a General Definition of the Coefficient of Determination. *Biometrika* 78: 691–692.

Nagy, Attila Sándor, and Miklós Pap

2004 Comparative Analysis of Dermatoglyphic Traits in Hungarian and Gypsy Populations. *Human Biology* 76(3): 383–400.

Neter, John

1996 *Applied Linear Statistical Models*. 4th ed. Chicago: Irwin.

Novembre, John, and Matthew Stephens

2008 Interpreting Principal Component Analyses of Spatial Population Genetic Variation. *Nature Genetics* 40(5): 1–4.

Oh, SeHee, MinDong Sung, Yumie Rhee, Namki Hong, and Yu Rang Park

2021 Evaluation of the Privacy Risks of Personal Health Identifiers and Quasi-Identifiers in a Distributed Research Network: Development and Validation Study. *JMIR Med Inform* 9(5): e24940.

Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, et al.

2022 *Vegan: Community Ecology Package*. <https://CRAN.R-project.org/package=vegan>.

Oliveira, Letitia Lowe

1978 Dermatoglyphics of University of Tennessee Students: Effects of Parental Age and Birth Order. PhD Dissertation, University of Tennessee. https://trace.tennessee.edu/utk_graddiss/4017.

Ousley, Stephen D.

1997 *The Quantitative Genetics of Epidermal Ridges Using Multivariate Maximum*

Likelihood Estimation. PhD Dissertation, University of Tennessee.
https://trace.tennessee.edu/utk_graddiss/7065.

Owsley, Douglas William

1978 Dermatoglyphic Variability and Asymmetry of Patients with Cleft Lip and Cleft Palate. PhD Dissertation, University of Tennessee. https://trace.tennessee.edu/utk_graddiss/4069.

Paliy, O., and V. Shankar

2016 Application of Multivariate Statistical Techniques in Microbial Ecology. *Molecular Ecology* 25(5): 1032–1057.

Pan, Yuchen, Zhi Xiao, Xianning Wang, and Daoli Yang

2017 A Multiple Support Vector Machine Approach to Stock Index Forecasting with Mixed Frequency Sampling. *Knowledge-Based Systems* 122: 90–102.

Paradis, E., and K. Schliep

2019 Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R. *Bioinformatics* 35: 526–528.

Pebesma, Edzer, Roger Bivand, Etienne Racine, et al.

2023 Sf: Simple Features for R. <https://cran.r-project.org/web/packages/sf/index.html>, accessed September 21, 2023.

Pedregosa, F., G. Varoquaux, A. Gramfort, et al.

2011 Scikit-Learn: Machine Learning in Python. *Sklearn.Dummy.DummyRegressor*. <https://scikit-learn/stable/modules/generated/sklearn.dummy.DummyRegressor.html>.

Pigliucci, Massimo

2010 Genotype-Phenotype Mapping and the End of the “genes as Blueprint” Metaphor. *Proceedings of the Royal Society B: Biological Sciences* 365: 557–566.

Plato, Chris C., James J. Cereghino, and Florence S. Steinberg

1975 The Dermatoglyphics of American Caucasians. *American Journal of Physical Anthropology* 42: 195–210.

Polcerová, Lenka, Richard L. Jantz, M. Králík, M. Chovancová, and Martin Čuta

2023 Sex Differences in Radioulnar Contrasts of the Finger Ridge Counts across 21 Human Population Samples. *Annals of Human Biology* 50(1). Taylor & Francis: 370–389.

R Core Team

2022 R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Reddy, B. Mohan, Dario A. Demarchi, S. Bharati, Vikrant Kumar, and Michael H. Crawford

2004 Patterns of Ethnic, Linguistic, and Geographic Heterogeneity of Palmar Interdigital Ridge Counts in the Indian Subcontinent. *Human Biology* 76(2): 211–228.

Rich, Benjamin

2023 Table1: Tables of Descriptive Statistics in HTML. <https://CRAN.R-project.org/package=table1>.

Ripley, Brian, Bill Venables, Douglas M. Bates, et al.

2023a MASS: Support Functions and Datasets for Venables and Ripley's MASS. <https://cran.r-project.org/web/packages/MASS/index.html>, accessed September 21, 2023.

2023b MASS: Support Functions and Datasets for Venables and Ripley's MASS. <https://cran.r-project.org/web/packages/MASS/>, accessed June 22, 2023.

Rissanen, J.

1978 Modeling by Shortest Data Description. *Automatica* 14(5): 465–471.

Rissanen, Jorma

1983 A Universal Prior for Integers and Estimation by Minimum Description Length. *Annals of Statistics* 11(2). Institute of Mathematical Statistics: 416–431.

Roseman, Charles C., and Timothy D. Weaver

2007 Molecules versus Morphology? Not for the Human Cranium. *BioEssays* 29: 1185–1188.

Rosenberg, Noah A., Jonathan K. Pritchard, James L. Weber, et al.

2002 Genetic Structure of Human Populations. *Science* 298: 2381–2385.

Sanchez-Fernandez, M., M. de-Prado-Cumplido, J. Arenas-Garcia, and F. Perez-Cruz

2004 SVM Multiregression for Nonlinear Channel Estimation in Multiple-Input Multiple-Output Systems. *IEEE Transactions on Signal Processing* 52(8): 2298–2307.

Schmidt, Amand F., and Chris Finan

2018 Linear Regression and the Normality Assumption. *Journal of Clinical Epidemiology* 98: 146–151.

Schneider, Astrid, Gerhard Hommel, and Maria Blettner

2010 Linear Regression Analysis. *Deutsches Ärzteblatt International* 107(44): 776–782.

Schrider, Daniel R., and Andrew D. Kern

2018 Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics* 34(4). Elsevier: 301–312.

SE, SAP

N.d. Durbin-Watson. Text. SAP SE, Dietmar-Hopp-Allee 16, 69190 Walldorf, Germany. https://help.sap.com/saphelp_SCM700_ehp02/helpdata/en/e8/3cc95360267614e10000000a174cb4/content.htm?no_cache=true, accessed July 7, 2023.

Sefidian, Amir Masoud

2022 A Guide on Regression Error Metrics (MSE, RMSE, MAE, MAPE, sMAPE, MPE) with

Python Code. Amir Masoud Sefidian - Sefidian Academy. <https://sefidian.com/2022/08/18/a-guide-on-regression-error-metrics-with-python-code/>, accessed June 30, 2023.

Sengupta, Asok bijoy, Badruddin Ahmed Bazmi, Subrata Sarkar, et al.
2013 A Cross Sectional Study of Dermatoglyphics and Dental Caries in Bengalee Children. Journal of Indian Society of Pedodontics and Preventive Dentistry 31(4): 245.

Sengupta, M., and B. Karmakar
2006 Genetics of Dermatoglyphic Asymmetry in Vaidyas of West Bengal, India. Human Biology 78(2): 199–214.

Shmueli, Galit
2010 To Explain or to Predict? Statistical Science 25(3). Institute of Mathematical Statistics: 289–310.

Siberchicot, Aurélie, Alice Julien-Laferrrière, Anne-Béatrice Dufour, Jean Thioulouse, and Stéphane Dray
2017 Adegraphics: An S4 Lattice-Based Package for the Representation of Multivariate Data. The R Journal 9(2): 198–212.

Siberchicot, Stéphane Dray and Aurélie, and with contributions from Jean Thioulouse Based on earlier work by Alice Julien-Laferrrière
2023 Adegraphics: An S4 Lattice-Based Package for the Representation of Multivariate Data. <https://cran.r-project.org/web/packages/adegraphics/>, accessed June 22, 2023.

Smola, Alex J., and Bernhard Schölkopf
2004 A Tutorial on Support Vector Regression. Statistics and Computing 14(3): 199–222.

Sokal, Robert R., Rosalind M. Harding, and Neal L. Oden
1989 Spatial Patterns of Human Gene Frequencies in Europe. American Journal of Physical Anthropology 80: 267–294.

Sokal, Robert R., Richard L. Jantz, and Barbara A. Thomson
1996 Dermatoglyphic Variation in Europe. American Journal of Physical Anthropology 100: 35–47.

Sokal, Robert R., and Paolo Menozzi
1982 Spatial Autocorrelations of HLA Frequencies in Europe Support Demic Diffusion of Early Farmers. The American Naturalist 119(1): 1–17.

South, Andy
2017 Rnaturalearthdata: World Vector Map Data from Natural Earth Used in “Rnaturalearth.” <https://cran.r-project.org/web/packages/rnaturalearthdata/index.html>, accessed September 21, 2023.

Spiess, Andrej-Nikolai, and Natalie Neumeyer

2010 An Evaluation of R² as an Inadequate Measure for Nonlinear Models in Pharmacological and Biochemical Research: A Monte Carlo Approach. *BMC Pharmacology* 10(1): 6.

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar

2019a Introduction. *In* Introduction to Data Mining, 2nd edition Pp. 1–22. Pearson.

2019b Data. *In* Introduction to Data Mining, 2nd edition Pp. 23–112. Pearson.

2019c Cluster Analysis: Basic Concepts and Algorithms. *In* Introduction to Data Mining, 2nd edition Pp. 525–612. Pearson.

2019d Classification: Basic Concepts and Techniques. *In* Introduction to Data Mining, 2nd edition Pp. 113–192. Pearson.

Tao, Qiqing, Koichiro Tamura, Fabia U. Battistuzzo, and Sudhir Kumar

2019 A Machine Learning Method for Detecting Autocorrelation of Evolutionary Rates in Large Phylogenies. *Molecular Biology and Evolution* 36(4): 811–824.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie

2001 Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society B* 63(2): 411–423.

Tuia, Devis, Jochem Verrelst, Luis Alonso, Fernando Perez-Cruz, and Gustavo Camps-Valls

2011 Multioutput Support Vector Regression for Remote Sensing Biophysical Parameter Estimation. *IEEE Geoscience and Remote Sensing Letters* 8(4): 804–808.

Turing, Alan

1952 The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 237(641): 37–72.

Veenstra, Timothy D.

2021 Omics in Systems Biology: Current Progress and Future Outlook. *PROTEOMICS* 21(3–4): 2000235.

Wertheim, Kasey

2011 Embryology and Morphology of Friction Ridge Skin. *In* The Fingerprint Sourcebook. Alan McRoberts, ed. Pp. 3-3: 3-26. National Institute of Justice.

<https://nij.ojp.gov/library/publications/fingerprint-sourcebook>.

Wickham, Hadley

2016 Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

<https://ggplot2.tidyverse.org>.

2020 Reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package. <https://cran.r-project.org/web/packages/reshape2/index.html>, accessed September 21, 2023.

Wickham, Hadley, Mara Averick, Jennifer Bryan, et al.

2019 Welcome to the Tidyverse. *Journal of Open Source Software* 4(43): 1686.

- Wickham, Hadley, Jim Hester, and Jennifer Bryan
2022 Readr: Read Rectangular Text Data. <https://CRAN.R-project.org/package=readr>.
- Wickham, Hadley, and RStudio
2023 Tidyverse: Easily Install and Load the “Tidyverse.” <https://cran.r-project.org/web/packages/tidyverse/>, accessed June 22, 2023.
- Wickham, Hadley, Dana Seidel, and RStudio
2022 Scales: Scale Functions for Visualization. <https://cran.r-project.org/web/packages/scales/>, accessed June 22, 2023.
- Wijerathne, Buddhika, Geetha K. Rathnayake, Shamila C. Adikari, et al.
2013 Sexual Dimorphism in Digital Dermatoglyphic Traits among Sinhalese People in Sri Lanka. *Journal of Physiological Anthropology* 32(27): 1–9.
- Wilke, Claus O.
2020 Cowplot: Streamlined Plot Theme and Plot Annotations for “Ggplot2.” <https://CRAN.R-project.org/package=cowplot>.
- Wu, Xindong, Vipin Kumar, J. Ross Quinlan, et al.
2008 Top 10 Algorithms in Data Mining. *Knowledge and Information Systems* 14(1): 1–37.
- Xing, Li, Mary L. Lesperance, and Xuekui Zhang
2020 Simultaneous Prediction of Multiple Outcomes Using Revised Stacking Algorithms. *Bioinformatics (Oxford, England)* 36(1): 65–72.
- Xu, Shuo, Xin An, Xiaodong Qiao, Lijun Zhu, and Lin Li
2013 Multi-Output Least-Squares Support Vector Regression Machines. *Pattern Recognition Letters* 34(9): 1078–1084.
- Yan, Yachen
2016 MLmetrics: Machine Learning Evaluation Metrics. <https://cran.r-project.org/web/packages/MLmetrics/>, accessed June 22, 2023.
- Yokota, Miyo
1997 Biological Relationships among Siberians: Craniometric, Serological, and Dermatoglyphic Approaches. PhD Dissertation, University of Tennessee. https://trace.tennessee.edu/utk_graddiss/4032.
- Zhang, Hai-Guo, Yao-Fong Chen, Ming Ding, et al.
2010 Dermatoglyphics from All Chinese Ethnic Groups Reveal Geographic Patterning. *PLoS ONE* 5(1): 1–12.
- Zhang, Xia, Zhaolong Ma, and Guimin Lv
2021 Applying and Assessing Multi-Output Support Vector Regression with Rainfall as

Additional Output for Monthly River Flow Forecasting. *Arabian Journal of Geosciences* 13(24): 1299.

Zhang, Zhongheng

2016 Model Building Strategy for Logistic Regression: Purposeful Selection. *Annals of Translational Medicine* 4(6): 111.

Zhu, Hao, Thomas Travison, Timothy Tsai, et al.

2021 kableExtra: Construct Complex Table with “kable” and Pipe Syntax. <https://cran.r-project.org/web/packages/kableExtra/>, accessed June 22, 2023.

APPENDICES

Appendix A Supplementary information and table, Chapter One

Common goodness-of-fit measures like the R^2 value depend on the number of explanatory variables in the model, and the R^2 value will automatically improve (increase) as more explanatory variables are added to the model, even if the variables consist of random noise (Kvalseth 1985). Further, with categorical explanatory or response variables or counts, the R^2 value may decrease compared to solely continuous numeric variables. If variables are not random, the R^2 value can be inflated, and models based on aggregate data like populations generally exhibit higher R^2 values than based on smaller groupings.

The R^2 goodness-of-fit metric, also called coefficient of determination, represents the proportion of variance in the response variable that is explained by the linear relationship between the explanatory variables and the response variables in the model and can be considered a model fit statistic and an effect size (Larose and Larose 2014:280). The R^2 value is calculated as one minus the ratio of the sums of squares of the residuals not captured by the model to the total sums of squares. The residual sums of squares are calculated as the squared difference for every individual between the observed values and the predicted values. The total sums of squares are calculated as the squared difference between the observed value for every individual and the mean of the observed values. Where $\bar{o}$ is the mean of the observed values,

$$\bar{o} = \frac{1}{n} \sum_{i=1}^n o_i$$

and the R^2 is defined as,

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

then the

$$R^2 = 1 - \frac{\sum_i (o_i - p_i)^2}{\sum_i (o_i - \bar{o})^2}$$

in terms of observed and predicted values.

The R^2 value usually ranges between zero and one with a higher R^2 value indicating better model fit, because more variance in the response variable is explained by the explanatory variables. Negative R^2 values can arise when the sums of squares of the residuals in the numerator is greater than the total sums of squares of the original response variable in the denominator. The negative R^2 values can be a result when the predicted values were derived from a different model-fitting procedure than the corresponding outcome observed values or when outliers may be driving the poor fit of the model. Should a negative R^2 value arise, the mean of the data fits the observed data better than the values predicted by the model, and the explanatory variables do not explain any additional variance in the response variable. The R^2 value may also be negative when linear regression is performed without including an intercept or when non-linear functions were used to fit the data (Kvalseth 1985).

To compare different model fit using R^2 , the models should be trained on the same data sets and tested on the same data sets, and the training and test sets should be split the same with the same pre-processing steps. The R^2 value can be influenced by specific characteristics of the data set and the model such that a high R^2 value can arise from a model overfit on the training data and that does not generalize well to new data, and a low R^2 can arise from overly simple models that underfit the training data but generalize well to new data. With a very high R^2 value, simplifying a model by removing redundant or uninformative explanatory variables or by using regularization techniques like LASSO to shrink coefficients of less informative explanatory variables toward zero can improve the generalization of the model to new data. High R^2 values when the model was applied to the test data can also occur if the test data size was small and not

representative of the population. When using the R^2 value to compare the fit of different models, the complexity of the model as well as the generalization to new data should be incorporated into the interpretation.

The coefficient of determination, R^2 , is not appropriate for non-linear models like Poisson and negative binomial (Kvalseth 1983; Spiess and Neumeyer 2010). Even though the R^2 value should not be calculated for nonlinear models, commercial statistical packages will still calculate the R^2 value for nonlinear models, often resulting in a negative R^2 value. A core assumption of the R^2 value is normal distribution of the errors, which may not hold for count data. Further, for zero-inflated models the R^2 value may not capture the zero-inflation component. Pseudo R^2 values can be used for assessing goodness-of-fit in nonlinear models and models with count data, but where the R^2 value indicates the proportion of variation in the response variable explained by the explanatory variables, pseudo R^2 values cannot be interpreted so clearly. Not all pseudo R^2 values reach a true maximum of one and have to be adjusted, and pseudo R^2 versions differ in their focus such as on the residuals or the log likelihood, so different versions of pseudo R^2 are not interchangeable or directly comparable (Cox and Snell 1989; Nagelkerke 1991).

The traditional R^2 value can be interpreted as improvement from the null model to the fitted model. The null model, or the intercept-only model without any explanatory variables, is represented in the R^2 equation denominator as the total sum of squares, calculated as the squared difference between the observed value for every individual and the mean of the observed values. The fitted model includes the explanatory variables and is represented in the R^2 equation numerator as the residual sum of squares, calculated as the squared difference for every individual between the observed values and each prediction of the response variable. Thus, the R^2 equation of one minus this ratio represents how much better the explanatory variables

estimate the variation in the response variable than just estimating the mean of the response variable, and a higher numerator results in a smaller ratio and overall higher R^2 value.

Table S1. Abbreviations used throughout this paper.

Abbreviation	Description
ARC	Absolute ridge-count
ANOVA	Analysis of variance
BMP	bone morphogenic protein
CTNNB1	catenin beta-1
DA.kmeans.2	cluster assignments from each k-means clustering attempt on Directional Asymmetry variables, based on an optimal $K = 2$
DA.kmeans.4	cluster assignments from each k-means clustering attempt on Directional Asymmetry variables, based on a dictated $K = 4$
DA.kmeans.10	cluster assignments from each k-means clustering attempt on Directional Asymmetry variables, based on an optimal $K = 10$
DA.kmeans.322	cluster assignments from each k-means clustering attempt on Directional Asymmetry variables, based on a dictated $K = 322$
dbMEM	Distance-based Moran's eigenvector map
DFA	discriminant function analysis
DKK2	Dickkopf Wnt signaling pathway inhibitor 2 gene
DOI	digital object identifier
EDAR	enhanced ectodysplasin-A receptor
EGFR	epidermal growth factor receptor
ETHGROUP	Ethnic group
FA.kmeans.2	cluster assignments from each k-means clustering attempt on Fluctuating Asymmetry variables, based on an optimal $K = 2$
FA.kmeans.4	cluster assignments from each k-means clustering attempt on Fluctuating Asymmetry variables, based on a dictated $K = 4$
FA.kmeans.8	cluster assignments from each k-means clustering attempt on Fluctuating Asymmetry variables, based on an optimal $K = 8$
FA.kmeans.322	cluster assignments from each k-means clustering attempt on Fluctuating Asymmetry variables, based on a dictated $K = 322$
FGF20	follicle growth factor 20
frc.kmeans.2	cluster assignments from each k-means clustering attempt on finger ridge-counts, based on an optimal $K = 2$
frc.kmeans.4	cluster assignments from each k-means clustering attempt on finger ridge-counts, based on a dictated $K = 4$
frc.kmeans.10	cluster assignments from each k-means clustering attempt on finger ridge-counts, based on an optimal $K = 10$
frc.kmeans.16	cluster assignments from each k-means clustering attempt on finger ridge-counts, based on an optimal $K = 16$
frc.kmeans.322	cluster assignments from each k-means clustering attempt on finger ridge-counts, based on a dictated $K = 322$
GLM	Generalized linear model
HLA	human leukocyte antigen
IDNO	Identification number
INVESTLOC	Investigator location

Table S1 Continued

Abbreviation	Description
ISAAC	Infrastructure for Scientific Applications and Advanced Computing
IQR	Interquartile range
<i>K</i>	number of clusters in k-means clusters
LAT	Latitude
LEF1	lymphoid enhancer-binding factor 1
Lfp1	Left finger pattern digit 1
Lfp2	Left finger pattern digit 2
Lfp3	Left finger pattern digit 3
Lfp4	Left finger pattern digit 4
Lfp5	Left finger pattern digit 5
LONG	Longitude
Lrc1	Left radial count digit 1
Lrc2	Left radial count digit 2
Lrc3	Left radial count digit 3
Lrc4	Left radial count digit 4
Lrc5	Left radial count digit 5
LRP4	low density lipoprotein receptor-related protein 4
Luc1	Left ulnar count digit 1
Luc2	Left ulnar count digit 2
Luc3	Left ulnar count digit 3
Luc4	Left ulnar count digit 4
Luc5	Left ulnar count digit 5
MAE	Mean absolute error
MAPE	Mean absolute percentage error
<i>N</i>	sample size, number of individuals in a sample
NA	Not available
NaN	Not a number
PCA	Principal component analysis
PCNM	Principal coordinates of neighbor matrices
PII	Pattern intensity index
RDA	Redundancy analysis
Rfp1	Right finger pattern digit 1
Rfp2	Right finger pattern digit 2
Rfp3	Right finger pattern digit 3
Rfp4	Right finger pattern digit 4
Rfp5	Right finger pattern digit 5
RMSE	Root mean squared error
Rrc1	Right radial count digit 1
Rrc2	Right radial count digit 2
Rrc3	Right radial count digit 3
Rrc4	Right radial count digit 4

Table S1 Continued

Abbreviation	Description
Rrc5	Right radial count digit 5
RSPO2	R-spondin 2
Ruc1	Right ulnar count digit 1
Ruc2	Right ulnar count digit 2
Ruc3	Right ulnar count digit 3
Ruc4	Right ulnar count digit 4
Ruc5	Right ulnar count digit 5
SAMPLEDECL	Sample declaration
SAMPLENO	Sample number
<i>Shh</i>	sonic hedgehog protein
SVR	Support vector regression
TGF α	transforming growth factor-alpha
TGFA	transforming growth factor alpha gene
TRC	Total ridge-count

Appendix B Supplementary information and tables, Chapter Three

Table S2. Abbreviation, Sample Number, Ethnic Group, and Region present in overall data set.

Rows indicates the first row in the overall data set for that set of attributes

Rows	Abbreviation	Sample Number	Ethnic Group	Region
1	AAB-1	43-	Rembranga	AUS
150	AAB-2	43*	Maningrida	AUS
442	ABE	165	Azerbaijanis	CAU
647	ABK	358	Abenaki	NAM
748	ABZ	205	Abazins-Tapanta	CAU
832	ADI	194	Andians	CAU
972	ADJ	204	Adjarians	CAU
1116	ADY	203	Adygs	CAU
1524	AEB	277	Aeta-Bataan	INS
1575	AEC	278	Aeta-Camarines	INS
1607	AEZ	279	Aeta-Zambales	INS
1705	AGU	207	Aguls	CAU
1853	AIG	54-	Aigram-Ayom	MEL
1884	AIN	-56	Ainu	EAS
1986	AIN-2	385	Ainu	EAS
2163	AKA	49-	Aka	AFR
2358	AKB	47-	Aka	AFR
2365	AKH	147	Abkhazians	CAU
2509	AKZ	174	Altai-Kizi	SIB
2658	ALE	154	Aleuts	SIB
2707	AOY	66-	Tandroy	AFR
2749	ARA-1	166	Arabs	NEA
2900	ARA-2	167	Arabs	NEA
3022	ARM-2	208	Armenians	CAU
3190	ARM-E	*54	Armenians	ASI
3267	ARM-W	-54	Armenians	ASI
3418	ASS	164	Assyrians	CAU
3526	ASY	65-	Antanosy	AFR
3572	ATA	343	Atacama	SAM
3701	ATH	-7	Athabaskan-Indian	NAM
3767	AUJ	169	Iyengars	ASI
3827	AVA	206	Avars	CAU
4011	BAB	27-	Babongo	AFR
4338	BAG	-97	Baga-Binari	AFR
4513	BAK	-59	Baka	AFR
4776	BAM	12-	Bamileke	AFR
4993	BAO-1	26-	Baoule	AFR
5815	BAS	100	Asua	AFR

Table S2 Continued

Rows	Abbreviation	Sample Number	Ethnic Group	Region
5940	BDJ	157	Badjuan-Pamir	CUS
6012	BEO	68-	Betsileo	AFR
6069	BHI	198	Bhils	ASI
6135	BHP	199	Bhils-Pavra	ASI
6183	BIE	-60	Bieno	AFR
6329	BKI	117	Bashkirs	USR
6395	BKO	-79	Bakola	AFR
6628	BKR	210	Balkars	CAU
6822	BLU	183	Balochi	CUS
6956	BOR	303	Borana	AFR
7108	BOZ	-77	Bozo	AFR
7242	BRE	355	Bretons	EUR
7479	BRJ	129	Buryats	SIB
7688	BSQ	249	Basques	EUR
7937	BSS	116	Besermians	USR
8015	BTY	238	Khakass-Beltyr	SIB
8155	BUL	83-	Bulgarians	USR
8287	BUN	36-	Papuans-Bundi	MEL
8674	BUR	308	Burji	AFR
8791	BUS	-94	!Kung	AFR
8912	BUS-2	387	!Kung	AFR
9150	BWA	-47	Bwaka	AFR
9472	BYE	82-	Belarusian	USR
9616	CAI	328	Kaingang	SAM
9683	CAR	37-	Karina-Carib	SAM
9709	CAS	-52	Cashinahua	SAM
9875	CCH	211	Chechens	CAU
10026	CHD-1	133	Chinese-Dungans	EAS
10113	CHD-2	134	Chinese-Dungans	EAS
10310	CHE	42-	Cherokee	NAM
10358	CHL	212	Chelkans	SIB
10434	CHO	-62	Chokwe	AFR
10618	CHU	120	Chuckchees	SIB
10818	COL	-73	Coloureds	AFR
11025	COR	357	Cornish	EUR
11275	CRO	232	Croatians	EUR
11480	CUA	28-	Cuanhama	AFR
11693	CUV	114	Chuvashes	USR
11831	CZE	330	Czechs	EUR
12071	DAL	283	Croatians	EUR

Table S2 Continued

Rows	Abbreviation	Sample Number	Ethnic Group	Region
12311	DAN	381	Danes	EUR
12485	DAR	213	Dargins	CAU
12623	DBA	98-	Baltic-Germans	USR
12736	DHU	179	Hungary-Germans	EUR
13236	DOG-2	14-	Dogon	AFR
13764	DOM	112	Magyar-Slovak	EUR
13929	DOT	119	Dolgans	SIB
14027	DUT-1	-96	Dutch	EUR
14248	DYO	13-	Jola	AFR
14497	E-DU	335	Germans	EUR
14701	E-KL	286	Germans	EUR
14920	E-LZ	2	Germans	EUR
15111	E-MC	363	Bavarians	EUR
15296	E-POW	349	Prussian-West-Pomeranians	EUR
15321	E-SI	351	Silesian-Germans	EUR
15363	E-TH	352	Thurungii	EUR
15476	E-WF	350	Westphalian	EUR
15521	EFE	-78	Efe	AFR
15723	EFE-2	50-	Efe	AFR
15842	EGY-1	31-	Egyptians-Arab	NEA
15946	EGY-2	31*	Egyptians	NEA
16052	EIS	74-	Easter-Island	OCE
16091	EKI-1	240	Evenk-Tunguses	SIB
16269	EKI-2	241	Evenk-Tunguses	SIB
16470	ELD	-90	British	EUR
17016	ENT	128	Entsi-Tundra	SIB
17088	EPA	75-	Easter-Island-Polynesians	OCE
17198	ESA	76-	Aappilattoq	NAM
17308	ESB	246	Inupiat-Barrow	NAM
17405	ESE	356	Ammassaliks	NAM
17600	ESF	250	Inuit-Kuujuaq	NAM
17789	ESH	247	Tikigagmiut-Hope	NAM
17870	ESK	-8	Alatna	NAM
17899	ESM	-9	Allakaket	NAM
17914	ESS	121	Chaplino-Yupik	SIB
18048	EST-1	-81	Estonians	USR
18290	EST-2	161	Estonians	USR
18428	ESU	77-	Upernavik	NAM
18527	ESW	248	Inupiat-Wainwright	NAM
18609	EVS-1	242	Evens	SIB

Table S2 Continued

Rows	Abbreviation	Sample Number	Ethnic Group	Region
18750	EVS-2	260	Evens	SIB
18923	FAL-1	15-	Fali-Toro	AFR
19377	FAL-2	15-	Fali-Kangou	AFR
19765	FAL-3	15-	Fali-Tinguelin	AFR
19838	FAO	21-	Faroese	EUR
20133	FIN	79-	Sami	EUR
20651	FLO	-58	Flores	INS
20841	FOU	16-	Fula	AFR
20960	FRA-3	340	French	EUR
21151	FRI	334	Frisians-East	EUR
21268	FU-T 3	-1	Germans	EUR
21431	GAR	56-	Garmend-Ayom	MEL
21445	GEO	173	Georgians	CAU
21591	GGZ	86-	Gagauz	USR
21711	GHI	-44	Walloon	EUR
21951	GIN	-63	Ginga	AFR
22139	GKU	390	Gciriku	AFR
22163	GLA	139	Giliaks	SIB
22256	GLS-1	140	Giliaks	SIB
22395	GLS-2	141	Giliaks	SIB
22498	GRE	-80	Greeks	EUR
22842	GRE-2	325	Greeks	EUR
22927	GRI	57-	Papuans-Giringiri	MEL
22965	GUA	329	Guarani	SAM
23015	GUR	233	Gurians	CAU
23119	HAG	-95	Hagahai	MEL
23243	HAN	-5	US-Whites	NAM
23334	HEH	-93	Hehe	AFR
23534	HRU	-89	Mamusi	MEL
23582	HUN	72-	Hungarians	EUR
23852	ICE	137	Icelanders	EUR
24043	IGU	214	Ingush	CAU
24197	IKU	153	Yezidi-Kurds	CAU
24272	ILO	275	Ilocano	INS
24292	IRE-1	326	Irish	EUR
24586	IRE-2	327	Irish	EUR
24622	IRU	135	Iranians	CUS
24795	ITI	376	Meru-Imenti	AFR
24895	ITK	163	Itelmen	SIB
25008	ITR	253	Italians	EUR

Table S2 Continued

Rows	Abbreviation	Sample Number	Ethnic Group	Region
25167	JAP-1	-72	Japanese	EAS
26149	JAP-3	386	Japanese	EAS
26347	JAV	271	Javanese	INS
26432	JBU	188	Bukharan-Jews	CUS
26619	JIV	324	Jivaro	SAM
26859	JNA	189	Kabardino-Balkar-Jews	CAU
26992	KAB	215	Kabardinians	CAU
27356	KAC	152	Kachins	SIB
27515	KAM	236	Kalmyks	CUS
27673	KAN	63-	Kana	MEL
27725	KAR	30-	Karen	EAS
27824	KAS	185	Kazakhs	SIB
27954	KCH	218	Karachay	CAU
28098	KER	244	Kerema	MEL
28187	KET	138	Kets	SIB
28277	KEY	313	Elgeyo	AFR
28377	KHA	127	Khanty	SIB
28432	KIA	365	Kikuyu-Kiambu	AFR
28536	KIR	10-	Cameroon-Bata	AFR
28715	KKK	182	Koryaks	SIB
28879	KKP	186	Karakalpaks	SIB
28973	KMD	175	Kumandins	SIB
29079	KNY	305	Sabaot-Kony	AFR
29139	KOB	336	Kobon	MEL
29154	KOK	346	Aland-Kokar	EUR
29238	KOP	115	Komi-Permyak	USR
29357	KOR-1	131	Koreans-North	EAS
29507	KOR-2	132	Koreans-North	EAS
29726	KOZ	122	Komi-Zyrians	USR
29872	KPS	315	Kipsigis	AFR
29969	KRD	184	Kurds	CUS
30065	KRG-1	187	Kyrgyz	SIB
30319	KRG-2	187	Kyrgyz	SIB
30371	KRL	85-	Karelians	USR
30516	KSS-1	93-	NE-Finns	EUR
30636	KSS-2	94-	NE-Finns	EUR
30781	KSS-3	95-	NE-Finns	EUR
30834	KUM	234	Kumyks	CAU
30992	KUR	18-	Kurumba-Fulse	AFR
31272	KWA	388	Kwangali	AFR

Table S2 Continued

Rows	Abbreviation	Sample Number	Ethnic Group	Region
31419	KYS	191	Kyzyl	SIB
57584	LAD	-64	Urban Blacks	AFR
31480	LAK	88-	Lapps	USR
31626	LAT	-82	Latvians	USR
31960	LAU	-75	Transvaal-Indians	ASI
32148	LAZ	219	Lazi	CAU
32293	LEM	46-	Mbuti	AFR
32326	LES	48-	Lese	AFR
32416	LGY	118	Lingayats	ASI
32466	LIT-1	-83	Lithuanians	USR
32669	LIT-2	160	Lithuanians	USR
32831	LKI	220	Laks	CAU
32967	LUI	-61	Luimbe	AFR
33132	MAM	11-	Mama	AFR
33322	MAN	51-	Mangbetu	AFR
33359	MAR-1	97-	Mari	USR
33457	MAR-2	96-	Mari	USR
33562	MAR-3	78-	Mari	USR
33668	MAR-4	113	Mari	USR
33824	MAY	272	Malaysians	INS
33849	MBK	392	Mbukushu	AFR
33853	MBU	389	Mbunza	AFR
33879	MCU	291	Makua	AFR
34056	MDB	243	Mandobo-North	MEL
34227	MEO	29-	Hmong	EAS
57479	MGE	221	Megrelians	CAU
34322	MIX	-53	Mixtecs	MES
34492	MKT	314	Marakwet	AFR
34597	MLS	341	Managalase	MEL
34646	MOM	59-	Mombasob-Ayom	MEL
34722	MOR	-46	Morrocans	AFR
34977	MOT	245	Motu	MEL
35624	MOV	87-	Moldovans	USR
35732	MRD	89-	Mordvins	USR
35882	MRV	262	Moravians	EUR
36082	MSI	143	Mansi	SIB
36198	MSS-1	310	Maasai-Kajiado	AFR
36310	MSS-2	311	Maasai-Kajiado	AFR
36406	MSS-3	312	Maasai-Narok	AFR
36450	MTH	223	Marathas	ASI

Table S2 Continued

Rows	Abbreviation	Sample Number	Ethnic Group	Region
36501	MTM	222	Mohevians	CAU
36645	MUB	373	Mwimbe-Muthambi	AFR
36803	MUK	316	Mukogodo	AFR
36851	MUS	342	Middle-Musa	MEL
37041	MYU	243	Muyu-Papuan	MEL
37130	NAN	235	Nanai	SIB
37267	NAS	254	Naskapi	NAM
37377	NDA	292	Ndau	AFR
37415	NDB	-66	Ndebele	AFR
37479	NDI	304	Nandi	AFR
37581	NEB	263	Brahmins	ASI
37631	NEF	144	Nentsi-Forest	SIB
37741	NEG	323	Nepalese-Gurkhas	ASI
38068	NET	209	Nentsi-Tundra	SIB
38203	NGA	126	Nganasans	SIB
38297	NGJ	-55	Amele	MEL
38493	NGW	317	Laikipiak-Ngwesi	AFR
38577	NHU	293	Nhungue	AFR
38610	NJA	294	Njanja	AFR
38645	NOR	364	Norwegians	EUR
38813	NSF-1	101	Lapps	EUR
38938	NSF-2	102	Lapps	EUR
39019	NSF-3	103	Lapps	EUR
39134	NSF-4	104	Lapps	EUR
39182	NSF-5	105	Lapps	EUR
39222	NUB-1	268	Nubian-Fedicca-Mohas	AFR
39377	NUB-2	269	Nubian-Arabized	AFR
39539	NUB-3	270	Nubian-Kunuz	AFR
39635	OES-2	44-	Austrians	EUR
40675	ORK	339	Orcadians	EUR
41320	ORO	201	Orochis	SIB
41392	OSS-1	224	Ossetians-North	CAU
41488	OSS-2	225	Ossetians-South	CAU
41541	OST	-43	Flemish	EUR
41938	PAP	361	Papuans	MEL
41706	PAP-1	32-	Papuans-Sandaun	MEL
41728	PAP-2	243	Mandobo-South	MEL
42006	PBO	197	Parsis	ASI
42068	PED	-65	Pedi	AFR
42245	PIN	111	Pinai	MEL

Table S2 Continued

Rows	Abbreviation	Sample Number	Ethnic Group	Region
42330	POK	307	Pokot	AFR
42426	POL-1	-99	Poles	EUR
42930	POL-3	259	Poles	EUR
43028	PSV	193	Pshavs	CAU
43074	QHL	252	Quechua-Lowland	SAM
43248	QUH	251	Quechua-Highland	SAM
43435	REH	-74	Rehoboth-Baster	AFR
43786	RHO	295	Ronga	AFR
43952	RND	302	Rendille	AFR
44102	ROS	258	Roskolniki	EUR
44317	RUS	80-	Russians	USR
44569	SAA	354	Sara	AFR
44709	SAG	151	Sagays	SIB
44841	SAN	-48	Sango	AFR
44842	SAR	332	Sardinians	EUR
45127	SAS	20-	Samo-South	AFR
45334	SAW	60-	Sawaiab-Ayom	MEL
45358	SBU	309	Samburu-Loikop	AFR
45460	SCB	-4	Germans	EUR
46020	SCO	338	Scots	EUR
46176	SDN	70-	Sakalava-North	AFR
46232	SEL	150	Selkups	SIB
46290	SEN	296	Sena	AFR
46338	SER	230	Serbians	EUR
46744	SHA	-69	Shangaan	AFR
46927	SHA-2	297	Shangaan	AFR
47004	SHN	298	Shona-Manyika	AFR
47030	SHO	177	Shors	SIB
47174	SHU	156	Shugnans	CUS
47258	SLG-1	255	Romani-Vinodol	EUR
47329	SLG-2	256	Romani-Gemer	EUR
47380	SLG-3	261	Romani-Czech	EUR
47449	SLO	257	Slovaks	EUR
47785	SOK	180	Croatians	EUR
47961	SOM	301	Somali	AFR
48192	SON	-84	Son-Koli	ASI
48513	SPA	-45	Spanish	EUR
48718	SPL	344	Spanish	EUR
48917	SPT	285	Spanish	EUR
49117	STL	168	Santhals	ASI

Table S2 Continued

Rows	Abbreviation	Sample Number	Ethnic Group	Region
49173	SUK	-91	Sukuma	AFR
49227	SUR	231	Trio-Indians	SAM
49303	SWE	345	Aland-Mariehamm	EUR
49453	SYU	391	Shambyu	AFR
49495	TAA	227	Tabasarans	CAU
49615	TAD	145	Tadjiks	CUS
49832	TAK	124	Crimean-Tatars	USR
49979	TAM	190	Tatars	USR
50040	TAS	192	Baraba-Tatars	SIB
50179	TAV	123	Volga-Tatars	USR
50372	TCH	228	Cherkess	CAU
50523	TEU	239	Teleuts	SIB
50625	TIM	-57	Timorese	INS
50790	TIN	61-	Tinam-Ayom	MEL
50799	TKM-1	146	Turkmen	CUS
51437	TLG	237	Telengit	SIB
51559	TLY	195	Talyshi	CAU
51729	TRK	171	Meskhetian-Turks	CUS
51901	TSH	229	Tsakhurs	CAU
52013	TSW	-67	Tswana	AFR
52194	TUB	178	Tubalars	SIB
52303	TUG-1	306	Tugen	AFR
52459	TUR-2	333	Turks	EUR
52536	TUR-3	347	Turks	EUR
52551	TUV	196	Tuvans	SIB
52754	TWA	45-	Twa	AFR
52769	UDG	202	Udegés	SIB
52832	UDI	181	Udis	CAU
52919	UDM	125	Udmurts	USR
53083	UDS	136	Urali-Idukki	ASI
53188	UKR	81-	Ukrainians	USR
53332	ULC	200	Ulchis	SIB
53431	USB	35-	US-Blacks	NAM
53732	USW	34-	US-Whites	NAM
54282	UYG	172	Uygurs	CUS
54462	UZB-1	162	Uzbeks	CUS
55034	UZB-2	162	Uzbeks	CUS
55118	VAC	73-	Croatians	EUR
55301	VAN	158	Vanches	CUS
55414	VEP	84-	Veps	USR

Table S2 Continued

Rows	Abbreviation	Sample Number	Ethnic Group	Region
55558	VIE	267	Vietnamese	EAS
55690	VIN	380	Garinagu	MES
55839	WAL	331	Welsh	EUR
56162	WAR	33-	Wardak	ASI
56176	XAL	130	Khalka-Mongols	EAS
56236	YAG	148	Yaghnobis	CUS
56333	YAK	142	Yakuts	SIB
56533	YAN	383	Yanomami	SAM
56733	YAZ	155	Yazgulyami	CUS
56836	YEA	149	Arabs-Yemenite	NEA
56932	YOR	-70	Yoruba	AFR
57112	YUK	382	Yukpa	SAM
57361	YUP	52-	Yupa-Indians	SAM
57380	YUW	62-	Yuwamb-Ayom	MEL
57418	ZCH	360	Swiss	EUR

Duplicated ID Numbers include: -43001 through and including -43072, -45180, -45346, -46004, -46006, -46007, -46013, -46015, -46021, -46023, -46024, -46025, -46036, -46040, -46042, -46043, -46045, -46050, -46054, -46061, -46062, -46064, -46065, -46070, -46073, -46081, -46086, -46097, -46098, -46099, -46100, -46103, -46107, -46108, -46109, -46113, -46115, -46118, -46119, -46126, -46128, -46130, -46131, -46135, -46137, -46138, -46139, -46143, -46145, -46146, -46147, -46149, -46153, -46154, -46156, -46157, -46158, -46161, -46164, -46165, -46167, -46169, -46170, -46171, -46176, -46181, -46182, -46184, -46185, -46191, -46205, -53003, -59286, -59289, -84010, -90229, 101204, 101206, 101207, 101208, 101209, 101210, 101211, 101212, 101213, 101216, 101218, 101219, 101220, 101221, 101223, 101232, 101233, 101234, 101235, 101236, 101237, 101239, 101242, 101245, 101246, 101247, 101250, 101261, 101501, 101662, 102241, 102302, 102303, 102304, 102305, 102306, 102307, 102309, 102311, 102312, 102313, 102319, 102322, 102328, 102329, 102331, 102338, 102347, 103006, 103012, 103014, 103041, 103042, 103066, 103070, 103072, 103074, 103084, 103142, 103400, 103404, 103405, 103406, 103407, 103410, 103411, 103414, 103415, 104170, 104181, 104182, 104183, 104184, 105002, 105162, 105167, 135135, 135136, 135137, 31-007, and 44-001 through and including 44-520.

The approximately 116 named Investigators are included here in plain text in alphabetical order by last name: M Akimova, Michael P Alpers, Mary Isabel Arrieta, Metin Atasu, F Auger, O Babakov, Tatiana Balujewa, Vera Bogdanova, Heinz Brehme, BJG Bringmann, Don R Brothwell, Bruinsma, M Copova, Gudrun Corvinus, G Davidova, N Debetz, Annie Masson Detourbet, Natalia A Dolinova, A Dubov, Nadezhda A Dubova, Sophie Ehrhardt, Aldur W Eriksson, V Ferak, Giovanni Floris, Henrik Forsius, K Fraser-Smith, A Roberto Frisancho, Helmuth Fuchs, M Fuciarelli, RA Geerdink, Robert Gessain, V Gochman, Martin Gusinde,

Gyula Gyenis, Vaclav Hajn, HJ Hammer, Robin G Harvey, G Hauser, Henriette L Heet, Georg Hera, W Herz, HW Hitzeroth, Johan Huizinga, Jiro Ikeda, Orazek Izmagulov, Lucienne Jakobi, Richard L Jantz, Carol L Jenkins, Francis E Johnston, Jorgen Balslev Jorgensen, PFJA Julien, Jungwirth, HW Jurgens, Anna Kalmus, Kenneth M Kensinger, A Kim, Kathryn Marianne Kolasa, Krstic, M Kuklik, I Ladaria, V Lebedjev, Legrand, W Lehmann, J Martin, Susan Mathew, Robert J Meier, Jasna Milicic, L Namsarainaidan, James V Neel, HH Noosten, Nyessen, Michio Okajima, Omsk University Expedition, Douglas William Owsley, Jens Ó. P. Pálsson, Kenneth R Parham, A Pentzos-Daponte, H Peter, Dolores Portabales, MF Pospisil, Miroslav Prokopec, Joelle Robert-Lamblin, Derek F Roberts, Peter J Rosa, Francisco Rothhammer, G Rykushina, R Ryman, N Rysnazakov, Heinrich Schade, Paul Schebesta, Hans-Georg Scheil, A Schevtschenko, V Schinkarenko, F Schwarzfischer, Zuzana Sieglöva, Daniela Sivakov, I Skrinjaric, Jan Hendrik Sypkens Smit, Eric Sunderland, K Syhymbajewa, Lidiya Ivanovna Tegako, Francois Twiesselmann, Adelaida G de Diaz Ungria, Valiunas, Hoan Thi Van, Van Huis, Stana Vrydagh-Laoureux, Geoffrey F Walker, AG de Wilde, Eike-Meinrad Winkler, Bernhard Wittwer, H Wojciechowska, Wood, Veronica Woolley, Miyo Yokota, and RM Yusupov.

Thirty-five Ethnic Groups contained finger patterns but not finger ridge-counts: Aeta-Bataan, Aeta-Camarines, Aeta-Zambales, Aigam-Ayom, Andians, Badjuan-Pamir, Balkars, Besermyans, Bwaka, Chuvashes, Garmend-Ayom, Giringiri-Ayom, Greeks (76 individuals only), Ilocano, Irish, Kana, Karakalpaks, Kazachs, Kyzyl, Managalase, Mandobo-North, Middle-Musa, Mombasob-Ayom, Muyu-Papuan, Naskapi, Papuans, Sawaiab-Ayom, Shugnans, Swiss, Tinam-Ayom, Turkmen, Vanches, Yakuts, Yazgulyami, and Yuwamb-Ayom.

Table S3. Software and algorithms required to execute code for dermatoglyphic analyses. Software package dependencies not included here

Software and algorithms	Source	Identifier
ade4 v1.7-22	Bougeard and Dray 2018	https://CRAN.R-project.org/package=ade4
adegraphics v1.0-18	Siberchicot et al. 2017	https://CRAN.R-project.org/package=adegraphics
adespatial v0.3-21	Dray et al. 2022	https://CRAN.R-project.org/package=adespatial
ape v5.7-1	Paradis and Schliep 2019	https://CRAN.R-project.org/package=ape
basemodels v1.1.0	Chen et al. 2023	https://CRAN.R-project.org/package=basemodels
caret v6.0-94	Kuhn 2022	https://CRAN.R-project.org/package=caret
cluster v2.1.4	Maechler et al. 2022	https://CRAN.R-project.org/package=cluster
clValid v0.7	Brock et al. 2008	https://CRAN.R-project.org/package=clValid
coefplot v1.2.8	Lander 2022	https://CRAN.R-project.org/package=coefplot
cowplot v1.1.1	Wilke 2020	https://CRAN.R-project.org/package=cowplot
e1071 v1.7-13	Meyer et al. 2022	https://CRAN.R-project.org/package=e1071
easystats v0.6.0	Lüdecke et al. 2022	https://CRAN.R-project.org/package=easystats
factoextra v1.0.7	Kassambara and Mundt 2020	https://CRAN.R-project.org/package=factoextra
geosphere v1.5-18	Hijmans 2022	https://CRAN.R-project.org/package=geosphere
ggplot2 v3.4.3	Wickham 2016	https://CRAN.R-project.org/package=ggplot2
ggpubr v0.6.0	Kassambara 2023	https://CRAN.R-project.org/package=ggpubr
ggthemes v4.2.4	Arnold 2021	https://CRAN.R-project.org/package=ggthemes
glmnet v4.1-8	Friedman et al. 2023	https://CRAN.R-project.org/package=glmnet
gridGraphics v0.5-1	Murrell and Wen 2020	https://CRAN.R-project.org/package=gridGraphics
kableExtra v1.3.4	Zhu et al. 2021	https://CRAN.R-project.org/package=kableExtra
kernlab v0.9-32	Karatzoglou et al. 2004	https://CRAN.R-project.org/package=kernlab
MASS v7.3-60	Ripley et al. 2023a	https://CRAN.R-project.org/package=MASS
MLmetrics v1.1.1	Yan 2016	https://CRAN.R-project.org/package=MLmetrics
nortest v1.0-4	Gross and Ligges 2015	https://CRAN.R-project.org/package=nortest
olsrr v0.5.3	Hebbali 2020	https://CRAN.R-project.org/package=olsrr
R v4.2.2	R Core Team	https://www.R-project.org/
readr v2.1.4	Wickham, Hester, and Bryan 2022	https://CRAN.R-project.org/package=readr
reshape2 v1.4.4	Wickham 2020	https://CRAN.R-project.org/package=reshape2

Table S3 Continued

Software and algorithms	Source	Identifier
rnaturalearth v0.3.4	Massicotte, South, and Hufkens 2023	https://CRAN.R-project.org/package=rnaturalearth
rnaturalearthdata v0.1.0	South 2017	https://CRAN.R-project.org/package=rnaturalearthdata
RStudio v2022.12.0.353	RStudio Team	http://www.rstudio.com/ or https://posit.co/
scales v1.2.1	Wickham, Seidel, and RStudio 2022	https://CRAN.R-project.org/package=scales
see v0.8.0	Lüdecke, Makowski, Patil, et al. 2023	https://CRAN.R-project.org/package=see
sf v1.0-14	Pebesma et al. 2023	https://CRAN.R-project.org/package=sf
spdep v1.2-8	Bivand 2022	https://CRAN.R-project.org/package=spdep
table1 v1.4.3	Rich 2023	https://CRAN.R-project.org/package=table1
tidyverse v2.0.0	Wickham et al. 2019	https://CRAN.R-project.org/package=tidyverse
vegan v2.6-4	Oksanen et al. 2022	https://CRAN.R-project.org/package=vegan
viridis v0.6.4	Garnier et al. 2021	https://CRAN.R-project.org/package=viridis
viridisLite v0.4.2	Garnier et al. 2021	https://CRAN.R-project.org/package=viridisLite

Table S4. Distribution of finger ridge-counts in overall data set compared to 10% subset

	Data		Subset Data	
	female (N=23882)	male (N=27781)	female (N=2336)	male (N=2829)
Left radial, digit 1				
Mean (SD)	14.3 (6.70)	16.2 (6.75)	14.3 (6.73)	16.1 (6.92)
Median [Min, Max]	15.0 [0, 34.0]	17.0 [0, 39.0]	15.0 [0, 31.0]	17.0 [0, 37.0]
Missing	2030 (8.5%)	2231 (8.0%)	189 (8.1%)	231 (8.2%)
Left ulnar, digit 1				
Mean (SD)	6.07 (7.92)	6.43 (8.26)	6.22 (7.91)	6.73 (8.26)
Median [Min, Max]	0 [0, 36.0]	0 [0, 37.0]	0 [0, 29.0]	0 [0, 32.0]
Missing	2031 (8.5%)	2233 (8.0%)	188 (8.0%)	233 (8.2%)
Left radial, digit 2				
Mean (SD)	8.75 (6.68)	9.52 (6.74)	8.87 (6.63)	9.53 (6.73)
Median [Min, Max]	10.0 [0, 34.0]	11.0 [0, 36.0]	10.0 [0, 30.0]	11.0 [0, 33.0]
Missing	2045 (8.6%)	2258 (8.1%)	189 (8.1%)	239 (8.4%)
Left ulnar, digit 2				
Mean (SD)	6.88 (8.00)	6.87 (7.94)	6.62 (7.90)	6.88 (7.94)
Median [Min, Max]	2.00 [0, 33.0]	2.00 [0, 34.0]	1.00 [0, 32.0]	2.00 [0, 29.0]
Missing	2042 (8.6%)	2242 (8.1%)	190 (8.1%)	235 (8.3%)
Left radial, digit 3				
Mean (SD)	11.6 (6.42)	12.7 (6.13)	11.5 (6.33)	12.6 (6.19)
Median [Min, Max]	13.0 [0, 40.0]	14.0 [0, 39.0]	13.0 [0, 29.0]	14.0 [0, 34.0]
Missing	2034 (8.5%)	2229 (8.0%)	189 (8.1%)	230 (8.1%)
Left ulnar, digit 3				
Mean (SD)	3.81 (7.02)	4.04 (7.21)	3.72 (6.98)	3.97 (7.19)
Median [Min, Max]	0 [0, 36.0]	0 [0, 36.0]	0 [0, 36.0]	0 [0, 31.0]
Missing	2021 (8.5%)	2211 (8.0%)	188 (8.0%)	228 (8.1%)
Left radial, digit 4				
Mean (SD)	15.4 (6.65)	16.4 (6.07)	15.4 (6.62)	16.2 (6.16)
Median [Min, Max]	16.0 [0, 46.0]	17.0 [0, 49.0]	16.0 [0, 37.0]	17.0 [0, 36.0]
Missing	2029 (8.5%)	2239 (8.1%)	191 (8.2%)	231 (8.2%)

Table S4 Continued

	Data		Subset Data	
	female (N=23882)	male (N=27781)	female (N=2336)	male (N=2829)
Left ulnar, digit 4				
Mean (SD)	5.76 (7.48)	6.19 (7.70)	5.74 (7.54)	6.02 (7.59)
Median [Min, Max]	0 [0, 32.0]	0 [0, 36.0]	0 [0, 29.0]	0 [0, 29.0]
Missing	2022 (8.5%)	2243 (8.1%)	189 (8.1%)	232 (8.2%)
Left radial, digit 5				
Mean (SD)	12.4 (5.37)	13.6 (4.91)	12.5 (5.30)	13.6 (5.01)
Median [Min, Max]	13.0 [0, 32.0]	14.0 [0, 30.0]	13.0 [0, 28.0]	14.0 [0, 30.0]
Missing	2057 (8.6%)	2234 (8.0%)	192 (8.2%)	231 (8.2%)
Left ulnar, digit 5				
Mean (SD)	1.41 (3.72)	1.59 (3.99)	1.41 (3.71)	1.54 (3.97)
Median [Min, Max]	0 [0, 25.0]	0 [0, 40.0]	0 [0, 19.0]	0 [0, 21.0]
Missing	2005 (8.4%)	2207 (7.9%)	185 (7.9%)	230 (8.1%)
Right radial, digit 1				
Mean (SD)	16.3 (6.63)	18.2 (6.65)	16.3 (6.56)	18.2 (6.73)
Median [Min, Max]	17.0 [0, 35.0]	19.0 [0, 38.0]	18.0 [0, 31.0]	19.0 [0, 36.0]
Missing	2037 (8.5%)	2234 (8.0%)	187 (8.0%)	236 (8.3%)
Right ulnar, digit 1				
Mean (SD)	6.70 (8.13)	8.46 (8.79)	6.59 (8.15)	8.56 (8.82)
Median [Min, Max]	0 [0, 35.0]	7.00 [0, 36.0]	0 [0, 30.0]	8.00 [0, 35.0]
Missing	2034 (8.5%)	2253 (8.1%)	189 (8.1%)	238 (8.4%)
Right radial, digit 2				
Mean (SD)	9.29 (6.57)	8.96 (6.76)	9.32 (6.59)	8.94 (6.74)
Median [Min, Max]	10.0 [0, 30.0]	10.0 [0, 32.0]	10.0 [0, 30.0]	10.0 [0, 29.0]
Missing	2047 (8.6%)	2258 (8.1%)	194 (8.3%)	242 (8.6%)
Right ulnar, digit 2				
Mean (SD)	7.04 (8.31)	8.15 (8.44)	6.83 (8.13)	8.06 (8.46)
Median [Min, Max]	0 [0, 35.0]	6.00 [0, 53.0]	0 [0, 35.0]	5.00 [0, 31.0]
Missing	2031 (8.5%)	2223 (8.0%)	187 (8.0%)	232 (8.2%)

Table S4 Continued

	Data		Subset Data	
	female (N=23882)	male (N=27781)	female (N=2336)	male (N=2829)
Right radial, digit 3				
Mean (SD)	11.6 (5.79)	12.0 (5.82)	11.6 (5.79)	11.8 (5.95)
Median [Min, Max]	12.0 [0, 42.0]	13.0 [0, 35.0]	12.0 [0, 29.0]	13.0 [0, 32.0]
Missing	2032 (8.5%)	2251 (8.1%)	189 (8.1%)	233 (8.2%)
Right ulnar, digit 3				
Mean (SD)	3.11 (6.59)	4.09 (7.27)	2.98 (6.44)	4.12 (7.25)
Median [Min, Max]	0 [0, 41.0]	0 [0, 37.0]	0 [0, 35.0]	0 [0, 32.0]
Missing	2014 (8.4%)	2206 (7.9%)	188 (8.0%)	228 (8.1%)
Right radial, digit 4				
Mean (SD)	15.5 (6.27)	16.0 (6.03)	15.4 (6.23)	15.9 (6.15)
Median [Min, Max]	16.0 [0, 39.0]	17.0 [0, 40.0]	16.0 [0, 39.0]	17.0 [0, 32.0]
Missing	2046 (8.6%)	2266 (8.2%)	192 (8.2%)	234 (8.3%)
Right ulnar, digit 4				
Mean (SD)	6.54 (7.65)	7.82 (7.96)	6.32 (7.59)	7.76 (7.85)
Median [Min, Max]	0 [0, 33.0]	7.00 [0, 37.0]	0 [0, 33.0]	7.00 [0, 31.0]
Missing	2017 (8.4%)	2224 (8.0%)	191 (8.2%)	231 (8.2%)
Right radial, digit 5				
Mean (SD)	12.2 (5.29)	13.3 (5.04)	12.3 (5.23)	13.2 (5.14)
Median [Min, Max]	13.0 [0, 31.0]	14.0 [0, 33.0]	13.0 [0, 27.0]	14.0 [0, 31.0]
Missing	2055 (8.6%)	2259 (8.1%)	193 (8.3%)	238 (8.4%)
Right ulnar, digit 5				
Mean (SD)	1.47 (3.74)	2.24 (4.62)	1.41 (3.67)	2.05 (4.40)
Median [Min, Max]	0 [0, 25.0]	0 [0, 30.0]	0 [0, 23.0]	0 [0, 21.0]
Missing	2005 (8.4%)	2201 (7.9%)	185 (7.9%)	229 (8.1%)
Third				
Mean (SD)	0.027 (0.382)	0.030 (0.447)	0.021 (0.298)	0.042 (0.497)
Median [Min, Max]	0 [0, 13.0]	0 [0, 25.0]	0 [0, 7.00]	0 [0, 10.0]
Missing	1932 (8.1%)	2122 (7.6%)	179 (7.7%)	221 (7.8%)

Table S5. Distribution of Directional Asymmetry variables separated by digit in overall data set compared to 10% subset

	Data		Subset Data	
	female (N=23882)	male (N=27781)	female (N=2336)	male (N=2829)
Directional Asymmetry, digit 1				
Mean (SD)	-1.80 (4.26)	-2.00 (4.27)	-1.81 (4.21)	-2.01 (4.31)
Median [Min, Max]	-2.00 [-25.0, 33.0]	-2.00 [-28.0, 20.0]	-2.00 [-19.0, 17.0]	-2.00 [-26.0, 19.0]
Missing	2106 (8.8%)	2342 (8.4%)	198 (8.5%)	247 (8.7%)
Directional Asymmetry, digit 2				
Mean (SD)	-0.545 (4.62)	-0.443 (4.65)	-0.423 (4.68)	-0.454 (4.58)
Median [Min, Max]	0 [-30.0, 27.0]	0 [-45.0, 25.0]	0 [-30.0, 20.0]	0 [-22.0, 19.0]
Missing	2126 (8.9%)	2369 (8.5%)	202 (8.6%)	256 (9.0%)
Directional Asymmetry, digit 3				
Mean (SD)	0.0774 (4.00)	0.514 (3.89)	0.109 (3.96)	0.484 (3.90)
Median [Min, Max]	0 [-34.0, 24.0]	0 [-20.0, 22.0]	0 [-18.0, 18.0]	0 [-17.0, 22.0]
Missing	2076 (8.7%)	2309 (8.3%)	192 (8.2%)	235 (8.3%)
Directional Asymmetry, digit 4				
Mean (SD)	-0.122 (3.73)	0.139 (3.57)	-0.149 (3.79)	0.125 (3.79)
Median [Min, Max]	0 [-30.0, 29.0]	0 [-22.0, 28.0]	0 [-30.0, 18.0]	0 [-19.0, 20.0]
Missing	2090 (8.8%)	2353 (8.5%)	200 (8.6%)	241 (8.5%)
Directional Asymmetry, digit 5				
Mean (SD)	0.207 (3.20)	0.267 (3.10)	0.109 (3.27)	0.323 (3.06)
Median [Min, Max]	0 [-17.0, 16.0]	0 [-18.0, 22.0]	0 [-15.0, 14.0]	0 [-11.0, 13.0]

Table S5 Continued

	Data		Subset Data	
	female (N=23882)	male (N=27781)	female (N=2336)	male (N=2829)
Missing	2102 (8.8%)	2315 (8.3%)	198 (8.5%)	243 (8.6%)

Table S6. Distribution of Fluctuating Asymmetry variables separated by digit in overall data set compared to 10% subset

	Data		Subset Data	
	female (N=23882)	male (N=27781)	female (N=2336)	male (N=2829)
Fluctuating Asymmetry, digit 1				
Mean (SD)	3.43 (3.10)	3.52 (3.13)	3.41 (3.06)	3.57 (3.14)
Median [Min, Max]	3.00 [0, 33.0]	3.00 [0, 28.0]	3.00 [0, 19.0]	3.00 [0, 26.0]
Missing	2106 (8.8%)	2342 (8.4%)	198 (8.5%)	247 (8.7%)
Fluctuating Asymmetry, digit 2				
Mean (SD)	3.30 (3.27)	3.33 (3.27)	3.32 (3.33)	3.28 (3.23)
Median [Min, Max]	2.00 [0, 30.0]	2.00 [0, 45.0]	2.00 [0, 30.0]	2.00 [0, 22.0]
Missing	2126 (8.9%)	2369 (8.5%)	202 (8.6%)	256 (9.0%)
Fluctuating Asymmetry, digit 3				
Mean (SD)	2.92 (2.73)	2.86 (2.69)	2.90 (2.70)	2.82 (2.74)
Median [Min, Max]	2.00 [0, 34.0]	2.00 [0, 22.0]	2.00 [0, 18.0]	2.00 [0, 22.0]
Missing	2076 (8.7%)	2309 (8.3%)	192 (8.2%)	235 (8.3%)

Table S6 Continued

	Data		Subset Data	
	female (N=23882)	male (N=27781)	female (N=2336)	male (N=2829)
Fluctuating Asymmetry, digit 4				
Mean (SD)	2.71 (2.57)	2.60 (2.45)	2.73 (2.62)	2.72 (2.64)
Median [Min, Max]	2.00 [0, 30.0]	2.00 [0, 28.0]	2.00 [0, 30.0]	2.00 [0, 20.0]
Missing	2090 (8.8%)	2353 (8.5%)	200 (8.6%)	241 (8.5%)
Fluctuating Asymmetry, digit 5				
Mean (SD)	2.38 (2.15)	2.32 (2.07)	2.41 (2.21)	2.34 (1.99)
Median [Min, Max]	2.00 [0, 17.0]	2.00 [0, 22.0]	2.00 [0, 15.0]	2.00 [0, 13.0]
Missing	2102 (8.8%)	2315 (8.3%)	198 (8.5%)	243 (8.6%)

Table S7. Different sample sizes in different PCA; Directional Asymmetry = 5 variables; finger ridge-counts = 21 variables; Fluctuating Asymmetry = 5 variables

Variables for PCA	Sample size
finger ridge-counts	46541
Directional Asymmetry	46549
Fluctuating Asymmetry	46549
finger ridge-counts females	21641
finger ridge-counts males	24864
Directional Asymmetry females	21501
Directional Asymmetry males	25048
Fluctuating Asymmetry females	21501
Fluctuating Asymmetry males	25048
finger ridge-counts Africa	10806
finger ridge-counts Americas	3522
finger ridge-counts Asia	15813
finger ridge-counts Europe	16400
Directional Asymmetry Africa	10809
Fluctuating Asymmetry Africa	10809
Directional Asymmetry Americas	3522
Fluctuating Asymmetry Americas	3522
Directional Asymmetry Asia	15816
Fluctuating Asymmetry Asia	15816
Directional Asymmetry Europe	16402
Fluctuating Asymmetry Europe	16402
finger ridge-counts Africa females	4537
finger ridge-counts Africa males	6269
finger ridge-counts Americas females	1825
finger ridge-counts Americas males	1697
finger ridge-counts Asia females	7316
finger ridge-counts Asia males	8497
finger ridge-counts Europe females	7821
finger ridge-counts Europe males	8579
Directional Asymmetry Africa females	4538
Directional Asymmetry Africa males	6271
Fluctuating Asymmetry Africa females	4537
Fluctuating Asymmetry Africa males	6269
Directional Asymmetry Americas females	1825
Directional Asymmetry America males	1697
Fluctuating Asymmetry Americas females	1825
Fluctuating Asymmetry Americas males	1697
Directional Asymmetry Asia females	7316
Directional Asymmetry Asia males	8500
Fluctuating Asymmetry Asia females	7316
Fluctuating Asymmetry Asia males	8500
Directional Asymmetry Europe females	7822

Table S7 Continued

Variables for PCA	Sample size
Directional Asymmetry Europe males	8580
Fluctuating Asymmetry Europe females	7822
Fluctuating Asymmetry Europe males	8580

Table S8. Summary of response variables TRC, ARC, and finger ridge-counts before any response variable-specific modification for GLM analysis and prediction

	Data		Training Data		Test Data	
	female (N=21499)	male (N=25042)	female (N=15063)	male (N=17517)	female (N=6436)	male (N=7525)
Total ridge-count						
Mean (SD)	134 (50.4)	145 (48.2)	134 (50.2)	145 (48.1)	135 (50.7)	145 (48.4)
Median [Min, Max]	140 [0, 320]	150 [0, 306]	140 [0, 320]	150 [0, 303]	140 [0, 308]	150 [0, 306]
Absolute ridge-count						
Mean (SD)	176 (88.7)	192 (89.1)	175 (88.3)	192 (88.9)	177 (89.5)	193 (89.7)
Median [Min, Max]	166 [0, 554]	184 [0, 531]	166 [0, 554]	184 [0, 522]	168 [0, 553]	184 [0, 531]
Left radial, digit 1						
Mean (SD)	14.2 (6.71)	16.2 (6.76)	14.2 (6.70)	16.1 (6.76)	14.3 (6.71)	16.2 (6.76)
Median [Min, Max]	15.0 [0, 34.0]	17.0 [0, 39.0]	15.0 [0, 34.0]	17.0 [0, 39.0]	15.0 [0, 34.0]	17.0 [0, 34.0]
Left ulnar, digit 1						
Mean (SD)	6.03 (7.91)	6.40 (8.24)	5.99 (7.89)	6.40 (8.23)	6.13 (7.96)	6.39 (8.26)
Median [Min, Max]	0 [0, 36.0]	0 [0, 37.0]	0 [0, 36.0]	0 [0, 35.0]	0 [0, 33.0]	0 [0, 37.0]
Left radial, digit 2						
Mean (SD)	8.72 (6.68)	9.47 (6.74)	8.70 (6.69)	9.47 (6.75)	8.76 (6.65)	9.48 (6.73)
Median [Min, Max]	10.0 [0, 34.0]	11.0 [0, 33.0]	10.0 [0, 32.0]	11.0 [0, 33.0]	10.0 [0, 34.0]	11.0 [0, 29.0]
Left ulnar, digit 2						
Mean (SD)	6.85 (7.99)	6.84 (7.93)	6.85 (7.98)	6.81 (7.94)	6.85 (8.02)	6.90 (7.93)
Median [Min, Max]	2.00 [0, 33.0]	2.00 [0, 34.0]	2.00 [0, 33.0]	2.00 [0, 34.0]	1.00 [0, 33.0]	2.00 [0, 31.0]

Table S8 Continued

	Data		Training Data		Test Data	
	female (N=21499)	male (N=25042)	female (N=15063)	male (N=17517)	female (N=6436)	male (N=7525)
Left radial, digit 3						
Mean (SD)	11.6 (6.43)	12.7 (6.15)	11.6 (6.39)	12.7 (6.13)	11.6 (6.50)	12.7 (6.18)
Median [Min, Max]	13.0 [0, 40.0]	14.0 [0, 39.0]	13.0 [0, 36.0]	14.0 [0, 39.0]	13.0 [0, 40.0]	14.0 [0, 37.0]
Left ulnar, digit 3						
Mean (SD)	3.77 (6.99)	4.00 (7.18)	3.75 (6.95)	3.95 (7.15)	3.81 (7.08)	4.10 (7.25)
Median [Min, Max]	0 [0, 36.0]	0 [0, 36.0]	0 [0, 36.0]	0 [0, 36.0]	0 [0, 34.0]	0 [0, 34.0]
Left radial, digit 4						
Mean (SD)	15.4 (6.67)	16.3 (6.08)	15.3 (6.64)	16.3 (6.08)	15.4 (6.72)	16.4 (6.07)
Median [Min, Max]	16.0 [0, 46.0]	17.0 [0, 49.0]	16.0 [0, 46.0]	17.0 [0, 49.0]	16.0 [0, 40.0]	17.0 [0, 41.0]
Left ulnar, digit 4						
Mean (SD)	5.73 (7.47)	6.15 (7.68)	5.71 (7.45)	6.10 (7.66)	5.75 (7.50)	6.26 (7.73)
Median [Min, Max]	0 [0, 32.0]	0 [0, 36.0]	0 [0, 32.0]	0 [0, 36.0]	0 [0, 31.0]	0 [0, 32.0]
Left radial, digit 5						
Mean (SD)	12.4 (5.38)	13.6 (4.92)	12.4 (5.38)	13.6 (4.92)	12.4 (5.37)	13.6 (4.93)
Median [Min, Max]	13.0 [0, 32.0]	14.0 [0, 30.0]	13.0 [0, 32.0]	14.0 [0, 30.0]	13.0 [0, 28.0]	14.0 [0, 29.0]
Left ulnar, digit 5						
Mean (SD)	1.41 (3.71)	1.57 (3.97)	1.42 (3.72)	1.56 (3.94)	1.37 (3.69)	1.61 (4.04)
Median [Min, Max]	0 [0, 25.0]	0 [0, 40.0]	0 [0, 24.0]	0 [0, 22.0]	0 [0, 25.0]	0 [0, 40.0]
Right radial, digit 1						

Table S8 Continued

	Data		Training Data		Test Data	
	female (N=21499)	male (N=25042)	female (N=15063)	male (N=17517)	female (N=6436)	male (N=7525)
Mean (SD)	16.3 (6.63)	18.2 (6.67)	16.3 (6.62)	18.2 (6.67)	16.3 (6.66)	18.2 (6.66)
Median [Min, Max]	17.0 [0, 35.0]	19.0 [0, 38.0]	17.0 [0, 35.0]	19.0 [0, 38.0]	18.0 [0, 35.0]	19.0 [0, 38.0]
Right ulnar, digit 1						
Mean (SD)	6.65 (8.12)	8.42 (8.79)	6.60 (8.09)	8.41 (8.77)	6.79 (8.19)	8.46 (8.82)
Median [Min, Max]	0 [0, 35.0]	7.00 [0, 36.0]	0 [0, 35.0]	7.00 [0, 36.0]	0 [0, 33.0]	7.00 [0, 35.0]
Right radial, digit 2						
Mean (SD)	9.27 (6.58)	8.92 (6.76)	9.28 (6.60)	8.91 (6.76)	9.25 (6.54)	8.94 (6.77)
Median [Min, Max]	10.0 [0, 30.0]	10.0 [0, 32.0]	10.0 [0, 30.0]	10.0 [0, 32.0]	10.0 [0, 28.0]	10.0 [0, 28.0]
Right ulnar, digit 2						
Mean (SD)	6.99 (8.29)	8.11 (8.42)	6.96 (8.27)	8.10 (8.42)	7.08 (8.35)	8.12 (8.43)
Median [Min, Max]	0 [0, 35.0]	6.00 [0, 36.0]	0 [0, 35.0]	6.00 [0, 35.0]	0 [0, 35.0]	6.00 [0, 36.0]
Right radial, digit 3						
Mean (SD)	11.6 (5.80)	12.0 (5.83)	11.6 (5.77)	12.0 (5.82)	11.7 (5.86)	12.0 (5.85)
Median [Min, Max]	12.0 [0, 42.0]	13.0 [0, 35.0]	12.0 [0, 33.0]	13.0 [0, 35.0]	12.0 [0, 42.0]	13.0 [0, 32.0]
Right ulnar, digit 3						
Mean (SD)	3.06 (6.55)	4.04 (7.24)	3.02 (6.49)	4.01 (7.20)	3.15 (6.70)	4.13 (7.32)
Median [Min, Max]	0 [0, 41.0]	0 [0, 37.0]	0 [0, 41.0]	0 [0, 37.0]	0 [0, 33.0]	0 [0, 37.0]
Right radial, digit 4						
Mean (SD)	15.4 (6.29)	16.0 (6.04)	15.4 (6.26)	16.0 (6.04)	15.5 (6.35)	16.1 (6.04)

Table S8 Continued

	Data		Training Data		Test Data	
	female (N=21499)	male (N=25042)	female (N=15063)	male (N=17517)	female (N=6436)	male (N=7525)
Median [Min, Max]	16.0 [0, 39.0]	17.0 [0, 40.0]	16.0 [0, 39.0]	17.0 [0, 39.0]	16.0 [0, 39.0]	17.0 [0, 40.0]
Right ulnar, digit 4						
Mean (SD)	6.50 (7.65)	7.77 (7.95)	6.45 (7.61)	7.74 (7.95)	6.62 (7.72)	7.85 (7.95)
Median [Min, Max]	0 [0, 33.0]	7.00 [0, 37.0]	0 [0, 33.0]	7.00 [0, 37.0]	0 [0, 31.0]	7.00 [0, 35.0]
Right radial, digit 5						
Mean (SD)	12.2 (5.30)	13.3 (5.05)	12.1 (5.32)	13.3 (5.04)	12.3 (5.28)	13.3 (5.07)
Median [Min, Max]	13.0 [0, 31.0]	14.0 [0, 33.0]	13.0 [0, 31.0]	14.0 [0, 33.0]	13.0 [0, 27.0]	14.0 [0, 30.0]
Right ulnar, digit 5						
Mean (SD)	1.46 (3.73)	2.22 (4.60)	1.43 (3.70)	2.20 (4.59)	1.52 (3.81)	2.26 (4.63)
Median [Min, Max]	0 [0, 25.0]	0 [0, 24.0]	0 [0, 24.0]	0 [0, 24.0]	0 [0, 25.0]	0 [0, 23.0]
Third						
Mean (SD)	0.0277 (0.384)	0.0308 (0.449)	0.0278 (0.390)	0.0326 (0.436)	0.0273 (0.367)	0.0267 (0.477)
Median [Min, Max]	0 [0, 13.0]	0 [0, 25.0]	0 [0, 13.0]	0 [0, 17.0]	0 [0, 9.00]	0 [0, 25.0]

Table S9. Summary of response variables Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry of diversity before any response variable-specific modification for GLM analysis and prediction

	Data		Training Data		Test Data	
	female (N=21501)	male (N=25048)	female (N=15042)	male (N=17543)	female (N=6459)	male (N=7505)
Directional Asymmetry, digit 1						
Mean (SD)	-1.80 (4.27)	-1.99 (4.28)	-1.79 (4.29)	-2.00 (4.29)	-1.83 (4.21)	-1.98 (4.25)
Median [Min, Max]	-2.00 [-25.0, 33.0]	-2.00 [-28.0, 20.0]	-2.00 [-24.0, 24.0]	-2.00 [-26.0, 19.0]	-2.00 [-25.0, 33.0]	-2.00 [-28.0, 20.0]
Directional Asymmetry, digit 2						
Mean (SD)	-0.547 (4.62)	-0.443 (4.65)	-0.524 (4.65)	-0.422 (4.65)	-0.603 (4.55)	-0.491 (4.63)
Median [Min, Max]	0 [-30.0, 27.0]	0 [-32.0, 25.0]	0 [-30.0, 27.0]	0 [-24.0, 25.0]	0 [-22.0, 21.0]	0 [-32.0, 24.0]
Directional Asymmetry, digit 3						
Mean (SD)	0.0695 (4.00)	0.507 (3.89)	0.0904 (4.01)	0.501 (3.88)	0.0209 (3.97)	0.521 (3.91)
Median [Min, Max]	0 [-34.0, 24.0]	0 [-20.0, 22.0]	0 [-34.0, 20.0]	0 [-20.0, 22.0]	0 [-31.0, 24.0]	0 [-18.0, 20.0]
Directional Asymmetry, digit 4						
Mean (SD)	-0.128 (3.74)	0.134 (3.57)	-0.157 (3.75)	0.133 (3.60)	-0.0610 (3.72)	0.136 (3.51)

Table S9 Continued

	Data		Training Data		Test Data	
	female (N=21501)	male (N=25048)	female (N=15042)	male (N=17543)	female (N=6459)	male (N=7505)
Median [Min, Max]	0 [-30.0, 29.0]	0 [-22.0, 28.0]	0 [-30.0, 29.0]	0 [-22.0, 28.0]	0 [-22.0, 20.0]	0 [-17.0, 21.0]
Directional Asymmetry, digit 5						
Mean (SD)	0.204 (3.20)	0.266 (3.09)	0.201 (3.21)	0.275 (3.09)	0.210 (3.16)	0.245 (3.11)
Median [Min, Max]	0 [-17.0, 16.0]	0 [-17.0, 22.0]	0 [-17.0, 16.0]	0 [-16.0, 22.0]	0 [-15.0, 16.0]	0 [-17.0, 15.0]
Fluctuating Asymmetry, digit 1						
Mean (SD)	3.44 (3.11)	3.53 (3.14)	3.44 (3.14)	3.54 (3.15)	3.44 (3.03)	3.50 (3.12)
Median [Min, Max]	3.00 [0, 33.0]	3.00 [0, 28.0]	3.00 [0, 24.0]	3.00 [0, 26.0]	3.00 [0, 33.0]	3.00 [0, 28.0]
Fluctuating Asymmetry, digit 2						
Mean (SD)	3.31 (3.27)	3.33 (3.27)	3.31 (3.31)	3.33 (3.27)	3.30 (3.19)	3.32 (3.26)
Median [Min, Max]	2.00 [0, 30.0]	2.00 [0, 32.0]	2.00 [0, 30.0]	2.00 [0, 25.0]	2.00 [0, 22.0]	2.00 [0, 32.0]
Fluctuating Asymmetry, digit 3						
Mean (SD)	2.93 (2.73)	2.86 (2.69)	2.93 (2.74)	2.86 (2.68)	2.91 (2.71)	2.87 (2.71)
Median [Min, Max]	2.00 [0, 34.0]	2.00 [0, 22.0]	2.00 [0, 34.0]	2.00 [0, 22.0]	2.00 [0, 31.0]	2.00 [0, 20.0]

Table S9 Continued

	Data		Training Data		Test Data	
	female (N=21501)	male (N=25048)	female (N=15042)	male (N=17543)	female (N=6459)	male (N=7505)
Fluctuating Asymmetry, digit 4						
Mean (SD)	2.71 (2.58)	2.60 (2.45)	2.73 (2.58)	2.61 (2.48)	2.69 (2.57)	2.59 (2.38)
Median [Min, Max]	2.00 [0, 30.0]	2.00 [0, 28.0]	2.00 [0, 30.0]	2.00 [0, 28.0]	2.00 [0, 22.0]	2.00 [0, 21.0]
Fluctuating Asymmetry, digit 5						
Mean (SD)	2.38 (2.15)	2.31 (2.07)	2.40 (2.15)	2.31 (2.07)	2.33 (2.15)	2.32 (2.08)
Median [Min, Max]	2.00 [0, 17.0]	2.00 [0, 22.0]	2.00 [0, 17.0]	2.00 [0, 22.0]	2.00 [0, 16.0]	2.00 [0, 17.0]
Index.Asymmetry						
Mean (SD)	8.25 (3.91)	8.17 (3.89)	8.29 (3.93)	8.17 (3.91)	8.17 (3.87)	8.16 (3.84)
Median [Min, Max]	7.55 [0, 43.0]	7.42 [0, 34.4]	7.55 [0, 43.0]	7.42 [0, 33.7]	7.48 [0, 35.0]	7.42 [0, 34.4]
Index.Diversity						
Mean (SD)	6.74 (3.09)	6.96 (3.19)	6.74 (3.09)	6.98 (3.20)	6.73 (3.08)	6.91 (3.19)
Median [Min, Max]	6.24 [0, 21.9]	6.43 [0, 23.4]	6.24 [0, 21.9]	6.43 [0, 23.1]	6.23 [0, 20.7]	6.40 [0, 23.4]
Index.Diversity.Holt						
Mean (SD)	3.93 (1.56)	4.02 (1.61)	3.94 (1.56)	4.03 (1.61)	3.92 (1.56)	4.00 (1.60)

Table S9 Continued

	Data		Training Data		Test Data	
	female (N=21501)	male (N=25048)	female (N=15042)	male (N=17543)	female (N=6459)	male (N=7505)
Median [Min, Max]	3.69 [0, 12.1]	3.74 [0, 12.0]	3.70 [0, 12.1]	3.74 [0, 11.8]	3.67 [0, 10.6]	3.71 [0, 12.0]

Table S10. Summary of response variables TRC and ARC before any response variable-specific modification for SVR analysis and prediction

	Data		Training Data		Validation Data		Test Data	
	female (N=23882)	male (N=27781)	female (N=12029)	male (N=13803)	female (N=5887)	male (N=7027)	female (N=5966)	male (N=6951)
Total ridge-count								
Mean (SD)	134 (50.4)	145 (48.2)	135 (50.6)	146 (48.5)	134 (50.8)	145 (47.7)	134 (49.6)	144 (48.2)
Median [Min, Max]	140 [0, 320]	150 [0, 306]	141 [0, 320]	151 [0, 303]	141 [0, 292]	149 [0, 306]	139 [0, 308]	148 [0, 297]
Missing	2381 (10.0%)	2733 (9.8%)	1186 (9.9%)	1327 (9.6%)	591 (10.0%)	697 (9.9%)	604 (10.1%)	709 (10.2%)
Absolute ridge-count								
Mean (SD)	176 (88.7)	192 (89.1)	177 (89.5)	194 (90.0)	175 (88.6)	192 (88.3)	174 (87.1)	189 (88.1)

Table S10 Continued

	Data		Training Data		Validation Data		Test Data	
	female (N=23882)	male (N=27781)	female (N=12029)	male (N=13803)	female (N=5887)	male (N=7027)	female (N=5966)	male (N=6951)
Median [Min, Max]	166 [0, 554]	184 [0, 531]	167 [0, 554]	186 [0, 522]	166 [0, 509]	183 [0, 531]	165 [0, 553]	181 [0, 513]
Missing	2381 (10.0%)	2733 (9.8%)	1186 (9.9%)	1327 (9.6%)	591 (10.0%)	697 (9.9%)	604 (10.1%)	709 (10.2%)

Table S11. Summary of response variables Directional Asymmetry and Fluctuating Asymmetry variables for each digit and indices of asymmetry of diversity before any response variable-specific modification for SVR analysis and prediction

	Data		Training Data		Validation Data		Test Data	
	female (N=23882)	male (N=27781)	female (N=12029)	male (N=13803)	female (N=5887)	male (N=7027)	female (N=5966)	male (N=6951)
Directional Asymmetry, digit 1								
Mean (SD)	-1.80 (4.26)	-2.00 (4.27)	-1.81 (4.28)	-2.03 (4.27)	-1.71 (4.17)	-2.05 (4.37)	-1.87 (4.31)	-1.88 (4.17)
Median [Min, Max]	-2.00 [- 25.0, 33.0]	-2.00 [- 28.0, 20.0]	-2.00 [- 24.0, 33.0]	-2.00 [- 28.0, 19.0]	-1.00 [- 22.0, 18.0]	-2.00 [- 26.0, 20.0]	-2.00 [- 25.0, 21.0]	-2.00 [- 26.0, 19.0]
Missing	2106 (8.8%)	2342 (8.4%)	1056 (8.8%)	1150 (8.3%)	521 (8.9%)	586 (8.3%)	529 (8.9%)	606 (8.7%)

Table S11 Continued

	Data		Training Data		Validation Data		Test Data	
	female (N=23882)	male (N=27781)	female (N=12029)	male (N=13803)	female (N=5887)	male (N=7027)	female (N=5966)	male (N=6951)
Directional Asymmetry, digit 2								
Mean (SD)	-0.545 (4.62)	-0.443 (4.65)	-0.542 (4.58)	-0.453 (4.69)	-0.578 (4.59)	-0.365 (4.55)	-0.518 (4.72)	-0.502 (4.67)
Median [Min, Max]	0 [-30.0, 27.0]	0 [-45.0, 25.0]	0 [-22.0, 27.0]	0 [-45.0, 24.0]	0 [-22.0, 23.0]	0 [-21.0, 25.0]	0 [-30.0, 21.0]	0 [-24.0, 20.0]
Missing	2126 (8.9%)	2369 (8.5%)	1066 (8.9%)	1148 (8.3%)	524 (8.9%)	602 (8.6%)	536 (9.0%)	619 (8.9%)
Directional Asymmetry, digit 3								
Mean (SD)	0.0774 (4.00)	0.514 (3.89)	0.0581 (3.95)	0.505 (3.89)	0.190 (4.03)	0.591 (3.83)	0.00551 (4.05)	0.454 (3.95)
Median [Min, Max]	0 [-34.0, 24.0]	0 [-20.0, 22.0]	0 [-34.0, 20.0]	0 [-20.0, 19.0]	0 [-18.0, 24.0]	0 [-17.0, 22.0]	0 [-19.0, 22.0]	0 [-20.0, 22.0]
Missing	2076 (8.7%)	2309 (8.3%)	1041 (8.7%)	1126 (8.2%)	514 (8.7%)	578 (8.2%)	521 (8.7%)	605 (8.7%)
Directional Asymmetry, digit 4								
Mean (SD)	-0.122 (3.73)	0.139 (3.57)	-0.118 (3.74)	0.153 (3.53)	-0.0882 (3.68)	0.170 (3.55)	-0.166 (3.79)	0.0793 (3.67)

Table S11 Continued

	Data		Training Data		Validation Data		Test Data	
	female (N=23882)	male (N=27781)	female (N=12029)	male (N=13803)	female (N=5887)	male (N=7027)	female (N=5966)	male (N=6951)
Median [Min, Max]	0 [-30.0, 29.0]	0 [-22.0, 28.0]	0 [-23.0, 29.0]	0 [-22.0, 28.0]	0 [-19.0, 23.0]	0 [-18.0, 17.0]	0 [-30.0, 18.0]	0 [-22.0, 23.0]
Missing	2090 (8.8%)	2353 (8.5%)	1047 (8.7%)	1154 (8.4%)	515 (8.7%)	594 (8.5%)	528 (8.9%)	605 (8.7%)
Directional Asymmetry, digit 5								
Mean (SD)	0.207 (3.20)	0.267 (3.10)	0.273 (3.20)	0.269 (3.09)	0.188 (3.16)	0.258 (3.11)	0.0932 (3.23)	0.272 (3.09)
Median [Min, Max]	0 [-17.0, 16.0]	0 [-18.0, 22.0]	0 [-17.0, 16.0]	0 [-18.0, 19.0]	0 [-14.0, 16.0]	0 [-15.0, 22.0]	0 [-15.0, 16.0]	0 [-17.0, 16.0]
Missing	2102 (8.8%)	2315 (8.3%)	1053 (8.8%)	1129 (8.2%)	514 (8.7%)	581 (8.3%)	535 (9.0%)	605 (8.7%)
Fluctuating Asymmetry, digit 1								
Mean (SD)	3.43 (3.10)	3.52 (3.13)	3.46 (3.11)	3.54 (3.13)	3.33 (3.03)	3.60 (3.22)	3.48 (3.15)	3.40 (3.05)
Median [Min, Max]	3.00 [0, 33.0]	3.00 [0, 28.0]	3.00 [0, 33.0]	3.00 [0, 28.0]	3.00 [0, 22.0]	3.00 [0, 26.0]	3.00 [0, 25.0]	3.00 [0, 26.0]
Missing	2106 (8.8%)	2342 (8.4%)	1056 (8.8%)	1150 (8.3%)	521 (8.9%)	586 (8.3%)	529 (8.9%)	606 (8.7%)

Table S11 Continued

	Data		Training Data		Validation Data		Test Data	
	female (N=23882)	male (N=27781)	female (N=12029)	male (N=13803)	female (N=5887)	male (N=7027)	female (N=5966)	male (N=6951)
Fluctuating Asymmetry, digit 2								
Mean (SD)	3.30 (3.27)	3.33 (3.27)	3.29 (3.23)	3.35 (3.31)	3.28 (3.26)	3.28 (3.18)	3.36 (3.35)	3.34 (3.30)
Median [Min, Max]	2.00 [0, 30.0]	2.00 [0, 45.0]	2.00 [0, 27.0]	2.00 [0, 45.0]	2.00 [0, 23.0]	2.00 [0, 25.0]	2.00 [0, 30.0]	2.00 [0, 24.0]
Missing	2126 (8.9%)	2369 (8.5%)	1066 (8.9%)	1148 (8.3%)	524 (8.9%)	602 (8.6%)	536 (9.0%)	619 (8.9%)
Fluctuating Asymmetry, digit 3								
Mean (SD)	2.92 (2.73)	2.86 (2.69)	2.90 (2.69)	2.87 (2.67)	2.94 (2.76)	2.83 (2.65)	2.95 (2.77)	2.87 (2.75)
Median [Min, Max]	2.00 [0, 34.0]	2.00 [0, 22.0]	2.00 [0, 34.0]	2.00 [0, 20.0]	2.00 [0, 24.0]	2.00 [0, 22.0]	2.00 [0, 22.0]	2.00 [0, 22.0]
Missing	2076 (8.7%)	2309 (8.3%)	1041 (8.7%)	1126 (8.2%)	514 (8.7%)	578 (8.2%)	521 (8.7%)	605 (8.7%)
Fluctuating Asymmetry, digit 4								
Mean (SD)	2.71 (2.57)	2.60 (2.45)	2.70 (2.58)	2.56 (2.44)	2.69 (2.51)	2.62 (2.39)	2.75 (2.61)	2.66 (2.52)

Table S11 Continued

	Data		Training Data		Validation Data		Test Data	
	female (N=23882)	male (N=27781)	female (N=12029)	male (N=13803)	female (N=5887)	male (N=7027)	female (N=5966)	male (N=6951)
Median [Min, Max]	2.00 [0, 30.0]	2.00 [0, 28.0]	2.00 [0, 29.0]	2.00 [0, 28.0]	2.00 [0, 23.0]	2.00 [0, 18.0]	2.00 [0, 30.0]	2.00 [0, 23.0]
Missing	2090 (8.8%)	2353 (8.5%)	1047 (8.7%)	1154 (8.4%)	515 (8.7%)	594 (8.5%)	528 (8.9%)	605 (8.7%)
Fluctuating Asymmetry, digit 5								
Mean (SD)	2.38 (2.15)	2.32 (2.07)	2.38 (2.15)	2.30 (2.08)	2.36 (2.11)	2.34 (2.06)	2.38 (2.18)	2.33 (2.06)
Median [Min, Max]	2.00 [0, 17.0]	2.00 [0, 22.0]	2.00 [0, 17.0]	2.00 [0, 19.0]	2.00 [0, 16.0]	2.00 [0, 22.0]	2.00 [0, 16.0]	2.00 [0, 17.0]
Missing	2102 (8.8%)	2315 (8.3%)	1053 (8.8%)	1129 (8.2%)	514 (8.7%)	581 (8.3%)	535 (9.0%)	605 (8.7%)
Index. Asymmetry								
Mean (SD)	8.25 (3.91)	8.17 (3.89)	8.24 (3.89)	8.16 (3.91)	8.18 (3.82)	8.19 (3.82)	8.35 (4.02)	8.17 (3.92)
Median [Min, Max]	7.55 [0, 43.0]	7.42 [0, 34.4]	7.55 [0, 42.3]	7.42 [0, 34.4]	7.48 [0, 30.1]	7.48 [0, 32.4]	7.55 [0, 43.0]	7.42 [0, 31.3]
Missing	2381 (10.0%)	2733 (9.8%)	1186 (9.9%)	1327 (9.6%)	591 (10.0%)	697 (9.9%)	604 (10.1%)	709 (10.2%)
Index. Diversity								

Table S11 Continued

	Data		Training Data		Validation Data		Test Data	
	female (N=23882)	male (N=27781)	female (N=12029)	male (N=13803)	female (N=5887)	male (N=7027)	female (N=5966)	male (N=6951)
Mean (SD)	6.74 (3.09)	6.96 (3.19)	6.75 (3.12)	6.94 (3.17)	6.68 (3.03)	6.95 (3.19)	6.77 (3.08)	7.00 (3.25)
Median [Min, Max]	6.24 [0, 21.9]	6.43 [0, 23.4]	6.24 [0, 21.5]	6.42 [0, 23.4]	6.20 [0, 21.9]	6.41 [0, 23.1]	6.29 [0, 20.2]	6.47 [0, 23.1]
Missing	2381 (10.0%)	2733 (9.8%)	1186 (9.9%)	1327 (9.6%)	591 (10.0%)	697 (9.9%)	604 (10.1%)	709 (10.2%)
Index. Diversity. Holt								
Mean (SD)	3.93 (1.56)	4.02 (1.61)	3.94 (1.57)	4.01 (1.60)	3.90 (1.53)	4.02 (1.60)	3.96 (1.56)	4.04 (1.64)
Median [Min, Max]	3.69 [0, 12.1]	3.74 [0, 12.0]	3.69 [0, 11.7]	3.72 [0, 12.0]	3.66 [0, 11.1]	3.74 [0, 11.8]	3.72 [0, 12.1]	3.77 [0, 11.7]
Missing	2381 (10.0%)	2733 (9.8%)	1186 (9.9%)	1327 (9.6%)	591 (10.0%)	697 (9.9%)	604 (10.1%)	709 (10.2%)

Appendix C Supplementary figures, Chapter Four

Appendix C.1 Supplementary figures for finger ridge-counts and summaries of ridge-counts

Appendix C.1.1 Supplementary figures for Spatial analysis via distance-based Moran's eigenvector maps (dbMEM)

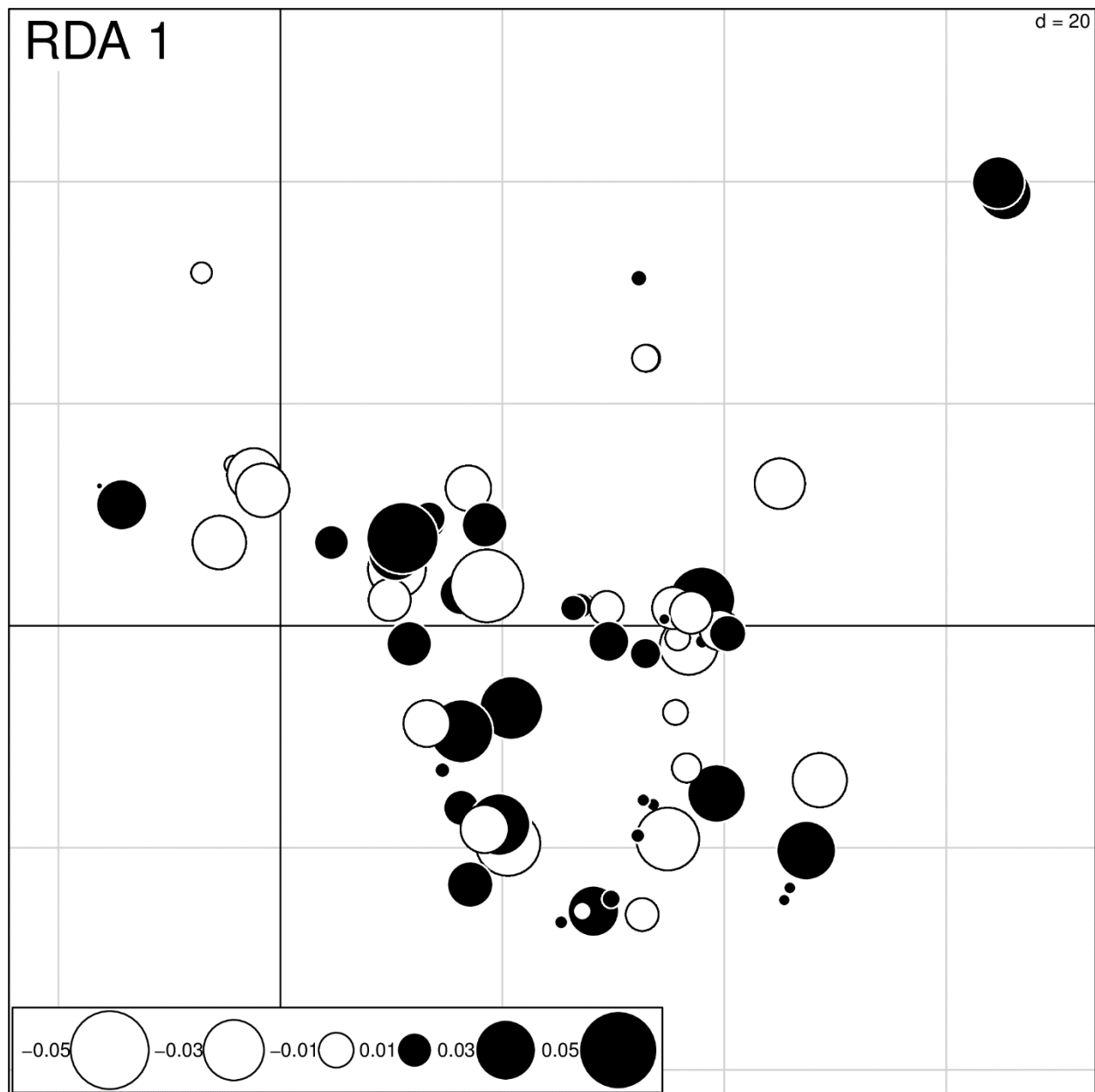


Figure S1. dbMEM analysis of detrended Landmass Africa finger patterns with the two significant dbMEM variables. Maps of the fitted site scores of the first canonical axis, which was significant at the $\alpha = 0.05$ threshold.

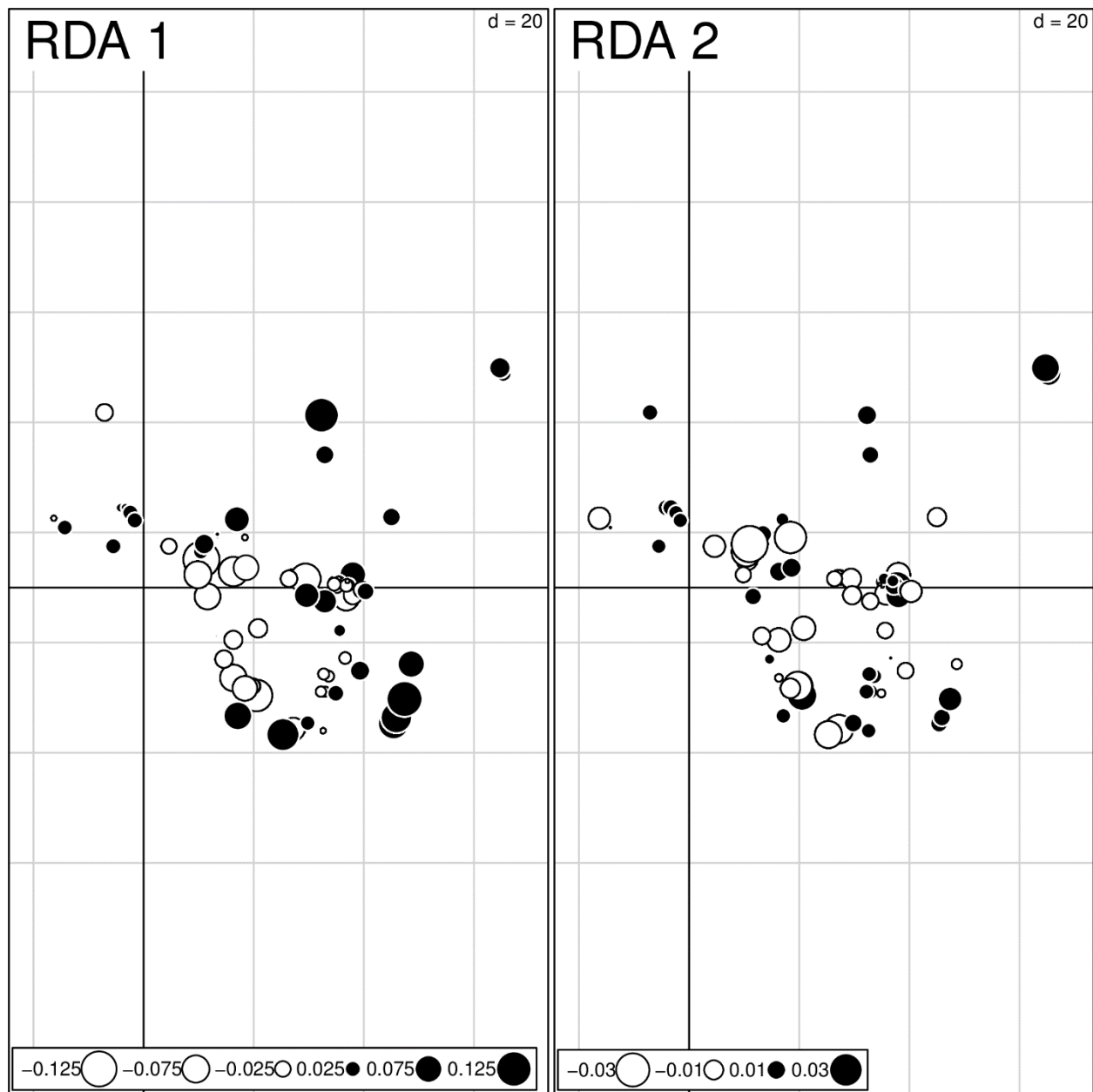


Figure S2. dbMEM analysis of detrended Landmass Africa finger ridge-counts with the six significant dbMEM variables. Maps of the fitted site scores of the first two canonical axes, which were significant at the $\alpha = 0.05$ threshold.

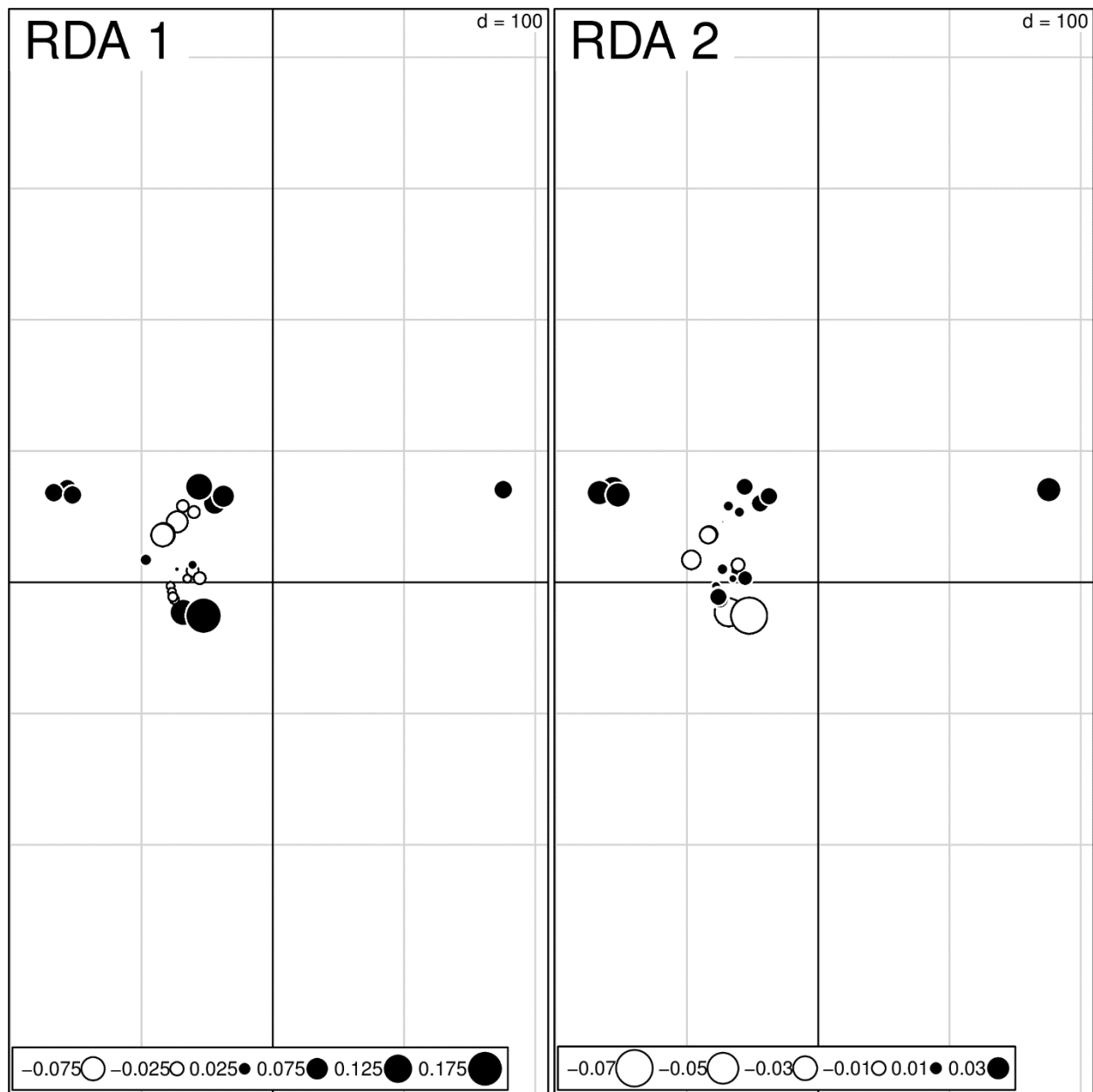


Figure S3. dbMEM analysis of detrended Landmass Americas finger patterns with the five significant dbMEM variables. Maps of the fitted site scores of the first two canonical axes, which were significant at the $\alpha = 0.05$ threshold.

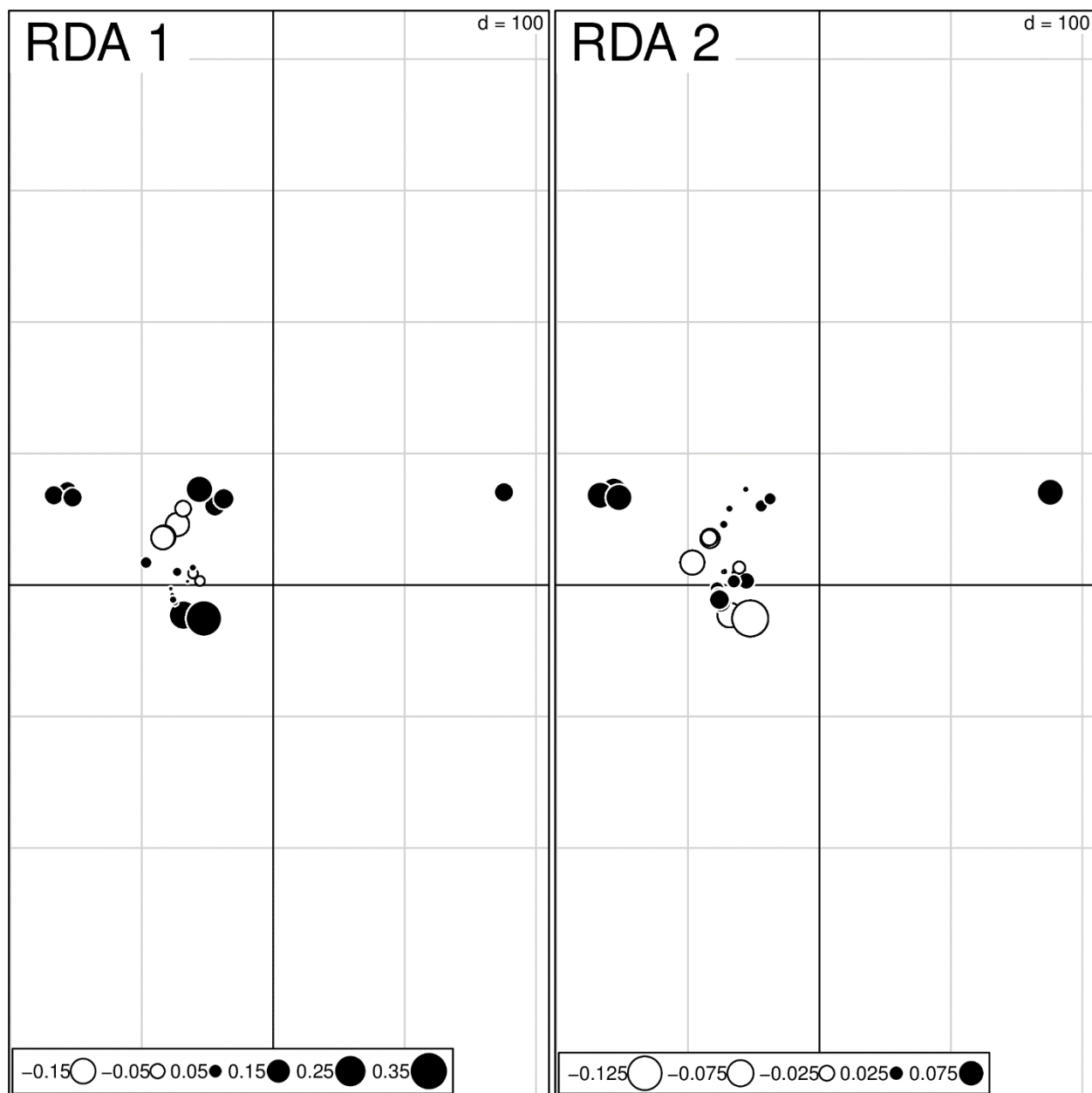


Figure S4. dbMEM analysis of detrended Landmass Americas finger ridge-counts with the five significant dbMEM variables. Maps of the fitted site scores of the first two canonical axes, which were significant at the $\alpha = 0.05$ threshold.

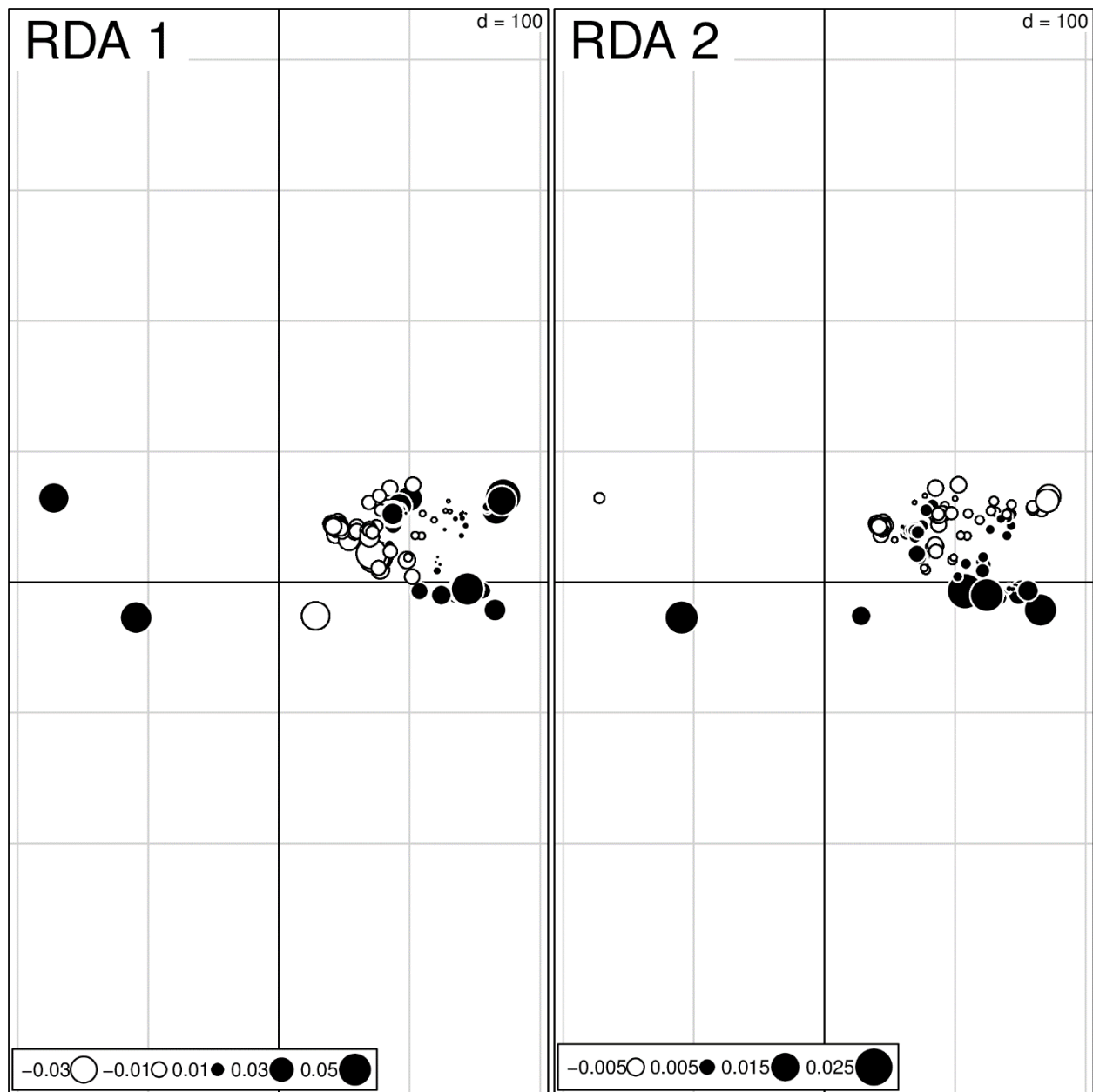


Figure S5. dbMEM analysis of detrended Landmass Asia finger patterns with the three significant dbMEM variables. Maps of the fitted site scores of the first two canonical axes, which were significant at the $\alpha = 0.05$ threshold.

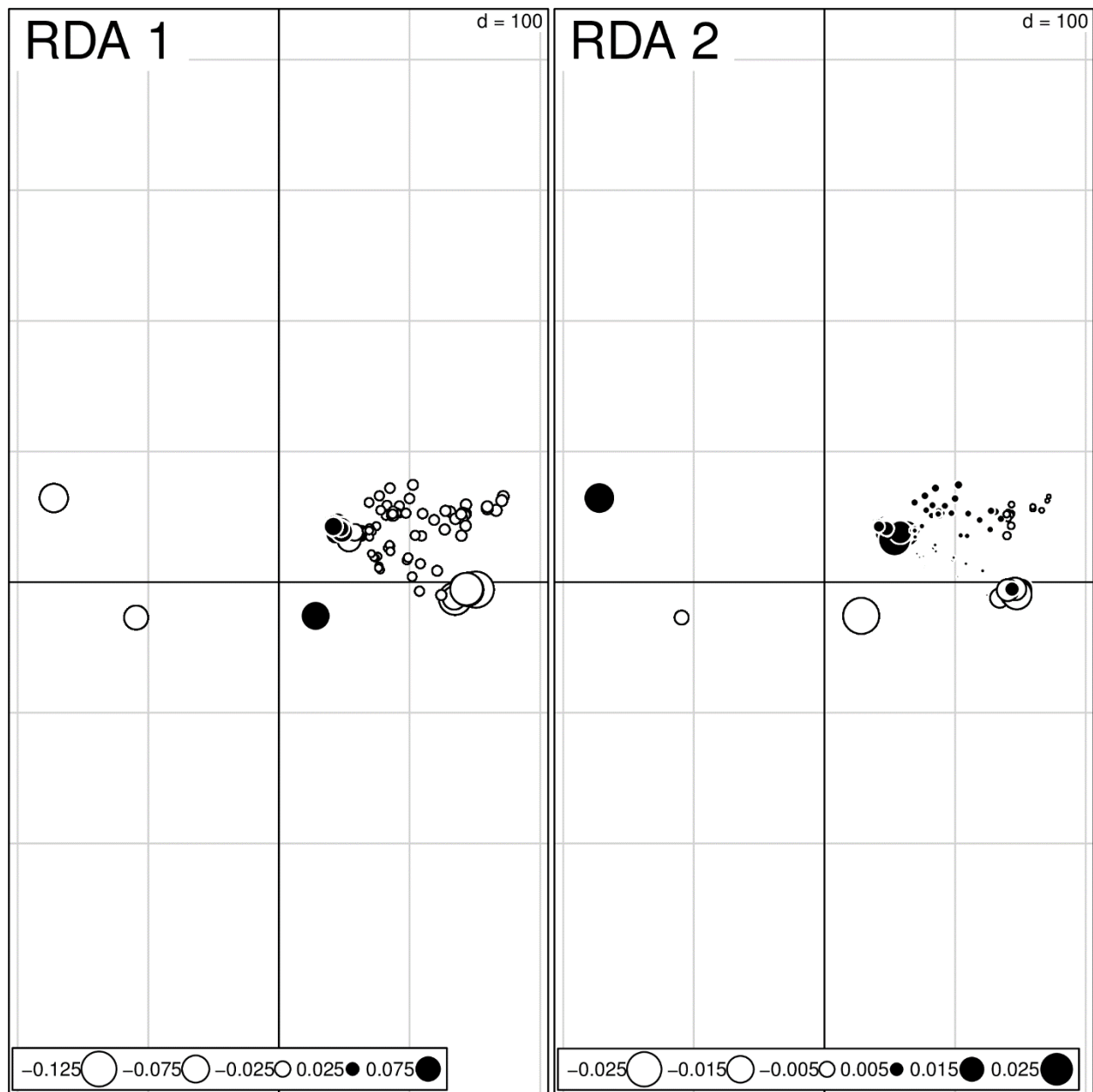


Figure S6. dbMEM analysis of detrended Landmass Asia finger ridge-counts with the two significant dbMEM variables. Maps of the fitted site scores of the first two canonical axes, which were significant at the $\alpha = 0.05$ threshold.

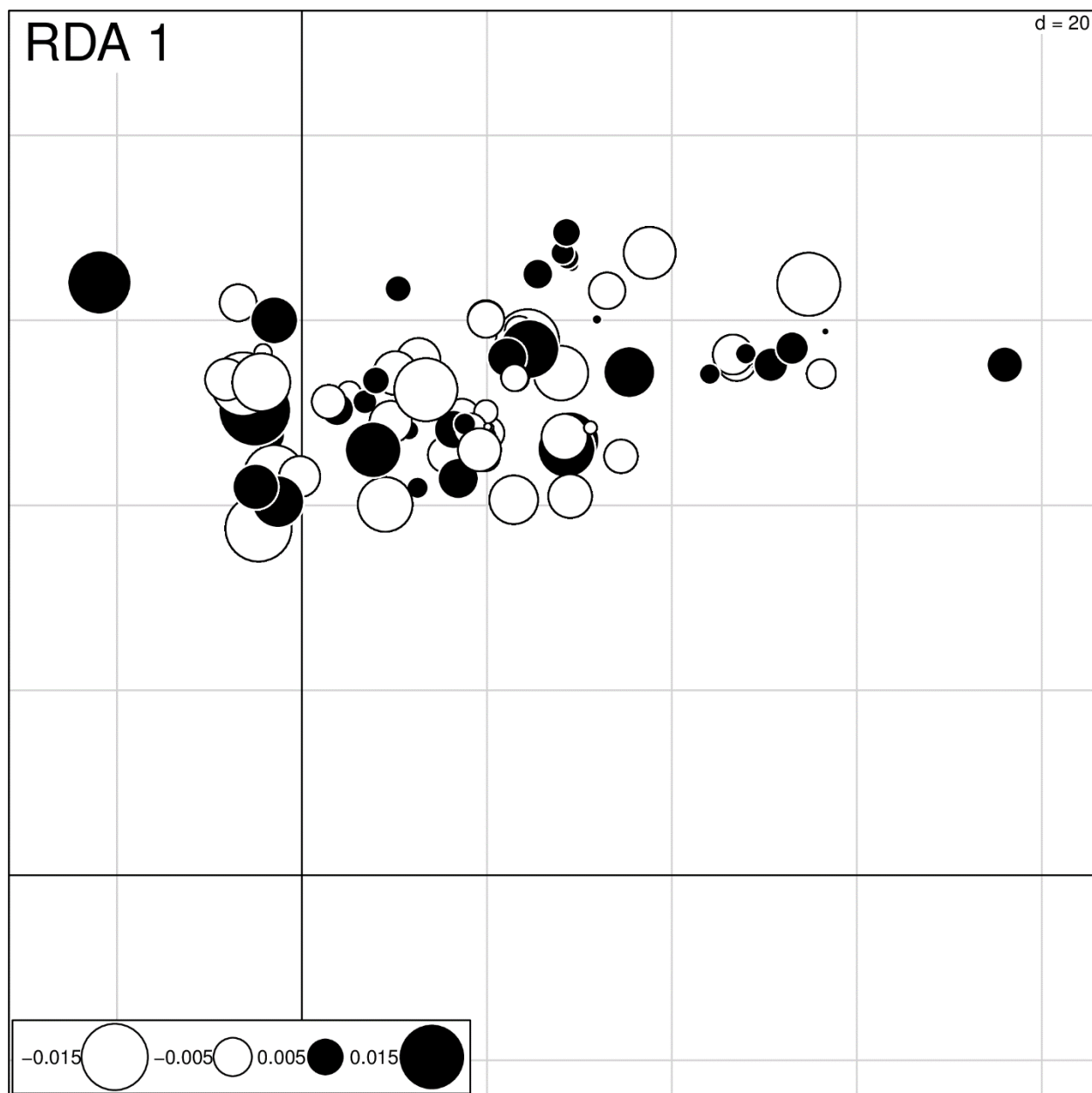


Figure S7. dbMEM analysis of detrended Landmass Europe finger patterns with the one significant dbMEM variable. Maps of the fitted site scores of the first canonical axis, which was significant at the $\alpha = 0.05$ threshold.

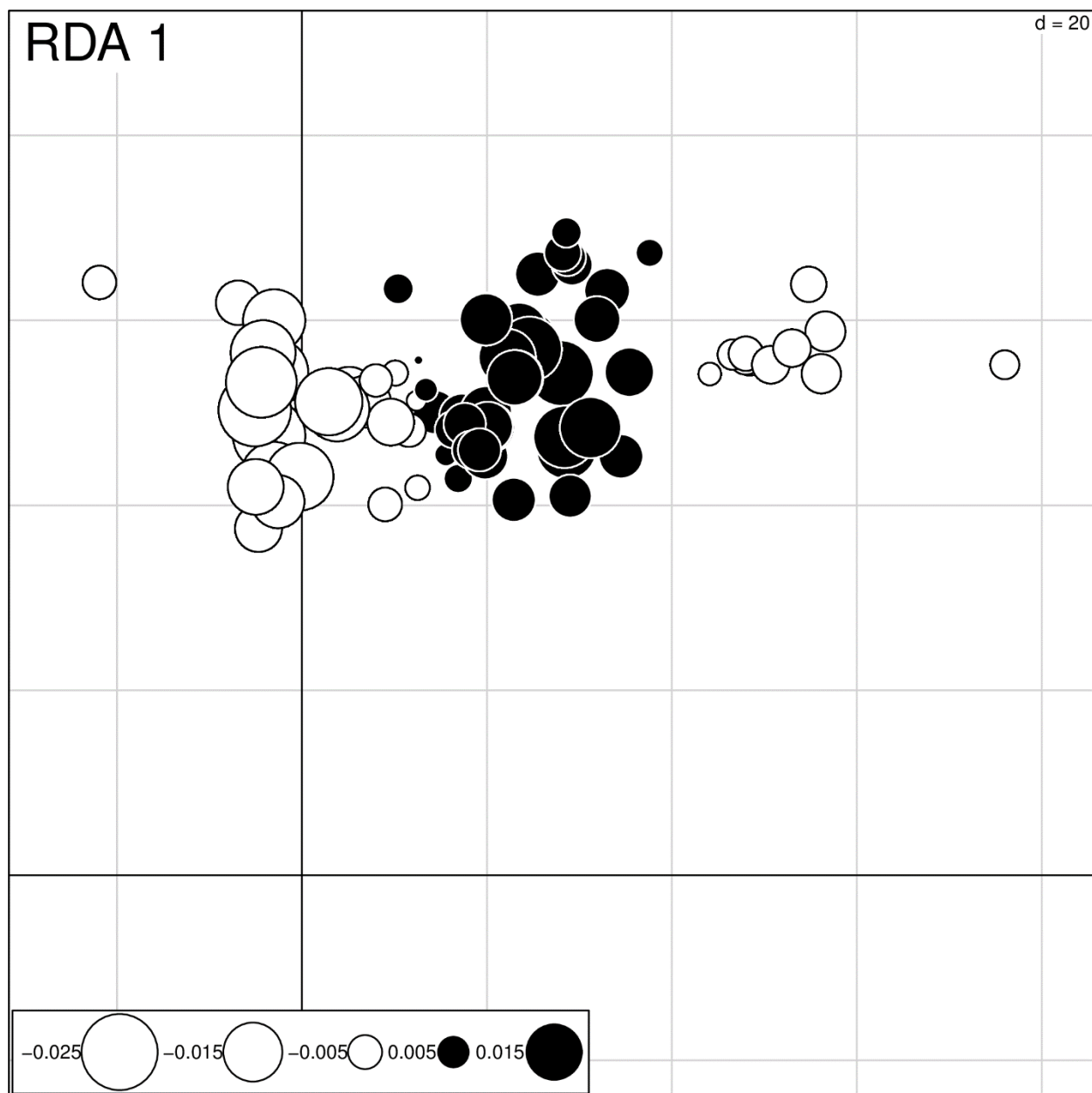


Figure S8. dbMEM analysis of detrended Landmass Europe finger ridge-counts with the one significant dbMEM variable. Maps of the fitted site scores of the first canonical axis, which was significant at the $\alpha = 0.05$ threshold.

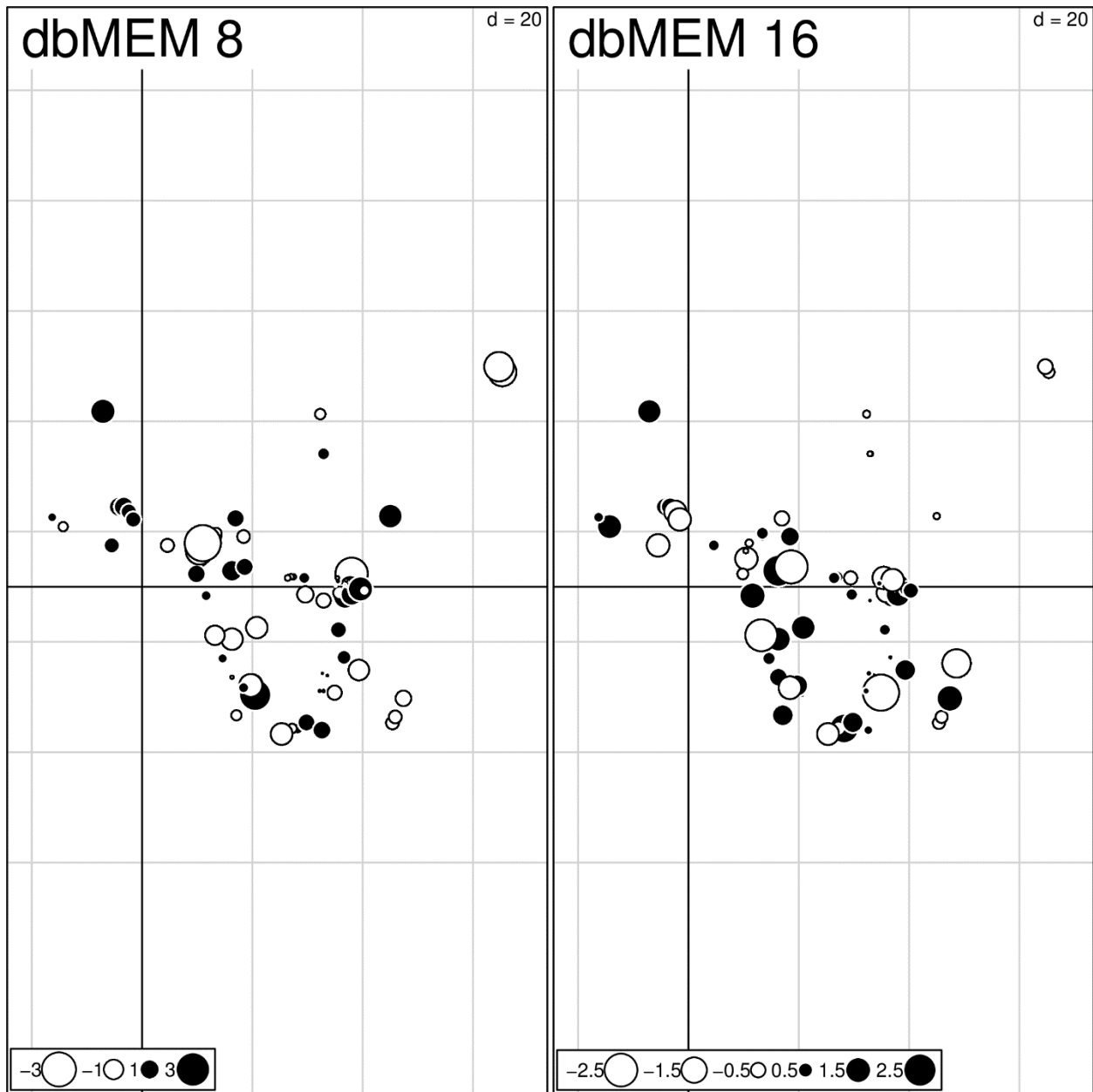


Figure S9. Map of two significant dbMEM variables for Landmass Africa finger patterns

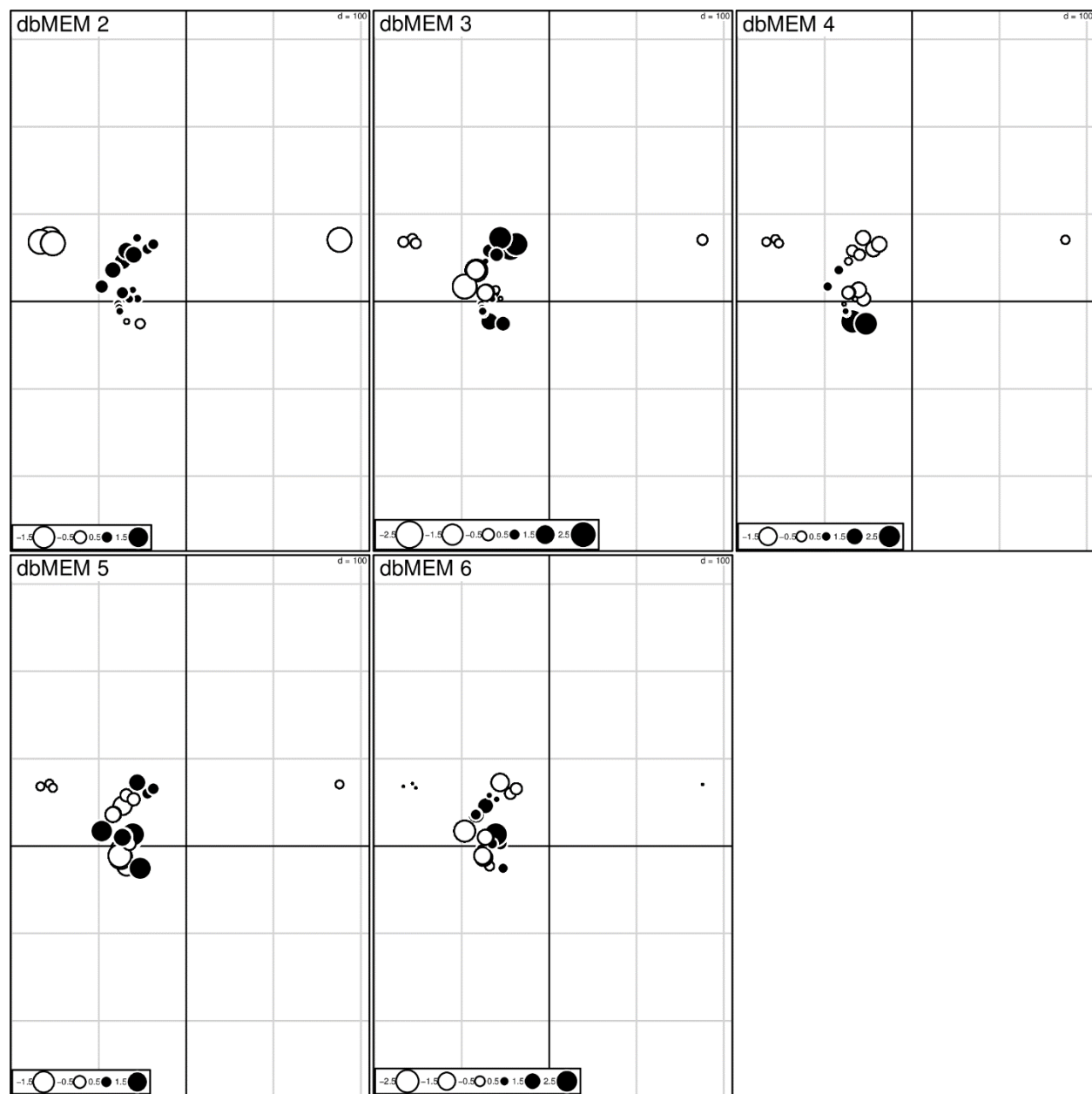


Figure S10. Map of five significant dbMEM variables for Landmass Americas finger patterns

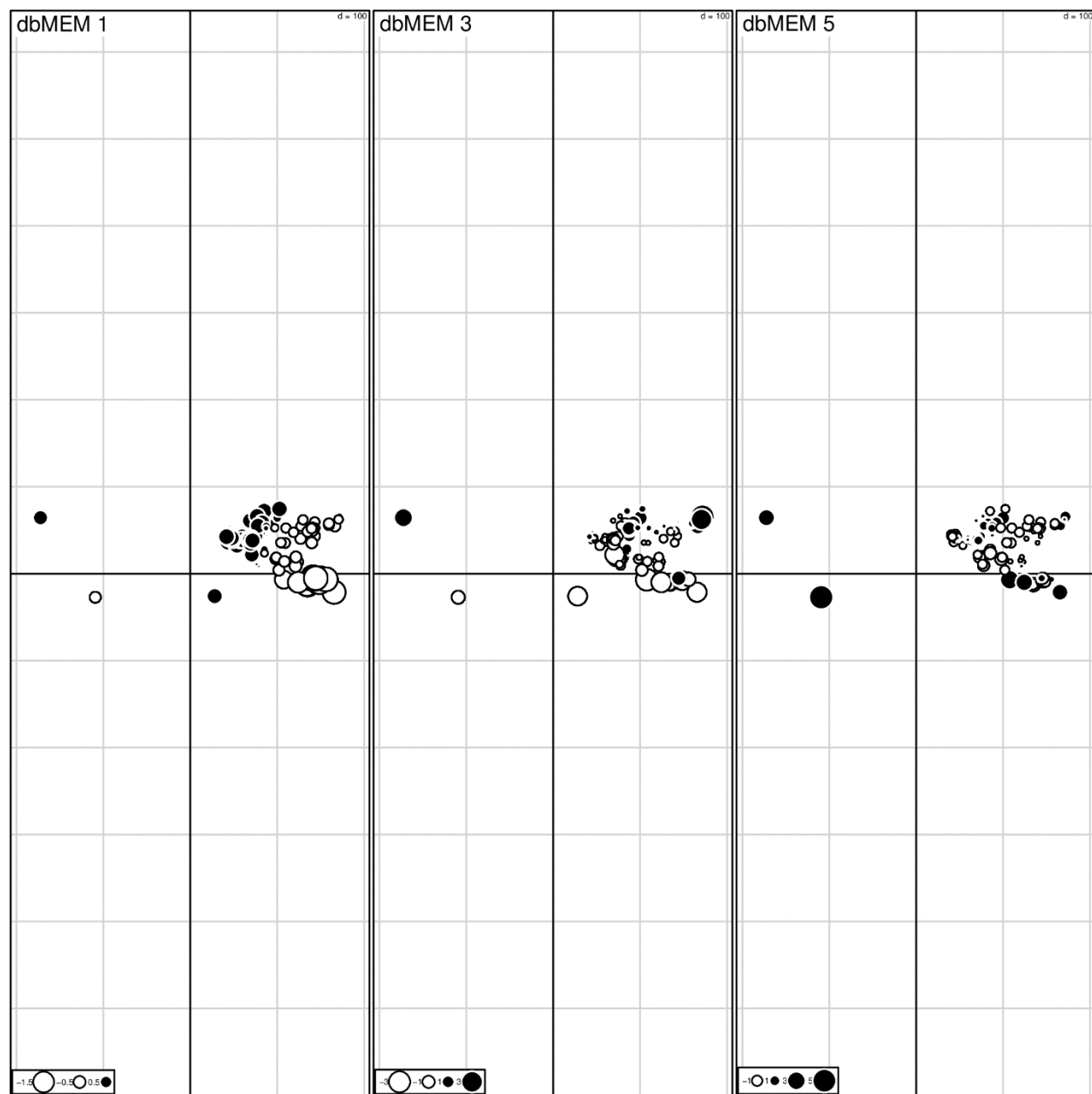


Figure S11. Map of three significant dbMEM variables for Landmass Asia finger patterns

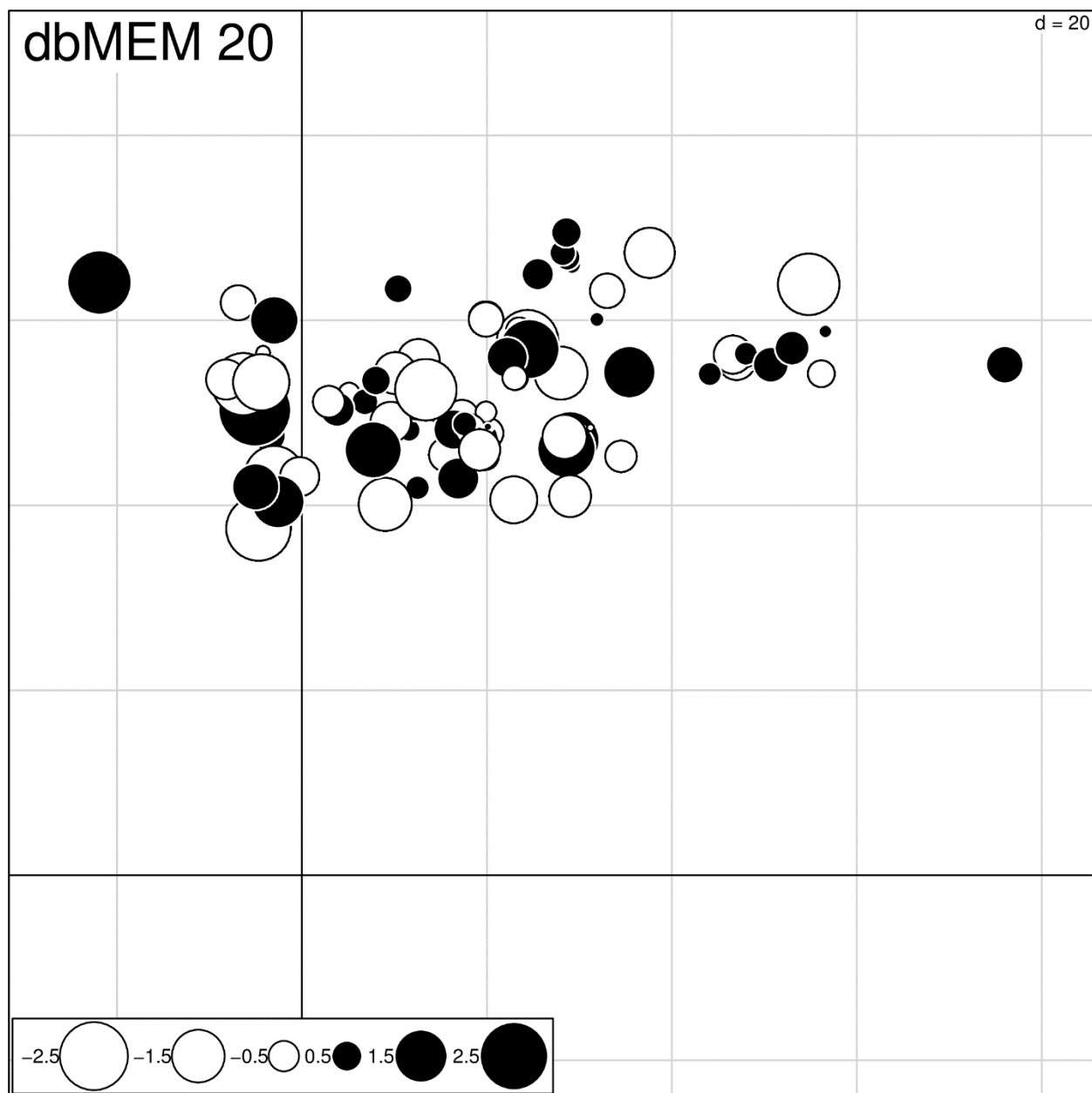


Figure S12. Map of one significant dbMEM variable for Landmass Europe finger patterns

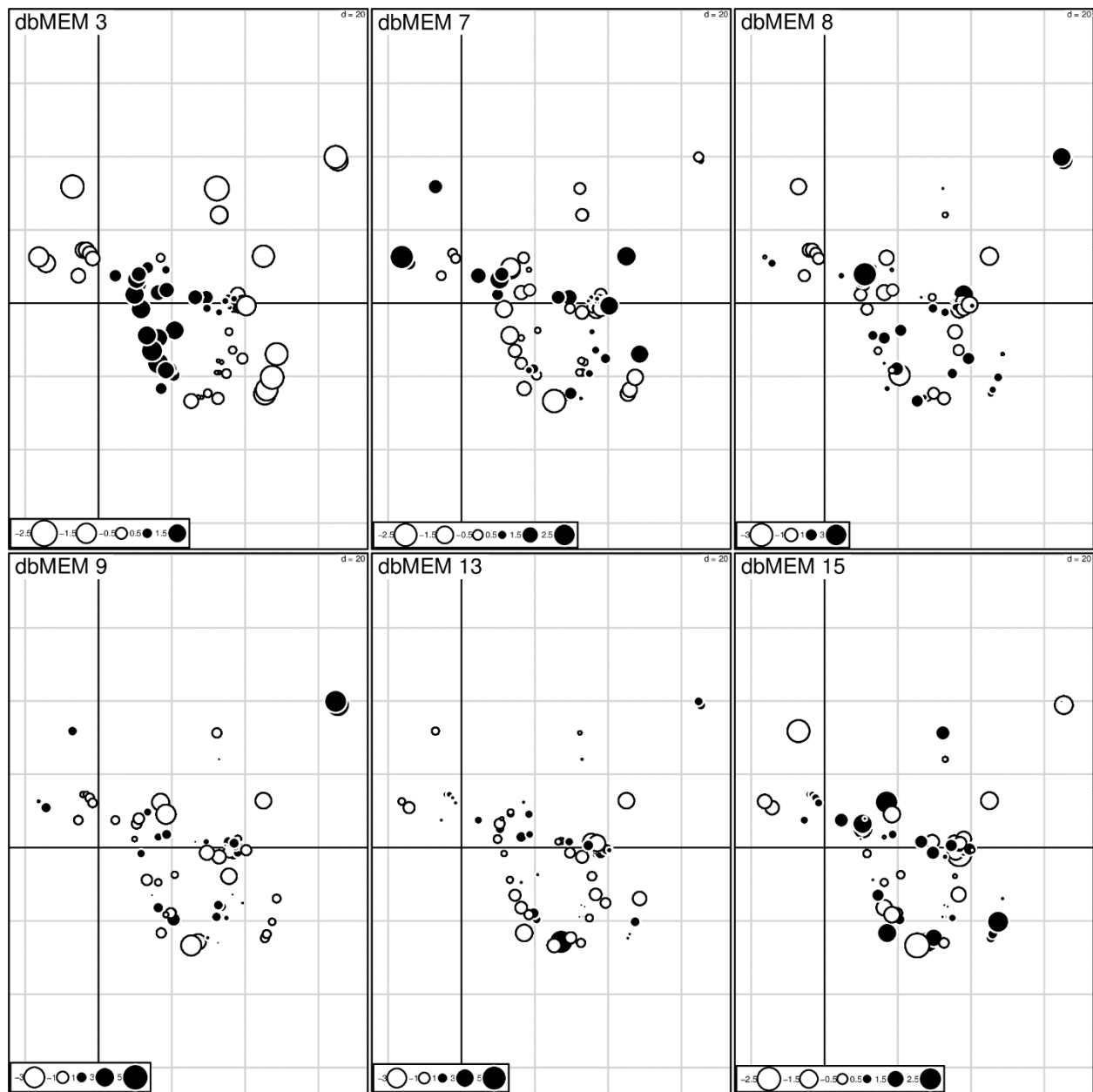


Figure S13. Map of six significant dbMEM variables for Landmass Africa finger ridge-counts

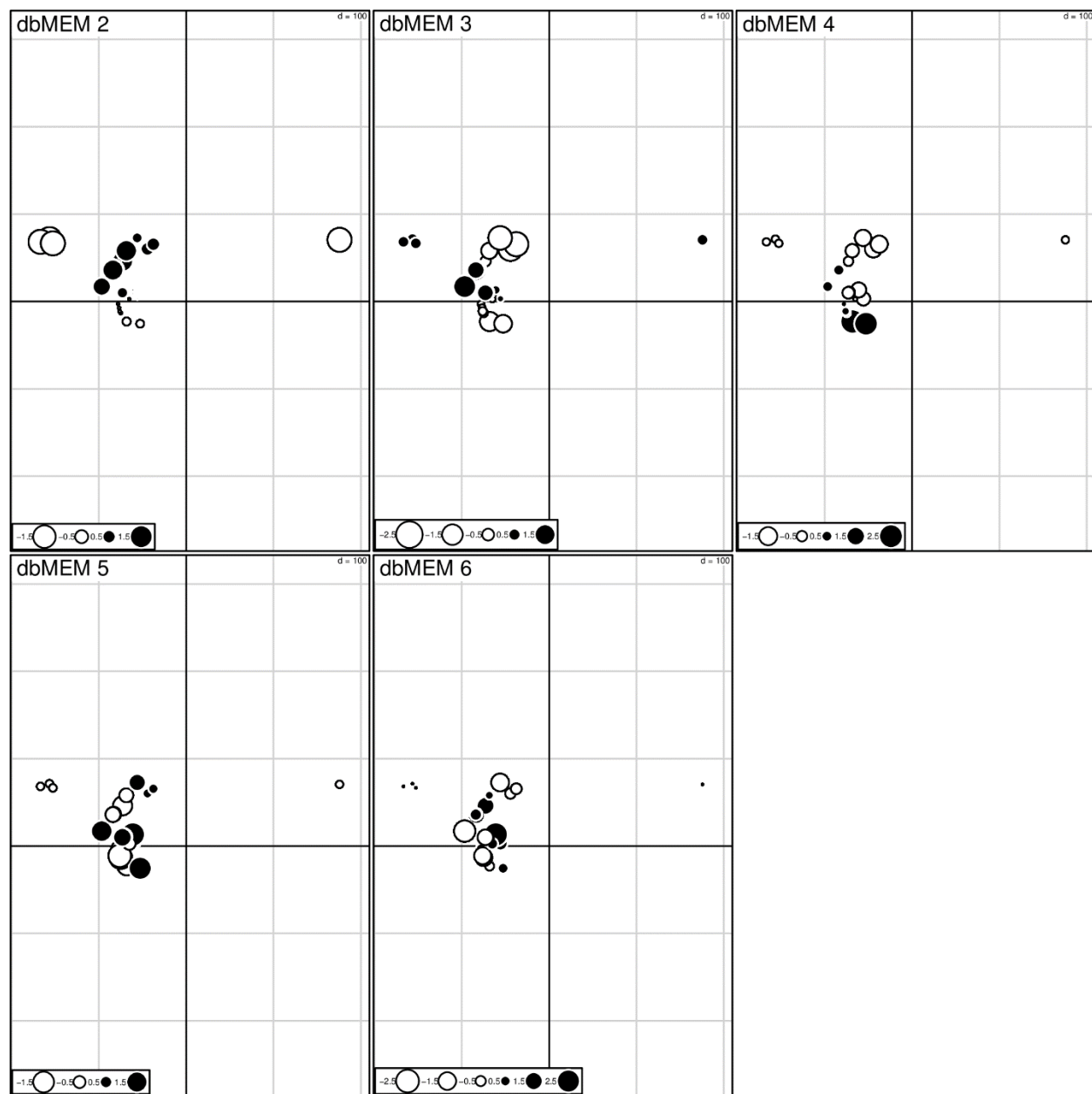


Figure S14. Map of five significant dbMEM variables for Landmass Americas finger ridge-counts

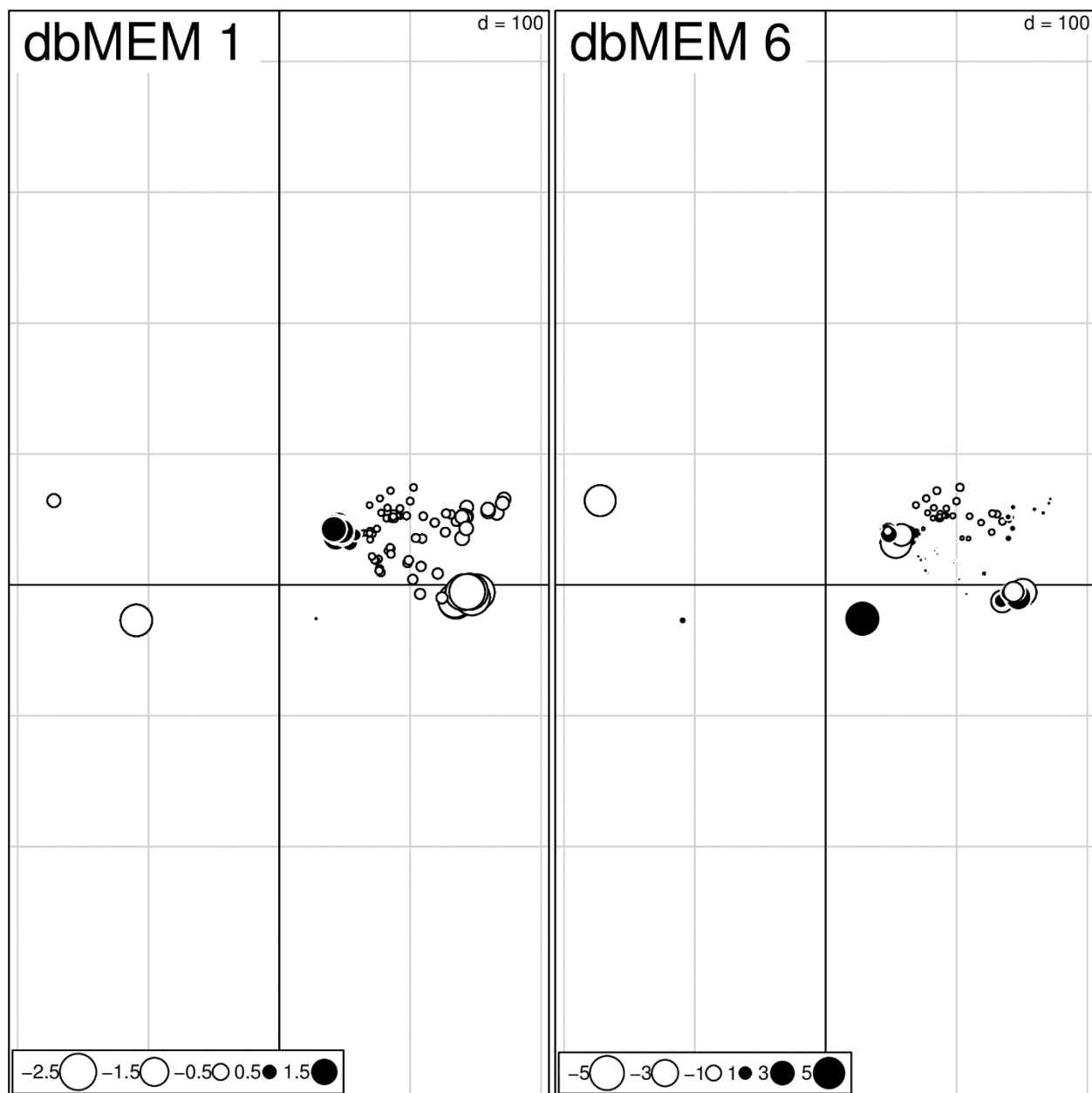


Figure S15. Map of two significant dbMEM variables for Landmass Asia finger ridge-counts

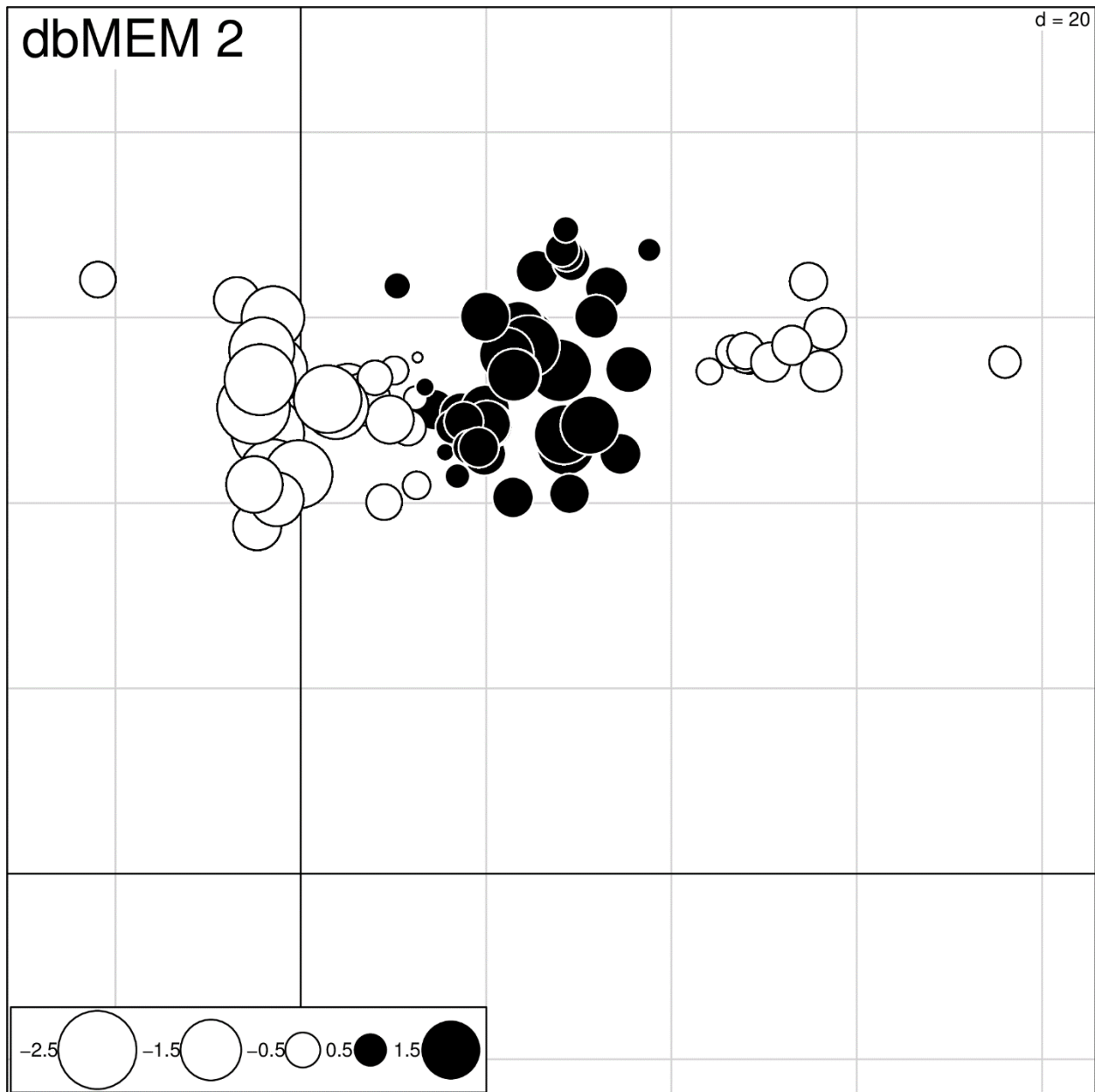
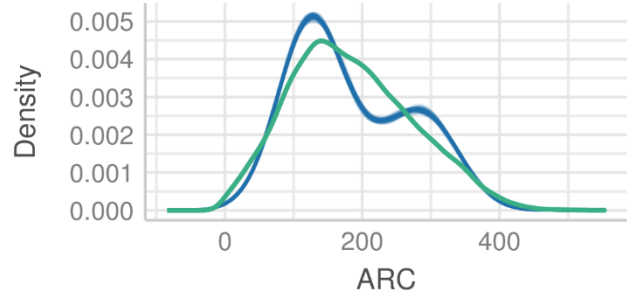
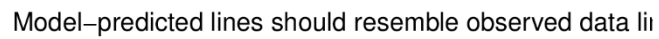


Figure S16. Map of one significant dbMEM variable for Landmass Europe finger ridge-counts

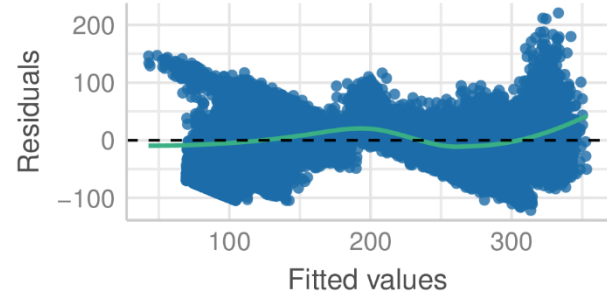
Appendix C.1.2 Supplementary figures for Baseline models for generalized linear modeling



— Observed data — Model-predicted data

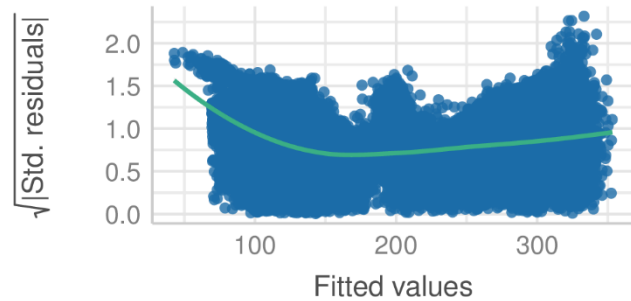
Linearity

Reference line should be flat and horizontal



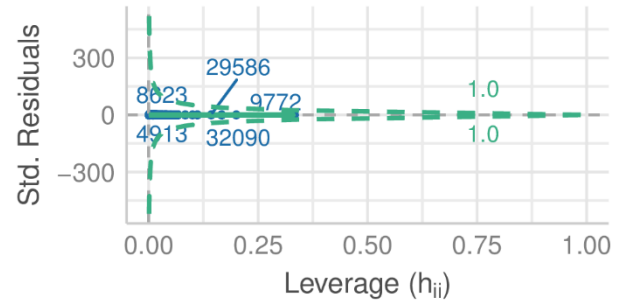
Homogeneity of Variance

Reference line should be flat and horizontal



Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5)

Normality of Residuals

Dots should fall along the line

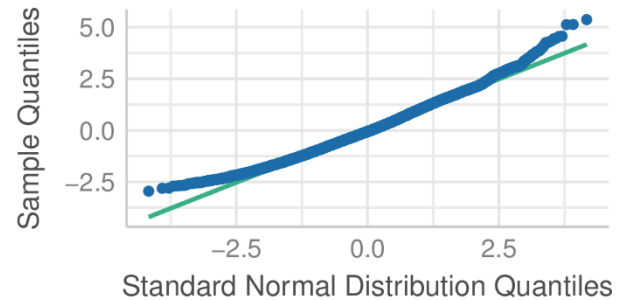


Figure S17. Linear model assumptions check for baseline linear model of ARC response variable, from GLM data set

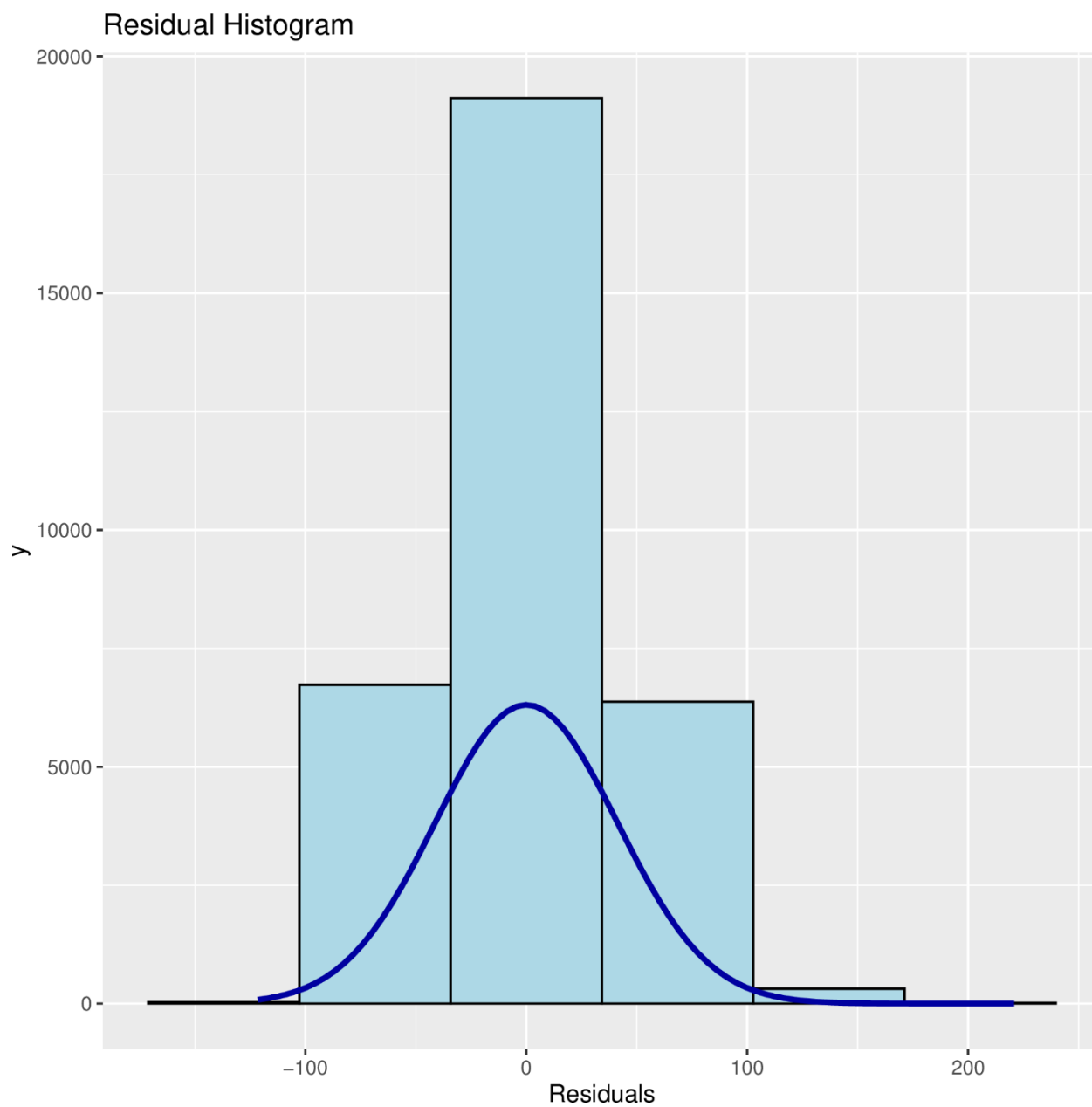
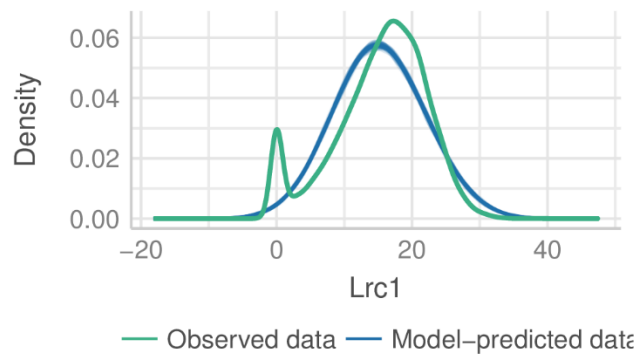


Figure S18. Histogram of residuals for baseline linear model of ARC response variable, from GLM data set

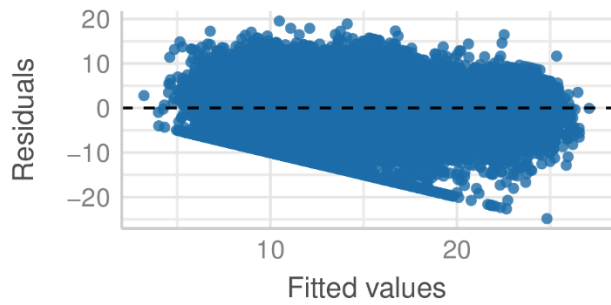
Posterior Predictive Check

Model-predicted lines should resemble observed data



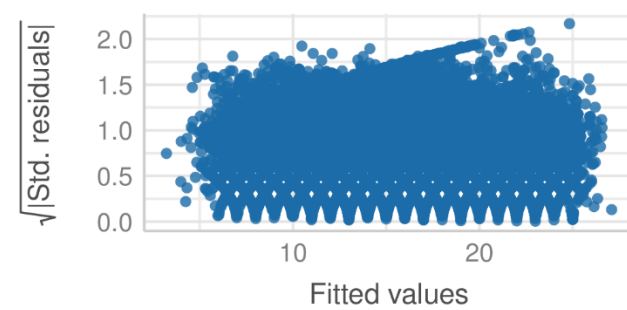
Linearity

Reference line should be flat and horizontal



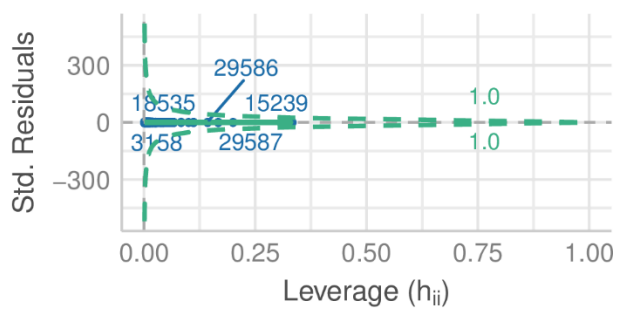
Homogeneity of Variance

Reference line should be flat and horizontal



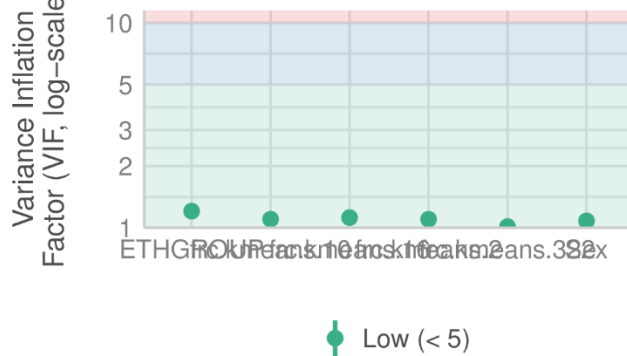
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

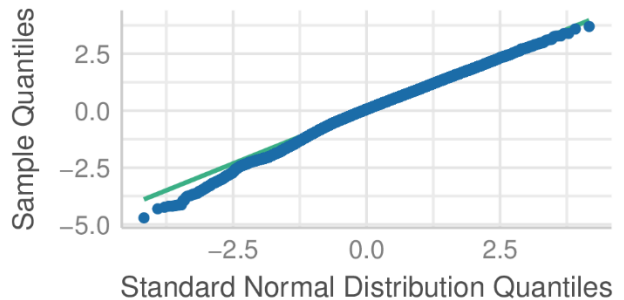


Figure S19. Linear model assumptions check for baseline linear model of Lrc1 response variable, from GLM data set

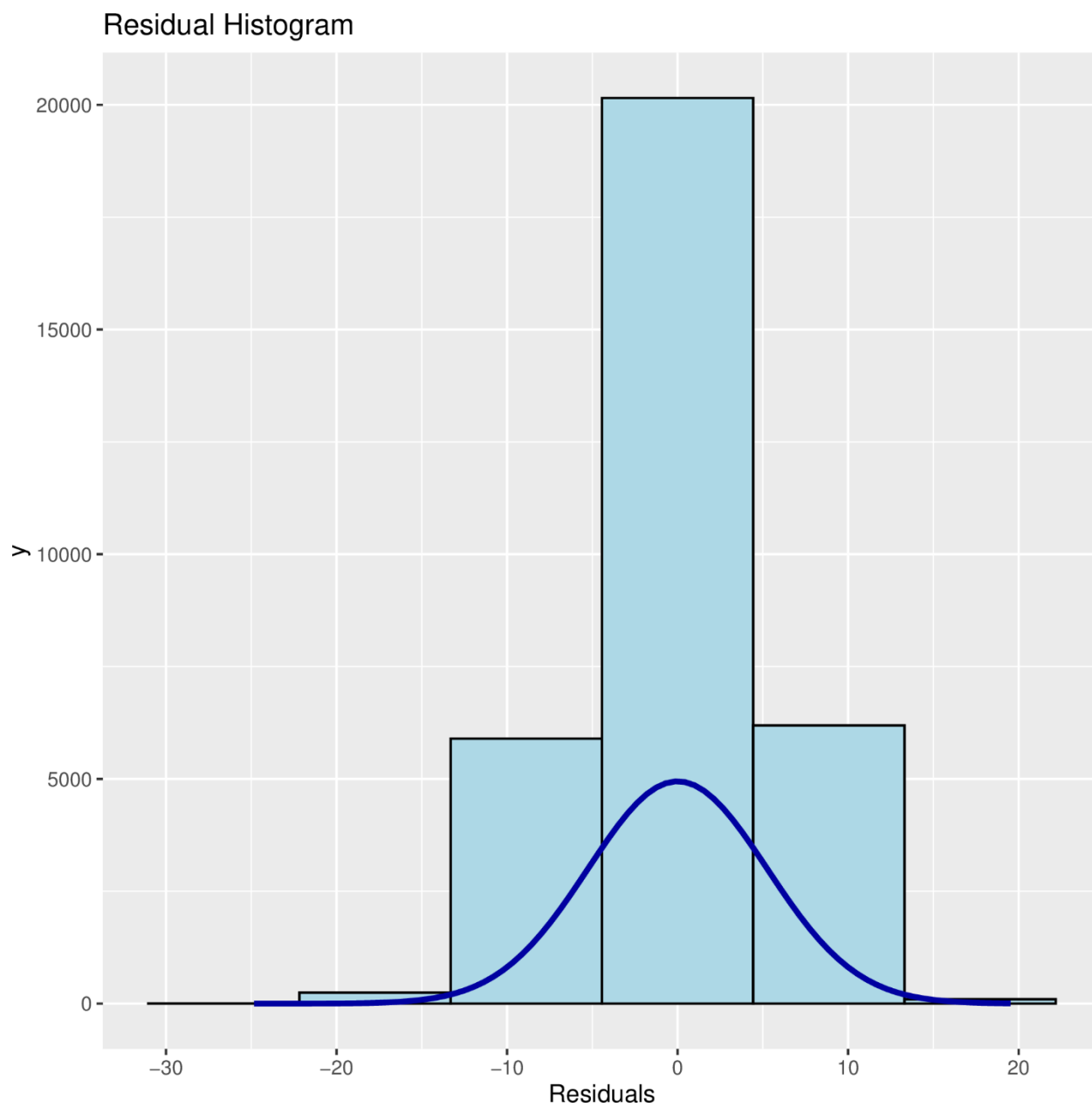
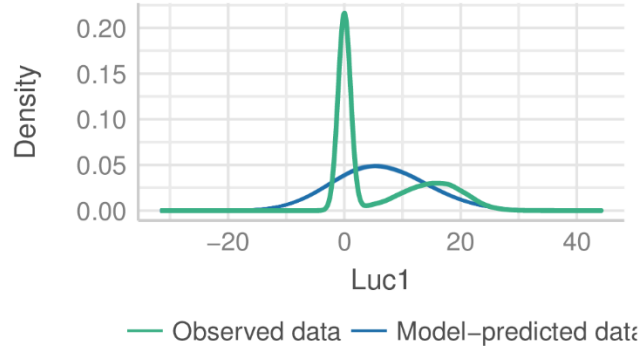


Figure S20. Histogram of residuals for baseline linear model of Lrc1 response variable, from GLM data set

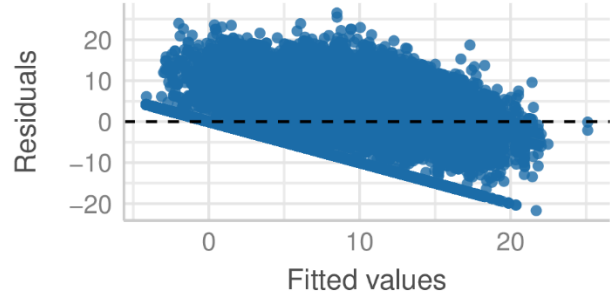
Posterior Predictive Check

Model-predicted lines should resemble observed data



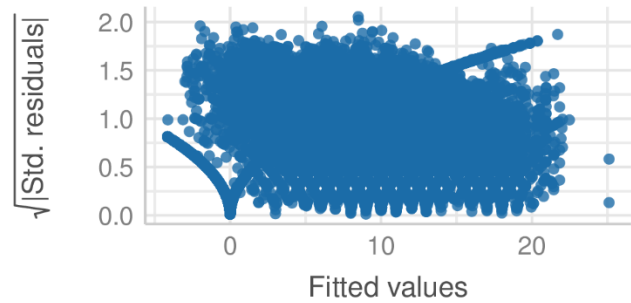
Linearity

Reference line should be flat and horizontal



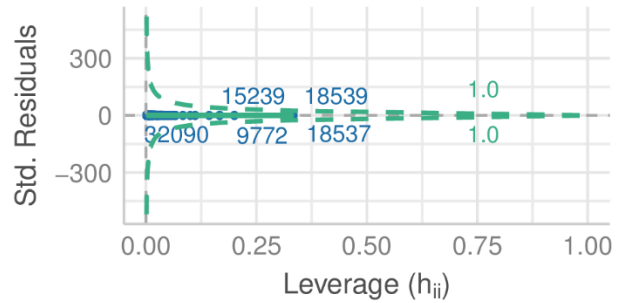
Homogeneity of Variance

Reference line should be flat and horizontal



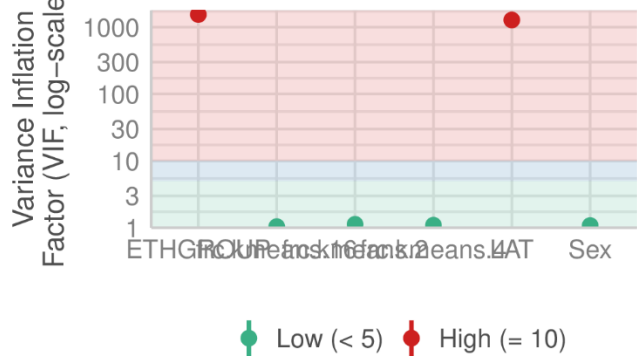
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

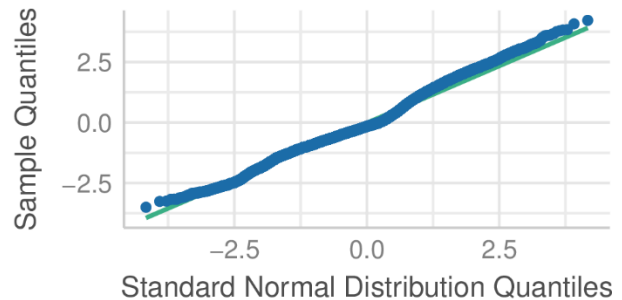


Figure S21. Linear model assumptions check for baseline linear model of Luc1 response variable, from GLM data set

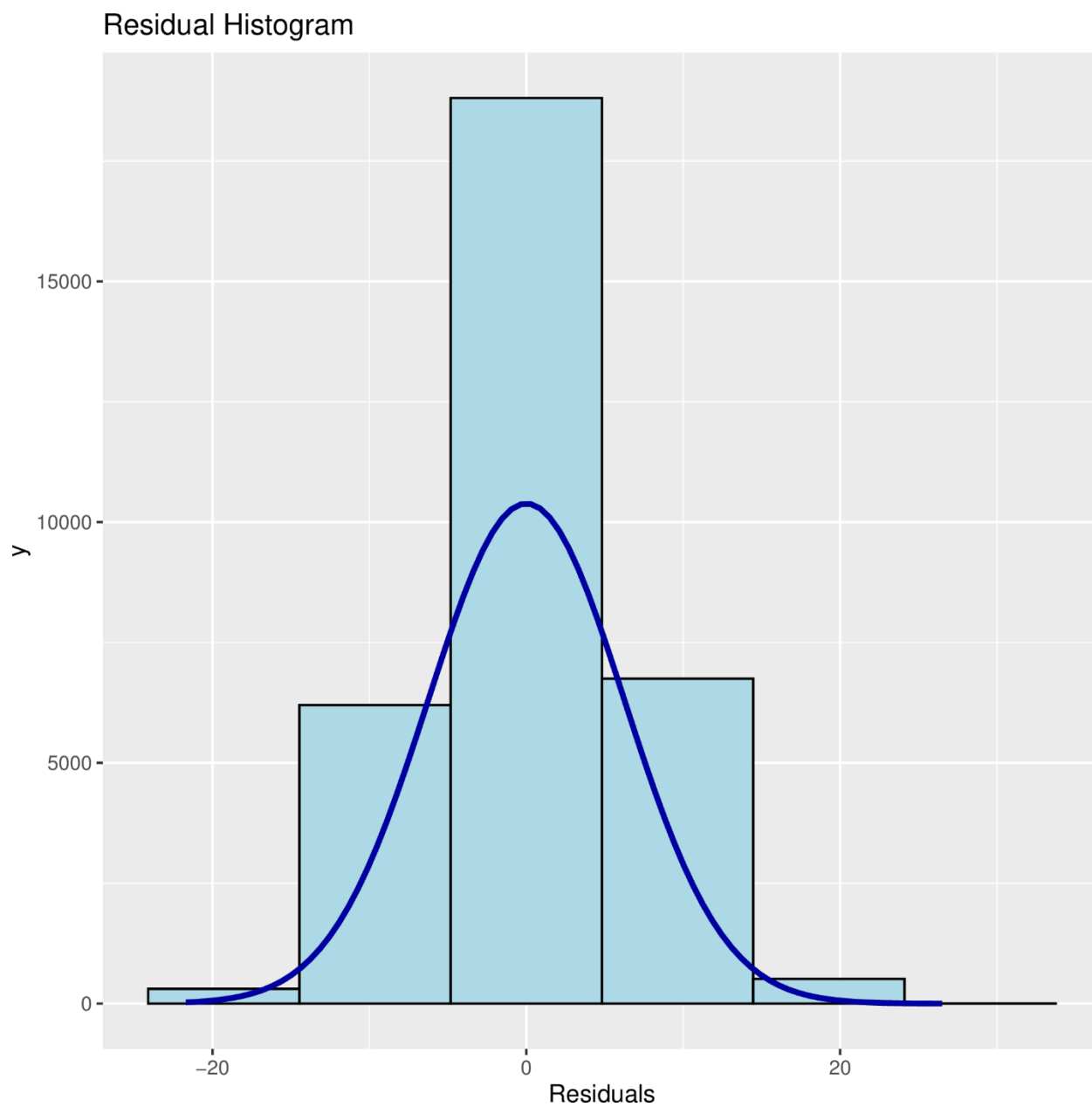
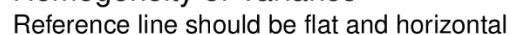
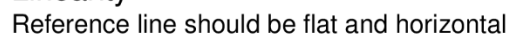
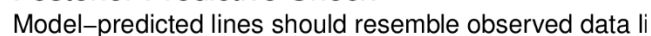
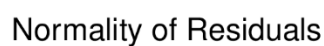
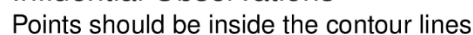


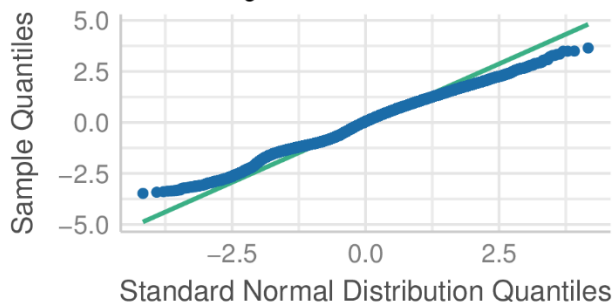
Figure S22. Histogram of residuals for baseline linear model of Luc1 response variable, from GLM data set



High collinearity (VIF) may inflate parameter uncertainty



Dots should fall along the line



470

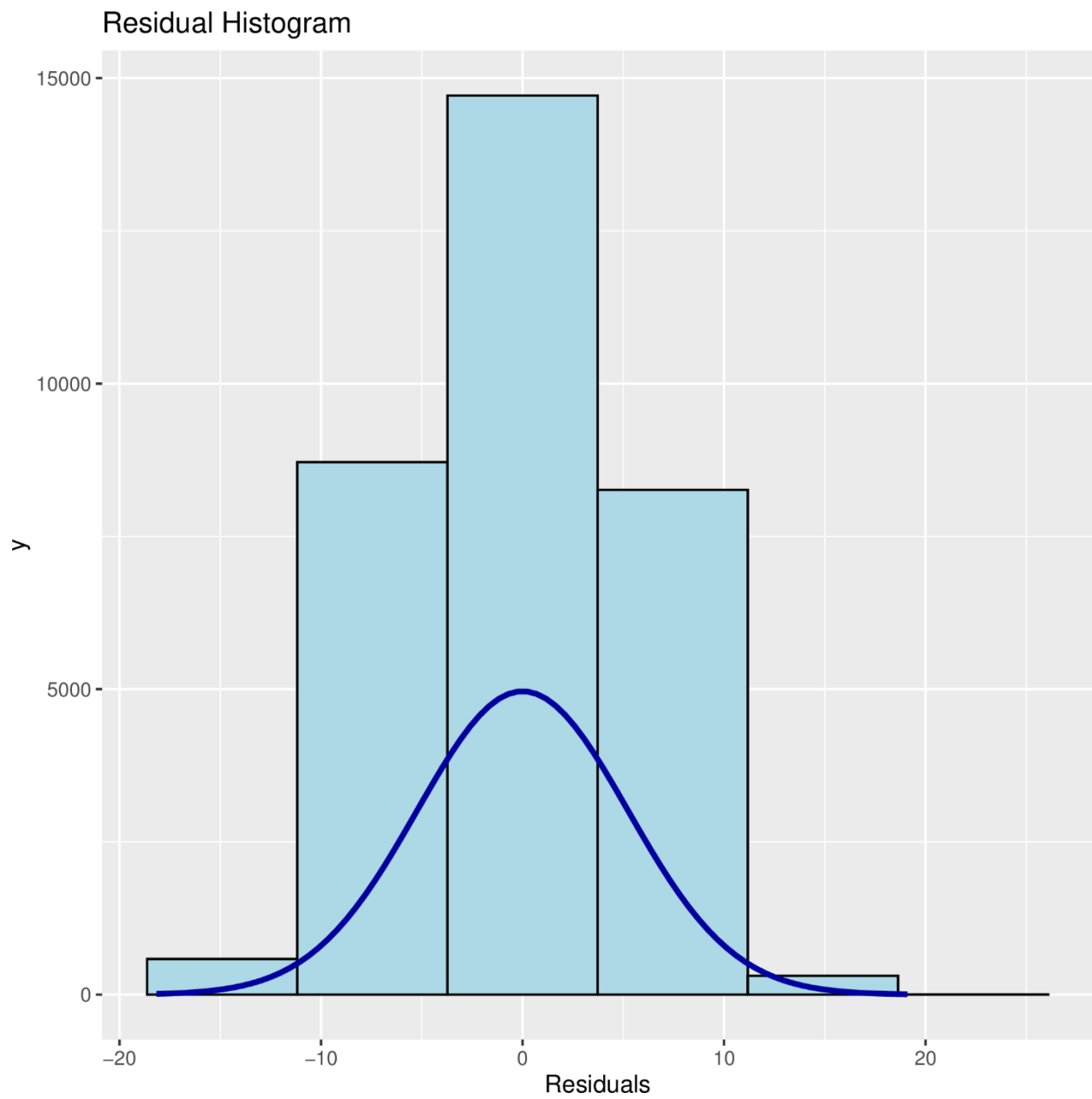
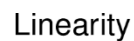


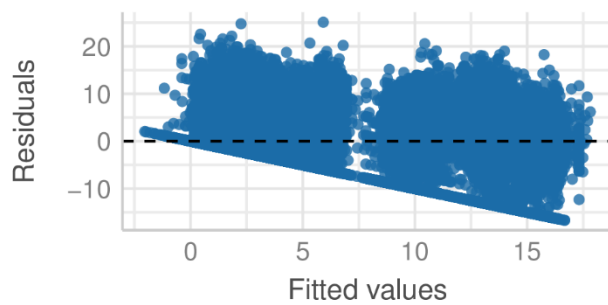
Figure S24. Histogram of residuals for baseline linear model of Lrc2 response variable, from GLM data set

Model-predicted lines should resemble observed data

li Reference line should be flat and horizontal

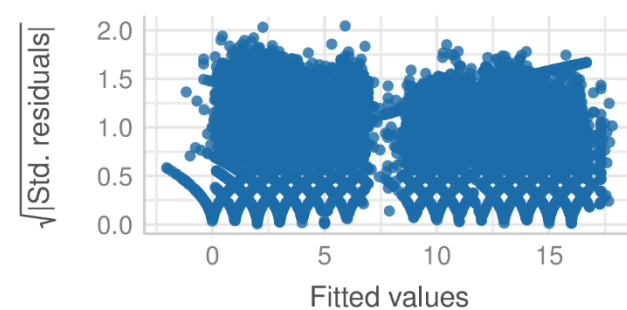


Reference line should be flat and horizontal



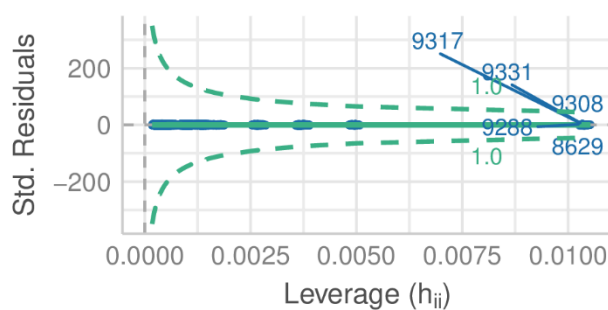
Reference line should be flat and horizontal

Reference line should be flat and horizontal



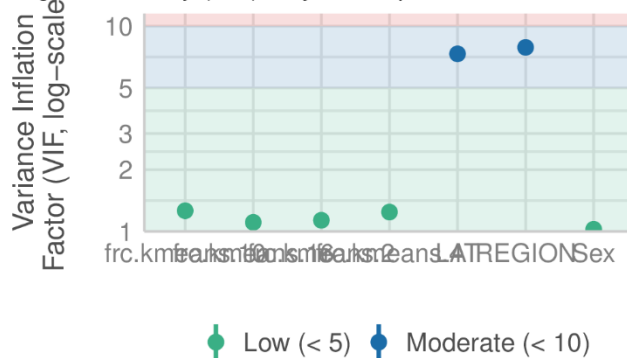
Points should be inside the contour lines

Points should be inside the contour lines



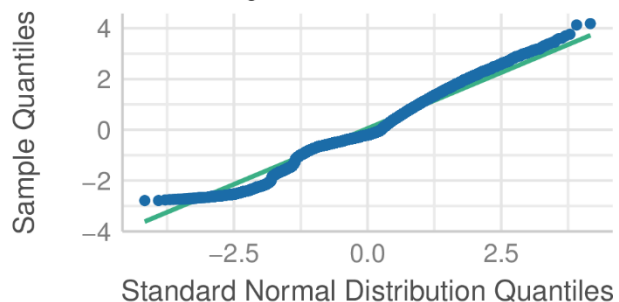
High collinearity (VIF) may inflate parameter uncertainty

High collinearity (VIF) may inflate parameter uncertainty



Dots should fall along the line

Dots should fall along the line



472

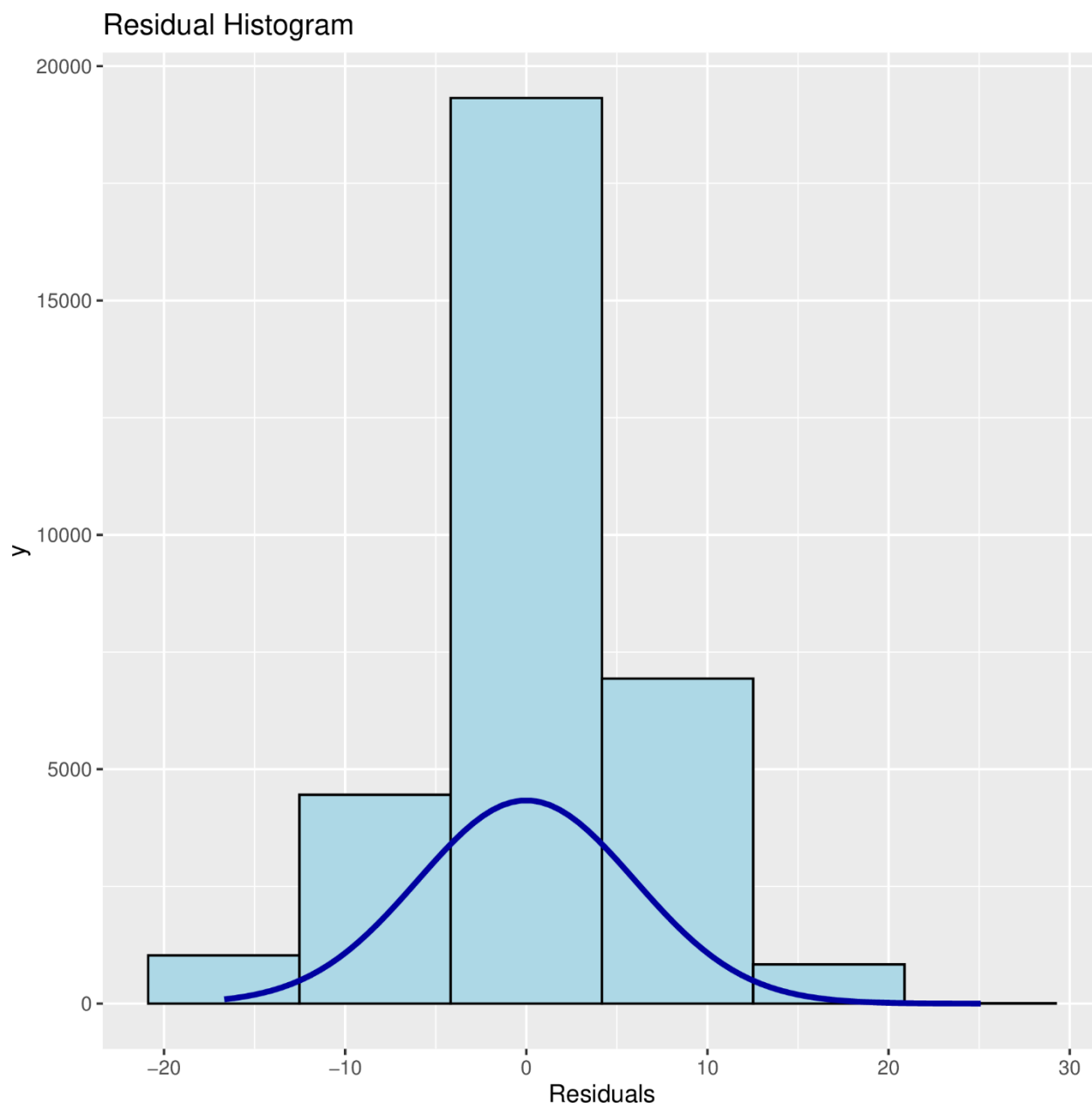
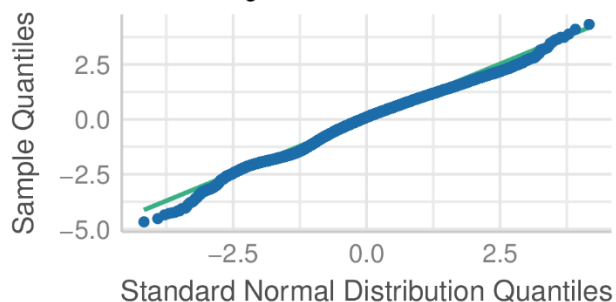
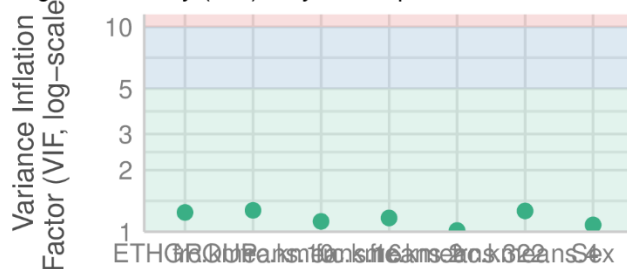
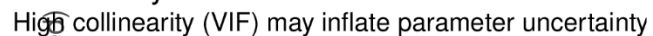
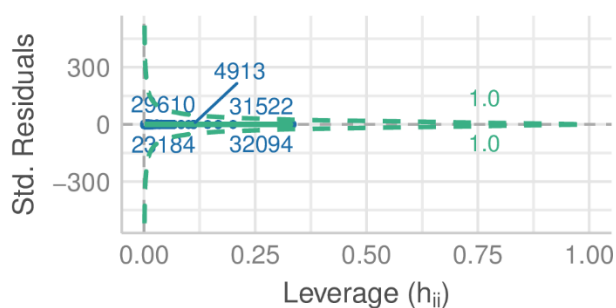
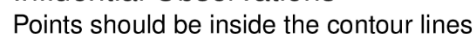
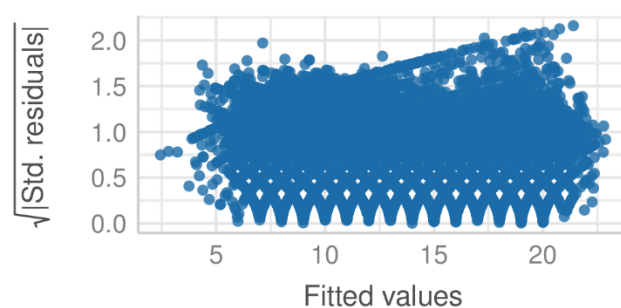
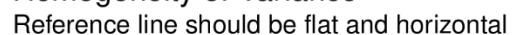
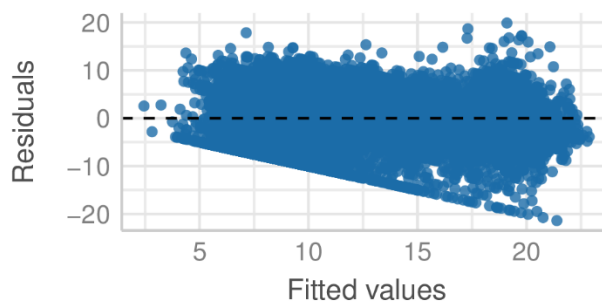
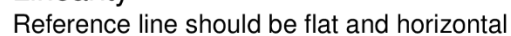
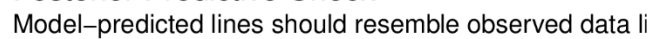


Figure S26. Histogram of residuals for baseline linear model of Luc2 response variable, from GLM data set



474

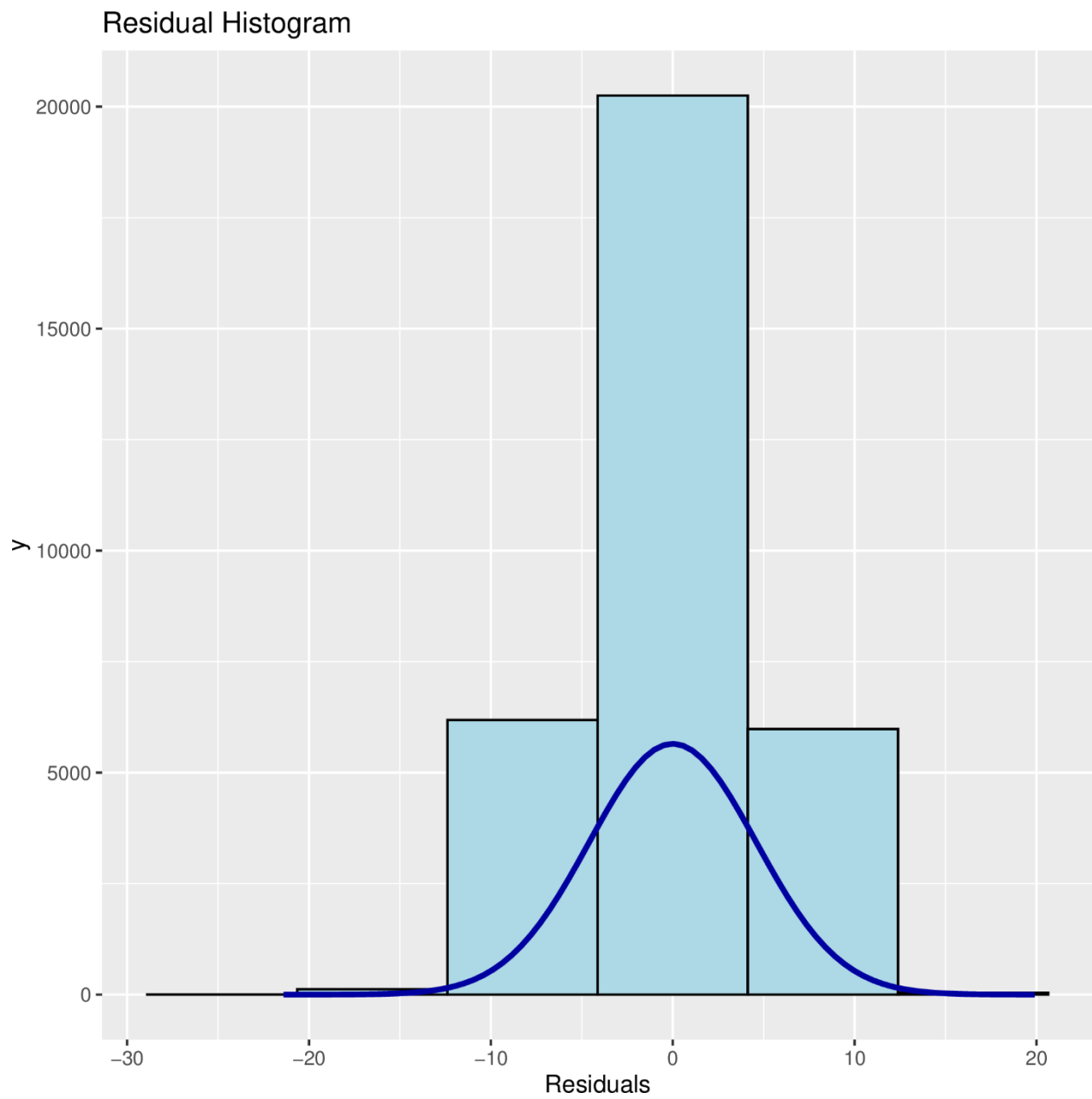


Figure S28. Histogram of residuals for baseline linear model of Lrc3 response variable, from GLM data set

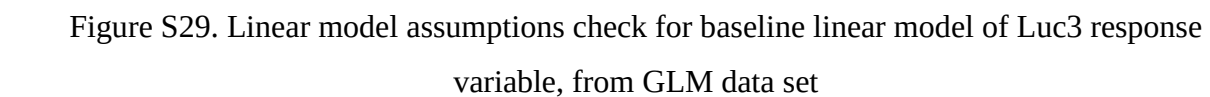
Model-predicted lines should resemble observed data



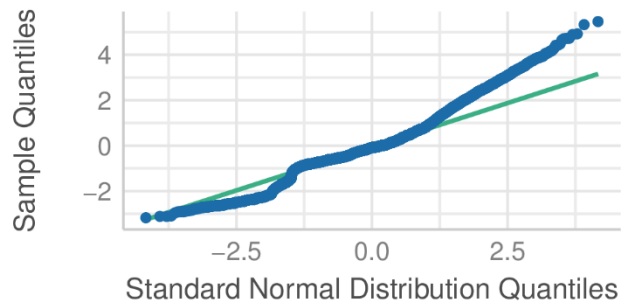
Reference line should be flat and horizontal



High collinearity (VIF) may inflate parameter uncertainty!



Dots should fall along the line



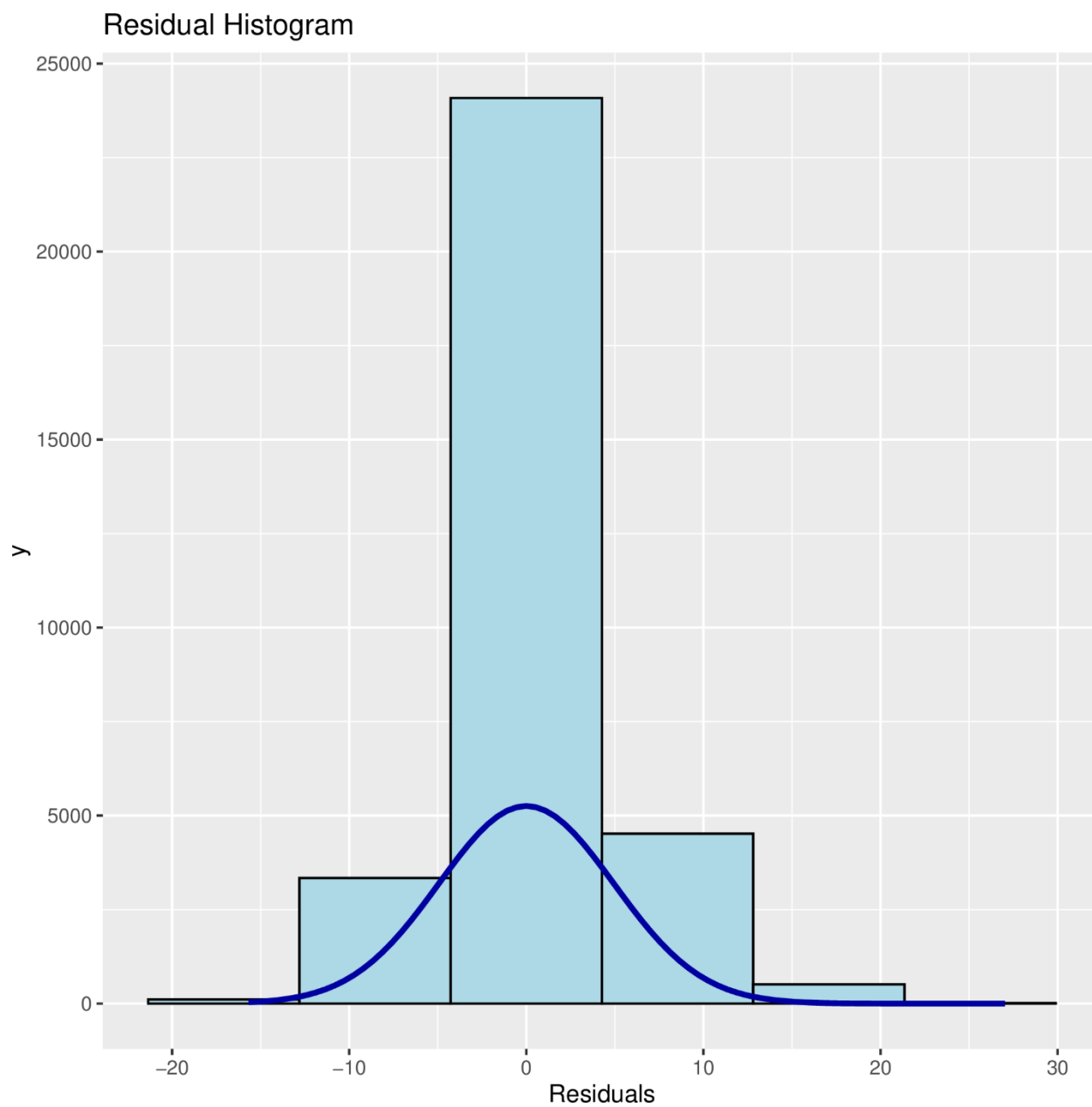
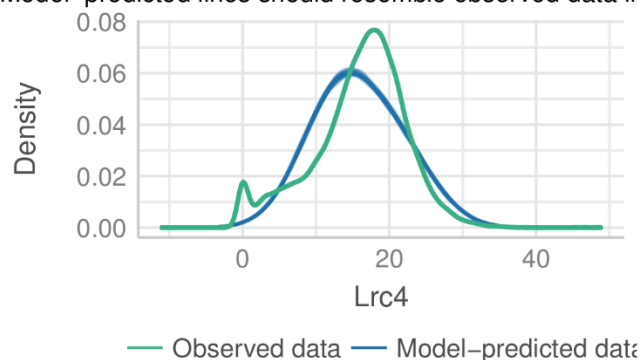
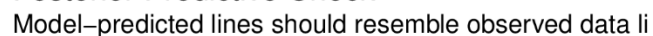
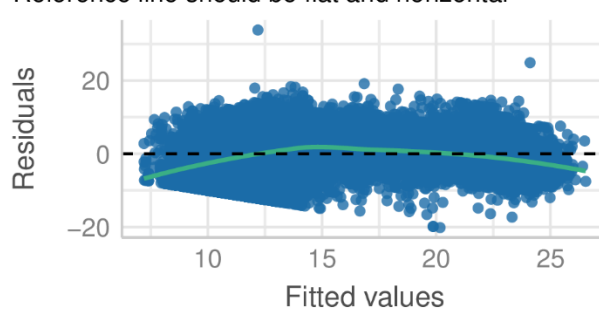


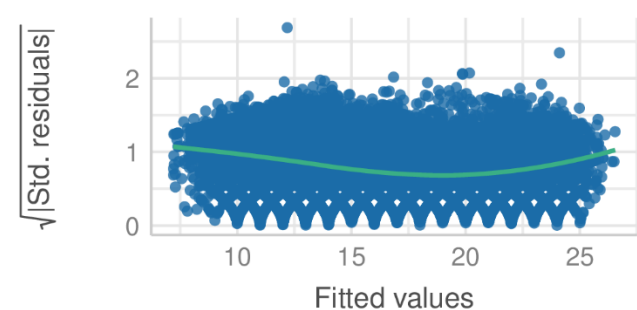
Figure S30. Histogram of residuals for baseline linear model of Luc3 response variable, from GLM data set



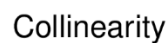
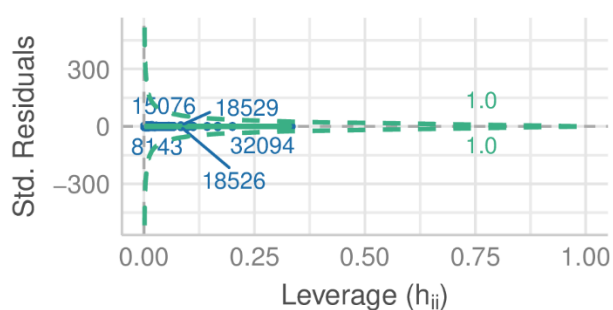
Reference line should be flat and horizontal



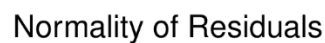
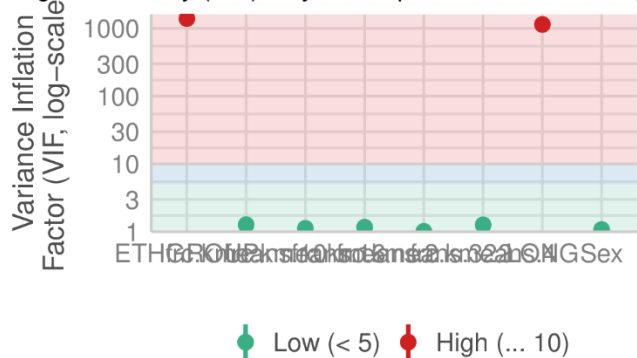
Reference line should be flat and horizontal



Points should be inside the contour lines



High collinearity (VIF) may inflate parameter uncertainty



Dots should fall along the line

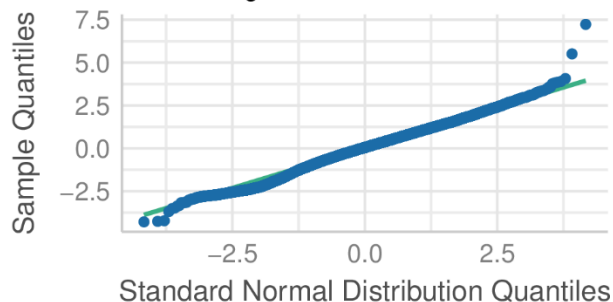


Figure S31. Linear model assumptions check for baseline linear model of Lrc4 response variable, from GLM data set

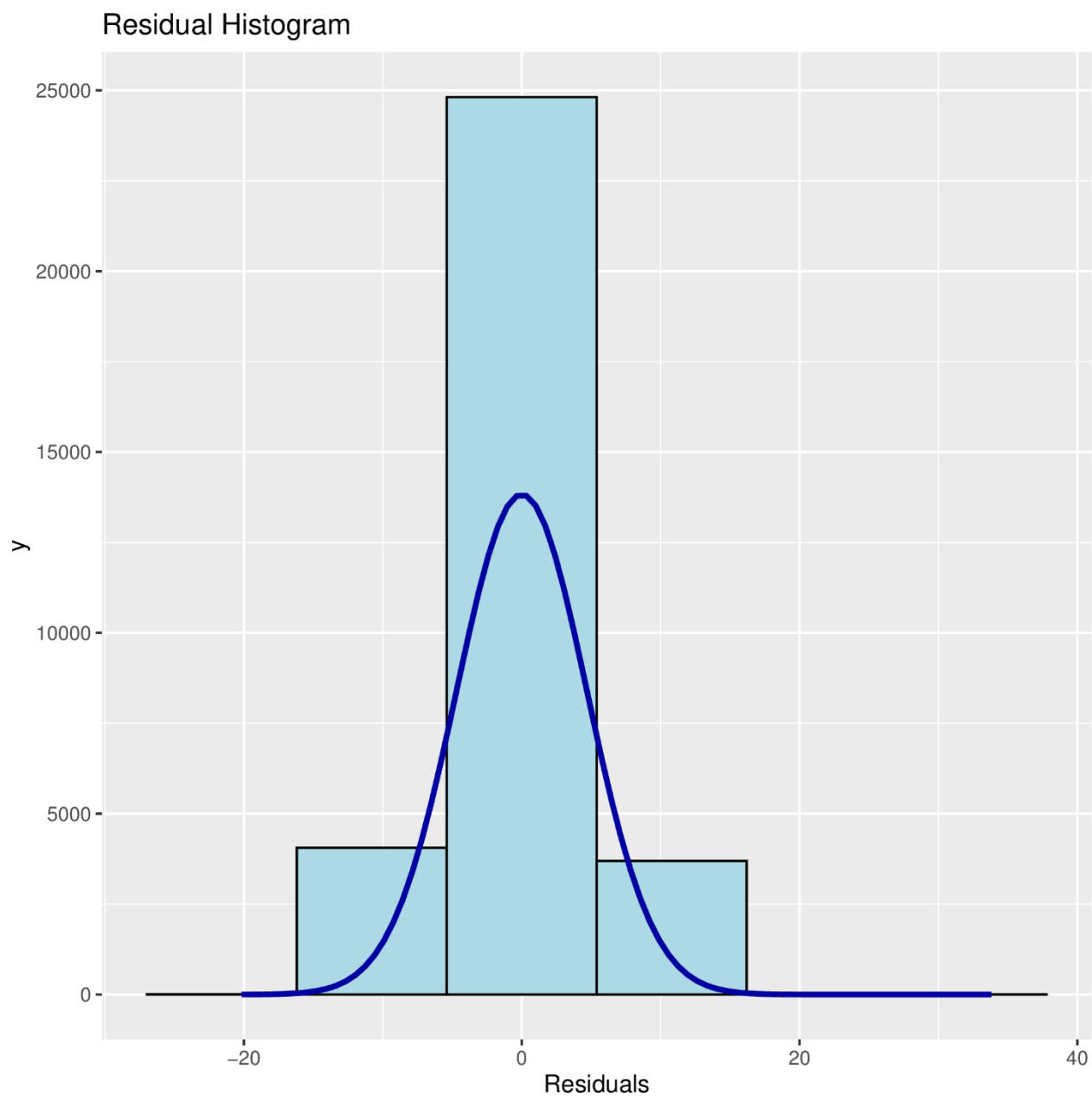


Figure S32. Histogram of residuals for baseline linear model of Lrc4 response variable, from GLM data set

Model-predicted lines should resemble observed data li



Reference line should be flat and horizontal



Fitted values

Leverage (h_{ii})

● Low (< 5)

Standard Normal Distribution Quantiles

variable, from GLM data set

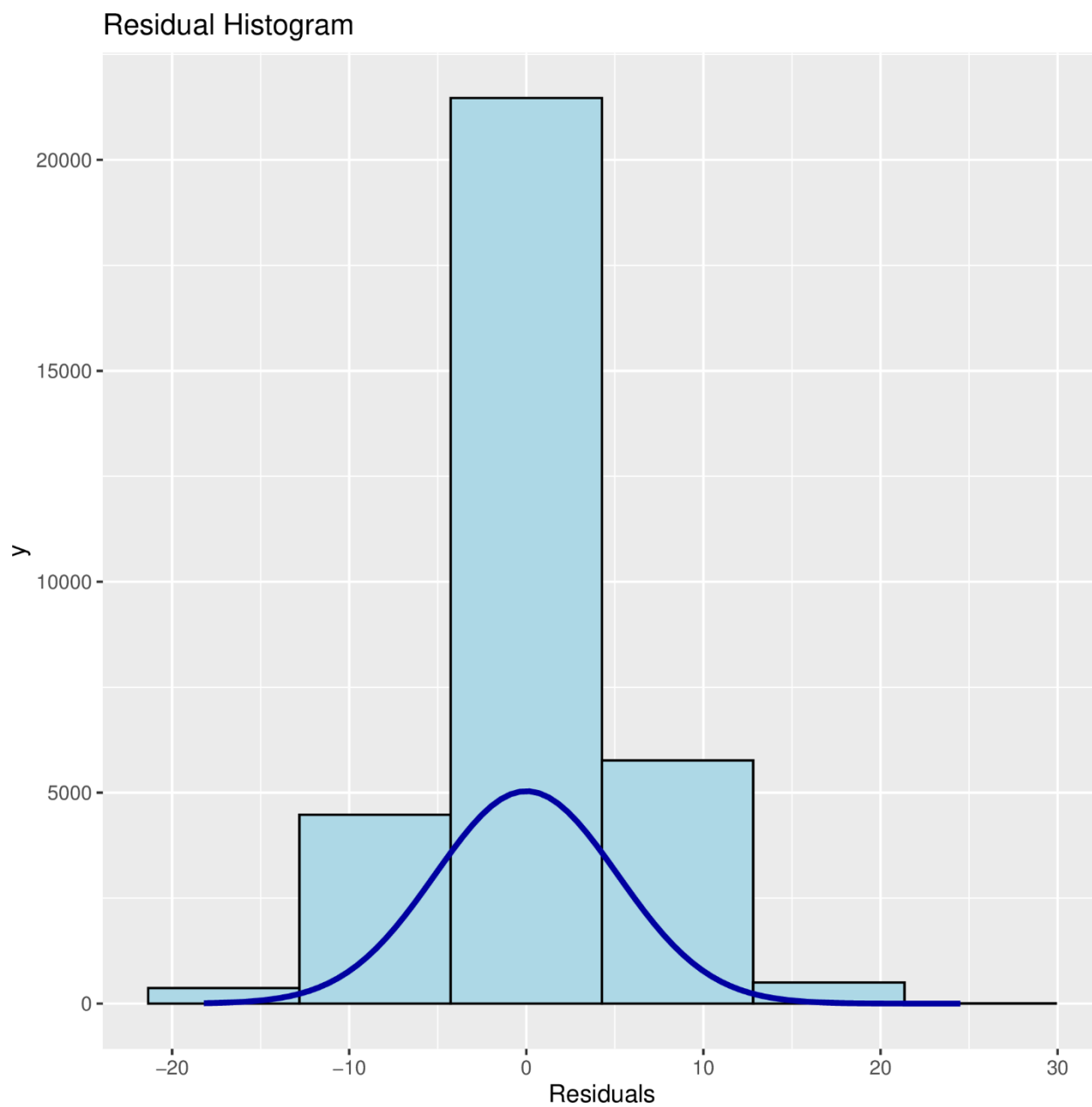
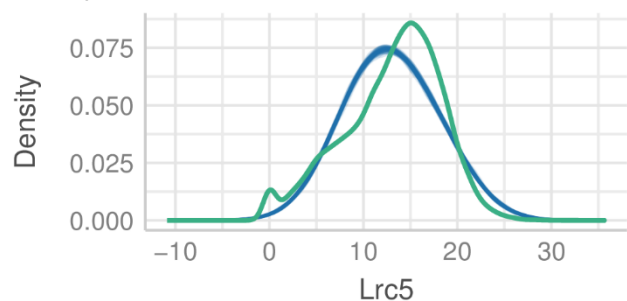


Figure S34. Histogram of residuals for baseline linear model of Luc4 response variable, from GLM data set

Posterior Predictive Check

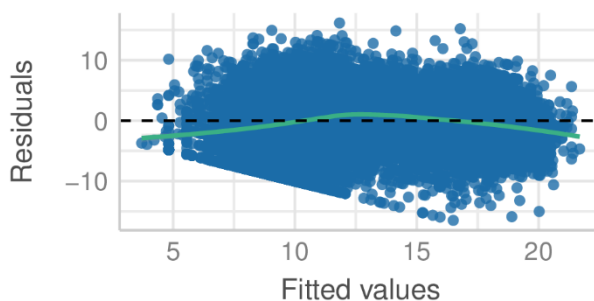
Model-predicted lines should resemble observed data lines



— Observed data — Model-predicted data

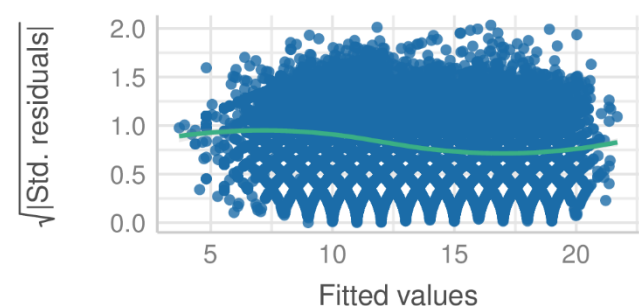
Linearity

Reference line should be flat and horizontal



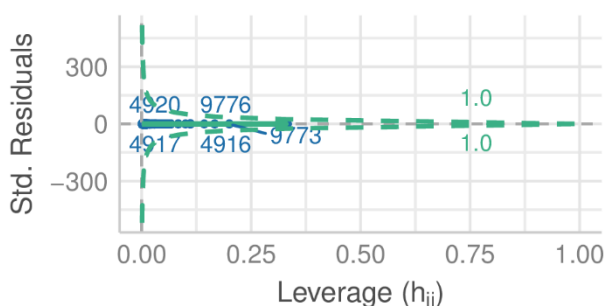
Homogeneity of Variance

Reference line should be flat and horizontal



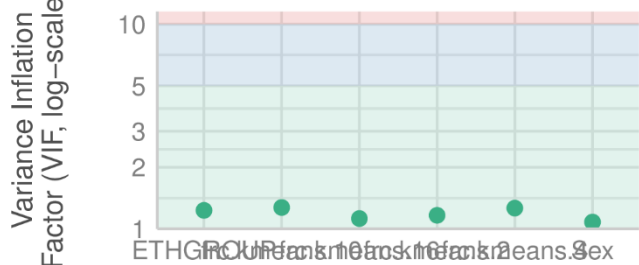
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5)

Normality of Residuals

Dots should fall along the line

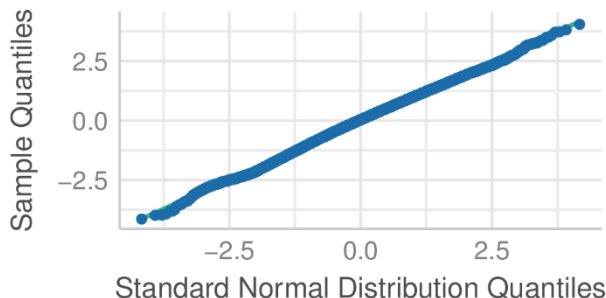


Figure S35. Linear model assumptions check for baseline linear model of Lrc5 response variable, from GLM data set

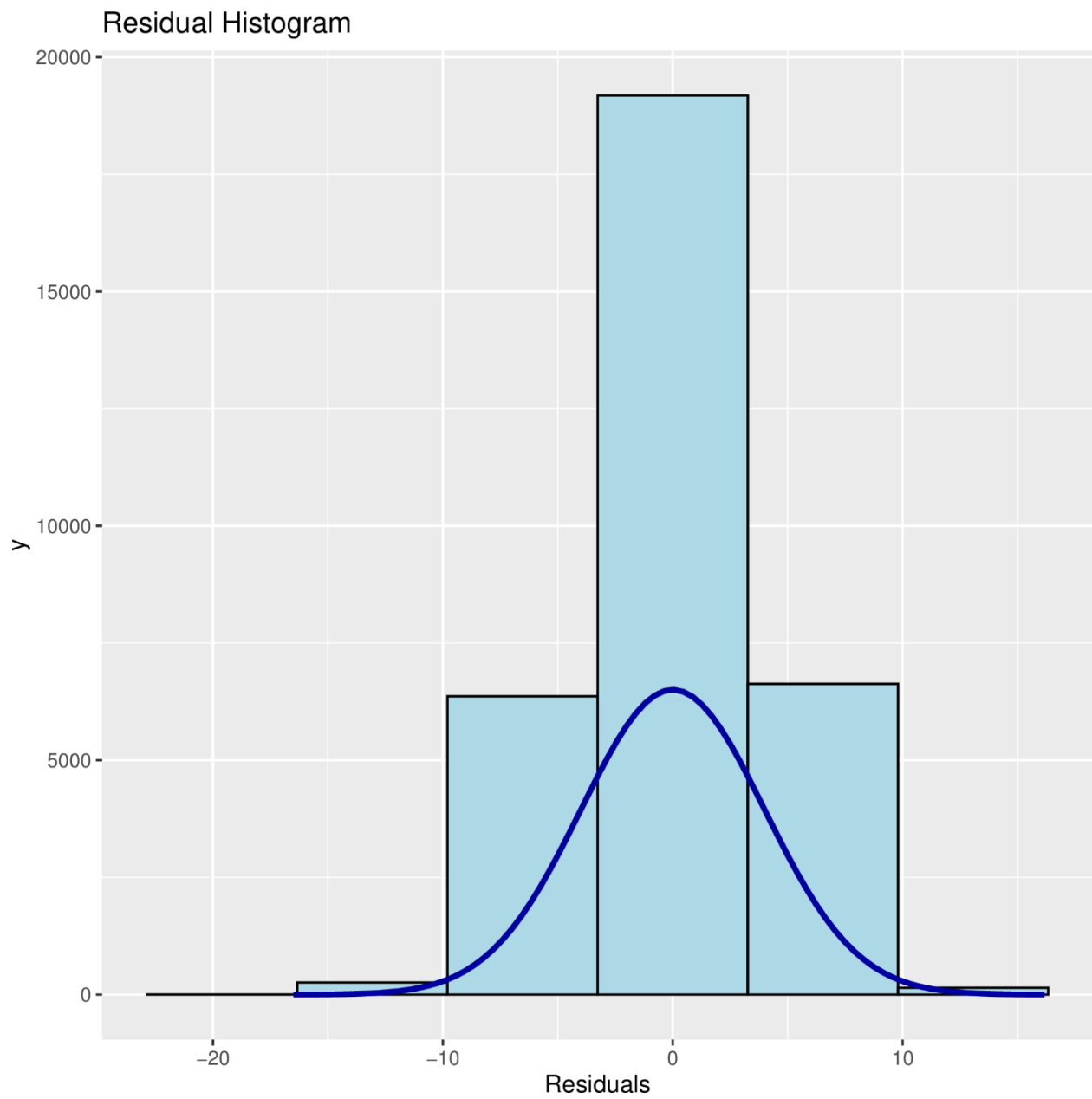


Figure S36. Histogram of residuals for baseline linear model of Lrc5 response variable, from GLM data set

Model-predicted lines should resemble observed data



Reference line should be flat and horizontal



Reference line should be flat and horizontal



Points should be inside the contour lines



High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5) ● Moderate (< 10)

Dots should fall along the line



variable, from GLM data set

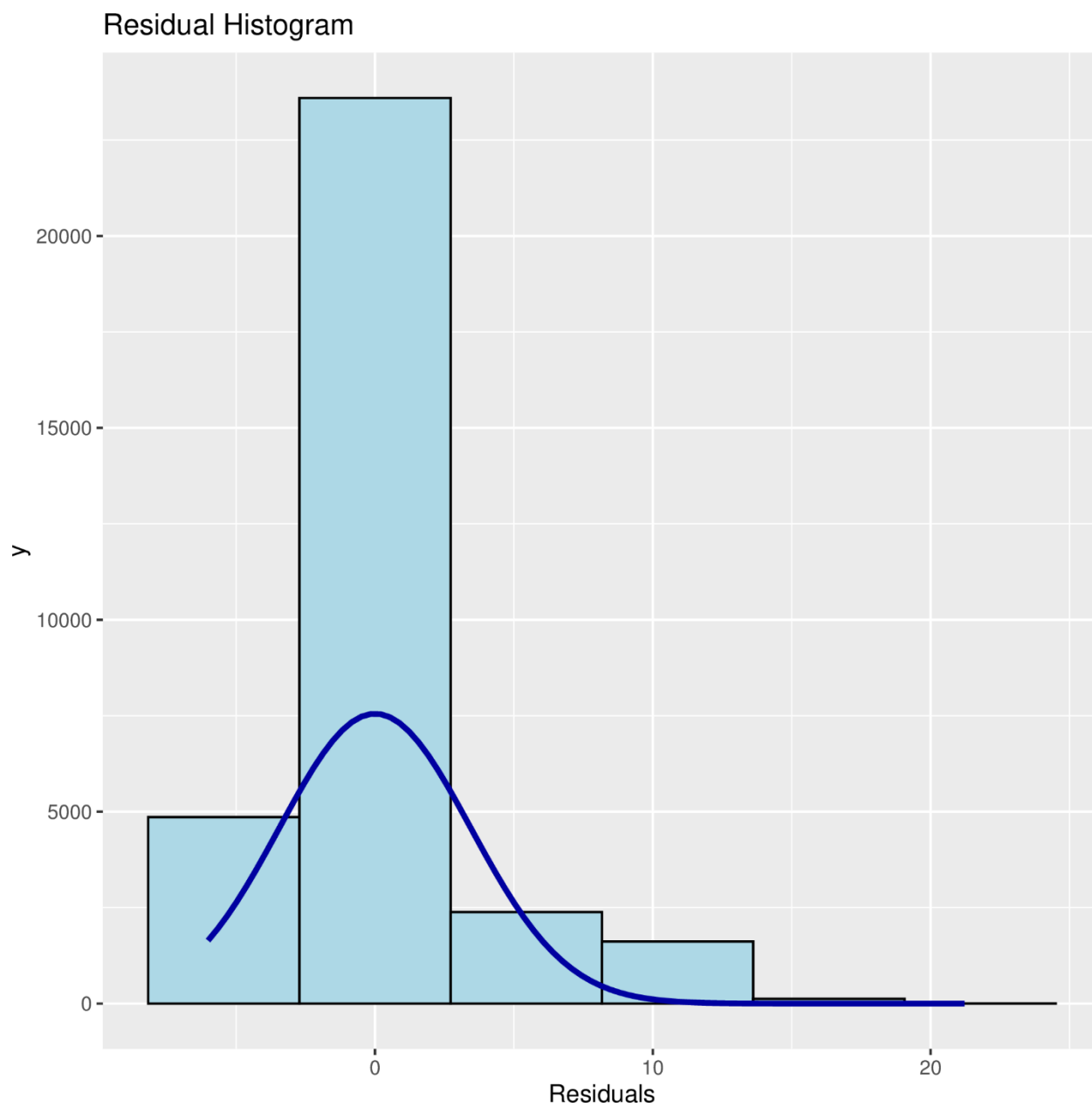
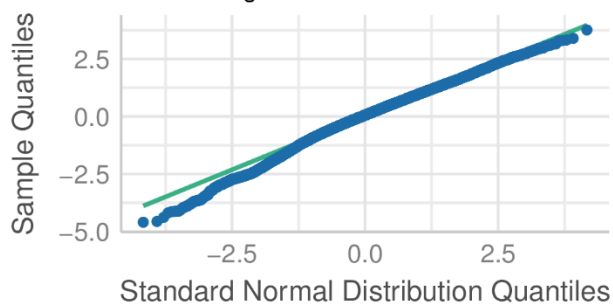
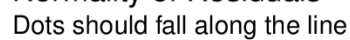
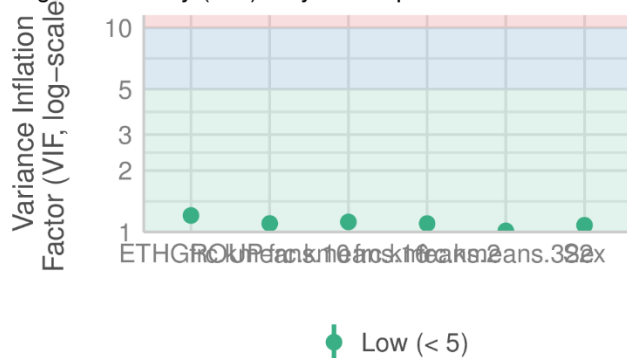
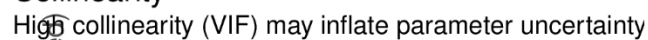
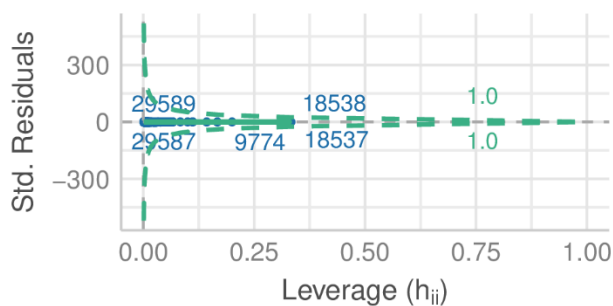
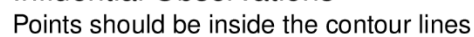
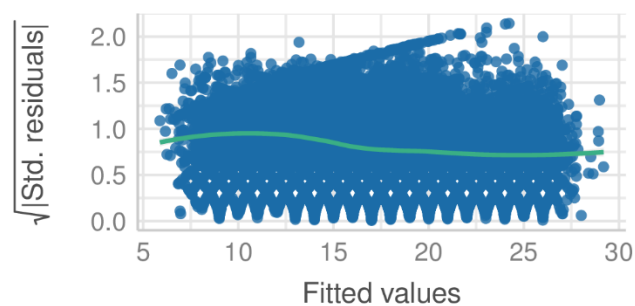
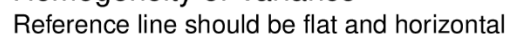
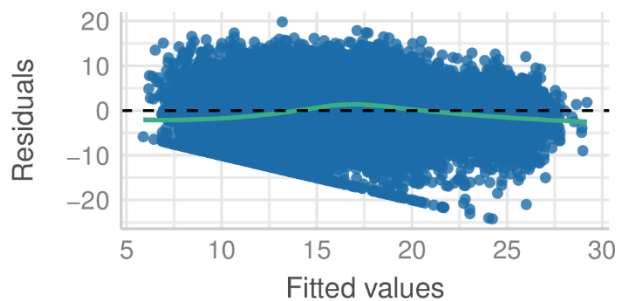
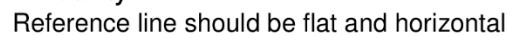
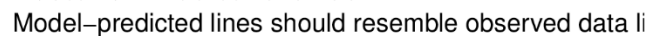


Figure S38. Histogram of residuals for baseline linear model of Luc5 response variable, from GLM data set



486

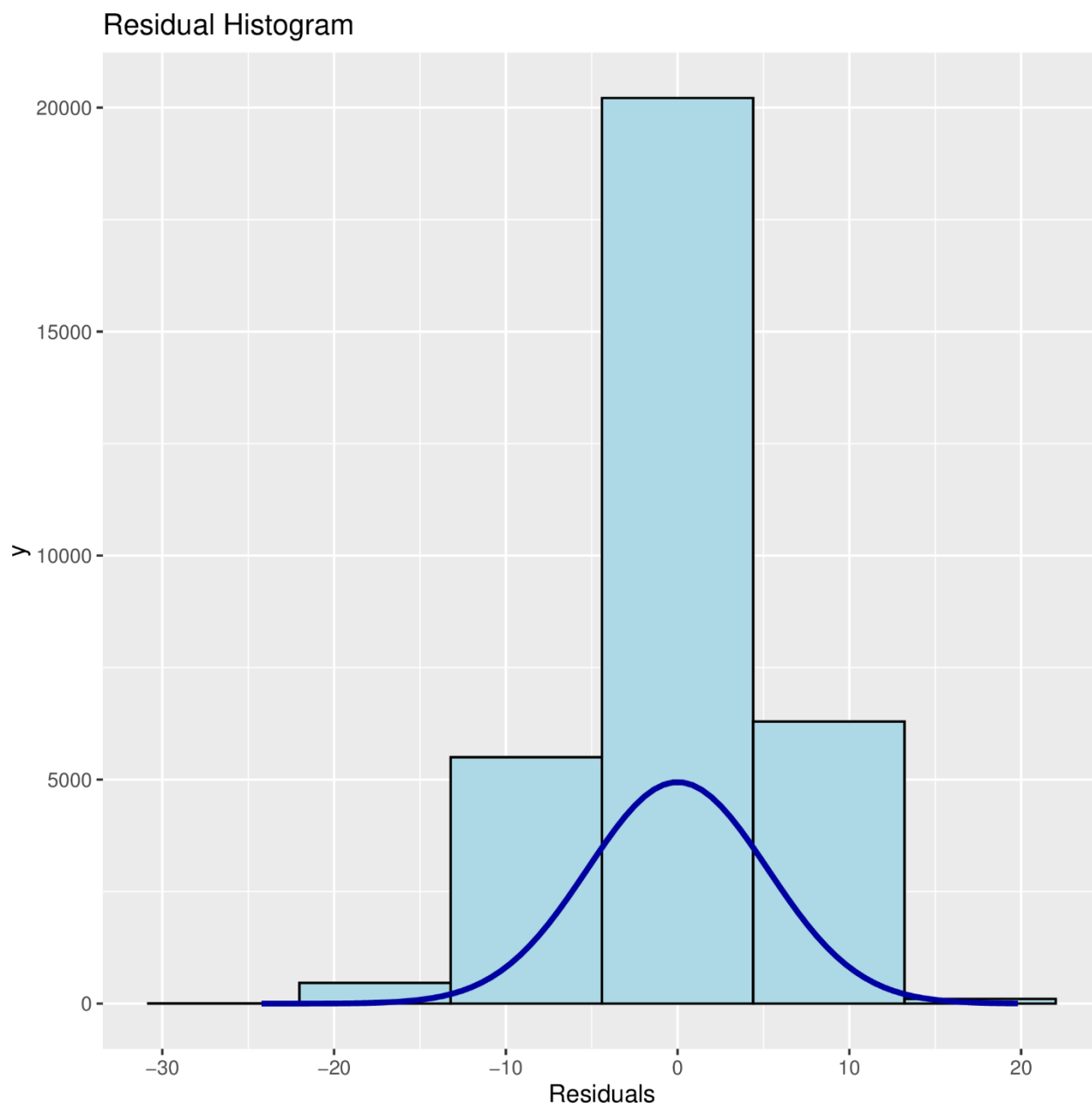


Figure S40. Histogram of residuals for baseline linear model of Rrc1 response variable, from GLM data set

Model-predicted lines should resemble observed data li



Reference line should be flat and horizontal



Fitted values

Leverage (h_{ii})

● Low (< 5)

Standard Normal Distribution Quantiles

variable, from GLM data set

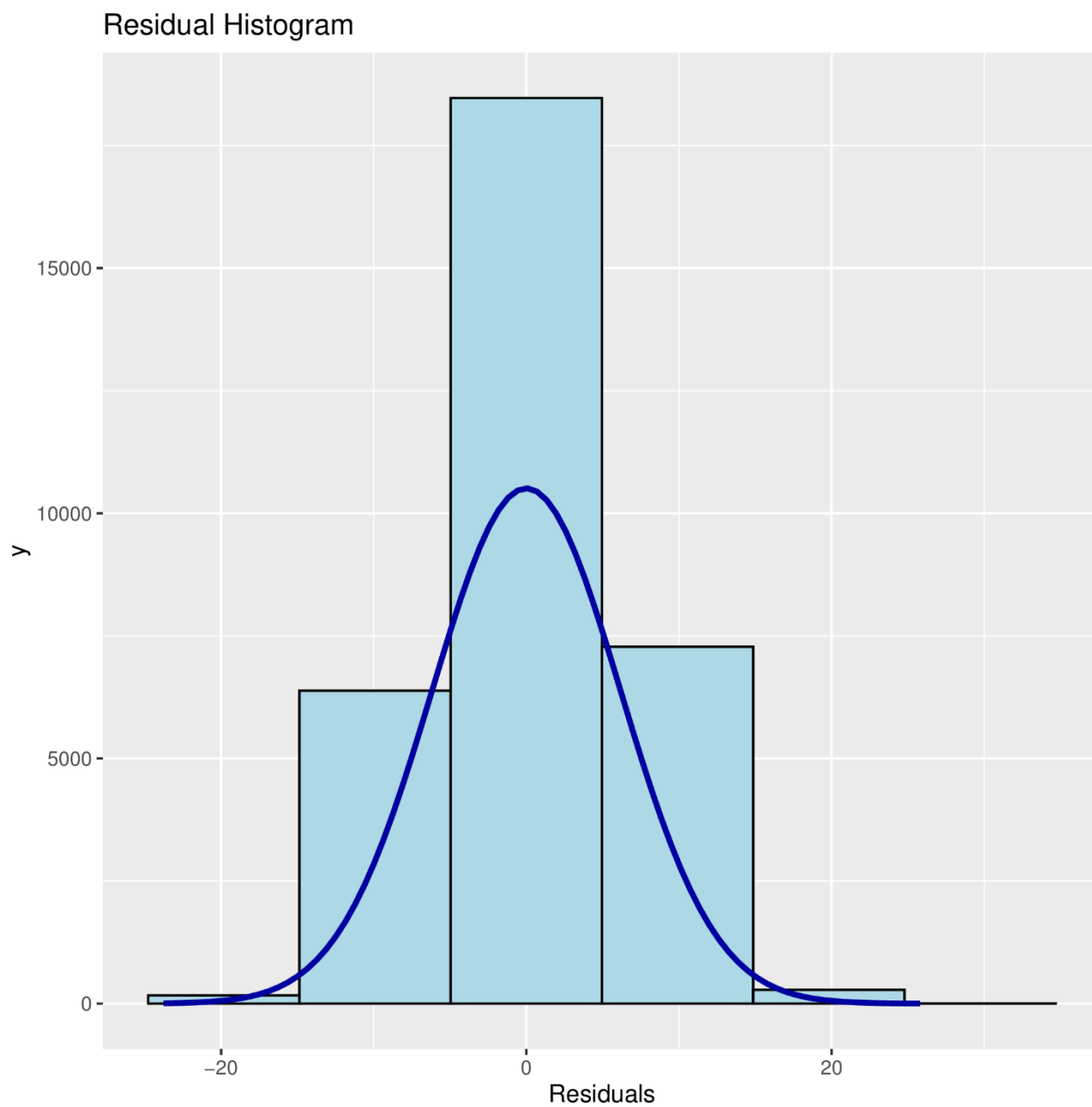
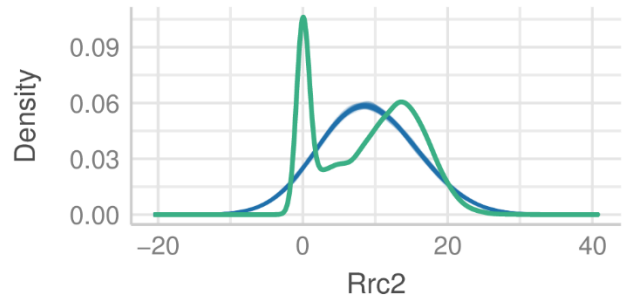


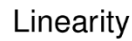
Figure S42. Histogram of residuals for baseline linear model of Ruc1 response variable, from GLM data set



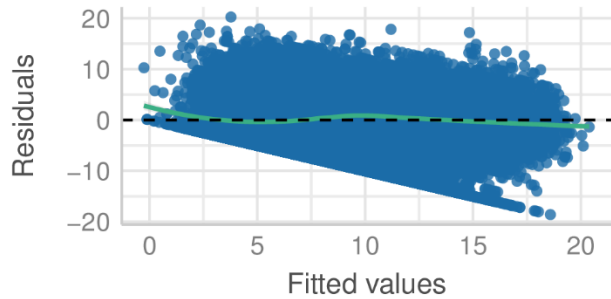
Model-predicted lines should resemble observed data li



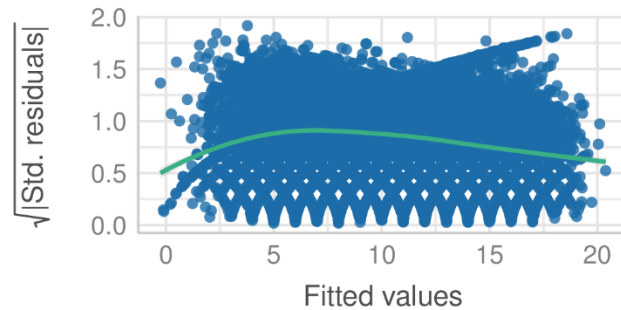
— Observed data — Model-predicted data



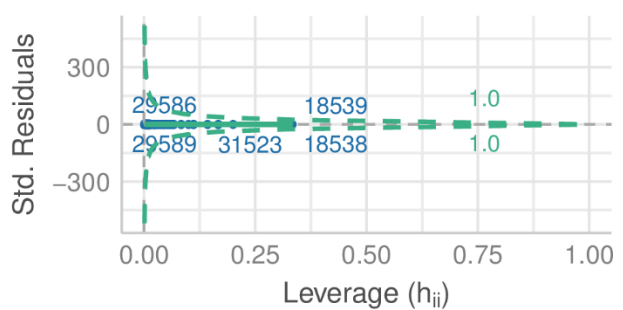
Reference line should be flat and horizontal



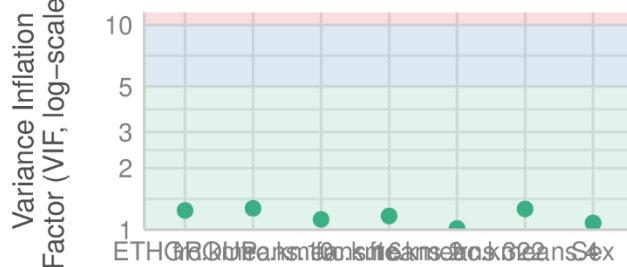
Reference line should be flat and horizontal



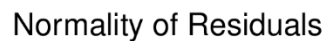
Points should be inside the contour lines



High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5)



Dots should fall along the line

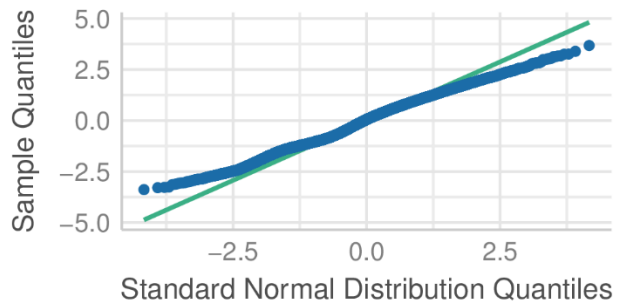


Figure S43. Linear model assumptions check for baseline linear model of Rrc2 response variable, from GLM data set

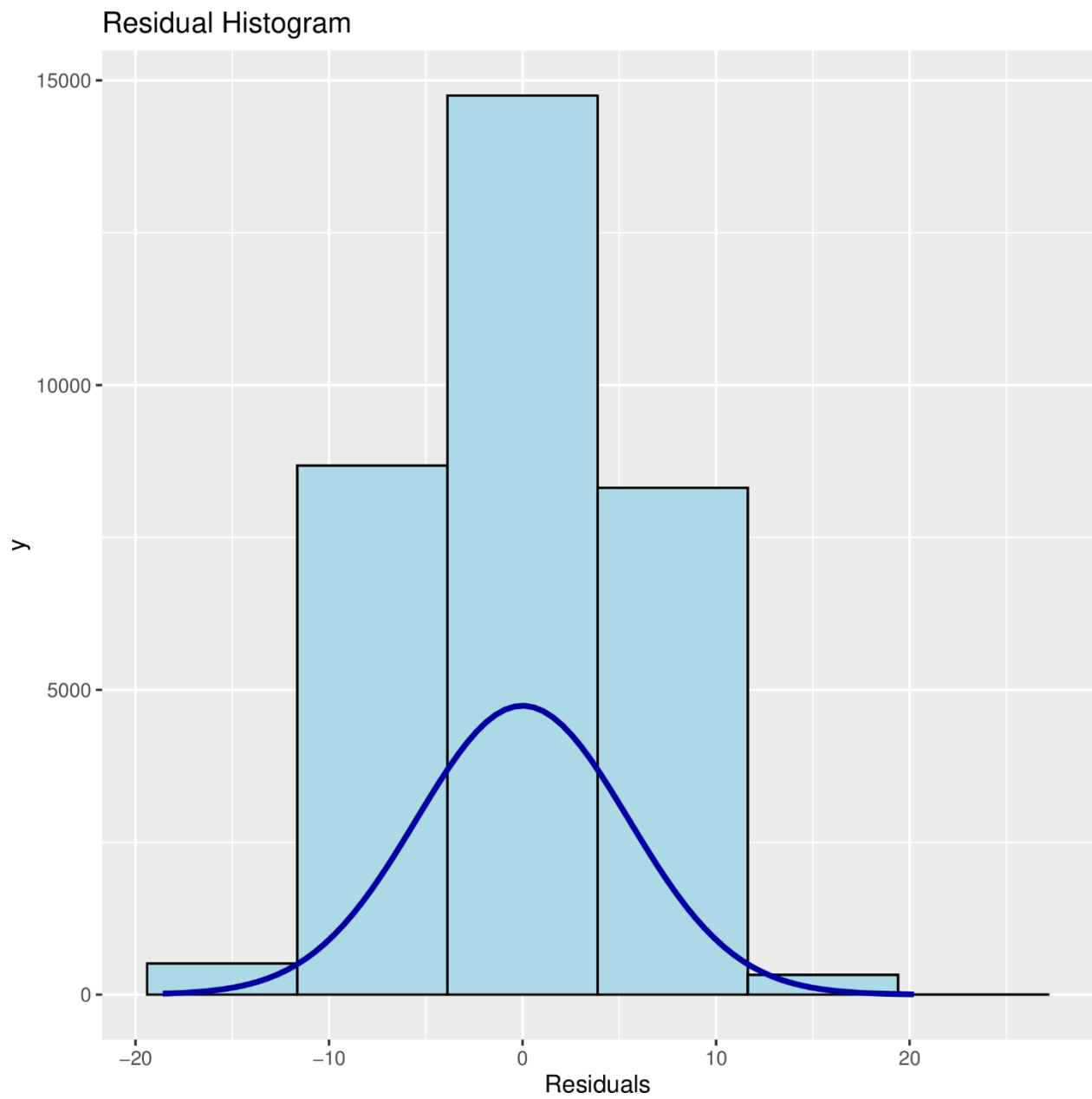
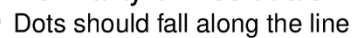
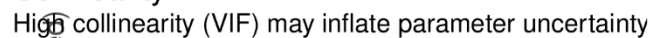
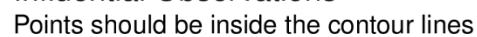
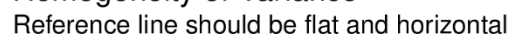
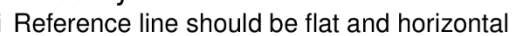
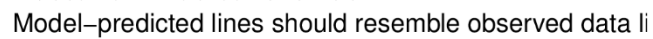


Figure S44. Histogram of residuals for baseline linear model of Rrc2 response variable, from GLM data set



variable, from GLM data set

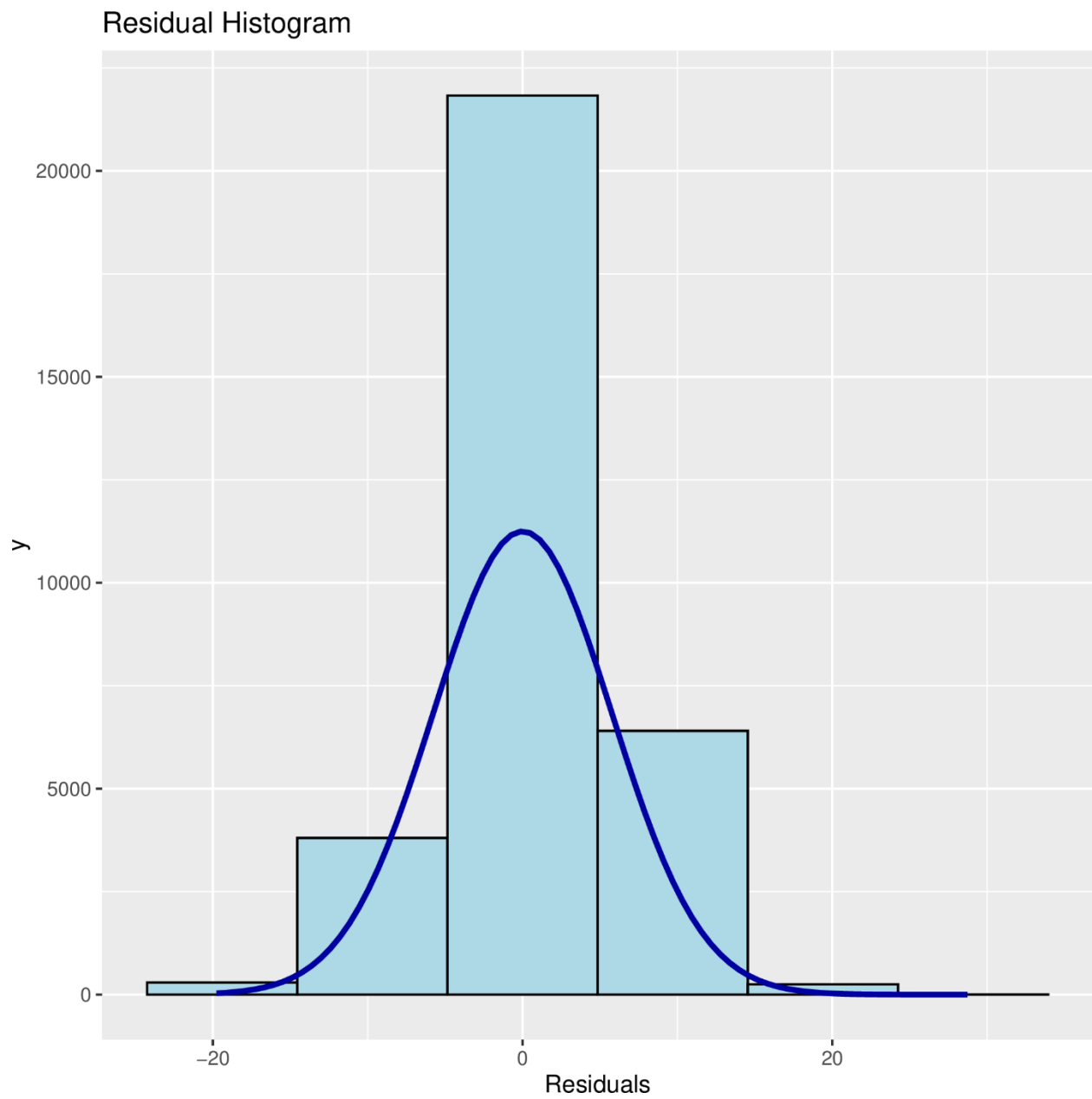
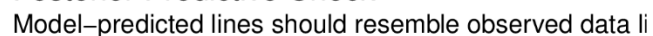


Figure S46. Histogram of residuals for baseline linear model of Ruc2 response variable, from GLM data set



Reference line should be flat and horizontal

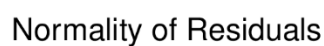
Reference line should be flat and horizontal



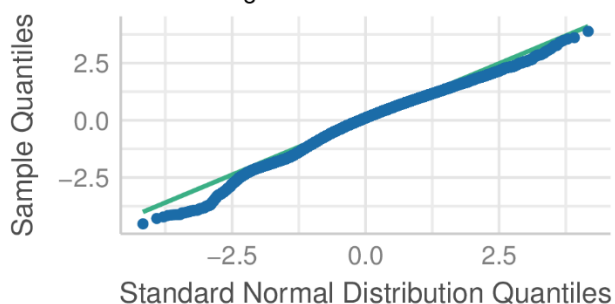
High collinearity (VIF) may inflate parameter uncertainty



Points should be inside the contour lines



Dots should fall along the line



494

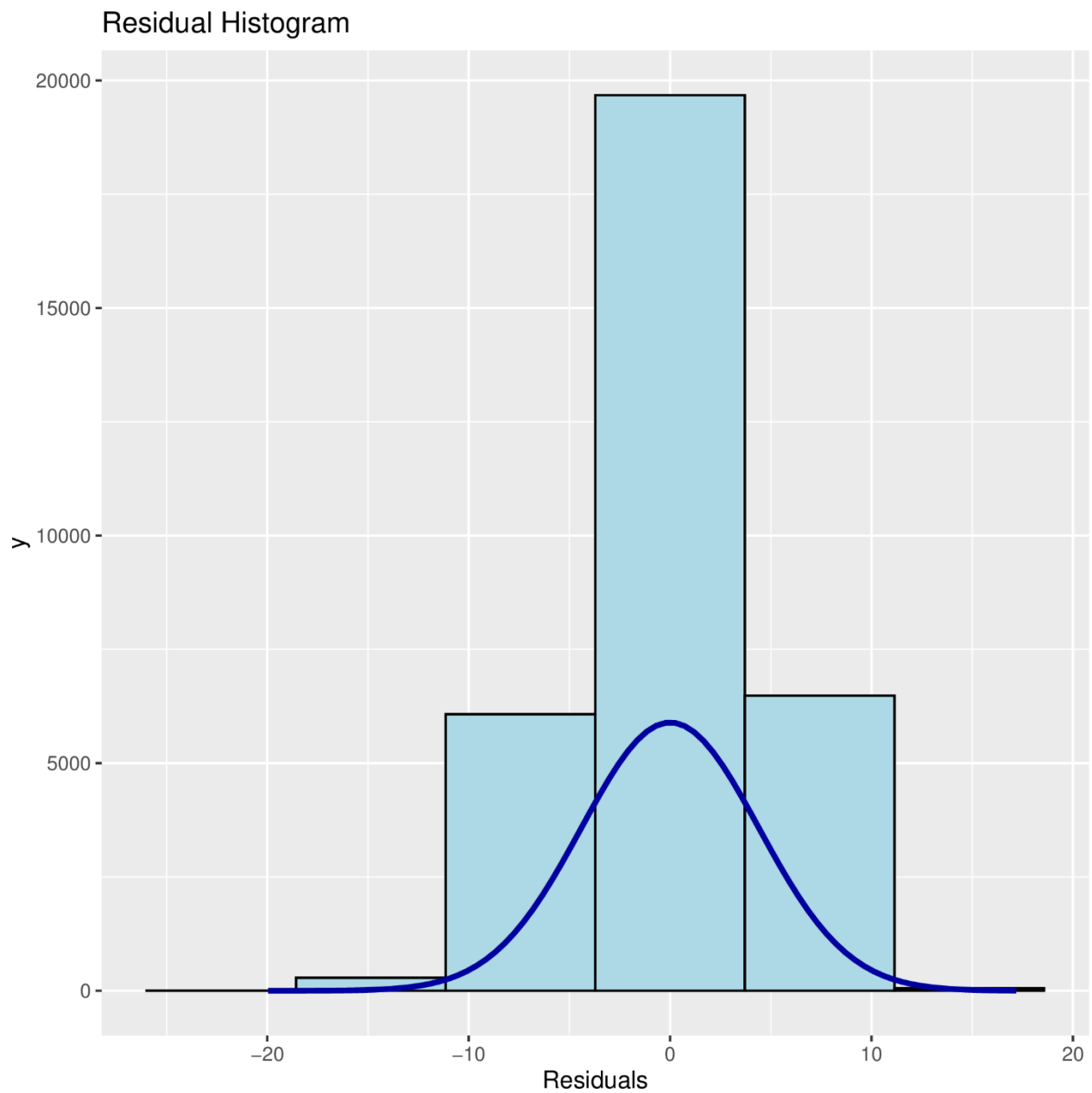


Figure S48. Histogram of residuals for baseline linear model of Rrc3 response variable, from GLM data set

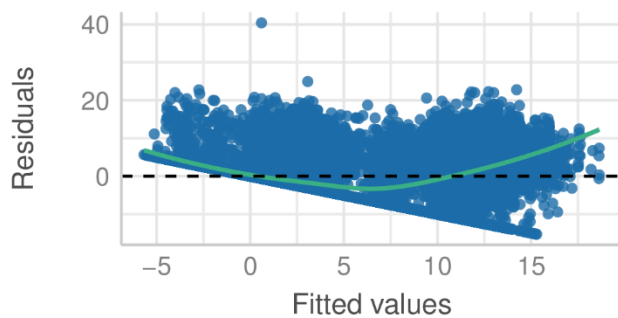
Posterior Predictive Check

Model-predicted lines should resemble observed data



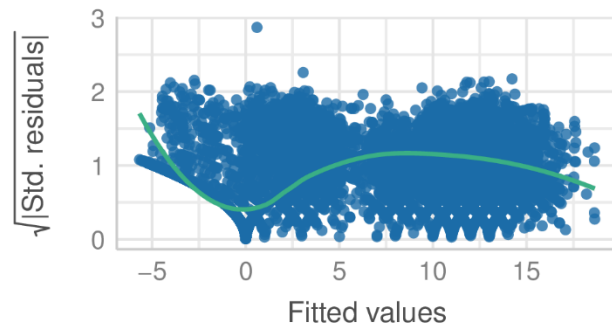
Linearity

Reference line should be flat and horizontal



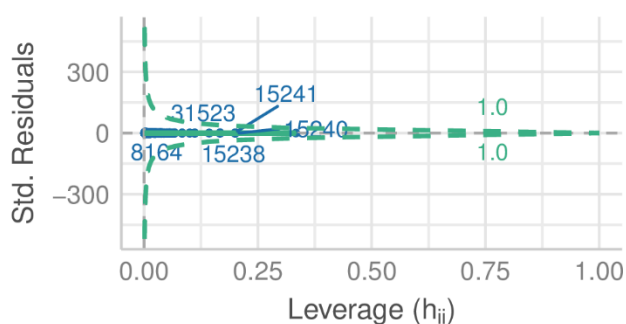
Homogeneity of Variance

Reference line should be flat and horizontal



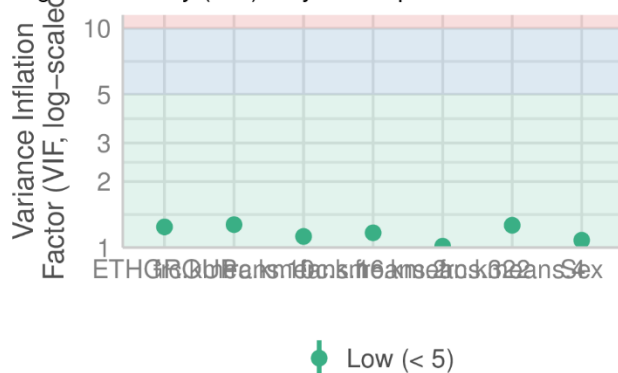
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

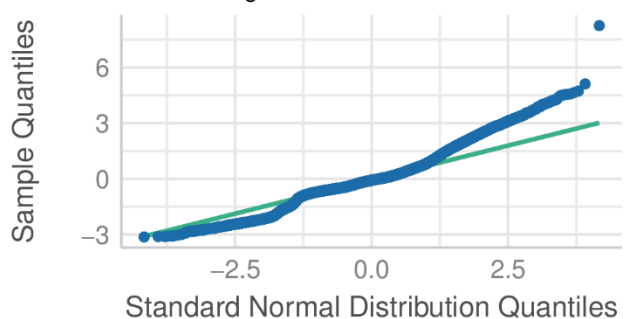


Figure S49. Linear model assumptions check for baseline linear model of Ruc3 response variable, from GLM data set

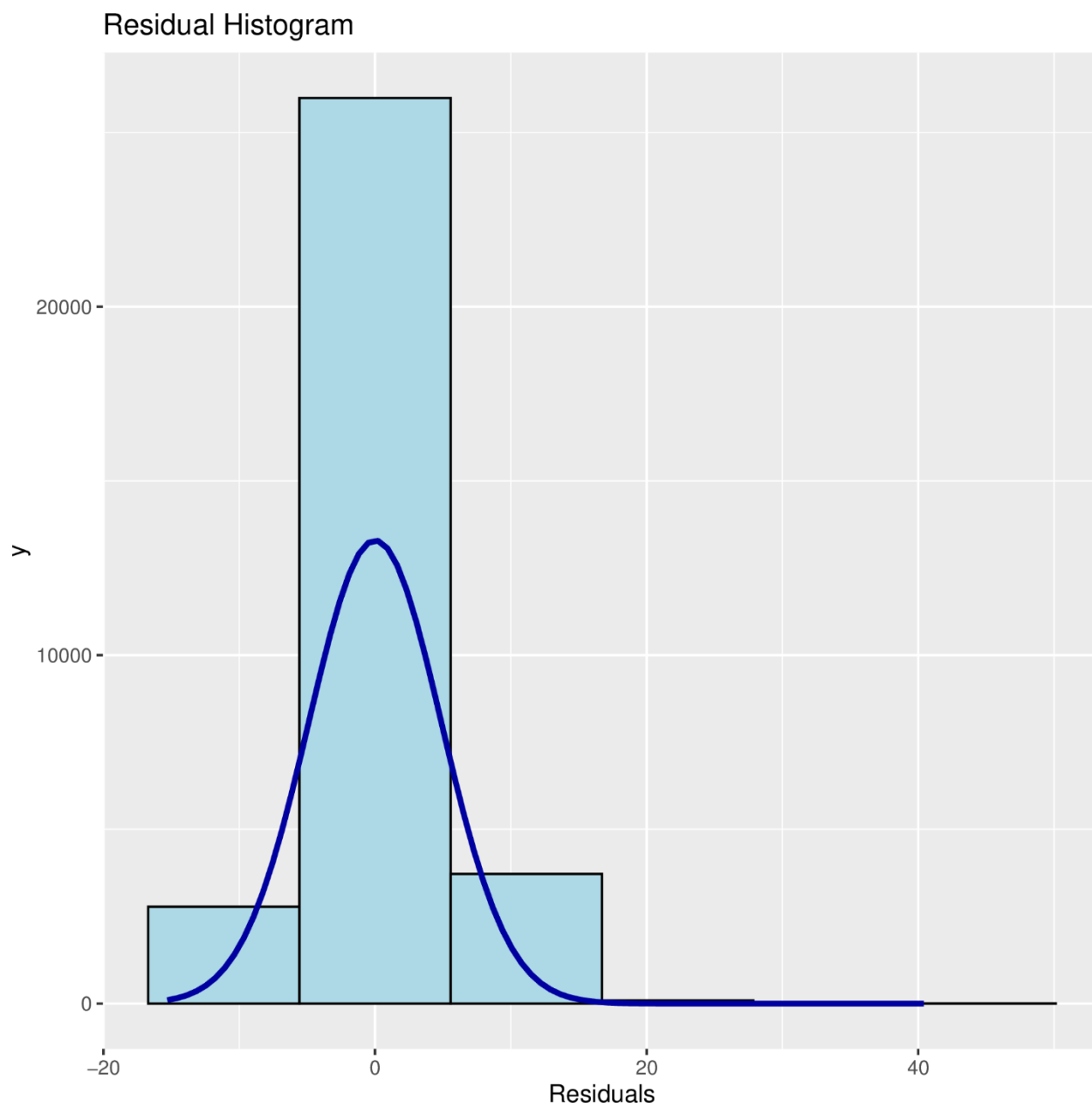
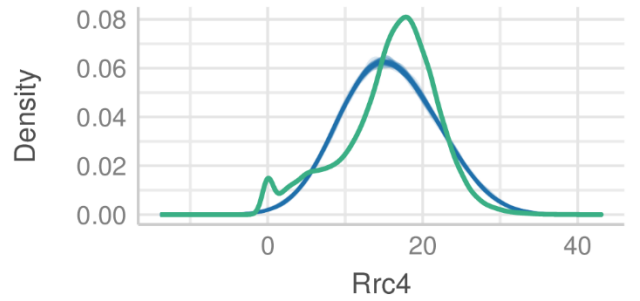


Figure S50. Histogram of residuals for baseline linear model of Ruc3 response variable, from GLM data set

Posterior Predictive Check

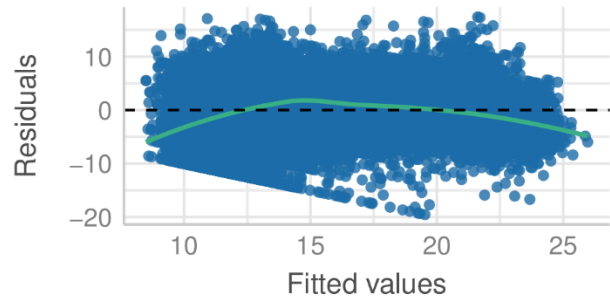
Model-predicted lines should resemble observed data li



— Observed data — Model-predicted data

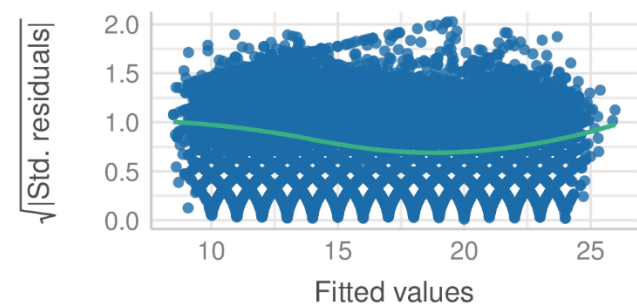
Linearity

Reference line should be flat and horizontal



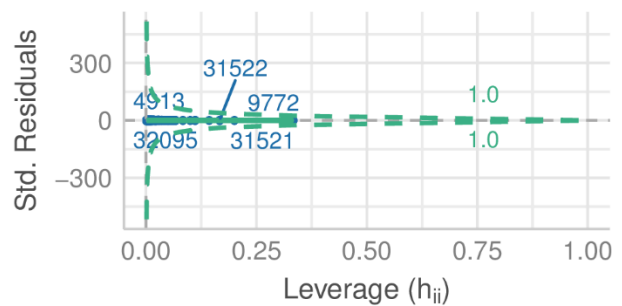
Homogeneity of Variance

Reference line should be flat and horizontal



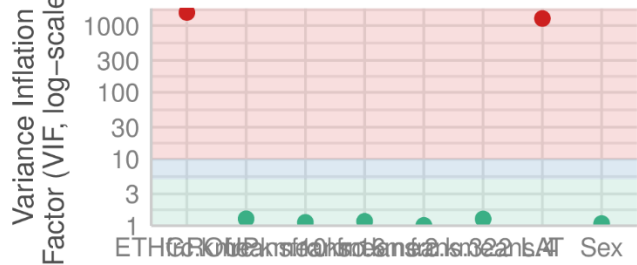
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5) ● High (... 10)

Normality of Residuals

Dots should fall along the line

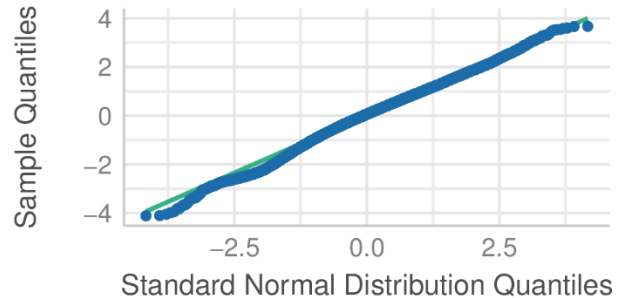


Figure S51. Linear model assumptions check for baseline linear model of Rrc4 response variable, from GLM data set

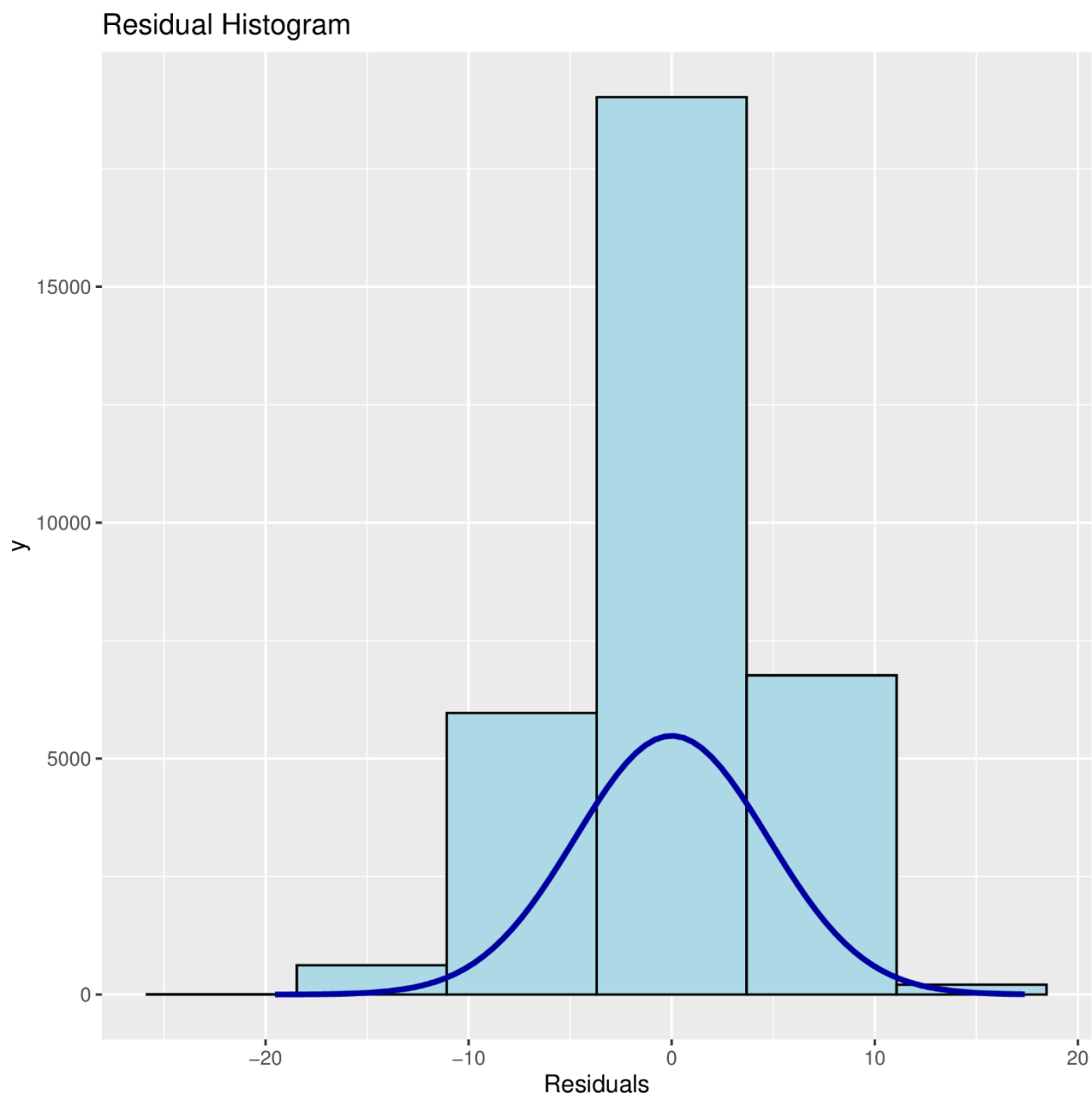


Figure S52. Histogram of residuals for baseline linear model of Rrc4 response variable, from GLM data set

Model-predicted lines should resemble observed data li



Reference line should be flat and horizontal

Calligraphy



Normality of Residuals

variable, from GLM data set

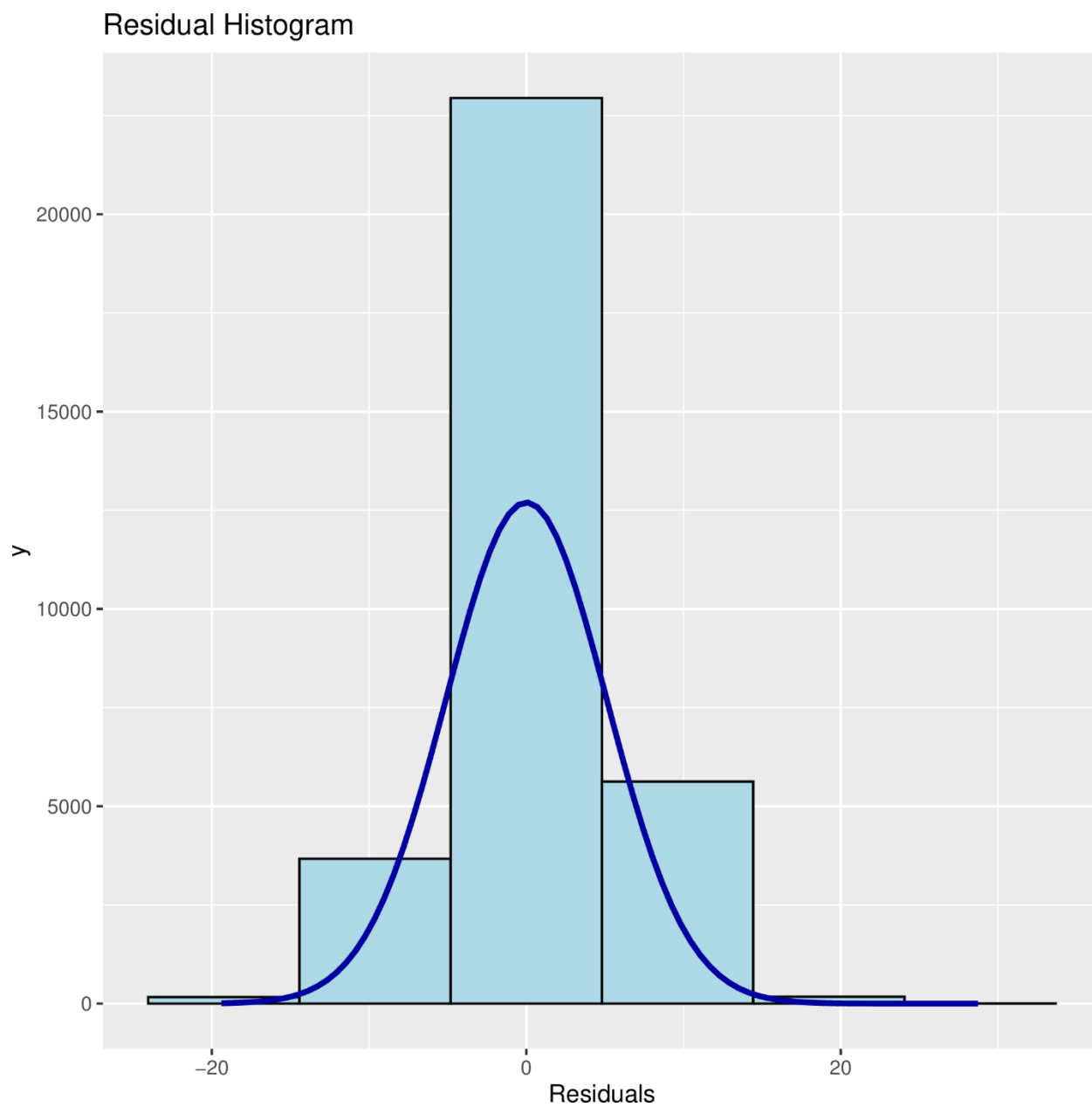
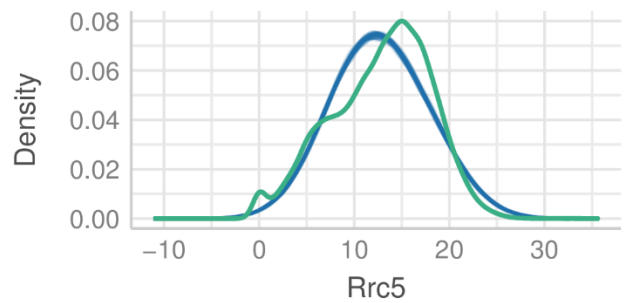


Figure S54. Histogram of residuals for baseline linear model of Ruc4 response variable, from GLM data set

Posterior Predictive Check

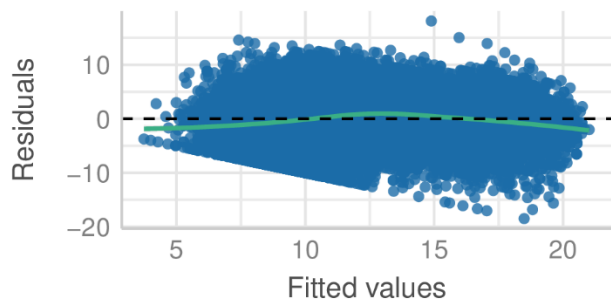
Model-predicted lines should resemble observed data



— Observed data — Model-predicted data

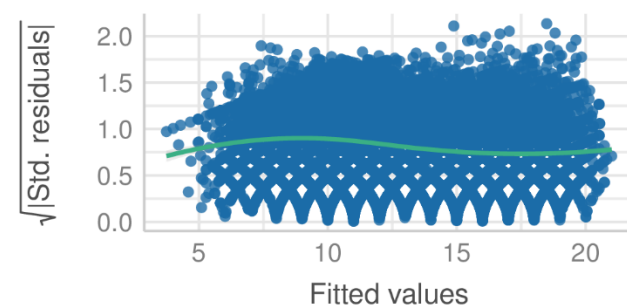
Linearity

Reference line should be flat and horizontal



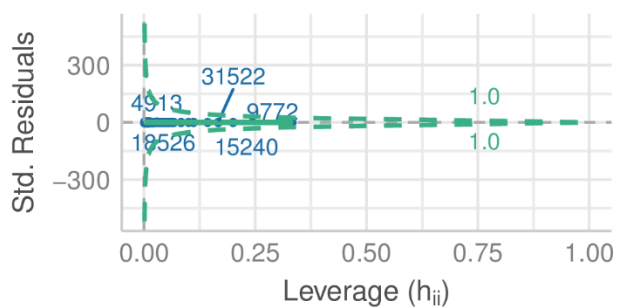
Homogeneity of Variance

Reference line should be flat and horizontal



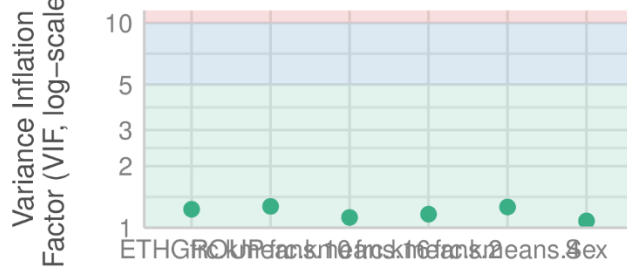
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5)

Normality of Residuals

Dots should fall along the line

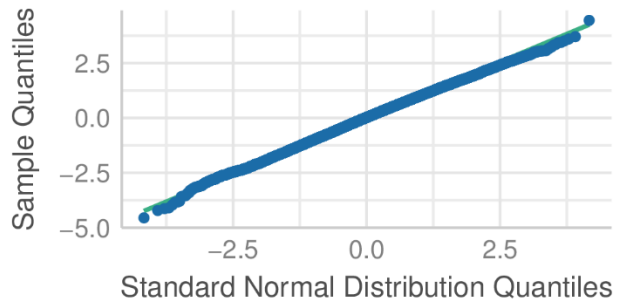


Figure S55. Linear model assumptions check for baseline linear model of Rrc5 response variable, from GLM data set

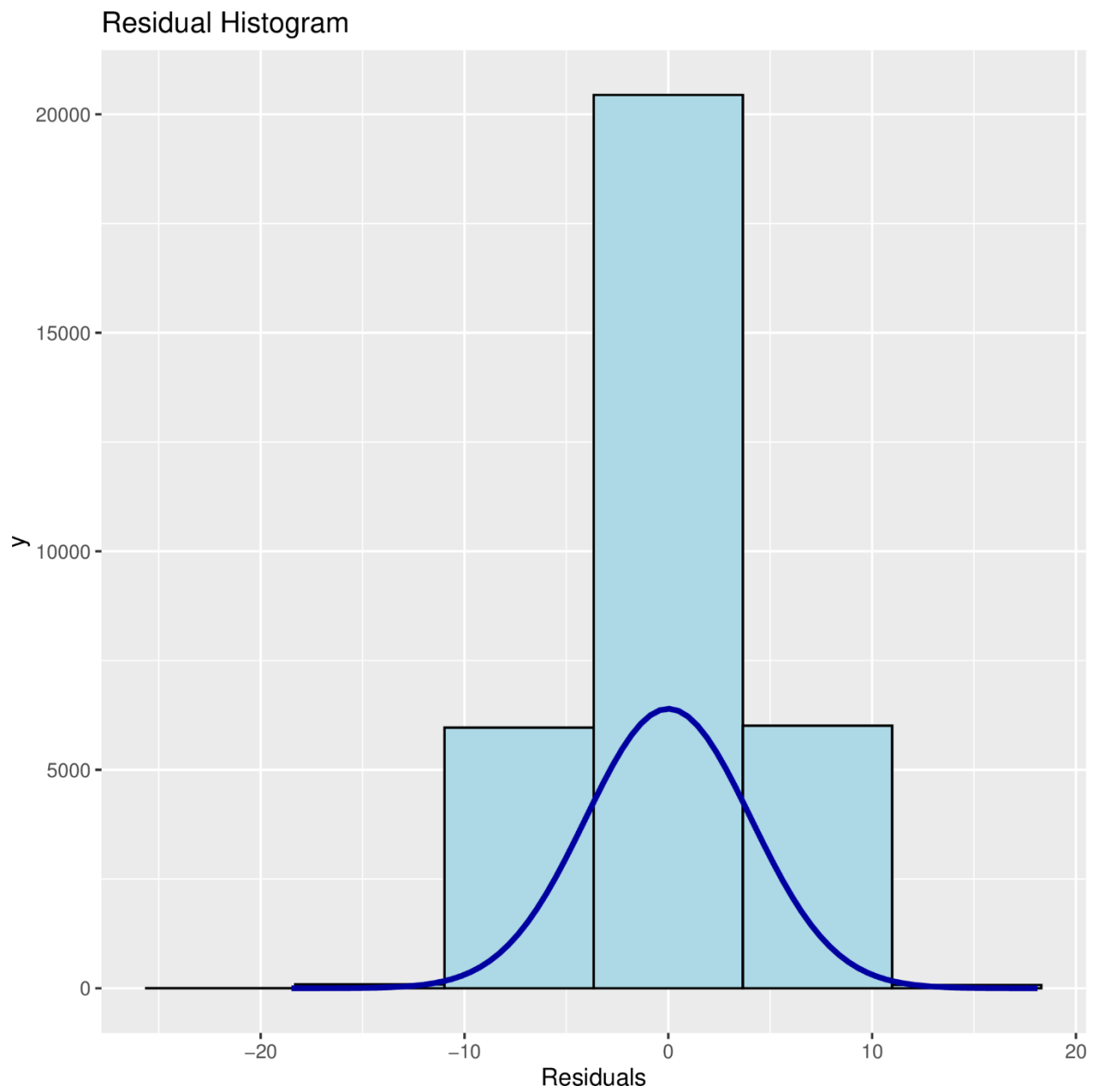
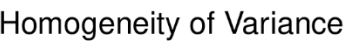


Figure S56. Histogram of residuals for baseline linear model of Rrc5 response variable, from GLM data set

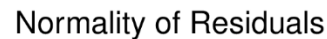
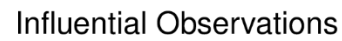
Model-predicted lines should resemble observed data li



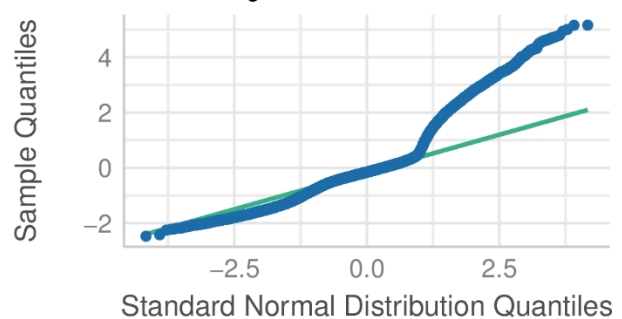
High collinearity (VIF) may inflate parameter uncertainty



Reference line should be flat and horizontal



Dots should fall along the line



504

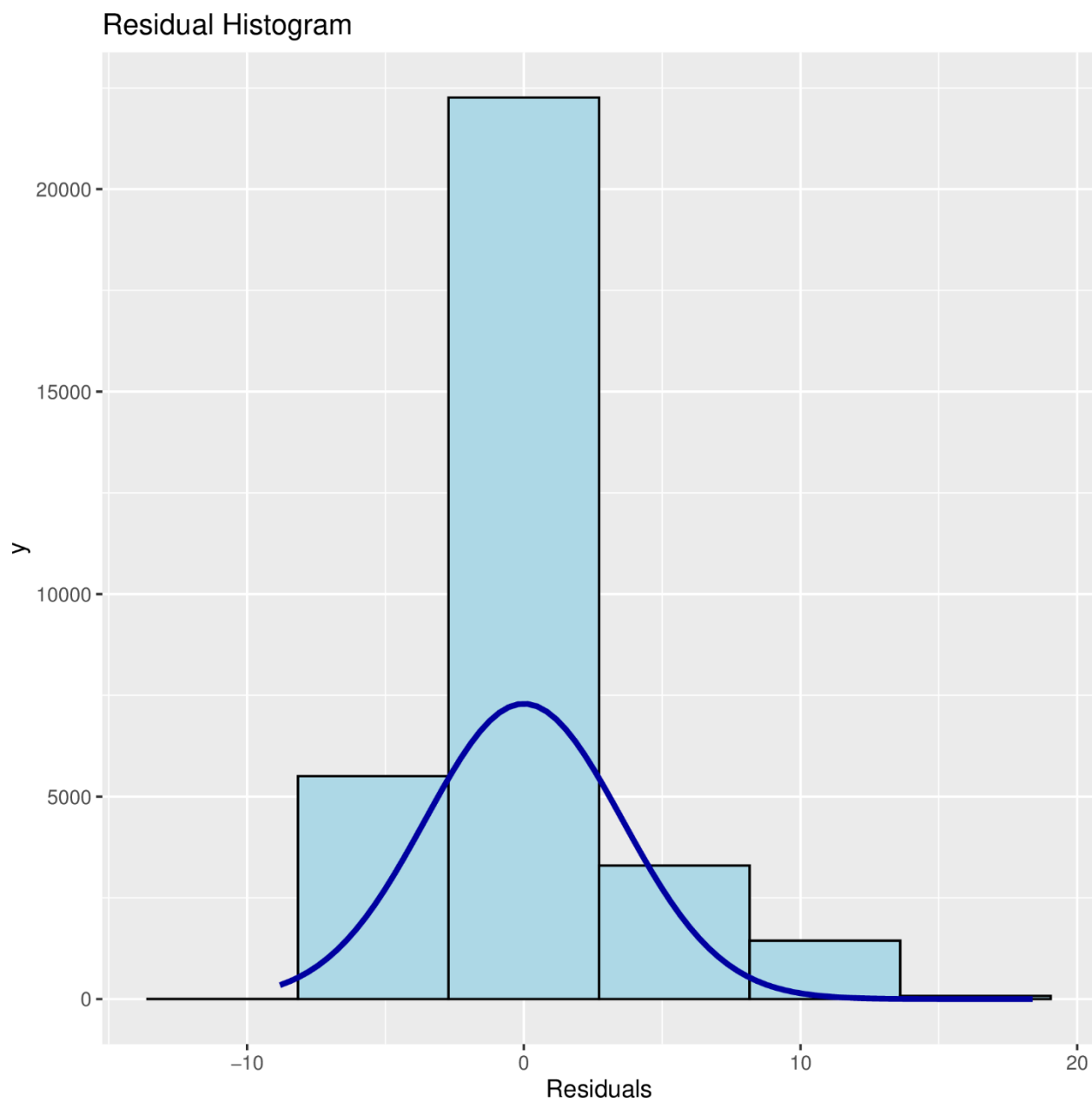


Figure S58. Histogram of residuals for baseline linear model of Ruc5 response variable, from GLM data set

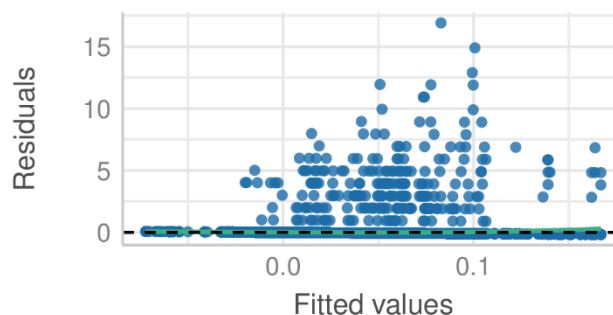
Posterior Predictive Check

Model-predicted lines should resemble observed data



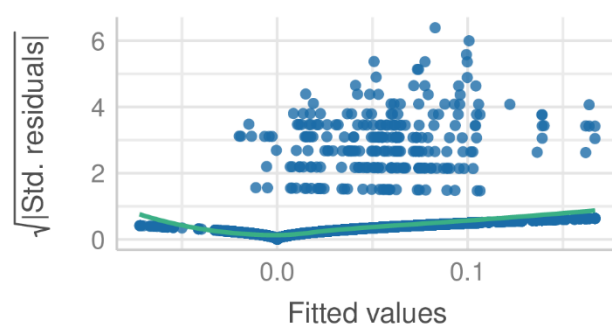
Linearity

Reference line should be flat and horizontal



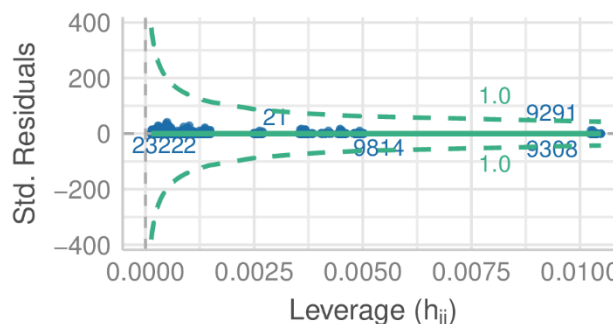
Homogeneity of Variance

Reference line should be flat and horizontal



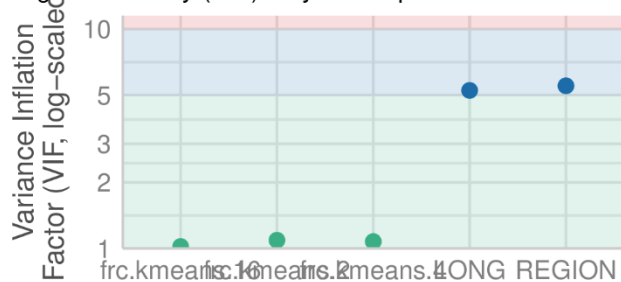
Influential Observations

Points should be inside the contour lines



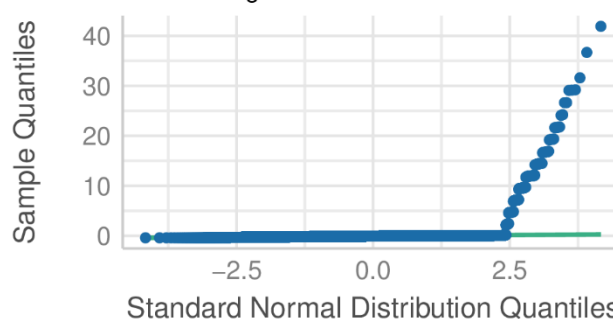
Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line



● Low (< 5) ● Moderate (< 10)

Figure S59. Linear model assumptions check for baseline linear model of Third response variable, from GLM data set

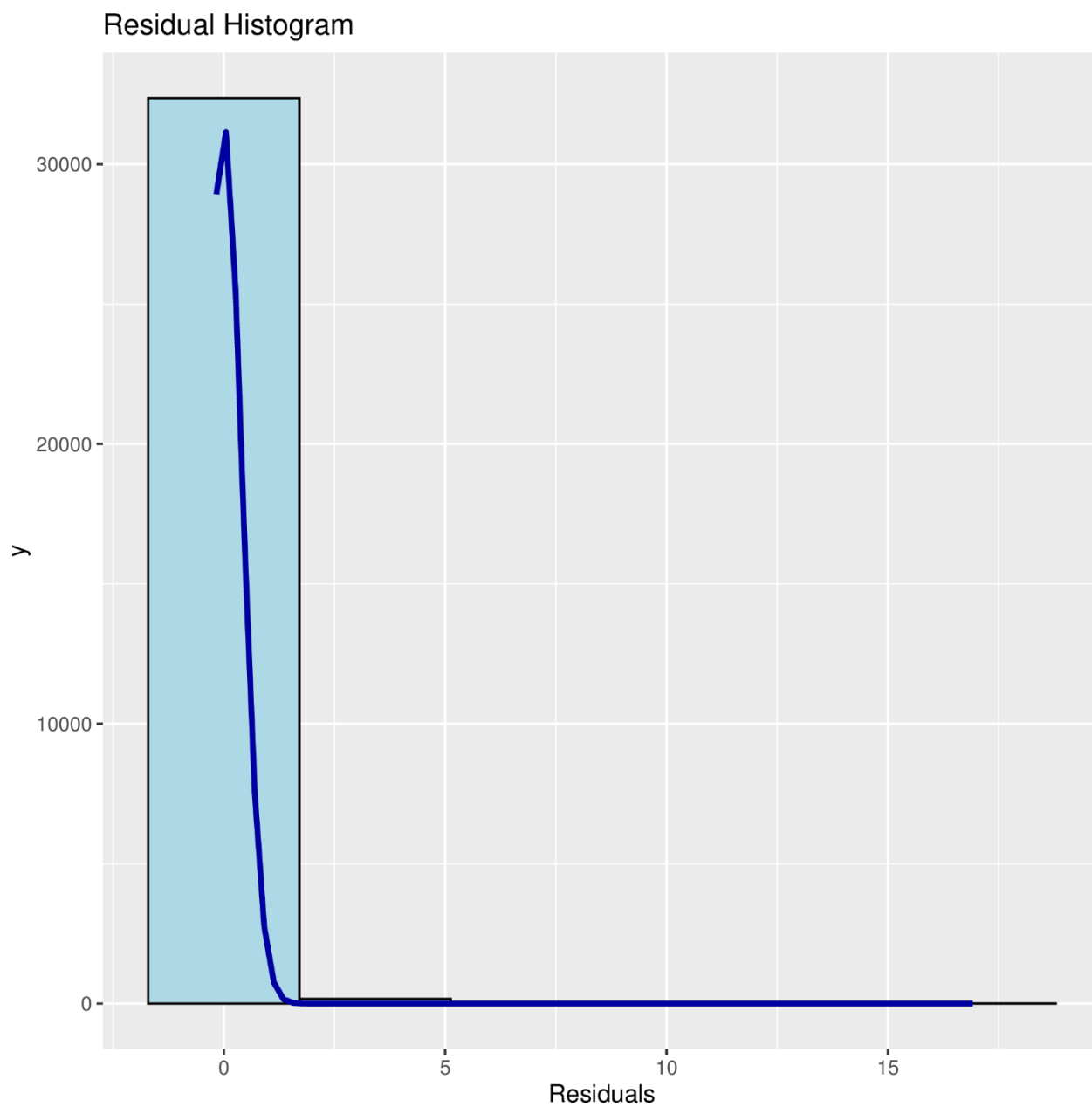
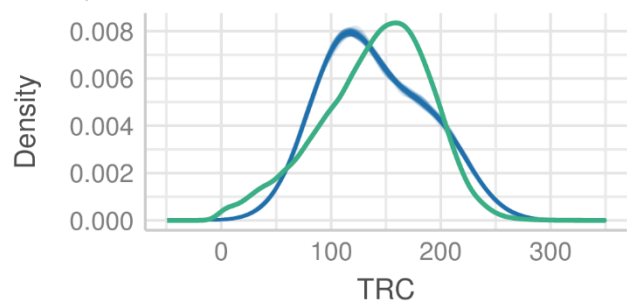
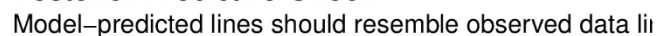


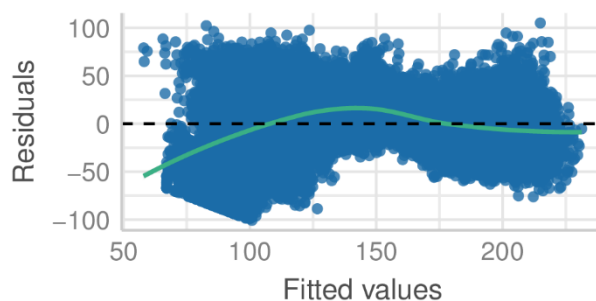
Figure S60. Histogram of residuals for baseline linear model of Third response variable, from GLM data set



— Observed data — Model-predicted data

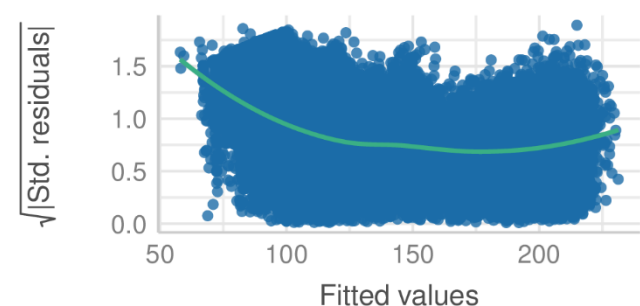
Linearity

Reference line should be flat and horizontal



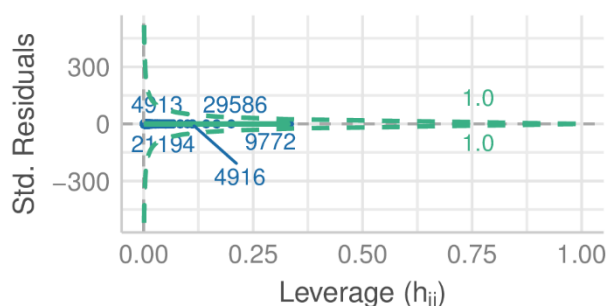
Homogeneity of Variance

Reference line should be flat and horizontal



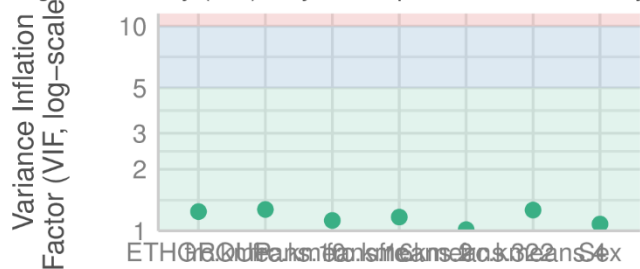
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5)

Normality of Residuals

Dots should fall along the line

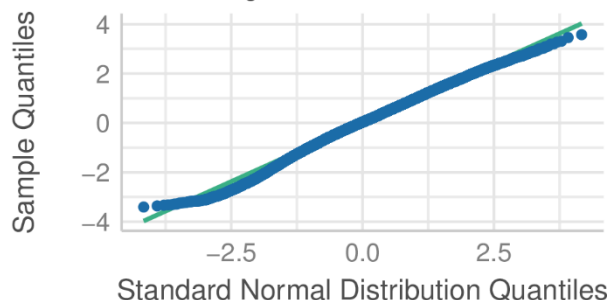


Figure S61. Linear model assumptions check for baseline linear model of TRC response variable, from GLM data set

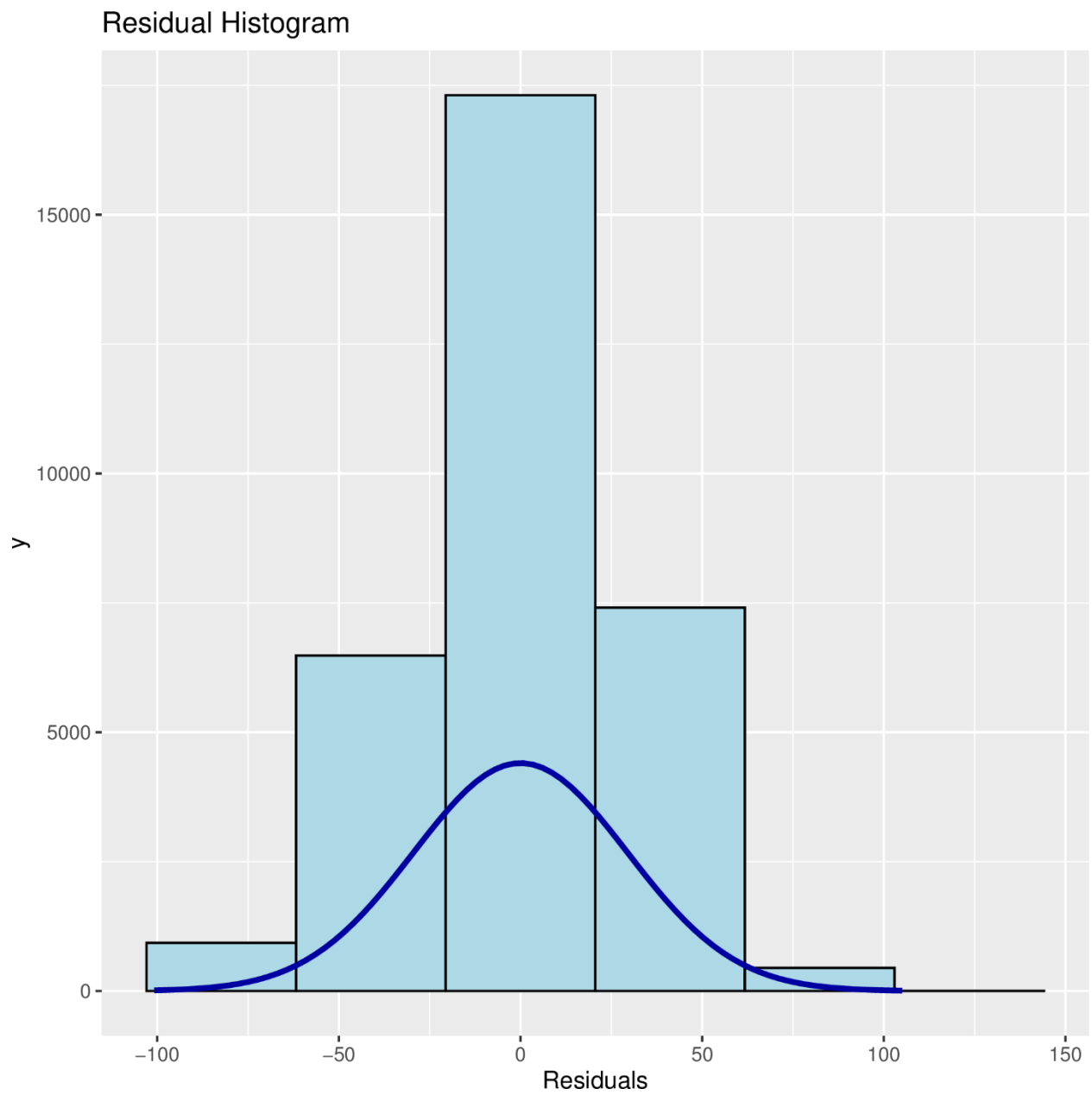
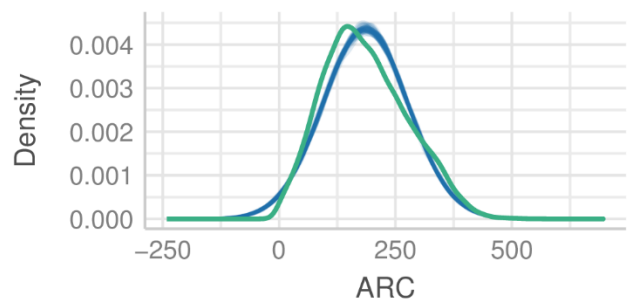


Figure S62. Histogram of residuals for baseline linear model of TRC response variable, from GLM data set

Appendix C.1.3 Supplementary figures for Baseline models for support vector regression

Posterior Predictive Check

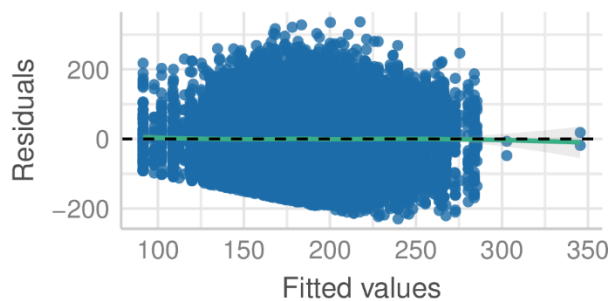
Model-predicted lines should resemble observed data



— Observed data — Model-predicted data

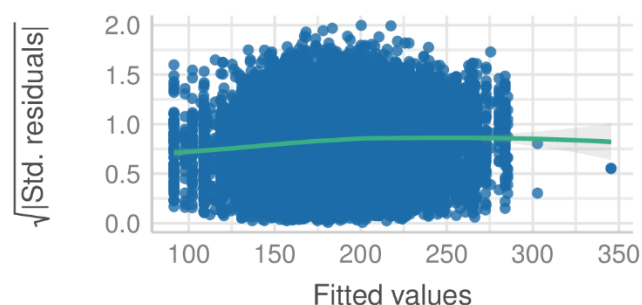
Linearity

Reference line should be flat and horizontal



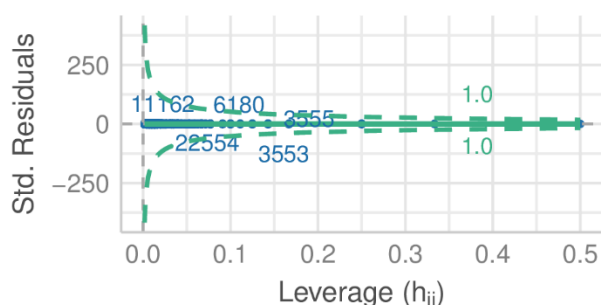
Homogeneity of Variance

Reference line should be flat and horizontal



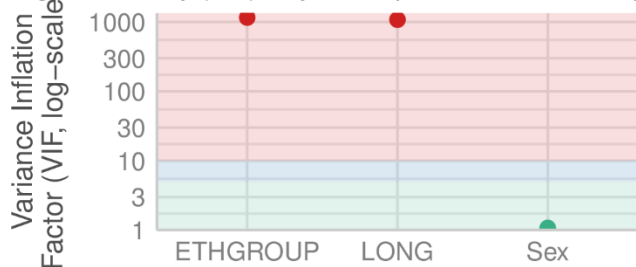
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5) ● High (... 10)

Normality of Residuals

Dots should fall along the line

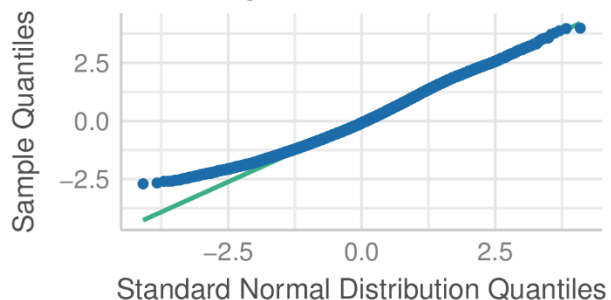


Figure S63. Linear model assumptions check for baseline linear model of ARC response variable, from SVR data set

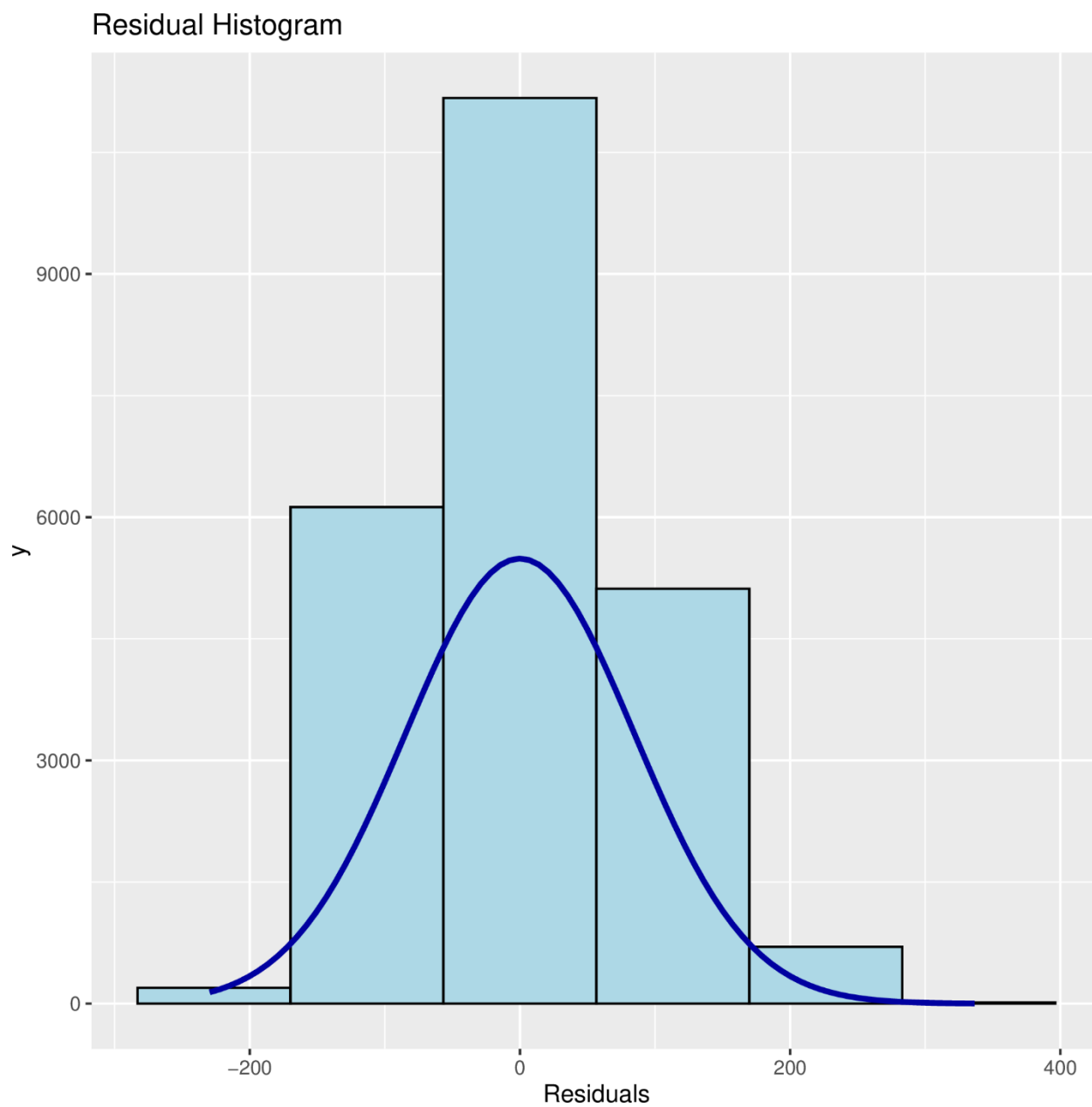
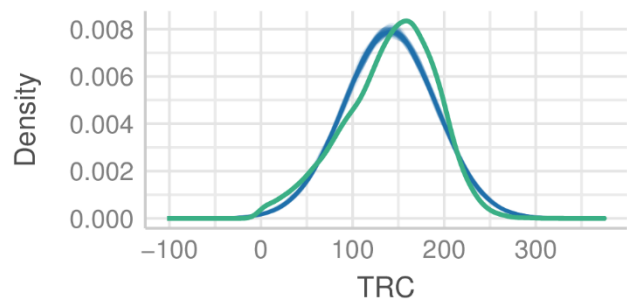


Figure S64. Histogram of residuals for baseline linear model of ARC response variable, from SVR data set

Posterior Predictive Check

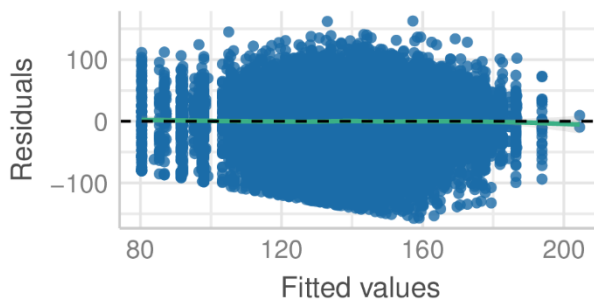
Model-predicted lines should resemble observed data



— Observed data — Model-predicted data

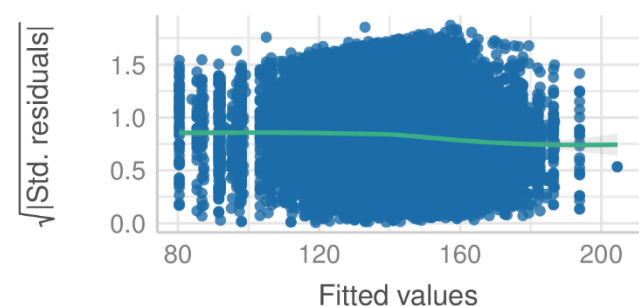
Linearity

Reference line should be flat and horizontal



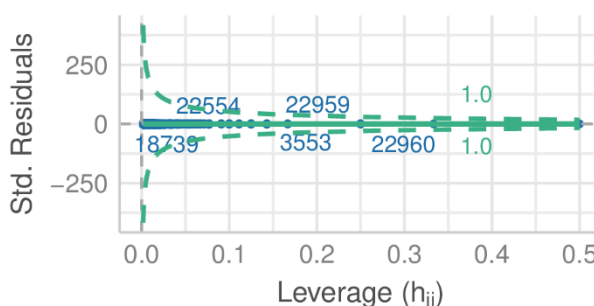
Homogeneity of Variance

Reference line should be flat and horizontal



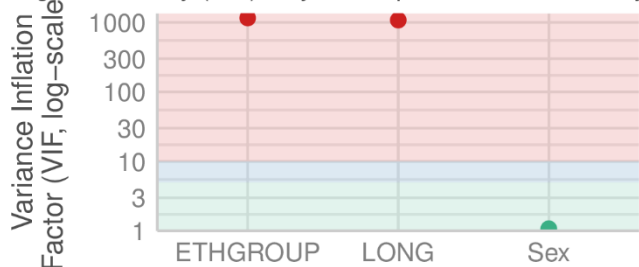
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5) ● High (... 10)

Normality of Residuals

Dots should fall along the line

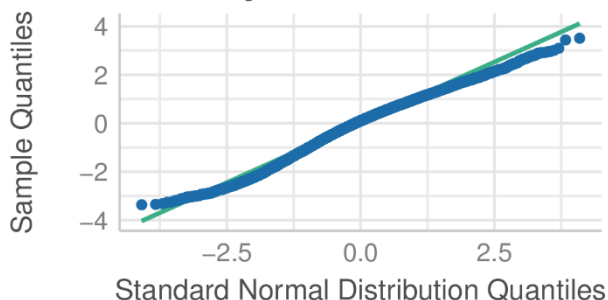


Figure S65. Linear model assumptions check for baseline linear model of TRC response variable, from SVR data set

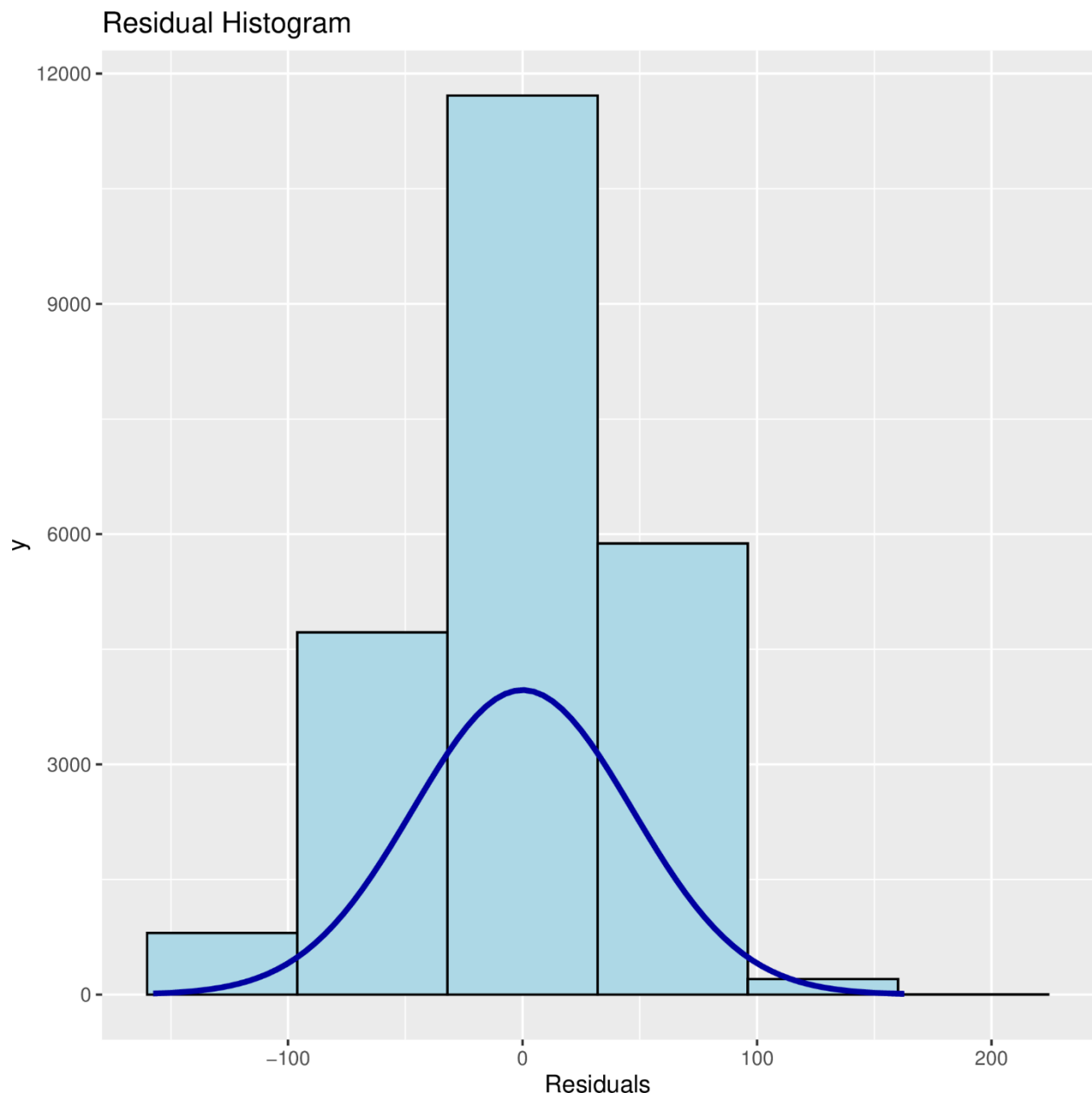


Figure S66. Histogram of residuals for baseline linear model of TRC response variable, from SVR data set

Appendix C.1.4 Supplementary figures for Support vector regression (SVR)

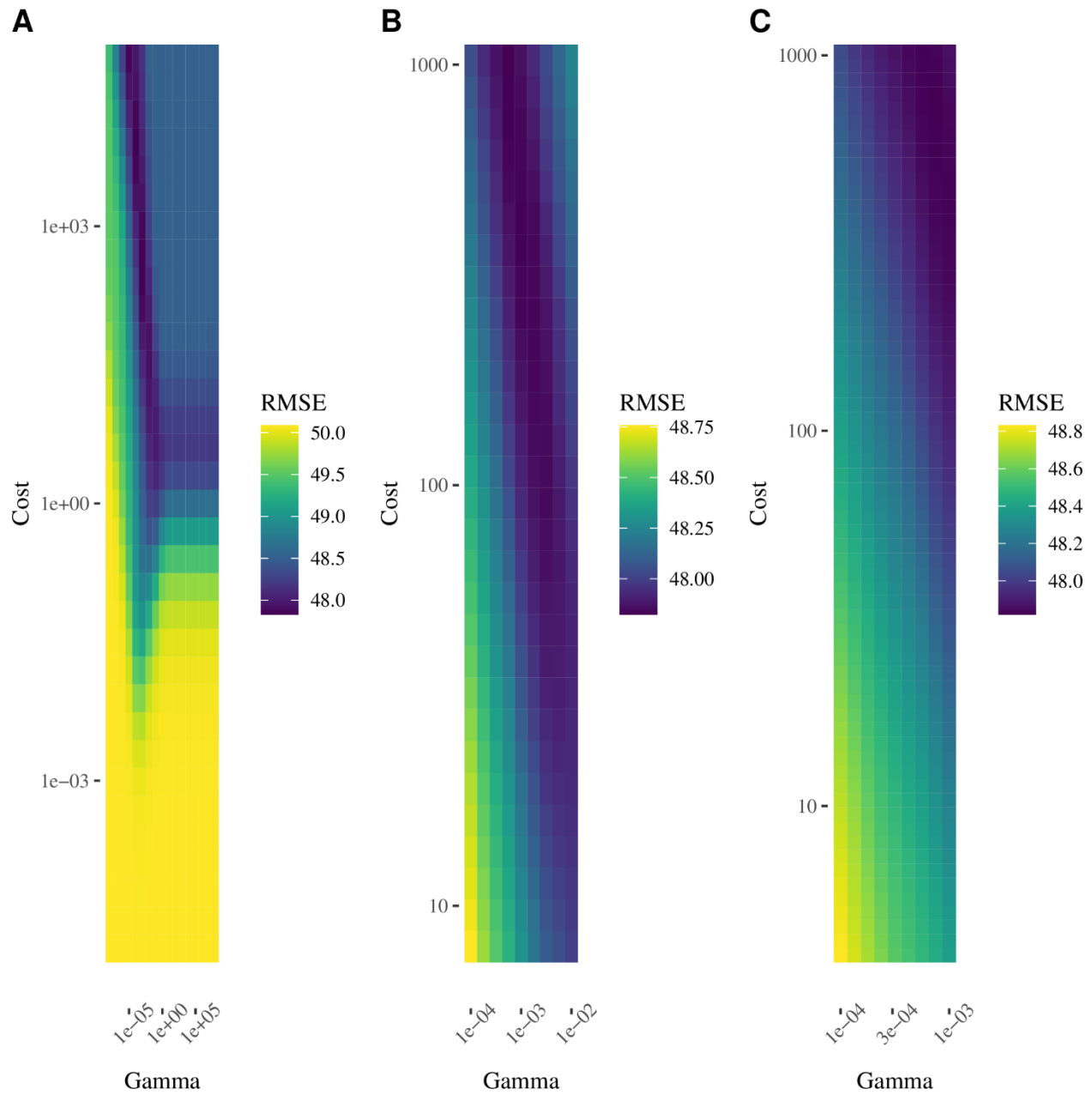


Figure S67. SVR TRC grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

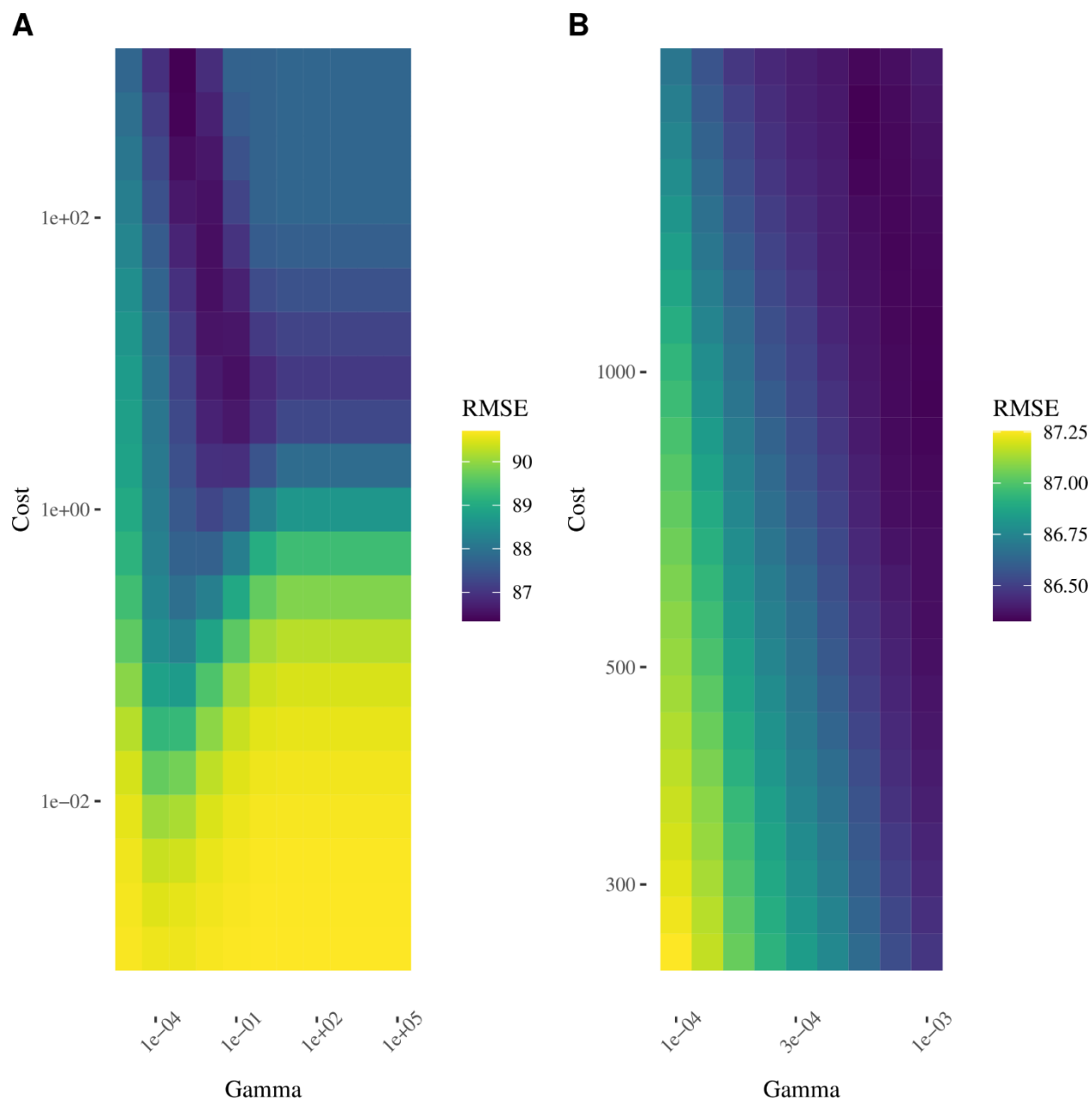


Figure S68. SVR ARC grid searches gamma vs cost. (A) coarse grid search and (B) fine grid search

Table S12. Change in SVR model performance through grid searching for TRC and ARC. The coarse, medium, and fine grid search performance metrics were calculated by comparing the model to validation data. The final model performance metrics were calculated by comparing the model to unseen test data. MAE = mean absolute error; MAPE = mean absolute percentage error; RMSE = root mean squared error

Response variable	Grid search	Hyperparameter range	MAE	RMSE	MAPE
Total ridge-count	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	37.656	47.634	0.263
	Medium, by 0.25	cost: $2^3 : 2^{10}$ γ : $10^{-4} : 10^{-1}$	37.676	47.632	0.263
	Fine, by 0.125	cost: $2^2 : 2^{11}$ γ : $10^{-4} : 10^{-3}$	37.682	47.640	0.263
	Final model	cost: 700 γ : 0.00075	37.643	47.570	0.262
Absolute ridge-count	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-10} : 2^{10}$ γ : $10^{-5} : 10^5$	68.612	85.137	0.393
	Medium, by 0.25	NA	NA	NA	NA
	Fine, by 0.125	cost: $2^8 : 2^{11}$ γ : $10^{-4} : 10^{-3}$	68.598	85.109	0.393
	Final model	cost: 1320 γ : 0.00075	68.272	84.423	0.391

Appendix C.2 Supplementary figures for directional asymmetry, fluctuating asymmetry, and indices of asymmetry and diversity

Appendix C.2.1 Supplementary figures for Spatial analysis via distance-based Moran's eigenvector maps (dbMEM)

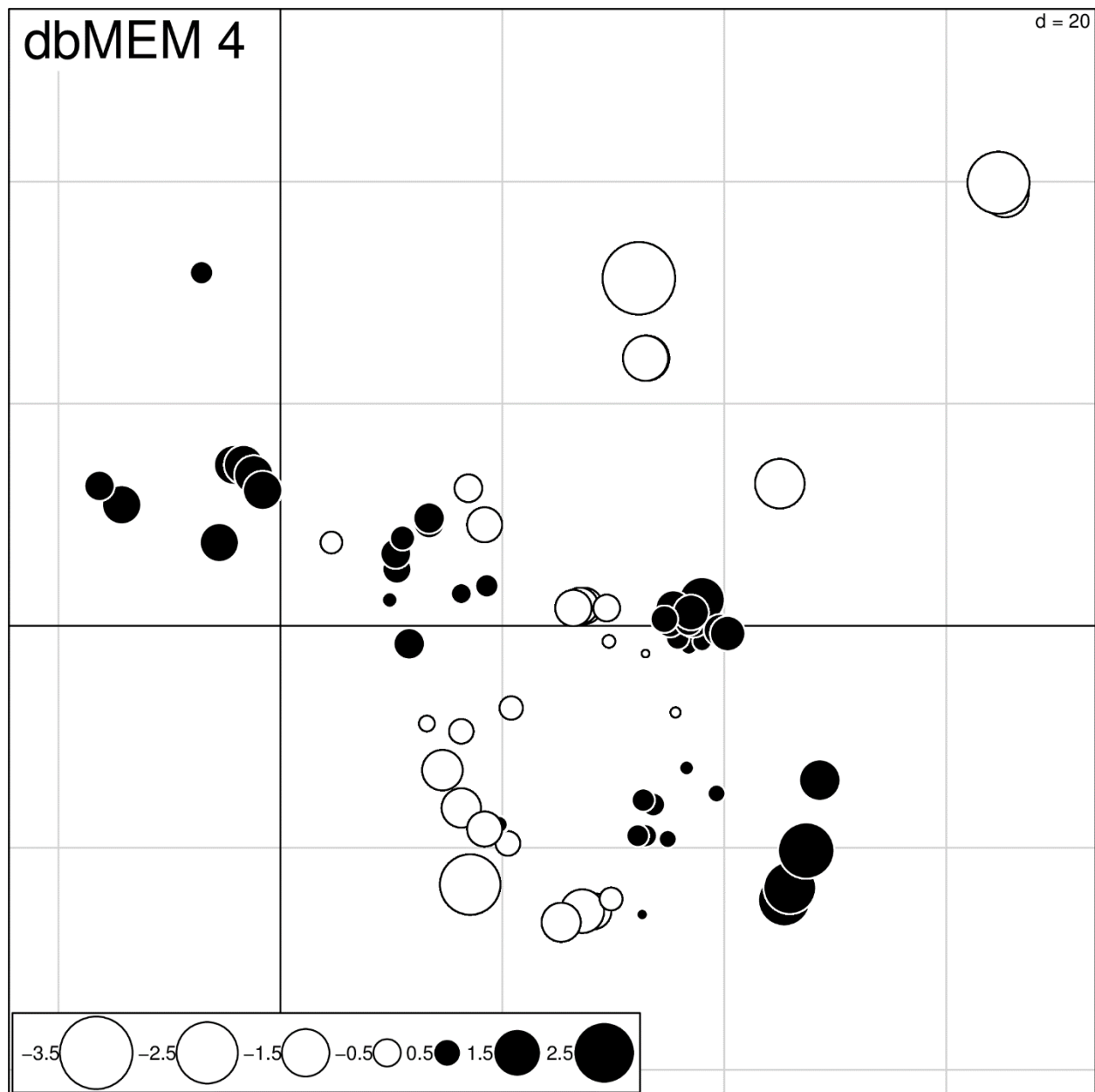


Figure S69. Map of the single significant dbMEM variable for Landmass Africa Directional Asymmetry

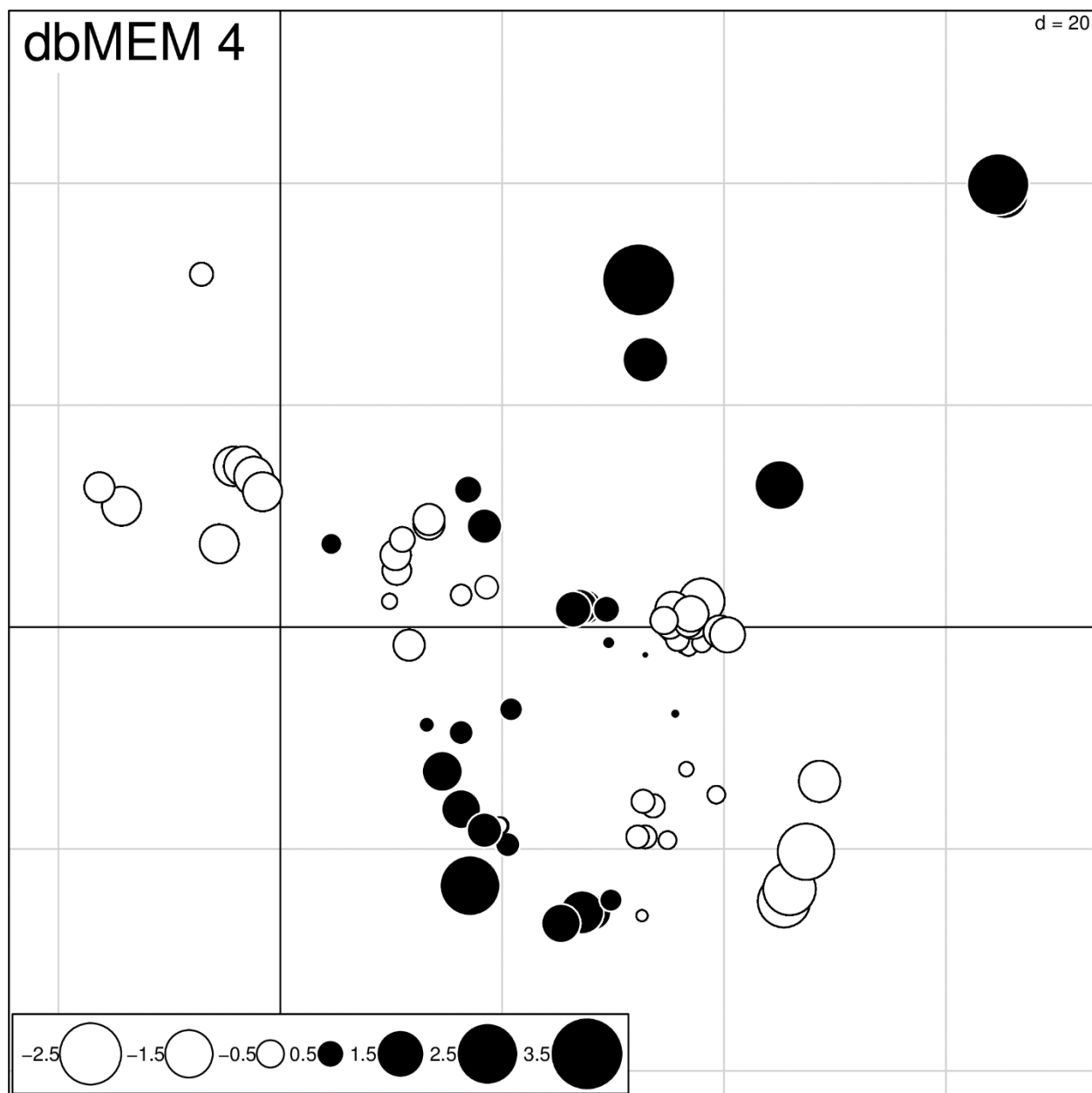


Figure S70. Map of the single significant dbMEM variable for Landmass Africa Fluctuating Asymmetry

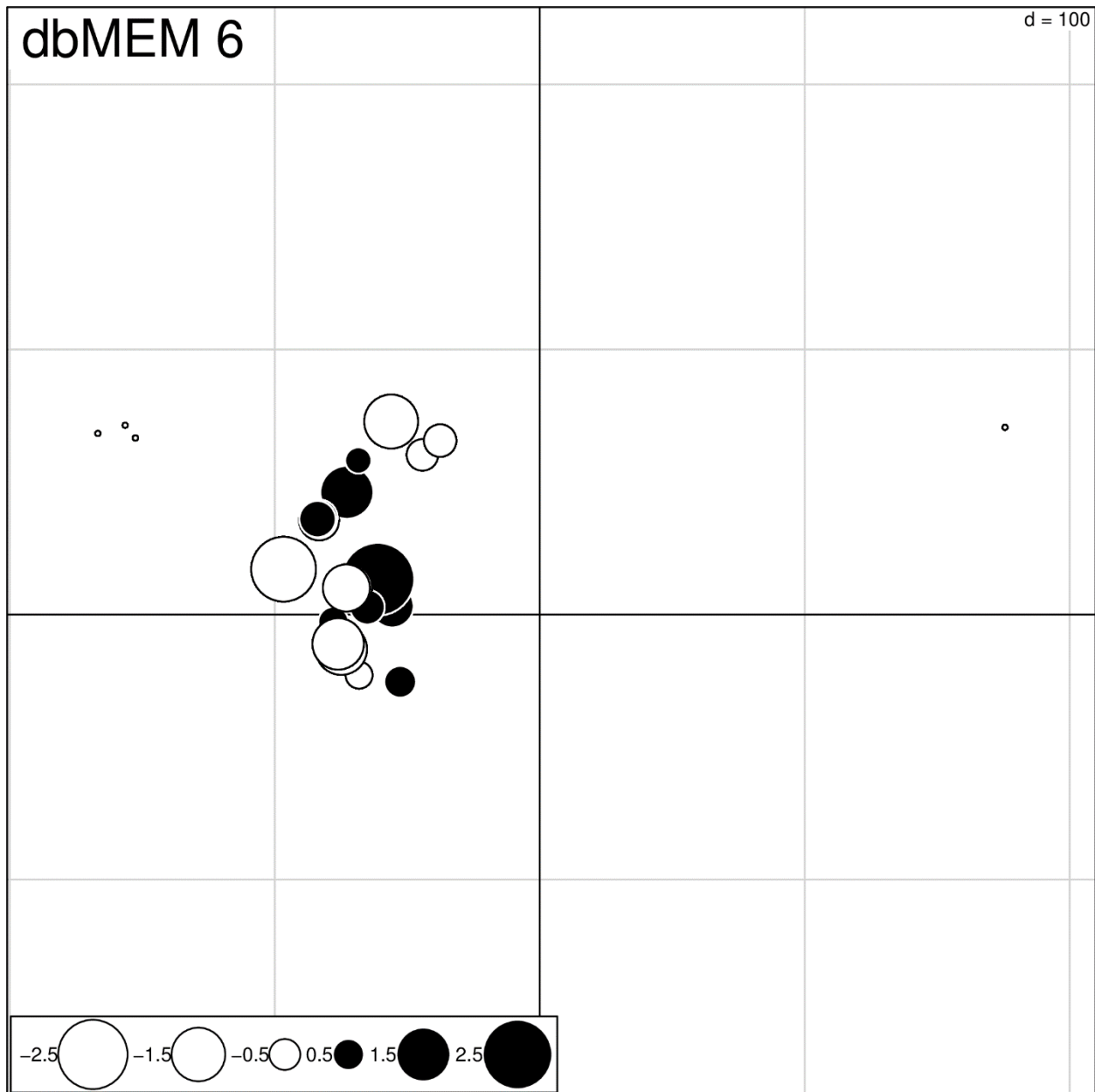


Figure S71. Map of the single significant dbMEM variable for Landmass Americas Directional Asymmetry

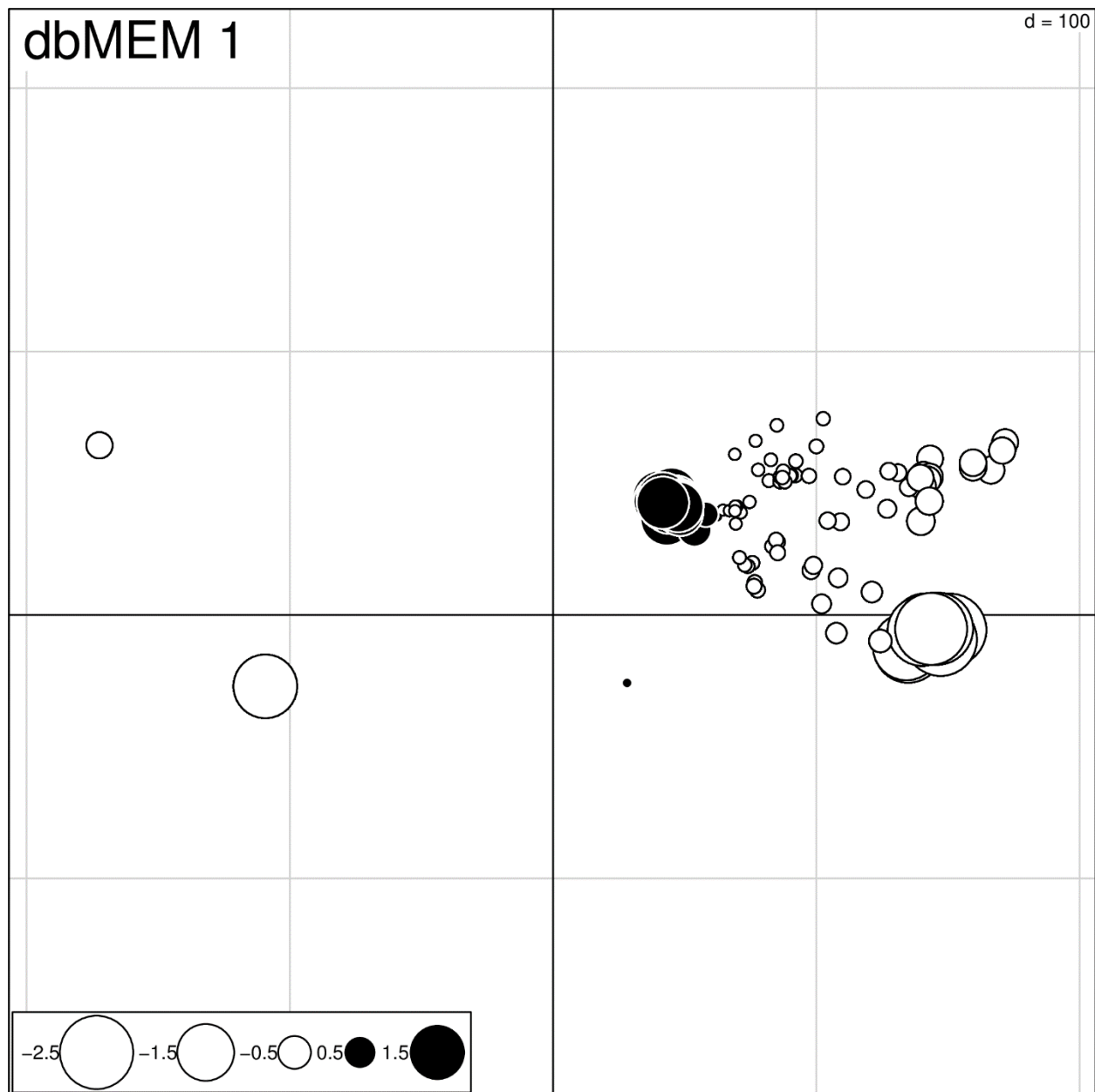


Figure S72. Map of the single significant dbMEM variable for Landmass Asia Directional Asymmetry

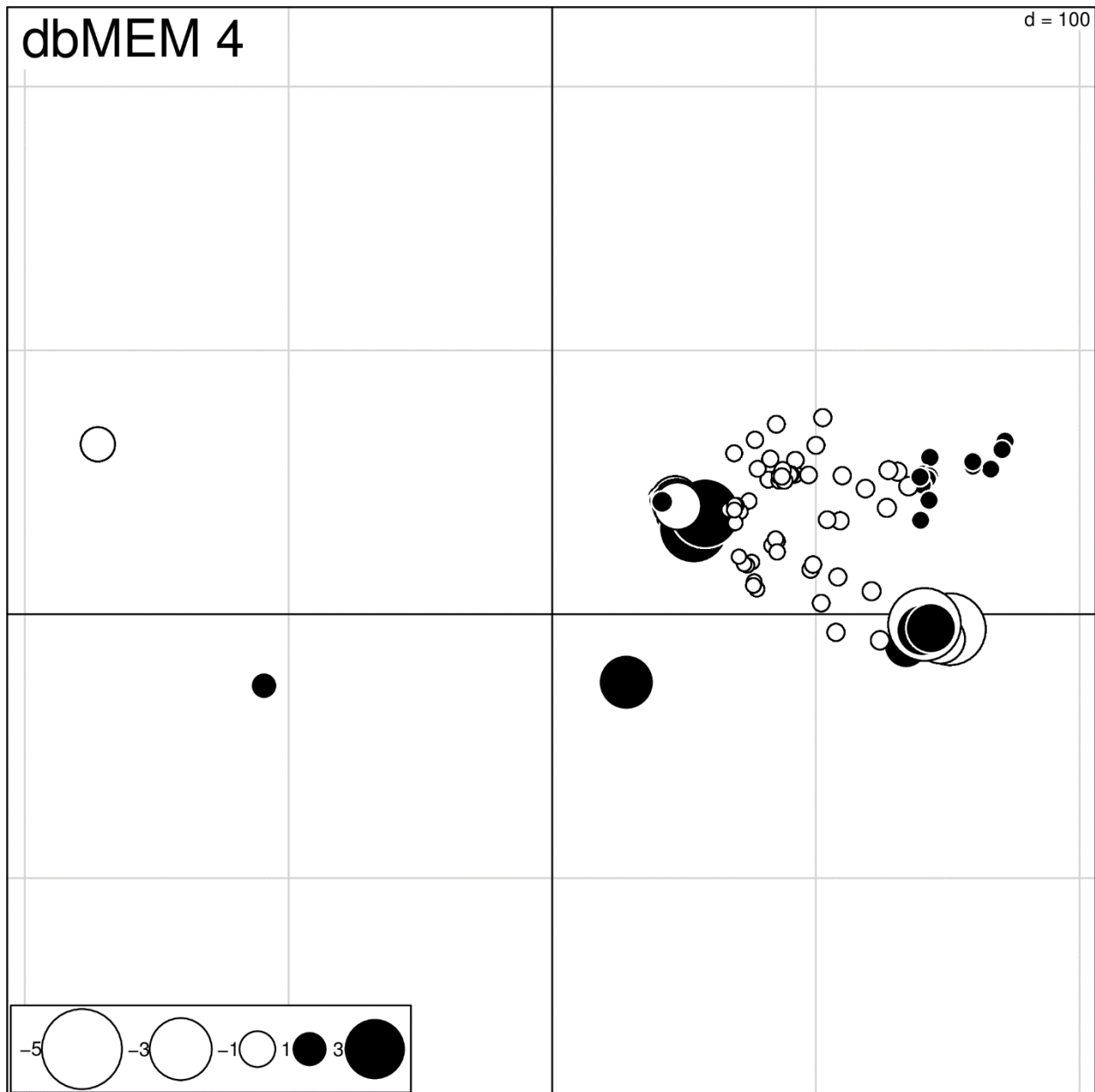


Figure S73. Map of the single significant dbMEM variable for Landmass Asia Fluctuating Asymmetry

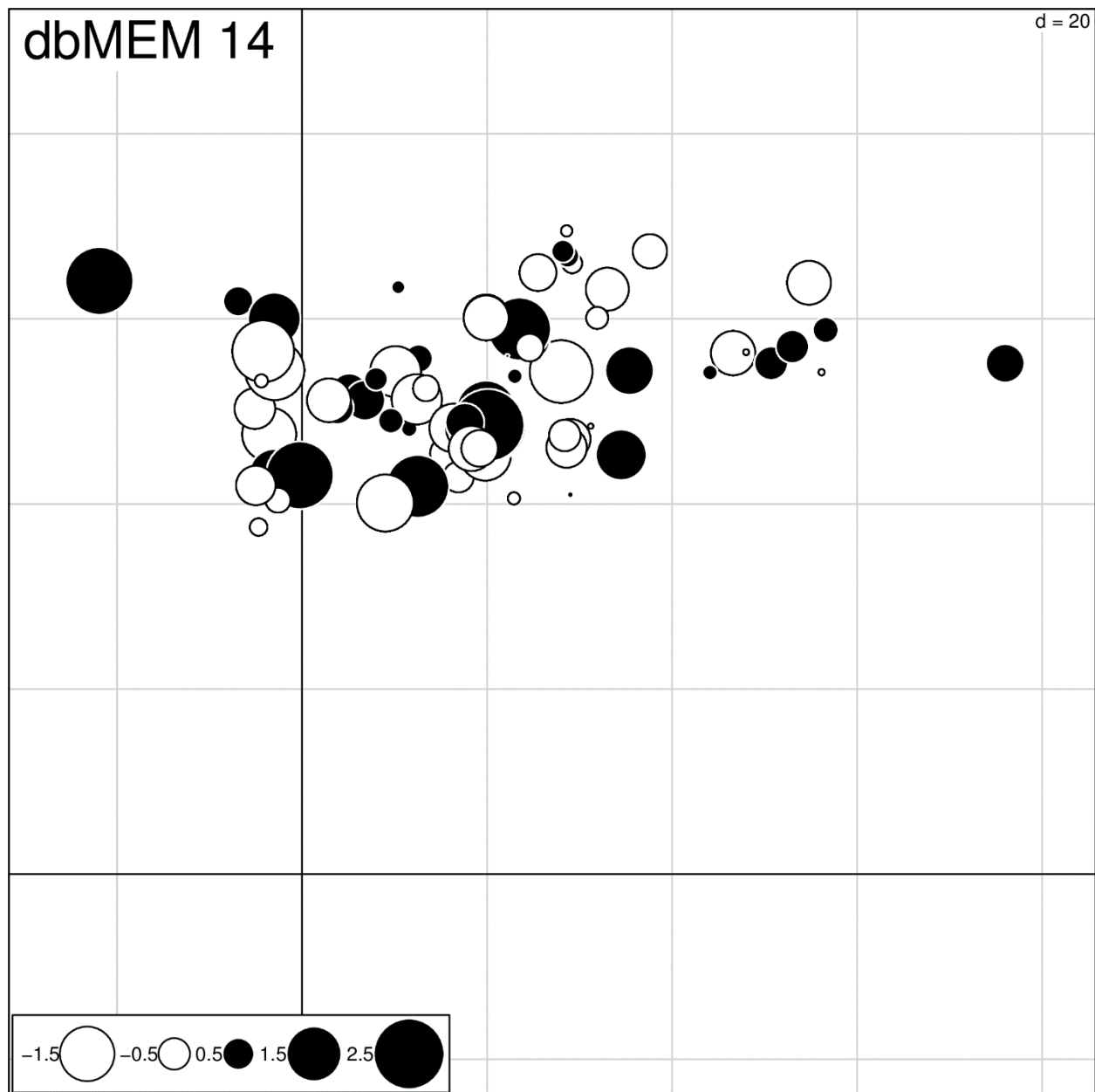


Figure S74. Map of the single significant dbMEM variable for Landmass Europe Directional Asymmetry

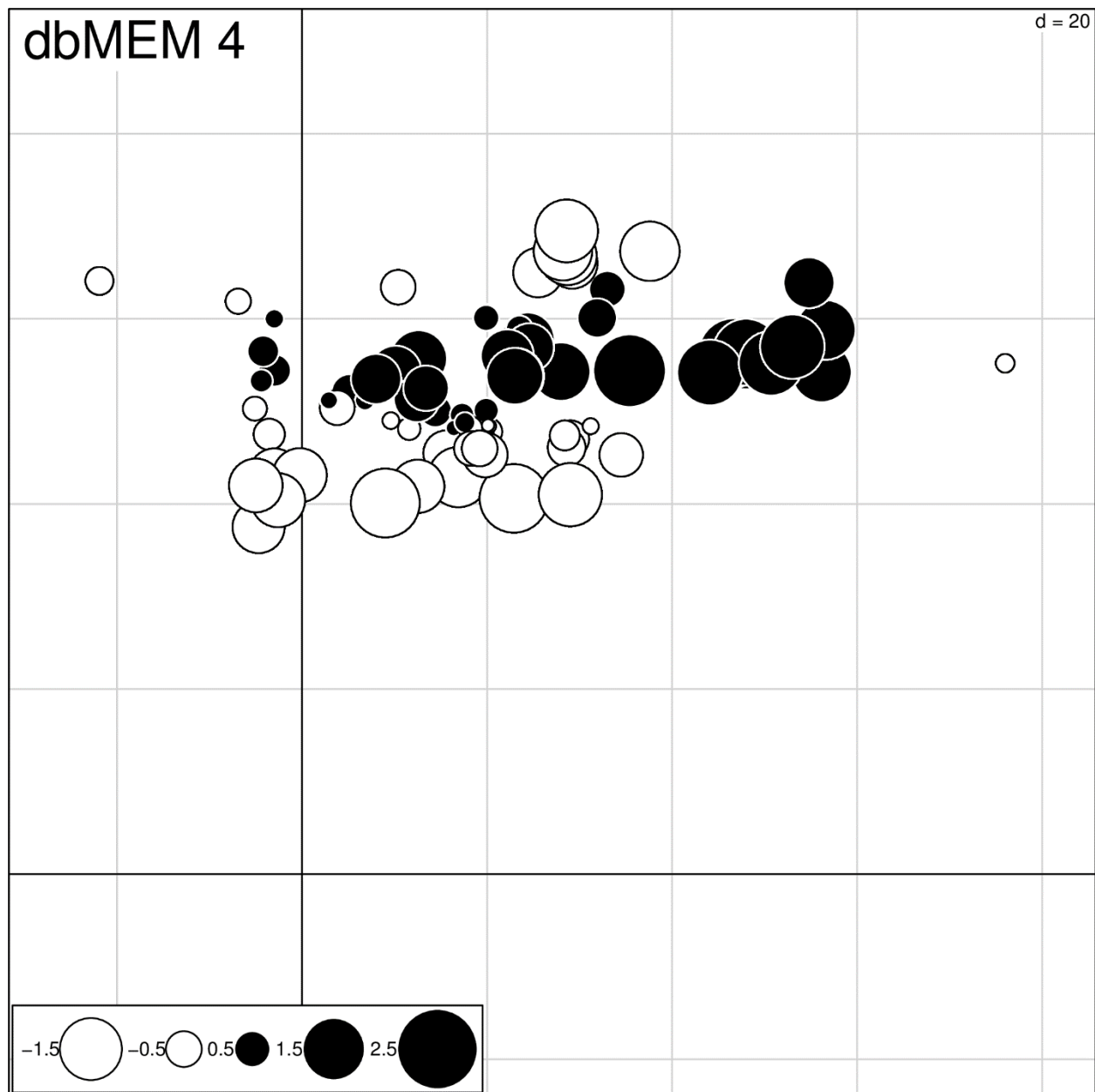
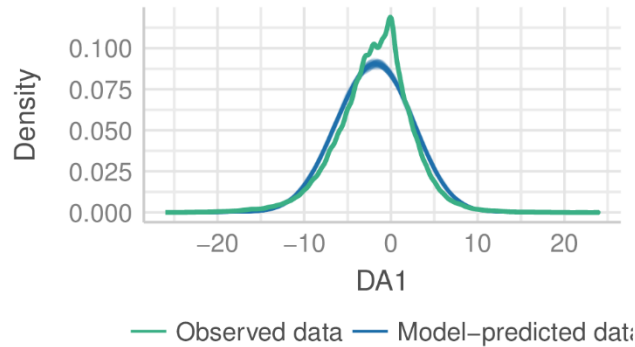


Figure S75. Map of the single significant dbMEM variable for Landmass Europe Fluctuating Asymmetry

Appendix C.2.2 Supplementary figures for Baseline models for generalized linear modeling

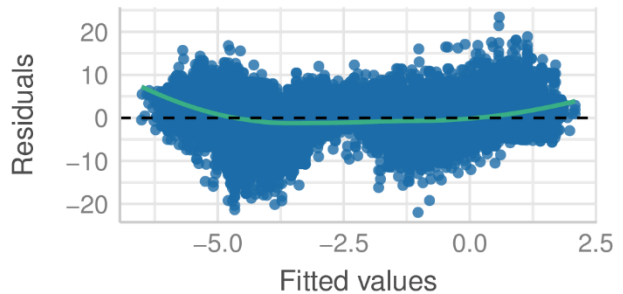
Posterior Predictive Check

Model-predicted lines should resemble observed data



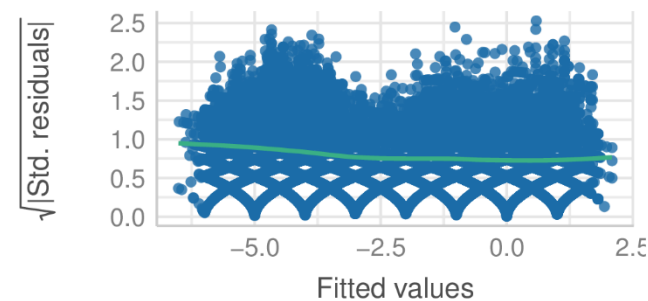
Linearity

Reference line should be flat and horizontal



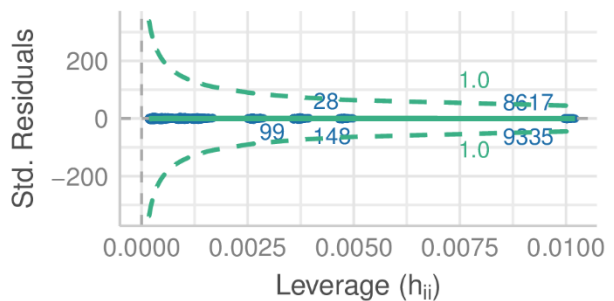
Homogeneity of Variance

Reference line should be flat and horizontal



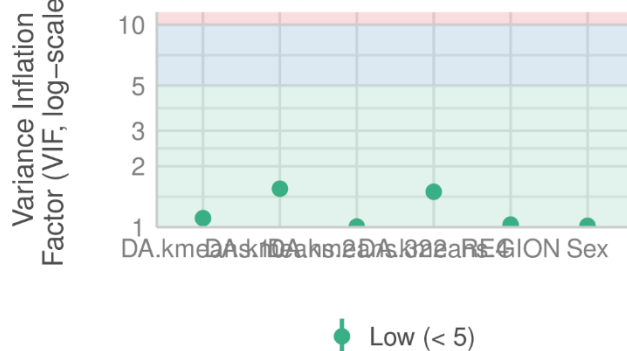
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

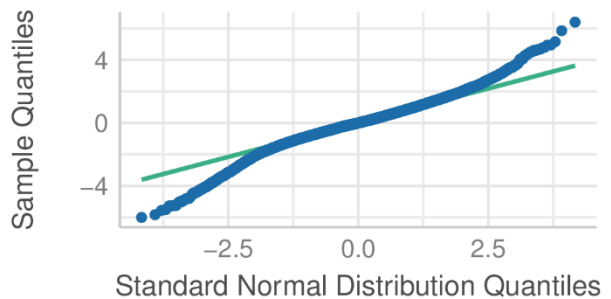


Figure S76. Linear model assumptions check for baseline linear model of DA1 response variable, from GLM data set

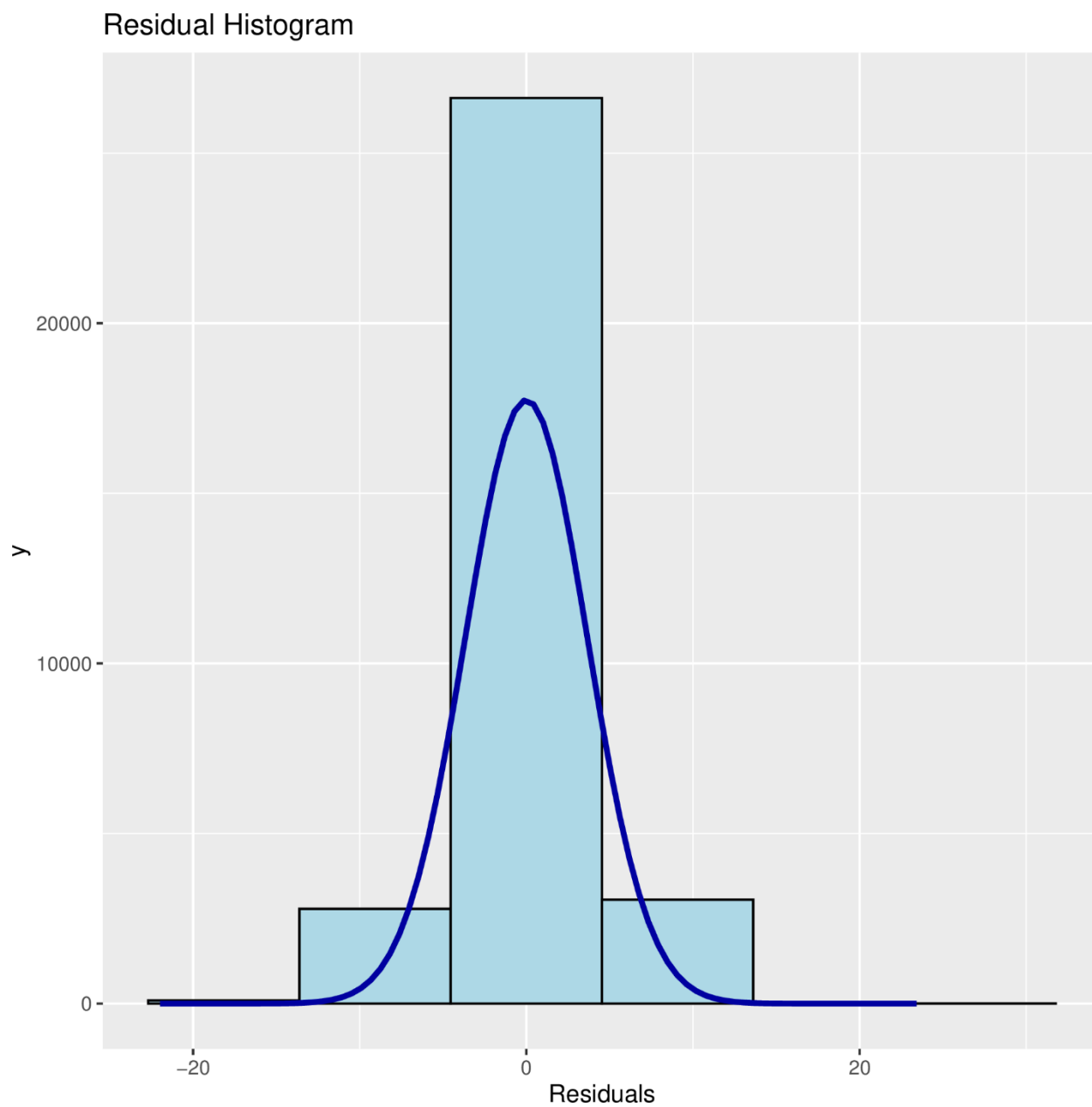
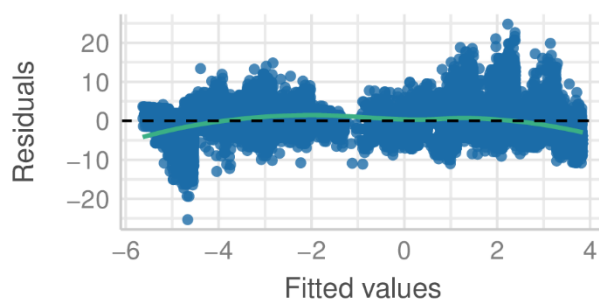


Figure S77. Histogram of residuals for baseline linear model of DA1 response variable, from GLM data set

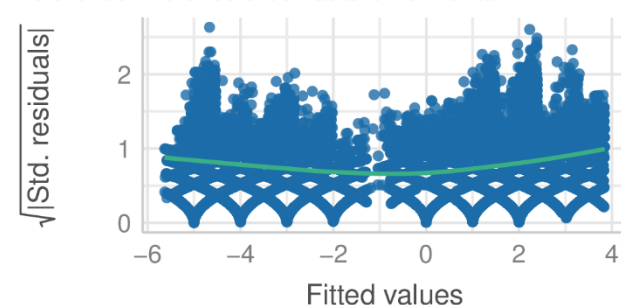
Model-predicted lines should resemble observed data li



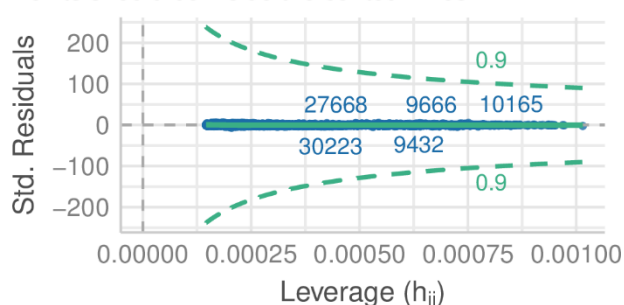
Reference line should be flat and horizontal



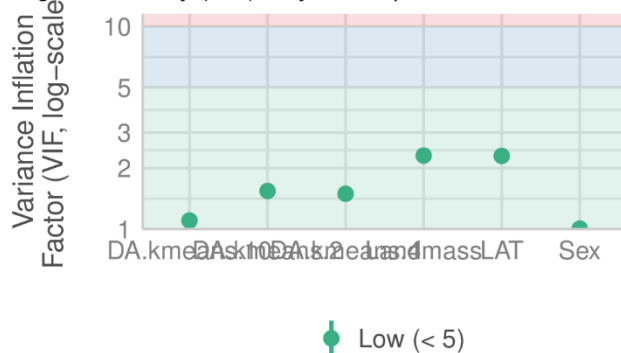
Reference line should be flat and horizontal



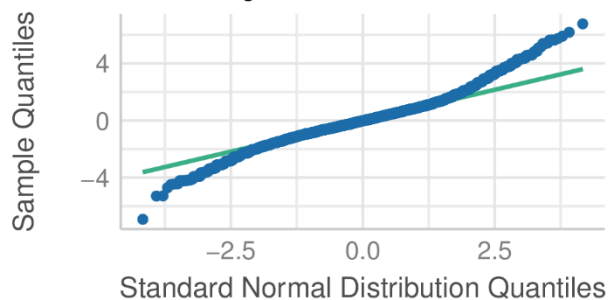
Points should be inside the contour lines



High collinearity (VIF) may inflate parameter uncertainty



Dots should fall along the line



variable, from GLM data set

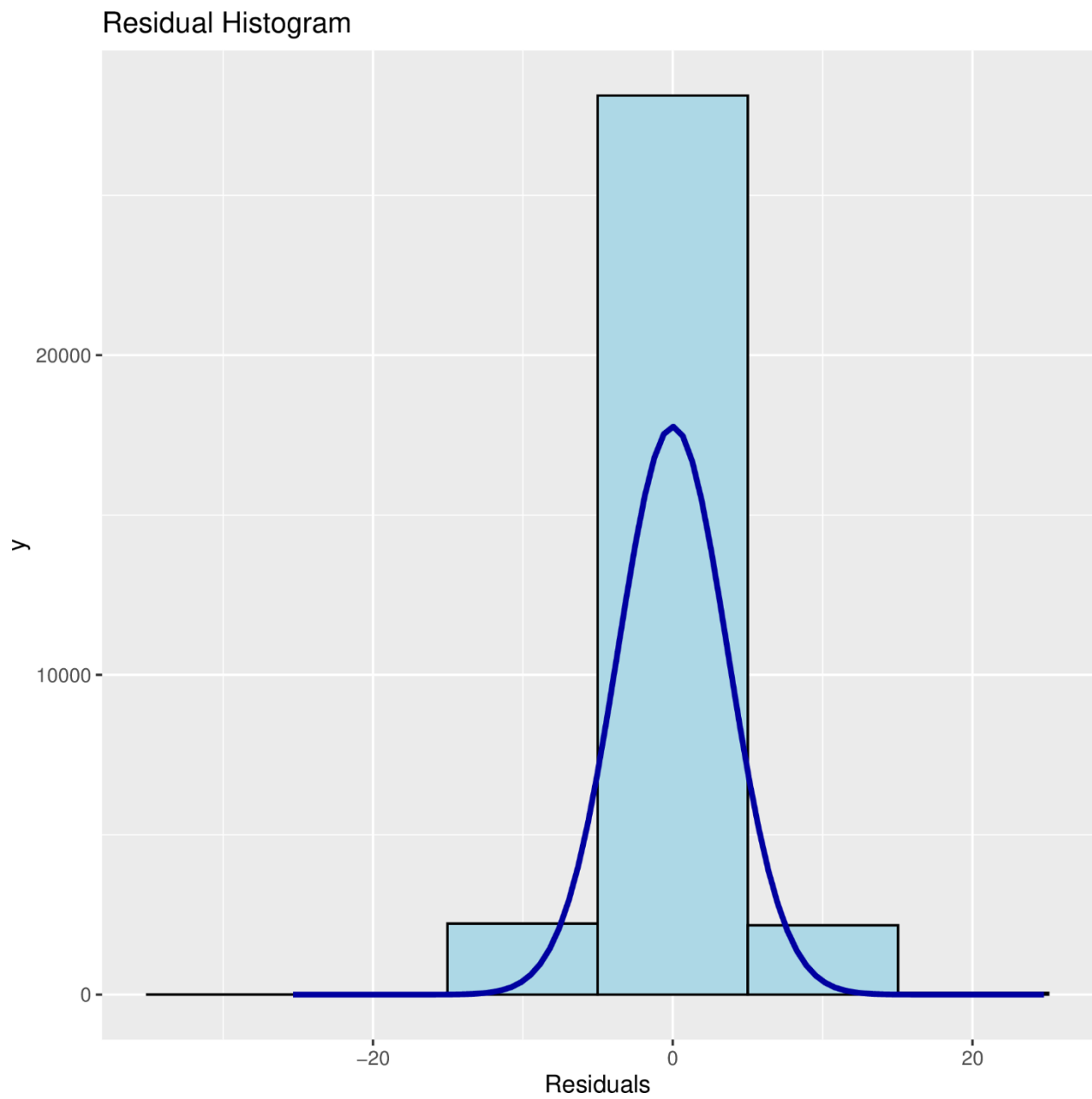


Figure S79. Histogram of residuals for baseline linear model of DA2 response variable, from GLM data set

Model-predicted lines should resemble observed data li



i Reference line should be flat and horizontal



Reference line should be flat and horizontal



Points should be inside the contour lines



High collinearity (VIF) may inflate parameter uncertainty



Dots should fall along the line



variable, from GLM data set

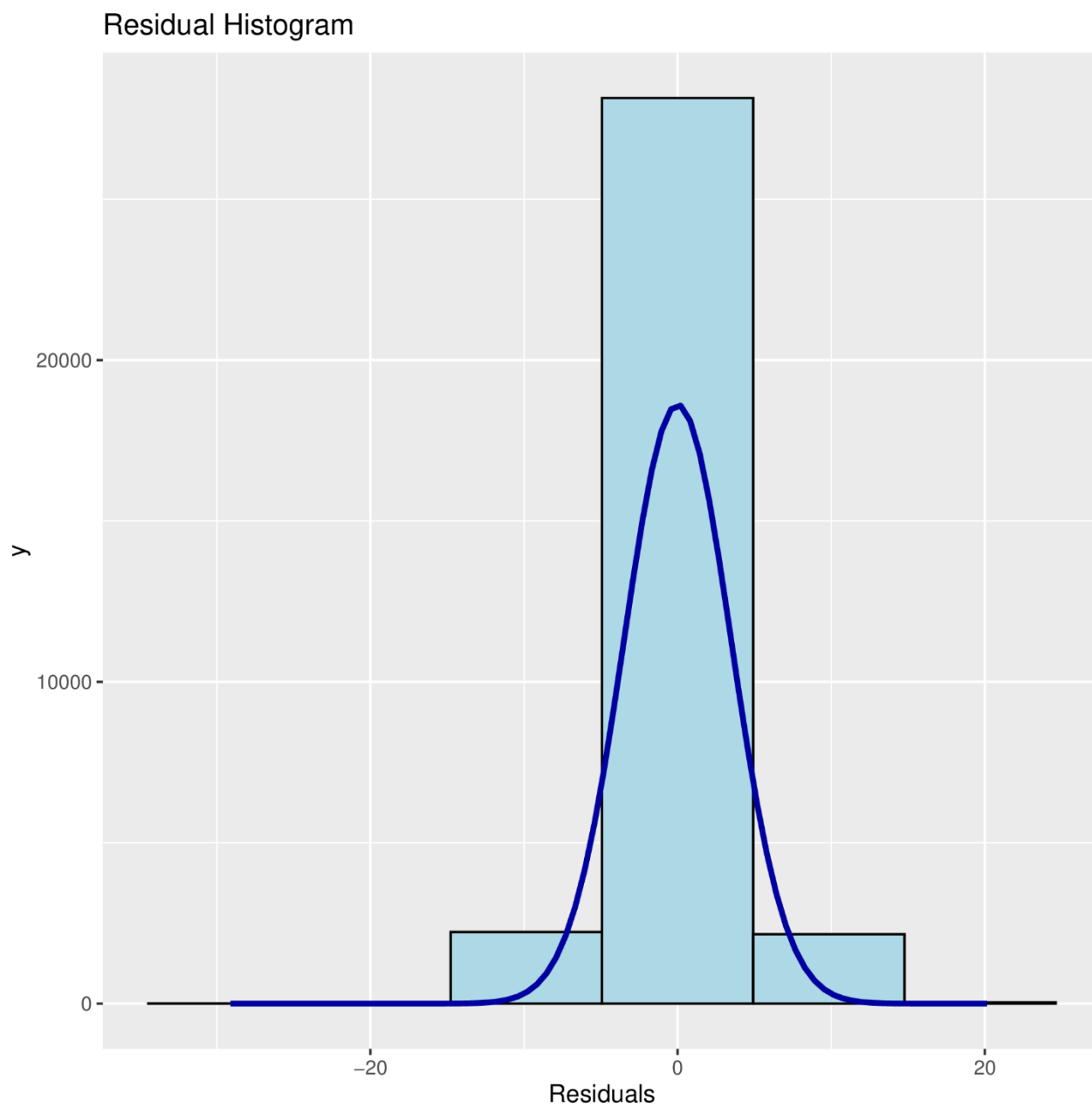
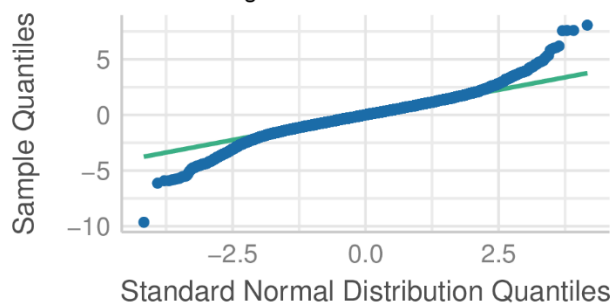
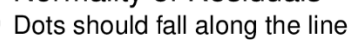
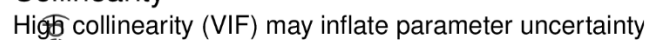
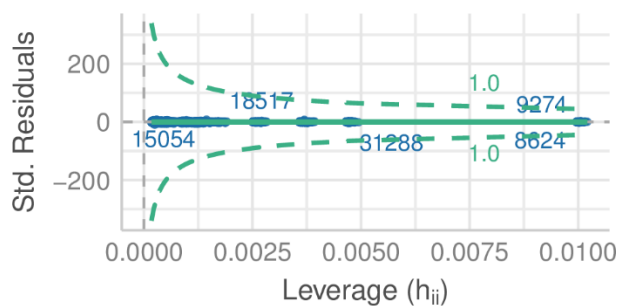
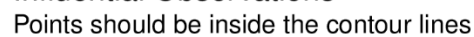
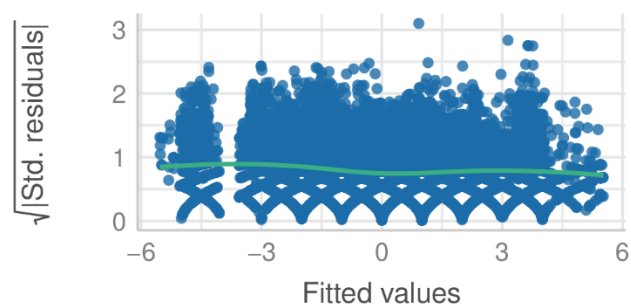
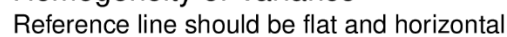
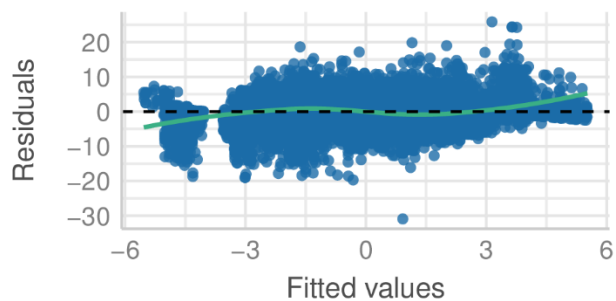
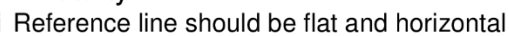
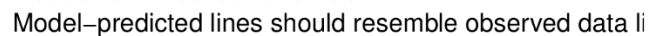


Figure S81. Histogram of residuals for baseline linear model of DA3 response variable, from GLM data set



532

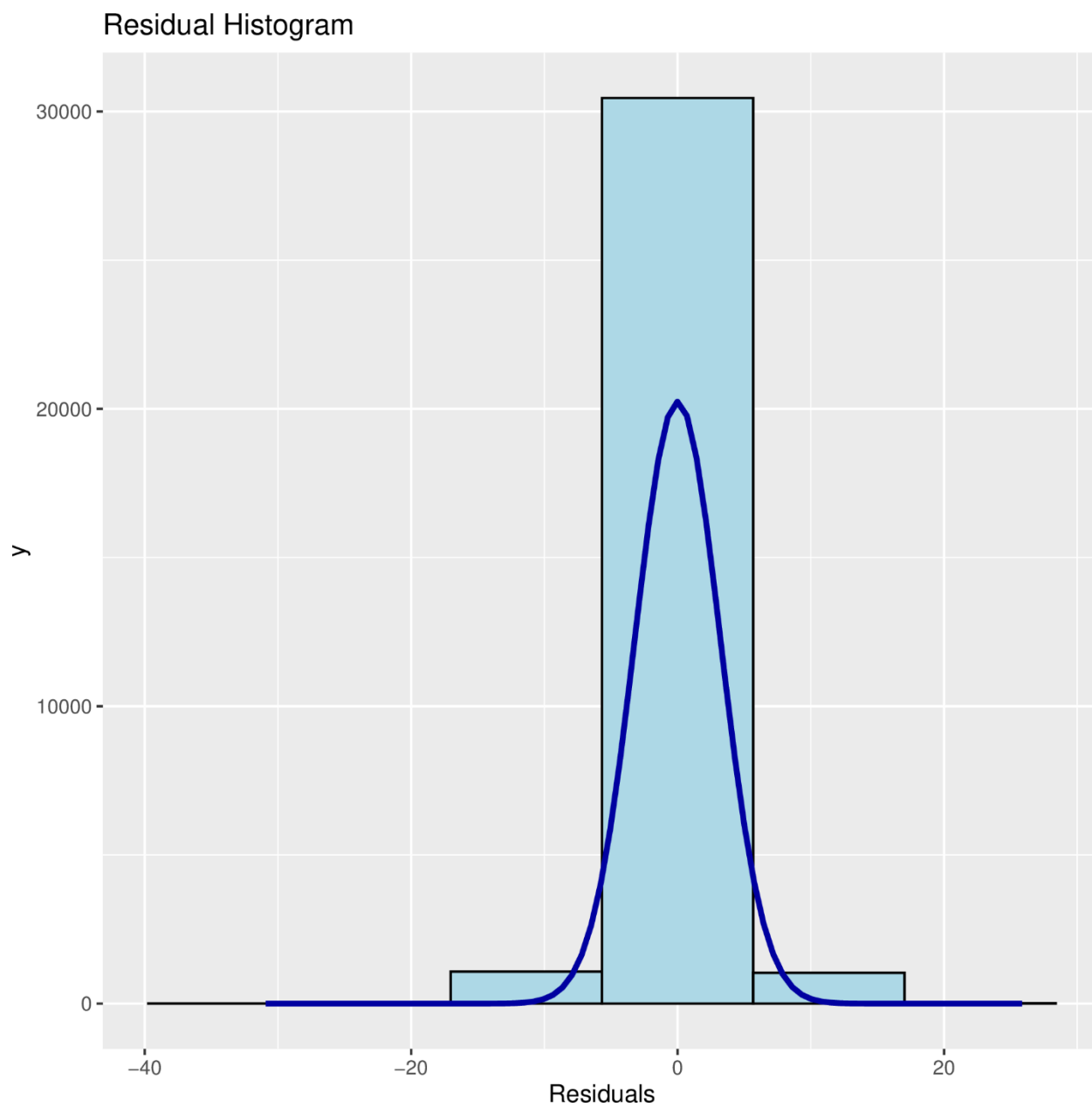


Figure S83. Histogram of residuals for baseline linear model of DA4 response variable, from GLM data set

Model-predicted lines should resemble observed data li



Reference line should be flat and horizontal



Reference line should be flat and horizontal



Points should be inside the contour lines



High collinearity (VIF) may inflate parameter uncertainty



Dots should fall along the line



variable, from GLM data set

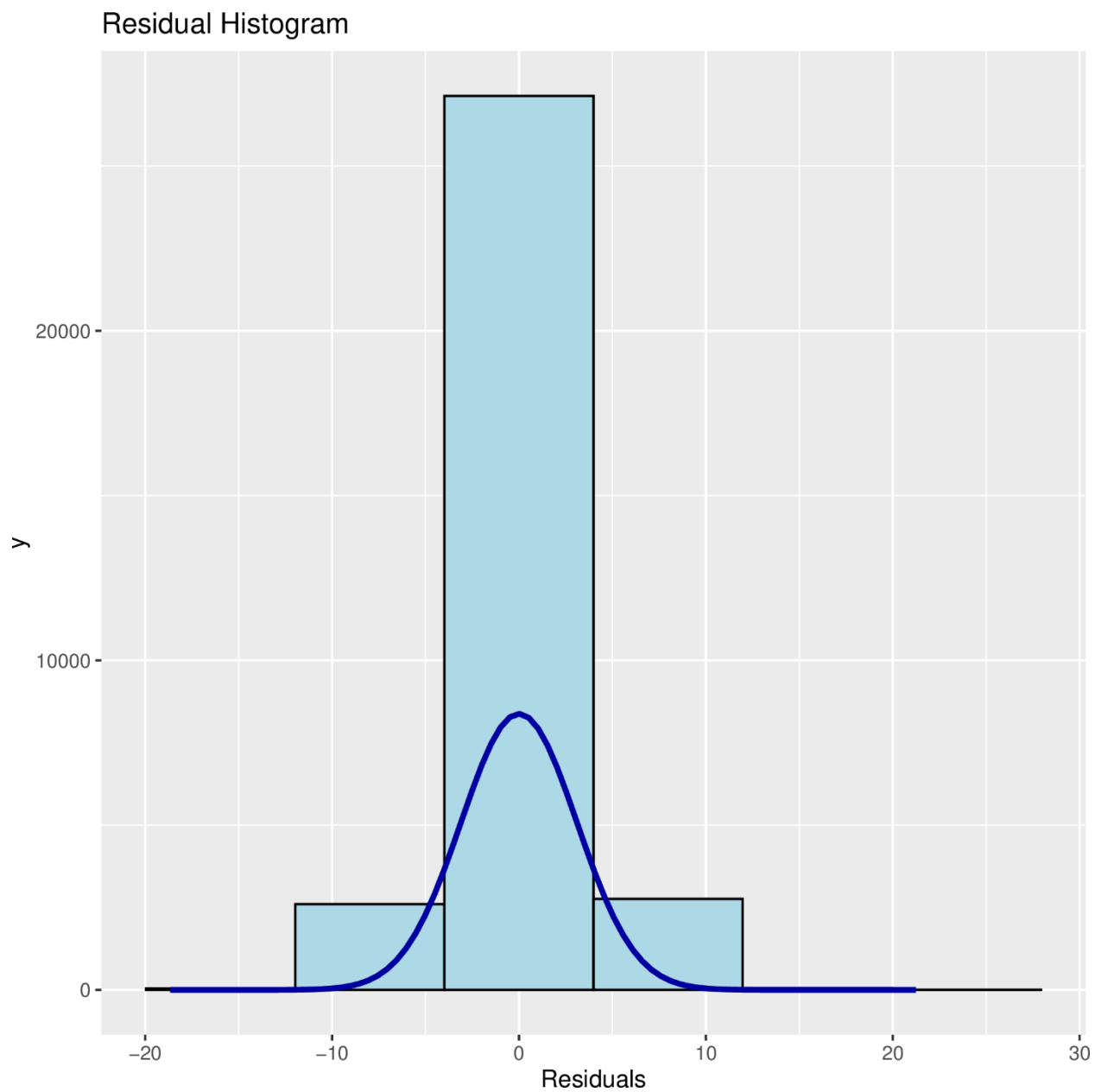
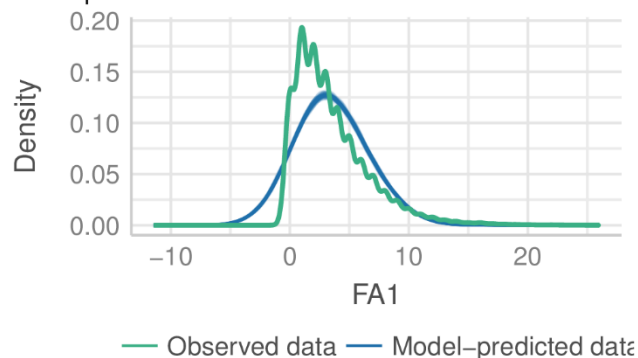


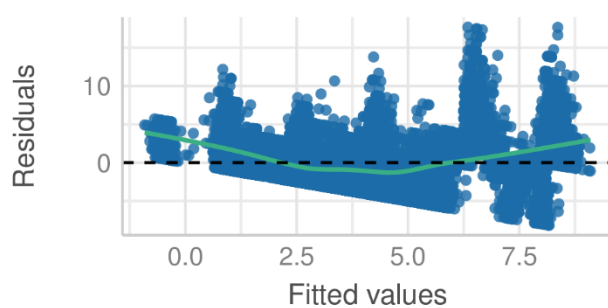
Figure S85. Histogram of residuals for baseline linear model of DA5 response variable, from GLM data set



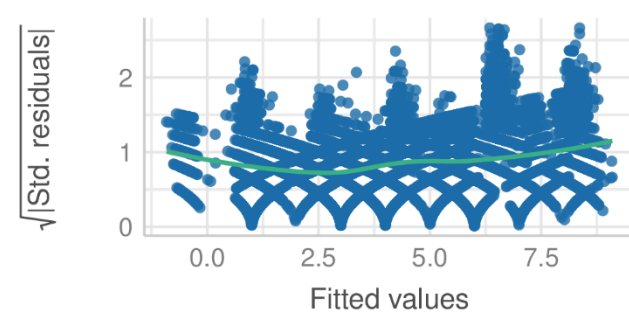
Model-predicted lines should resemble observed data li Reference line should be flat and horizontal



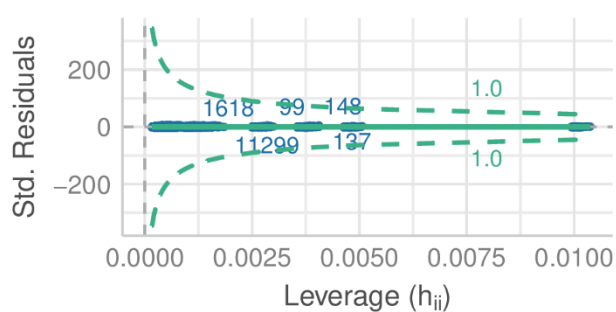
Reference line should be flat and horizontal



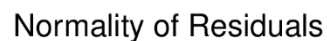
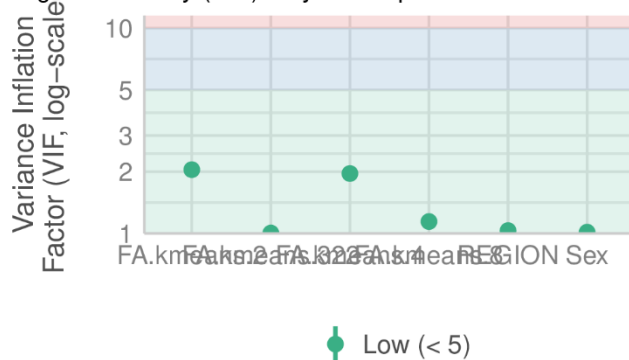
Reference line should be flat and horizontal



Points should be inside the contour lines



High collinearity (VIF) may inflate parameter uncertainty



Dots should fall along the line

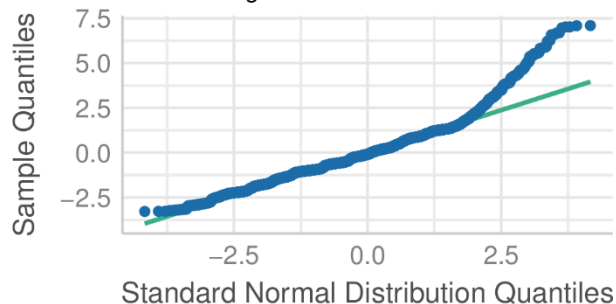


Figure S86. Linear model assumptions check for baseline linear model of FA1 response variable,
from GLM data set

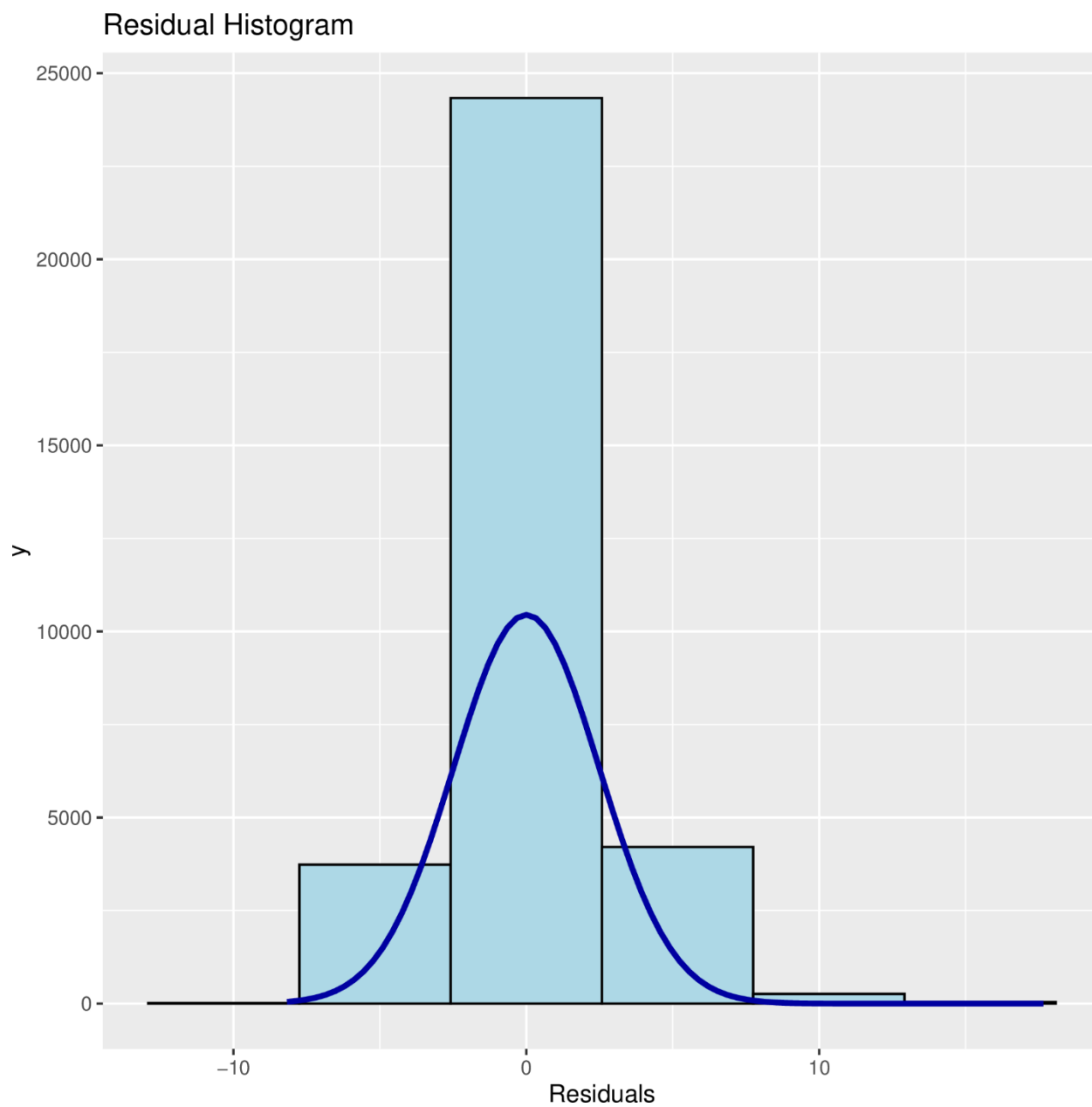
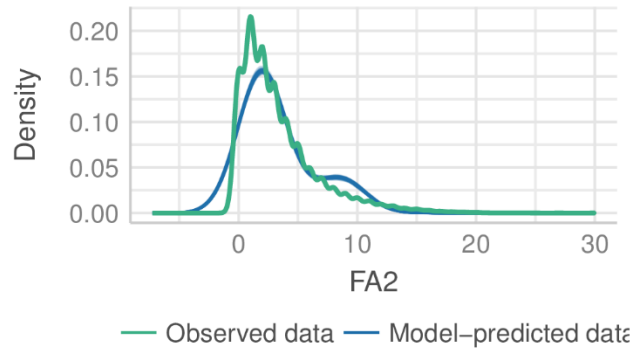


Figure S87. Histogram of residuals for baseline linear model of FA1 response variable, from GLM data set

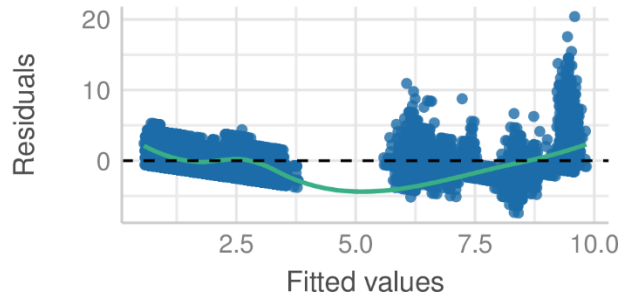
Posterior Predictive Check

Model-predicted lines should resemble observed data li



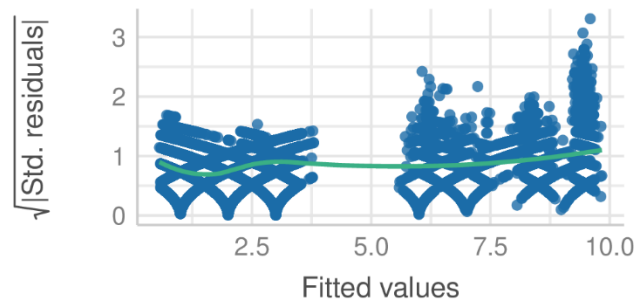
Linearity

Reference line should be flat and horizontal



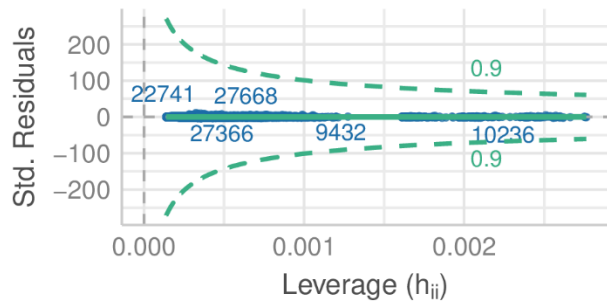
Homogeneity of Variance

Reference line should be flat and horizontal



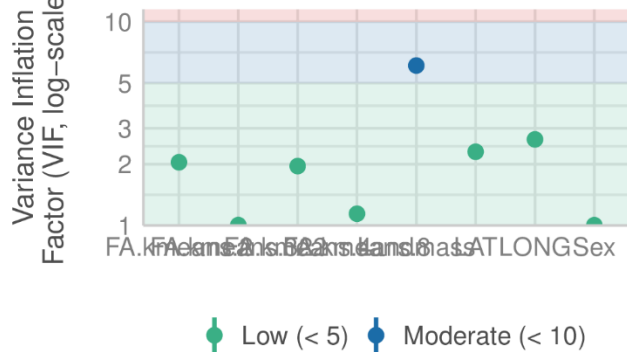
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

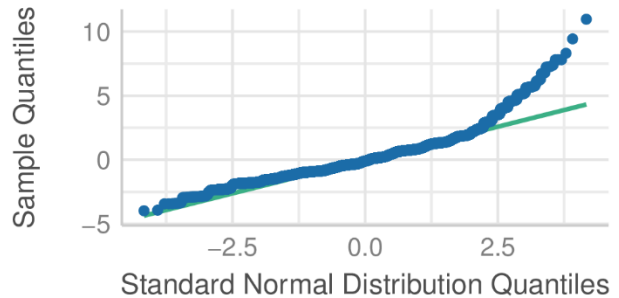


Figure S88. Linear model assumptions check for baseline linear model of FA2 response variable, from GLM data set

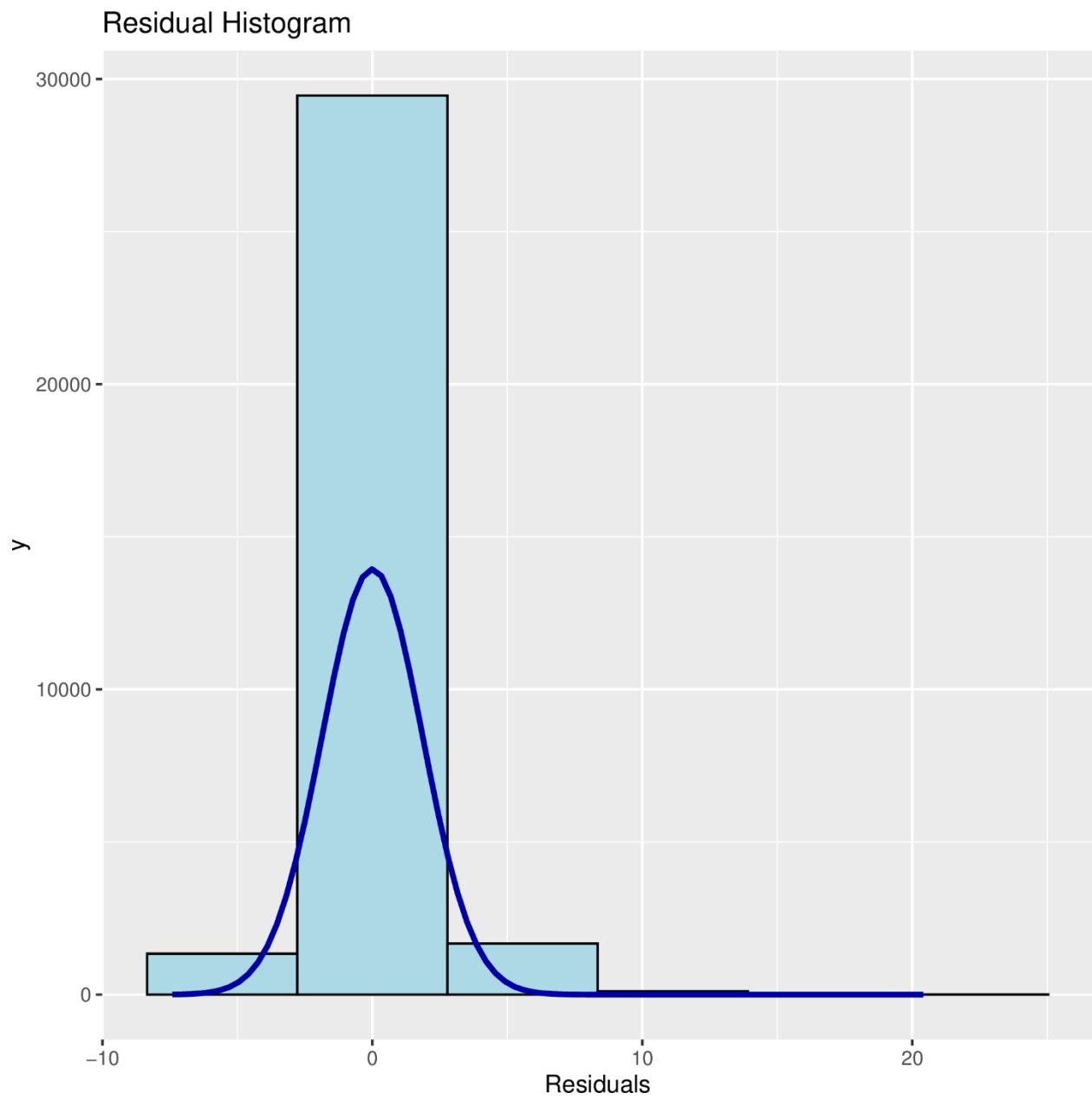
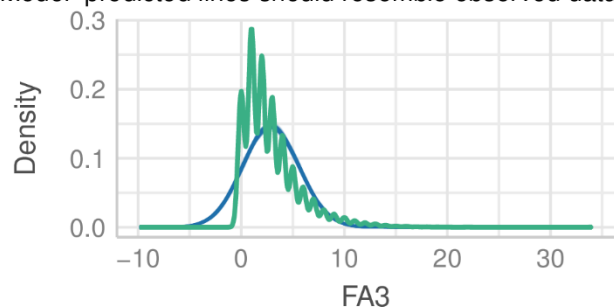


Figure S89. Histogram of residuals for baseline linear model of FA2 response variable, from GLM data set

Posterior Predictive Check

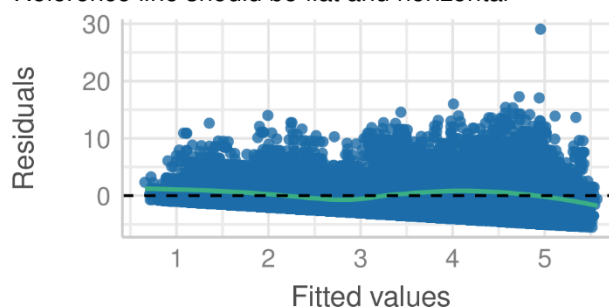
Model-predicted lines should resemble observed data



— Observed data — Model-predicted data

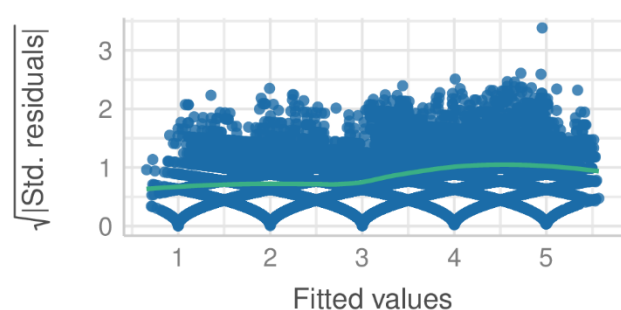
Linearity

Reference line should be flat and horizontal



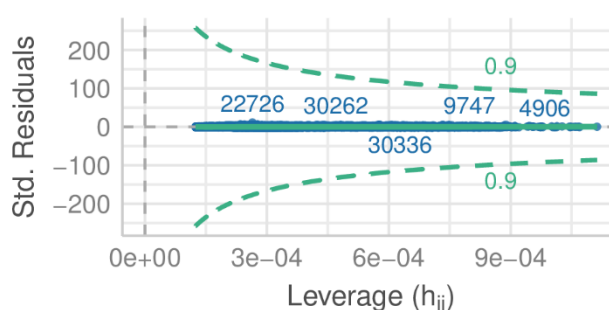
Homogeneity of Variance

Reference line should be flat and horizontal



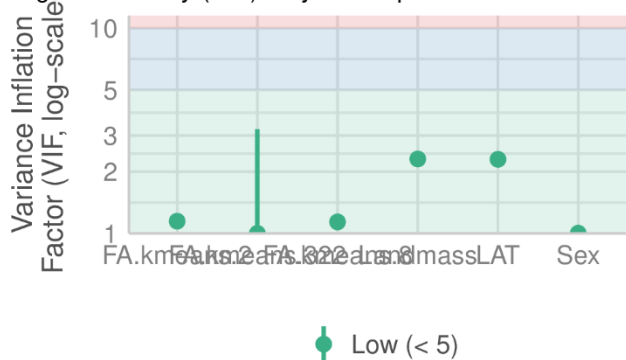
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5)

Normality of Residuals

Dots should fall along the line

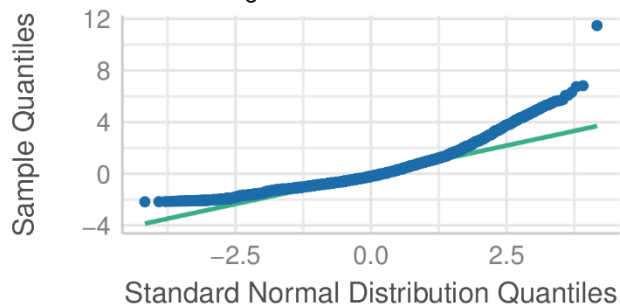


Figure S90. Linear model assumptions check for baseline linear model of FA3 response variable, from GLM data set

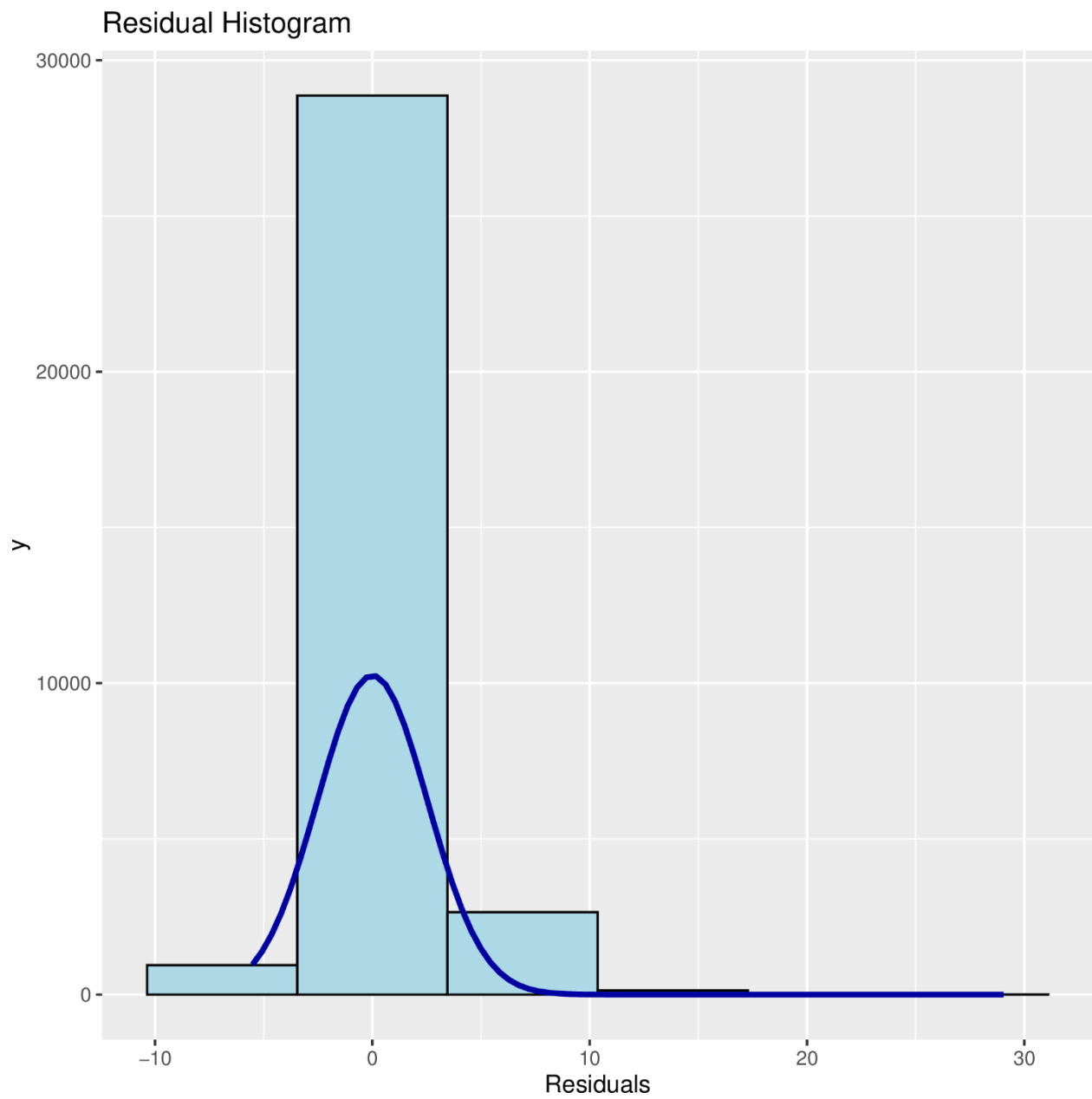
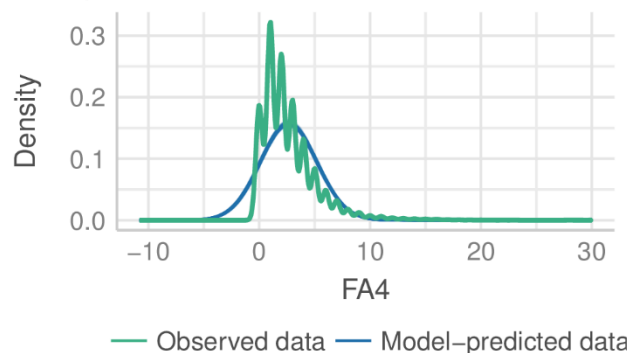


Figure S91. Histogram of residuals for baseline linear model of FA3 response variable, from GLM data set

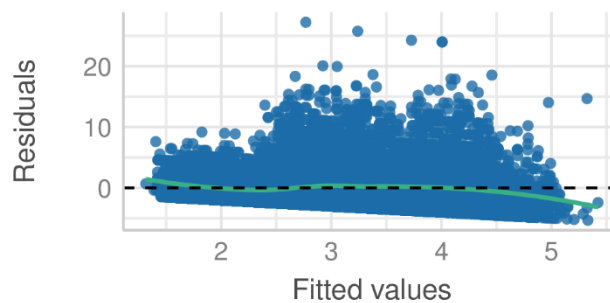
Posterior Predictive Check

Model-predicted lines should resemble observed data



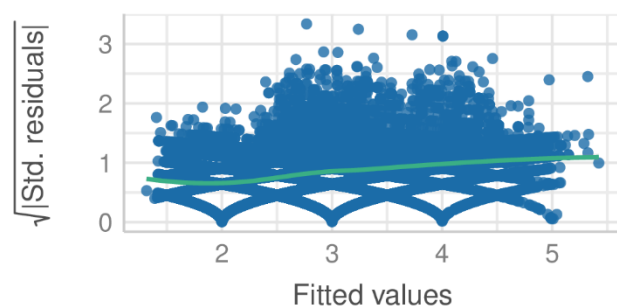
Linearity

Reference line should be flat and horizontal



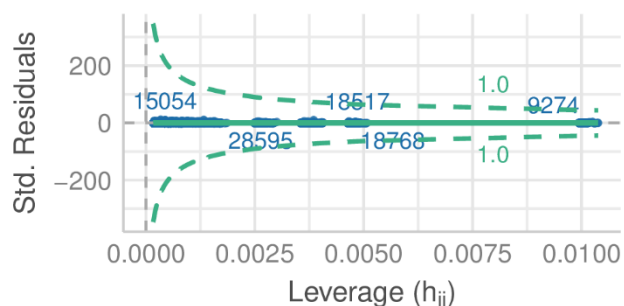
Homogeneity of Variance

Reference line should be flat and horizontal



Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



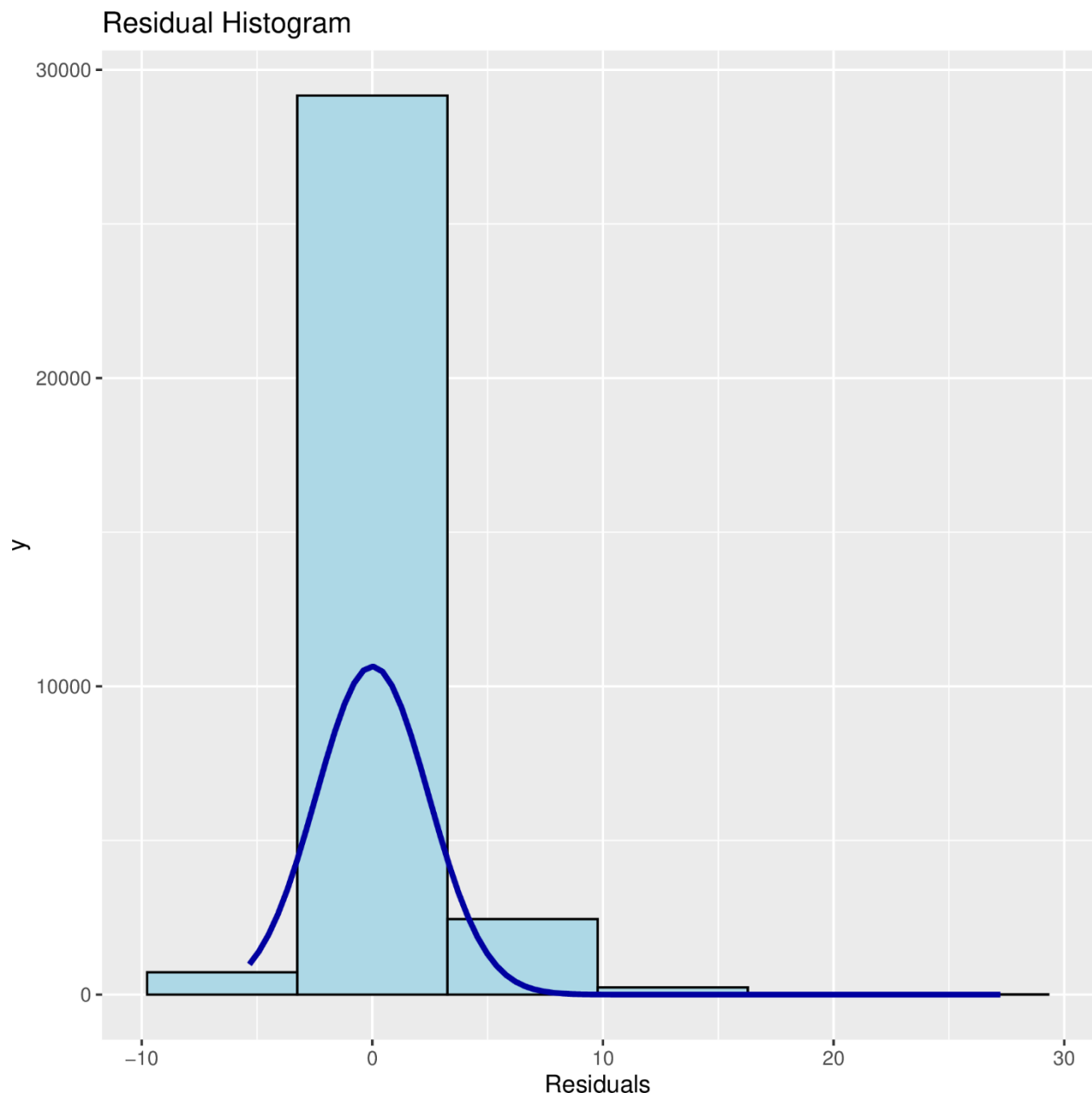
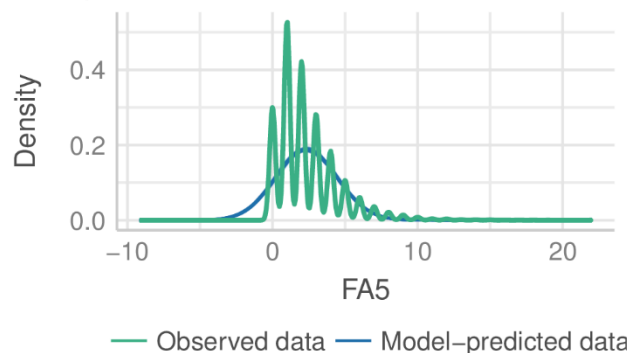


Figure S93. Histogram of residuals for baseline linear model of FA4 response variable, from GLM data set

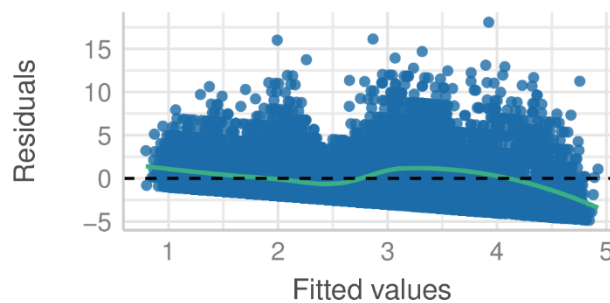
Posterior Predictive Check

Model-predicted lines should resemble observed data



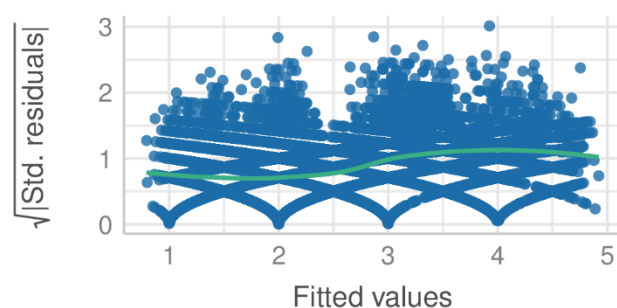
Linearity

Reference line should be flat and horizontal



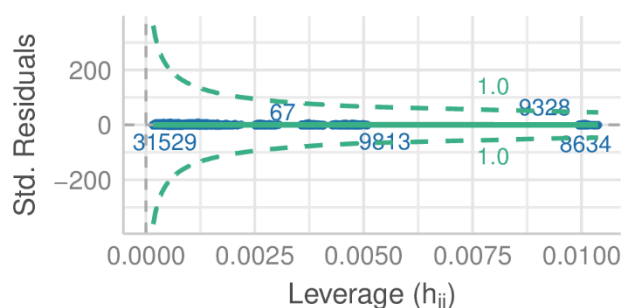
Homogeneity of Variance

Reference line should be flat and horizontal



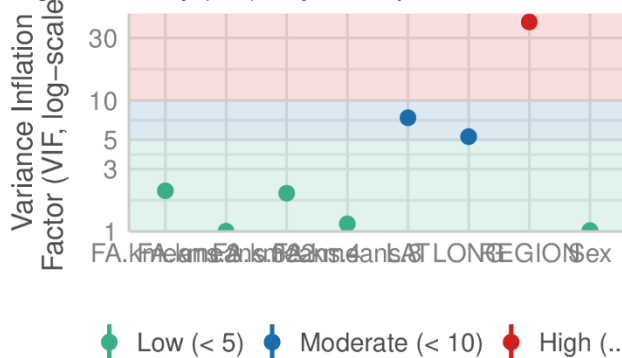
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

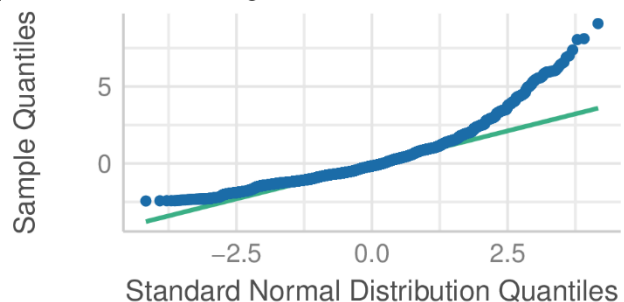


Figure S94. Linear model assumptions check for baseline linear model of FA5 response variable, from GLM data set

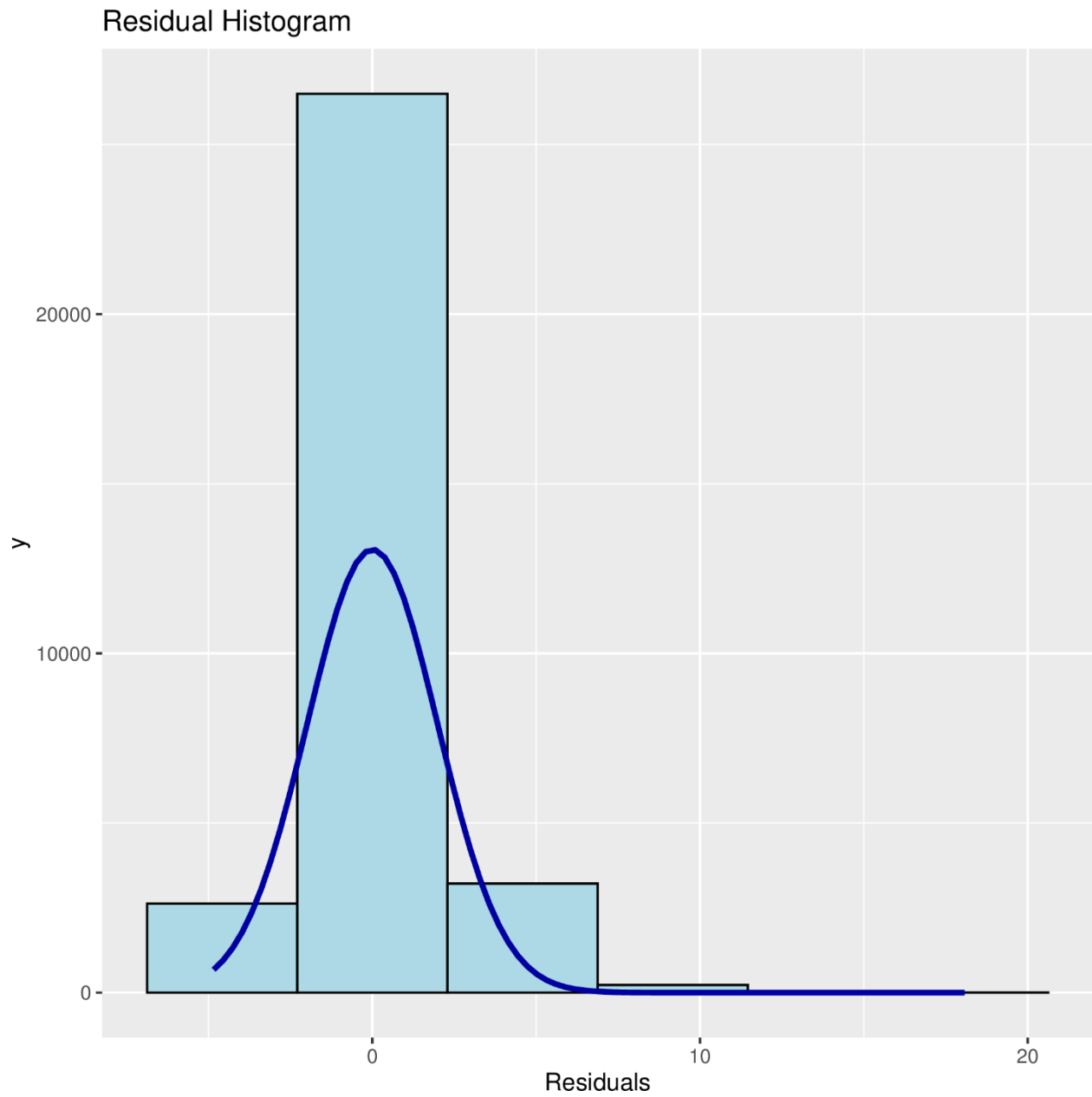
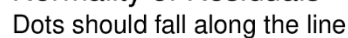
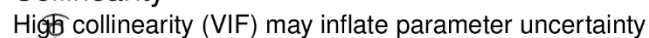
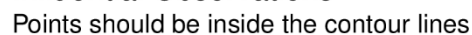
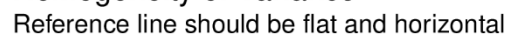
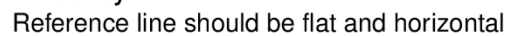
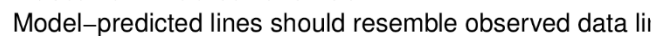


Figure S95. Histogram of residuals for baseline linear model of FA5 response variable, from GLM data set



response variable, from GLM data set

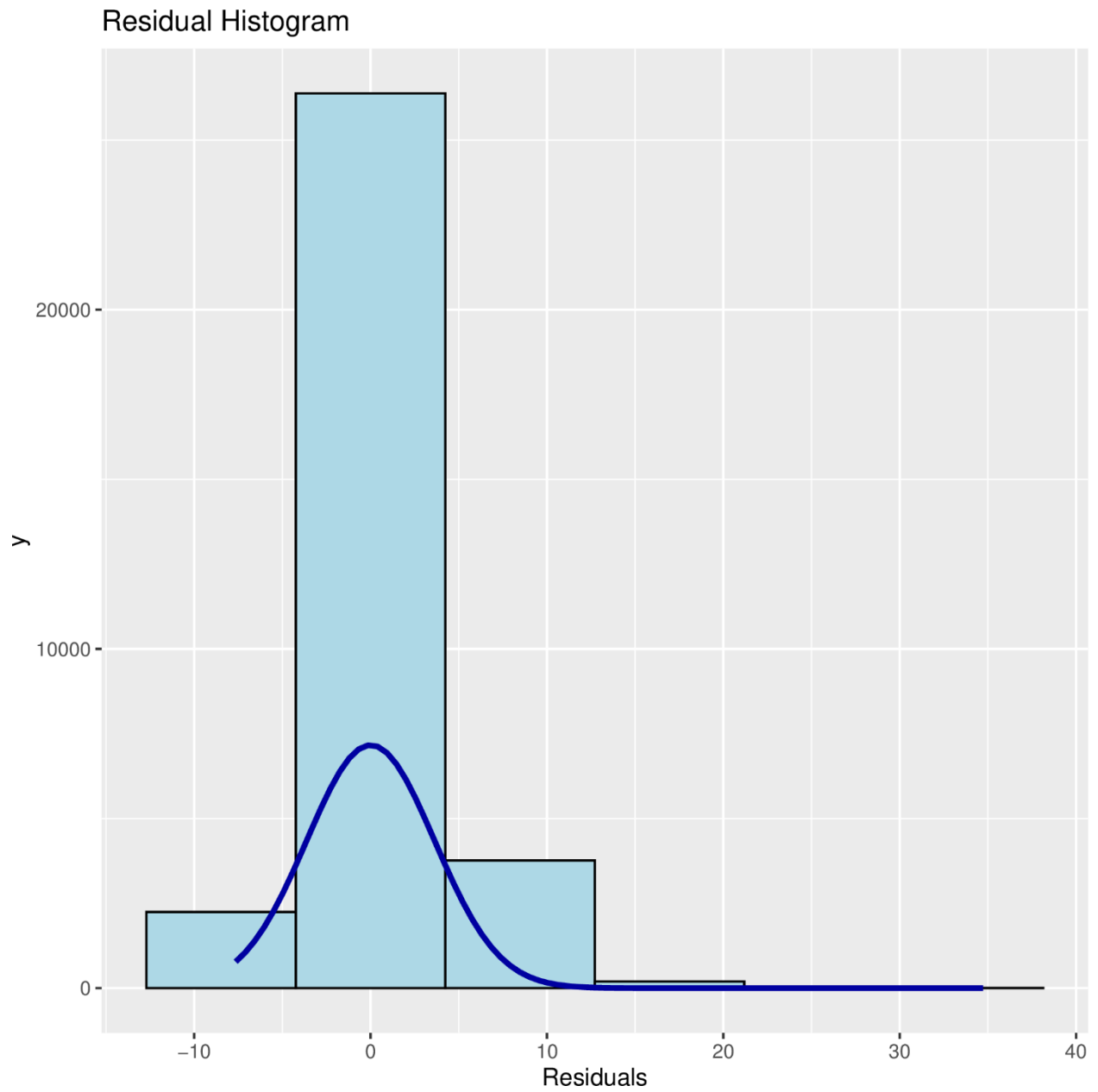


Figure S97. Histogram of residuals for baseline linear model of Index.Asymmetry response variable, from GLM data set

Model-predicted lines should resemble observed data lines



Reference line should be flat and horizontal

 $\sqrt{|\text{Std. residuals}|}$

Std. Residuals

Variance Inflation Factor (VIF, log-scale)

Sample Quantiles

● Low (< 5)

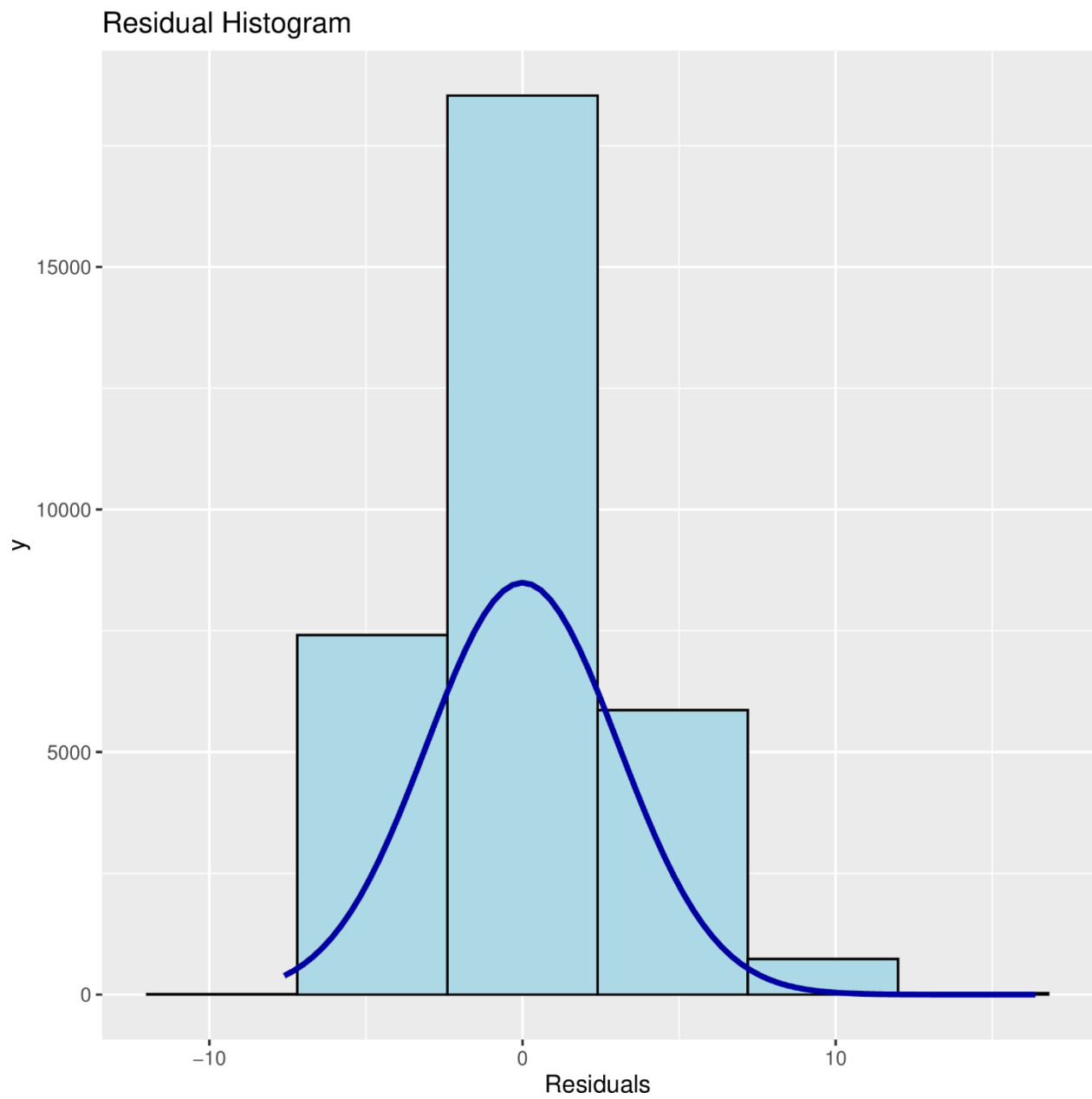
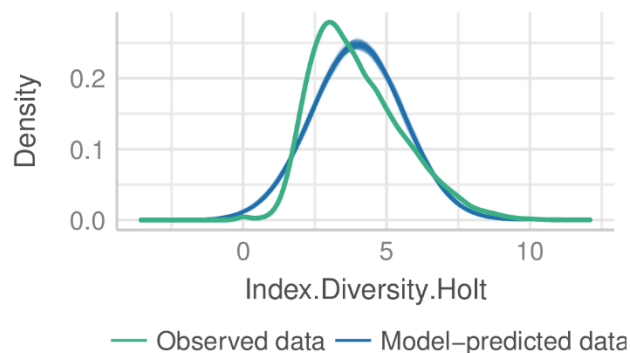


Figure S99. Histogram of residuals for baseline linear model of Index.Diversity response variable, from GLM data set

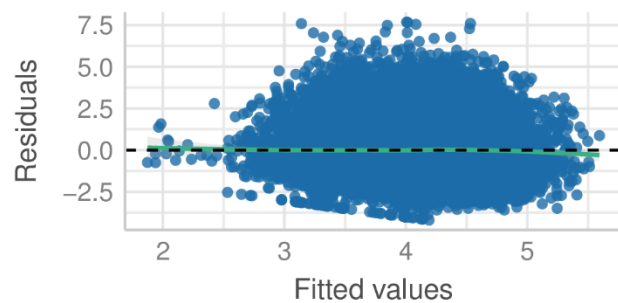
Posterior Predictive Check

Model-predicted lines should resemble observed data



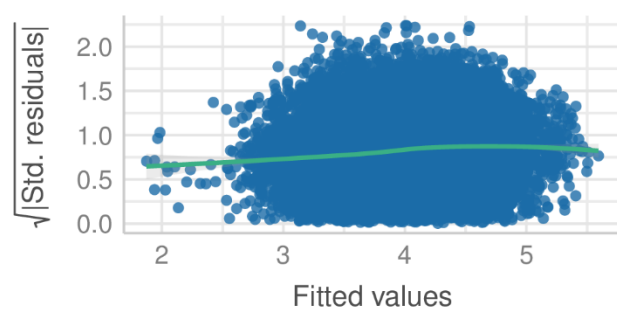
Linearity

Reference line should be flat and horizontal



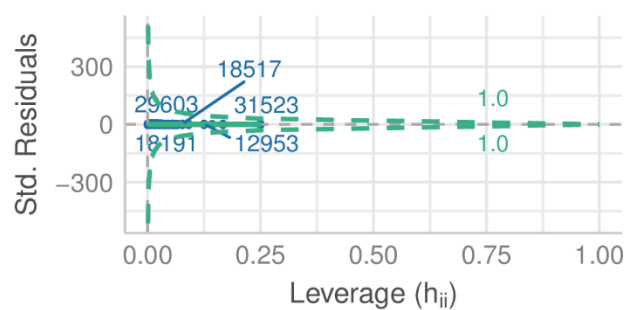
Homogeneity of Variance

Reference line should be flat and horizontal



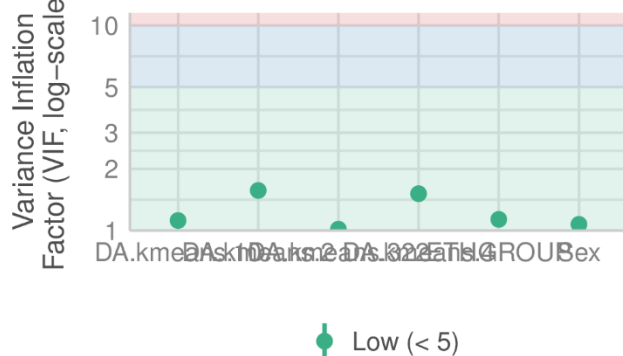
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

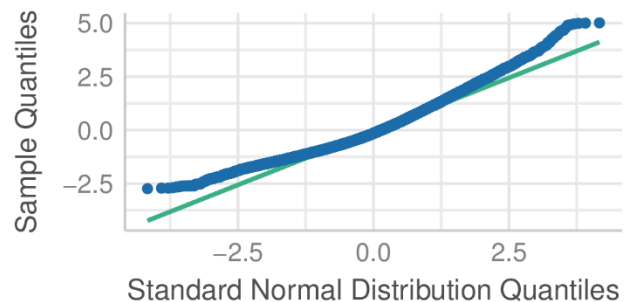


Figure S100. Linear model assumptions check for baseline linear model of Index.Diversity.Holt response variable, from GLM data set

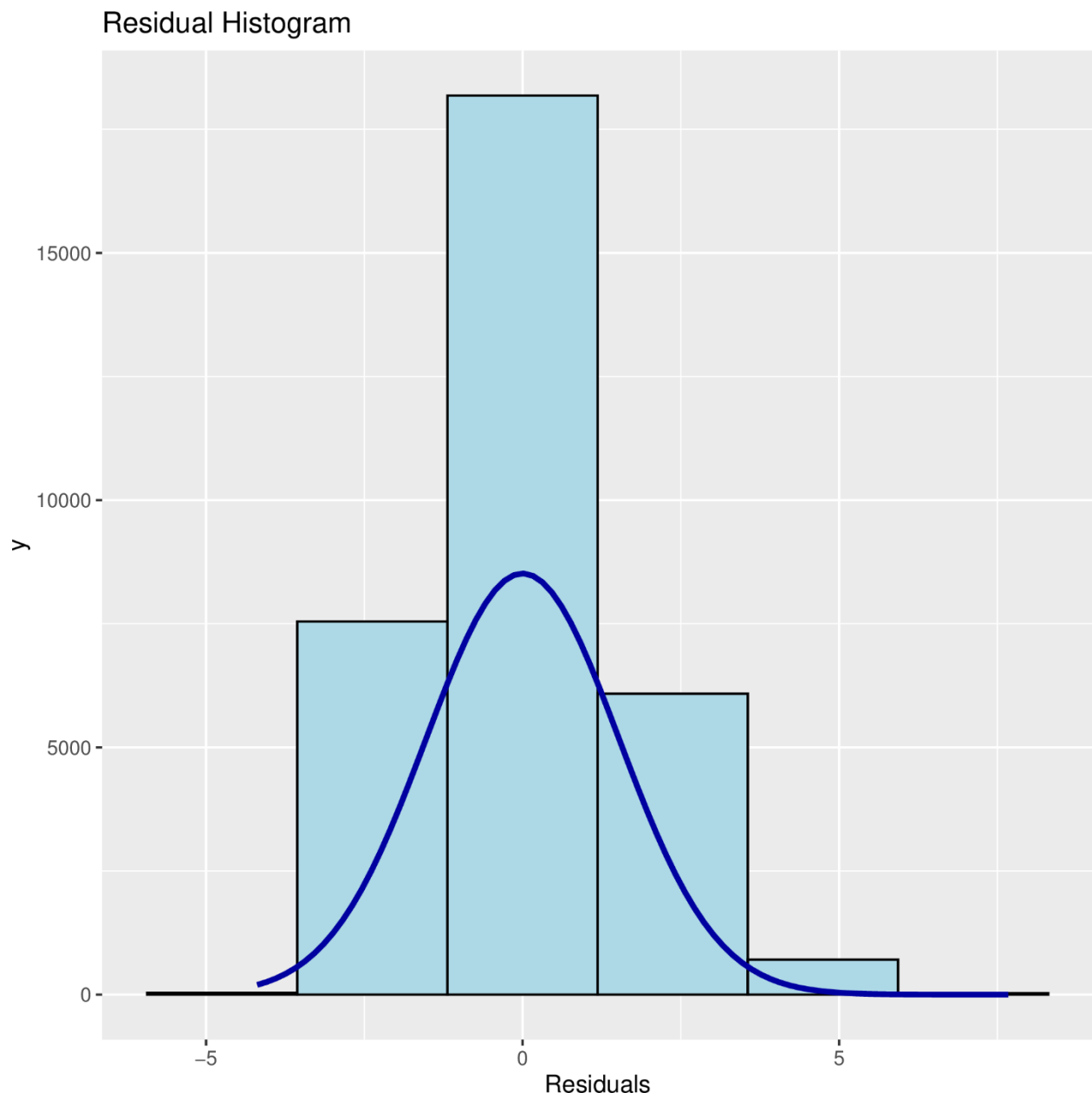
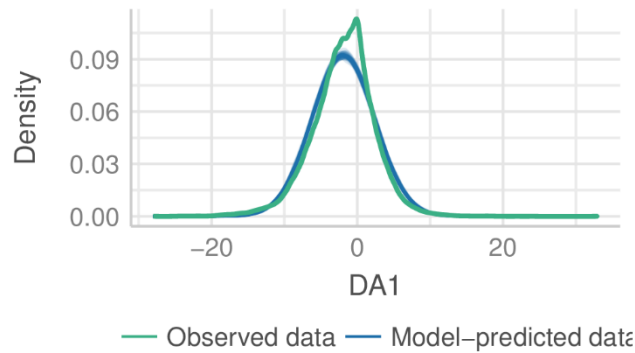


Figure S101. Histogram of residuals for baseline linear model of Index.Diversity.Holt response variable, from GLM data set

Appendix C.2.3 Supplementary figures for Baseline models for support vector regression

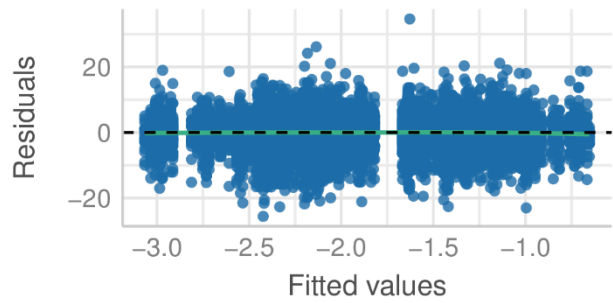
Posterior Predictive Check

Model-predicted lines should resemble observed data li



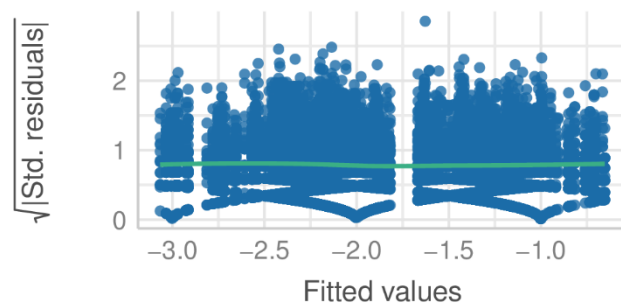
Linearity

Reference line should be flat and horizontal



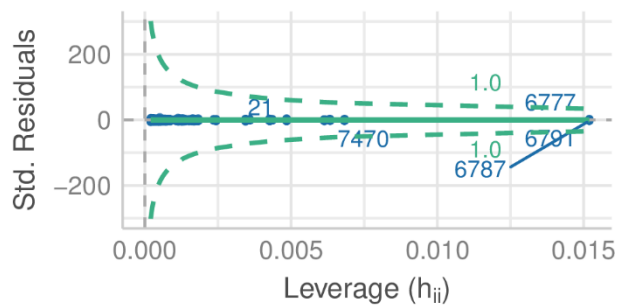
Homogeneity of Variance

Reference line should be flat and horizontal



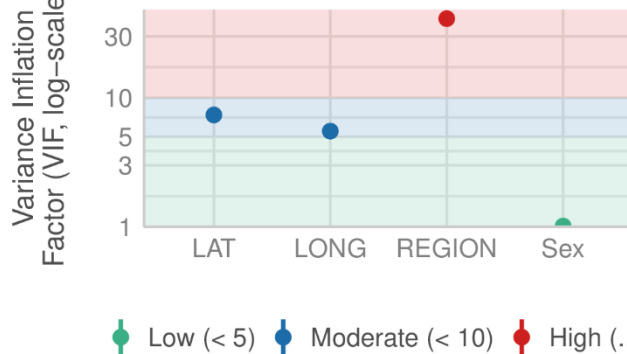
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

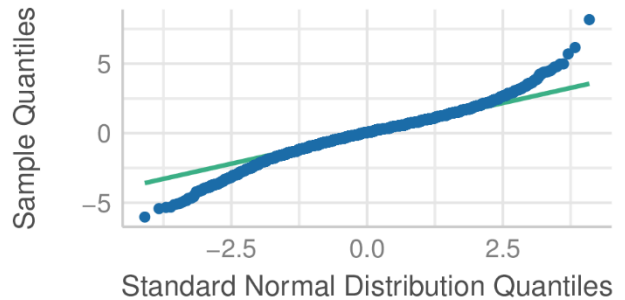


Figure S102. Linear model assumptions check for baseline linear model of DA1 response variable, from SVR data set

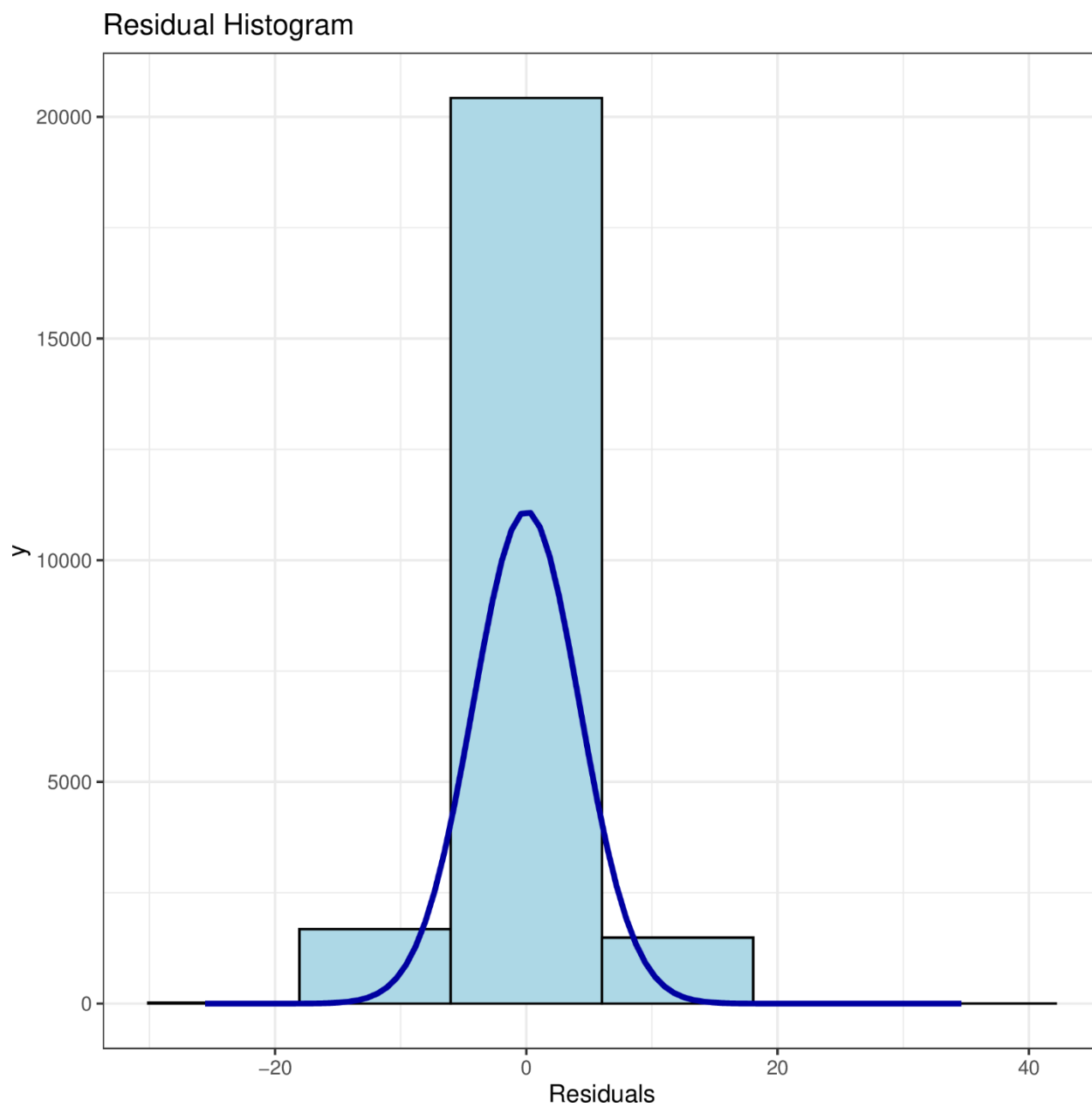
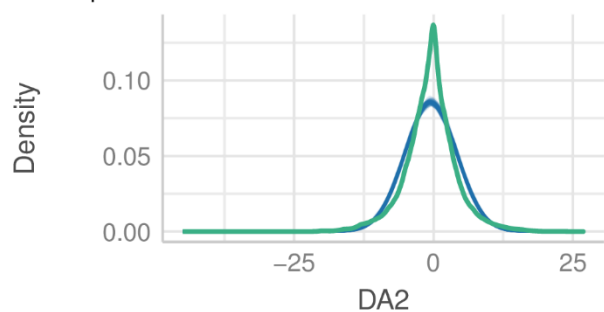


Figure S103. Histogram of residuals for baseline linear model of DA1 response variable, from SVR data set

Posterior Predictive Check

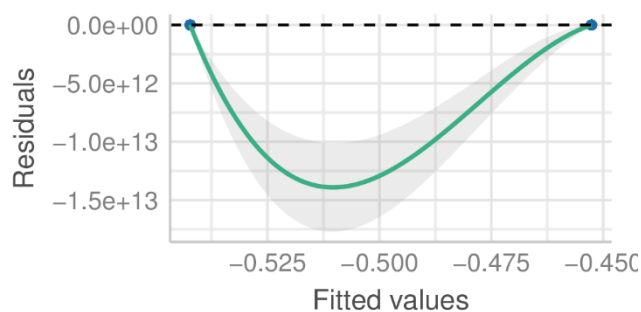
Model-predicted lines should resemble observed data



— Observed data — Model-predicted density

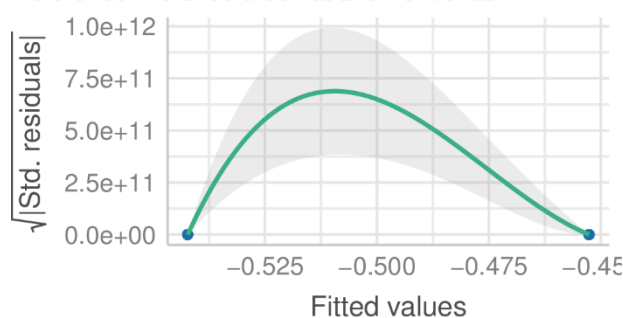
Linearity

Reference line should be flat and horizontal



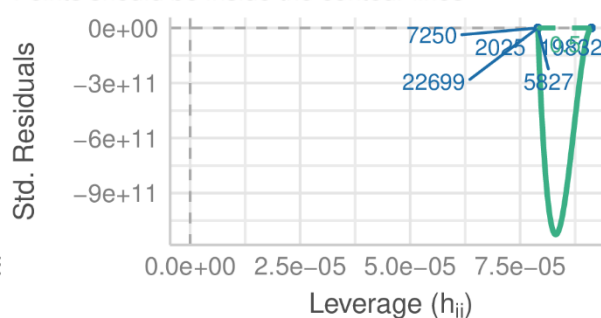
Homogeneity of Variance

Reference line should be flat and horizontal



Influential Observations

Points should be inside the contour lines



Normality of Residuals

Dots should fall along the line

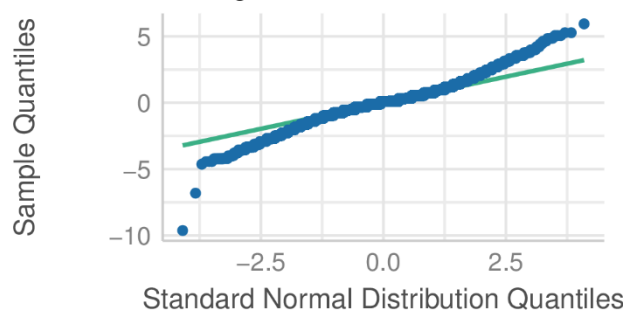


Figure S104. Linear model assumptions check for baseline linear model of DA2 response variable, from SVR data set

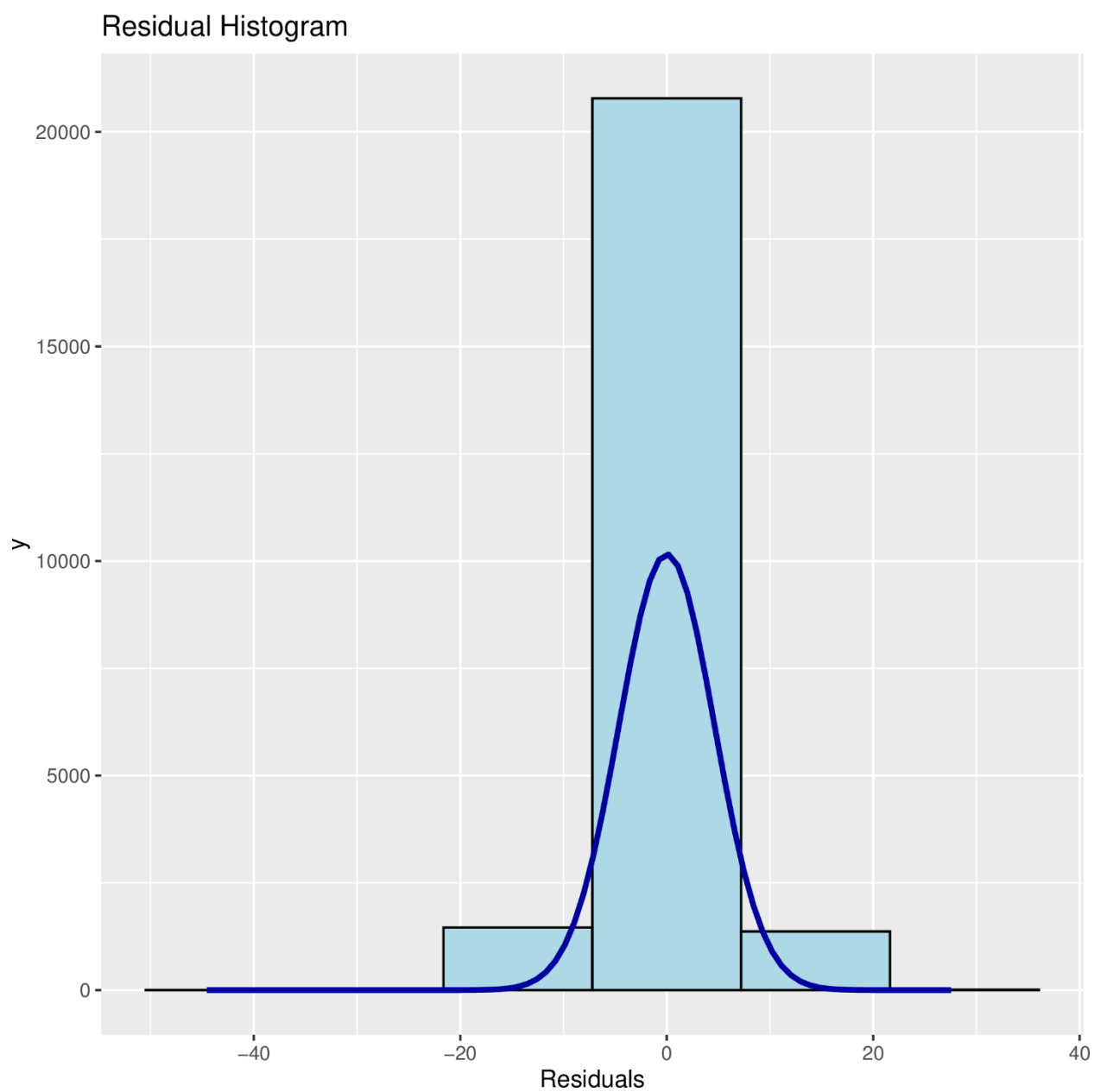
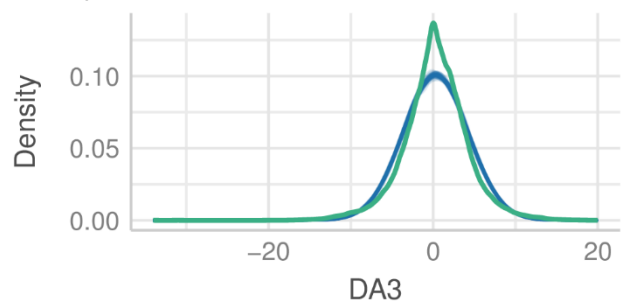


Figure S105. Histogram of residuals for baseline linear model of DA2 response variable, from SVR data set

Posterior Predictive Check

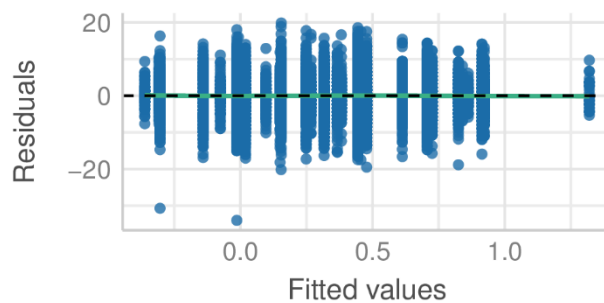
Model-predicted lines should resemble observed data li



— Observed data — Model-predicted data

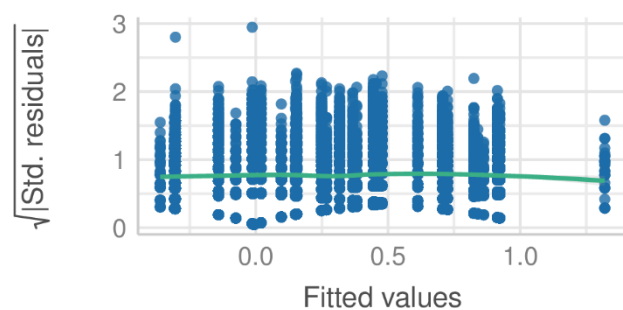
Linearity

Reference line should be flat and horizontal



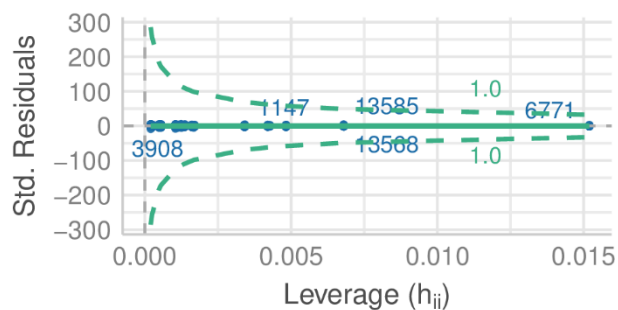
Homogeneity of Variance

Reference line should be flat and horizontal



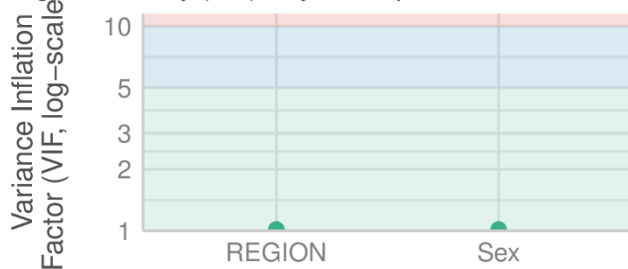
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5)

Normality of Residuals

Dots should fall along the line

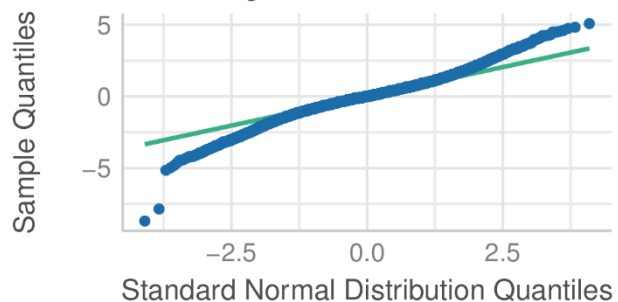


Figure S106. Linear model assumptions check for baseline linear model of DA3 response variable, from SVR data set

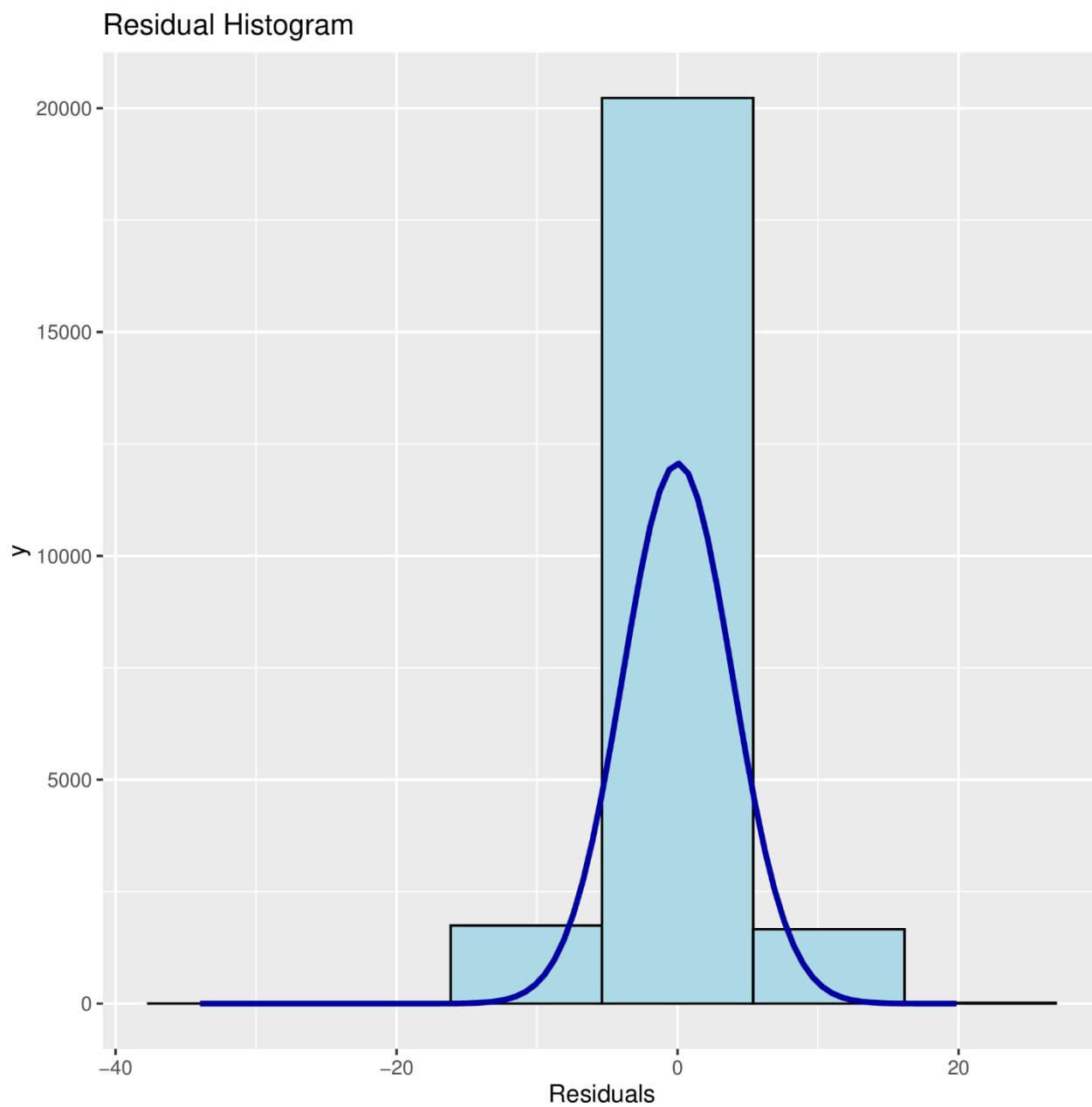
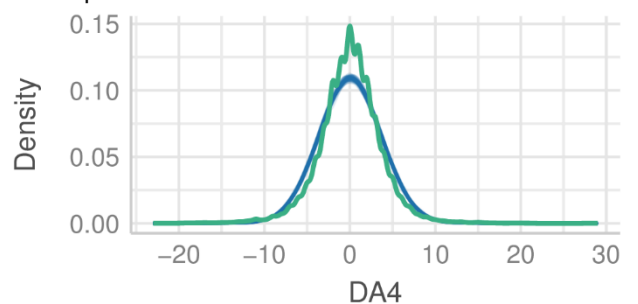


Figure S107. Histogram of residuals for baseline linear model of DA3 response variable, from SVR data set

Posterior Predictive Check

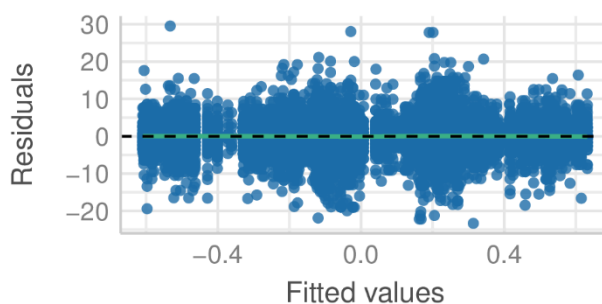
Model-predicted lines should resemble observed data



— Observed data — Model-predicted data

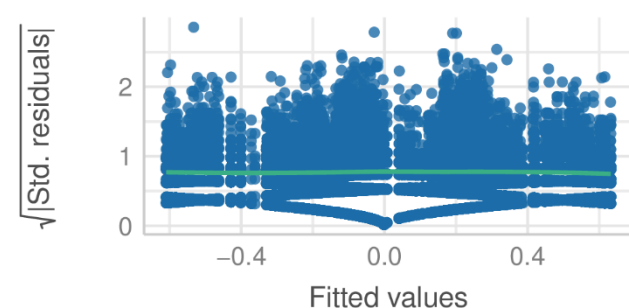
Linearity

Reference line should be flat and horizontal



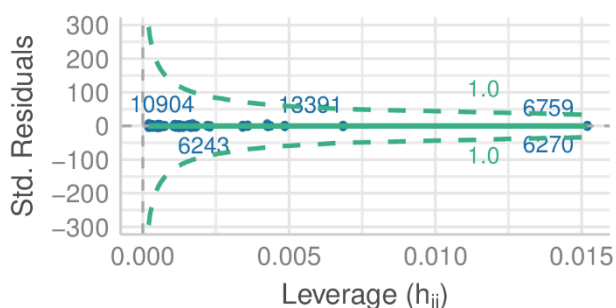
Homogeneity of Variance

Reference line should be flat and horizontal



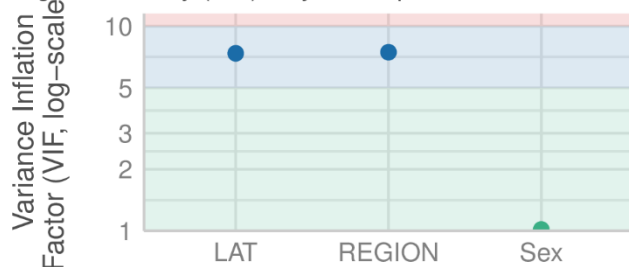
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5) ● Moderate (< 10)

Normality of Residuals

Dots should fall along the line

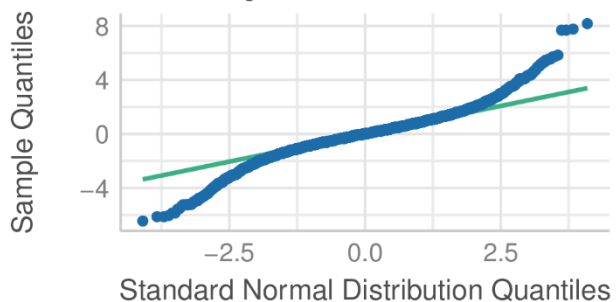


Figure S108. Linear model assumptions check for baseline linear model of DA4 response variable, from SVR data set

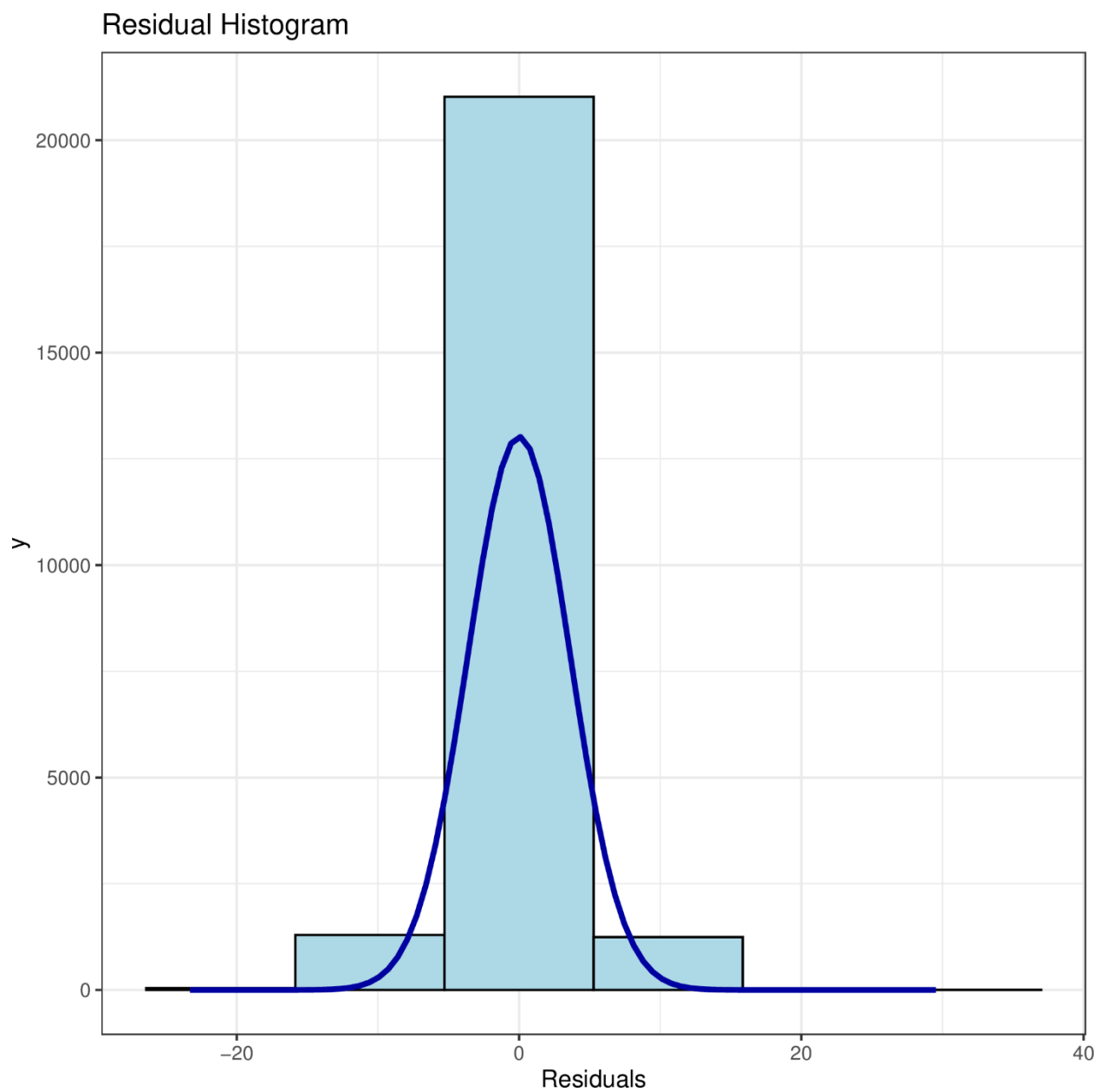
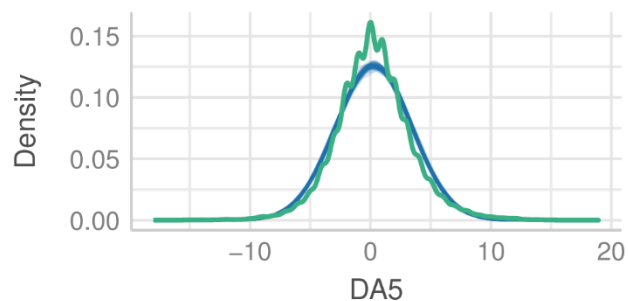


Figure S109. Histogram of residuals for baseline linear model of DA4 response variable, from SVR data set

Posterior Predictive Check

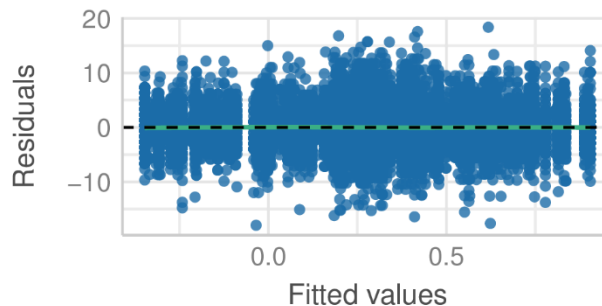
Model-predicted lines should resemble observed data



— Observed data — Model-predicted data

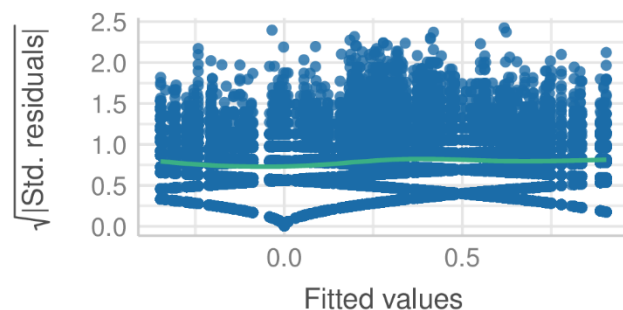
Linearity

Reference line should be flat and horizontal



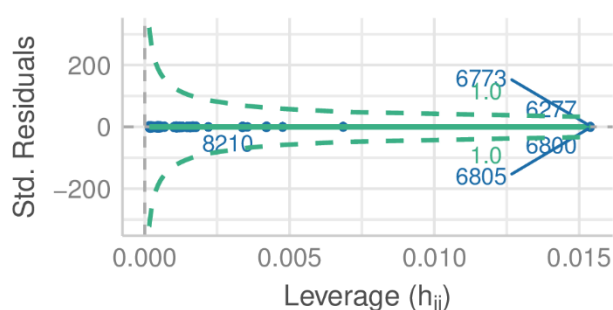
Homogeneity of Variance

Reference line should be flat and horizontal



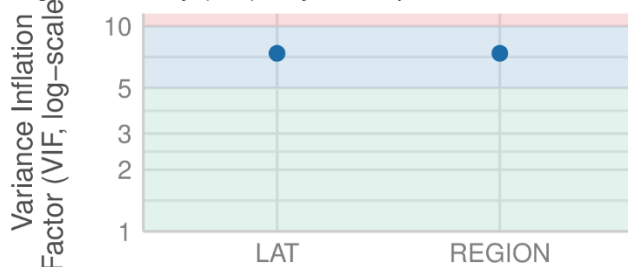
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



◆ Moderate (< 10)

Normality of Residuals

Dots should fall along the line

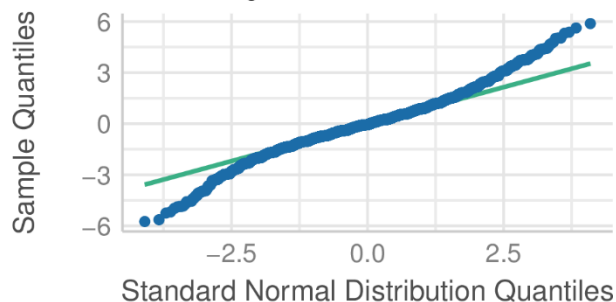


Figure S110. Linear model assumptions check for baseline linear model of DA5 response variable, from SVR data set

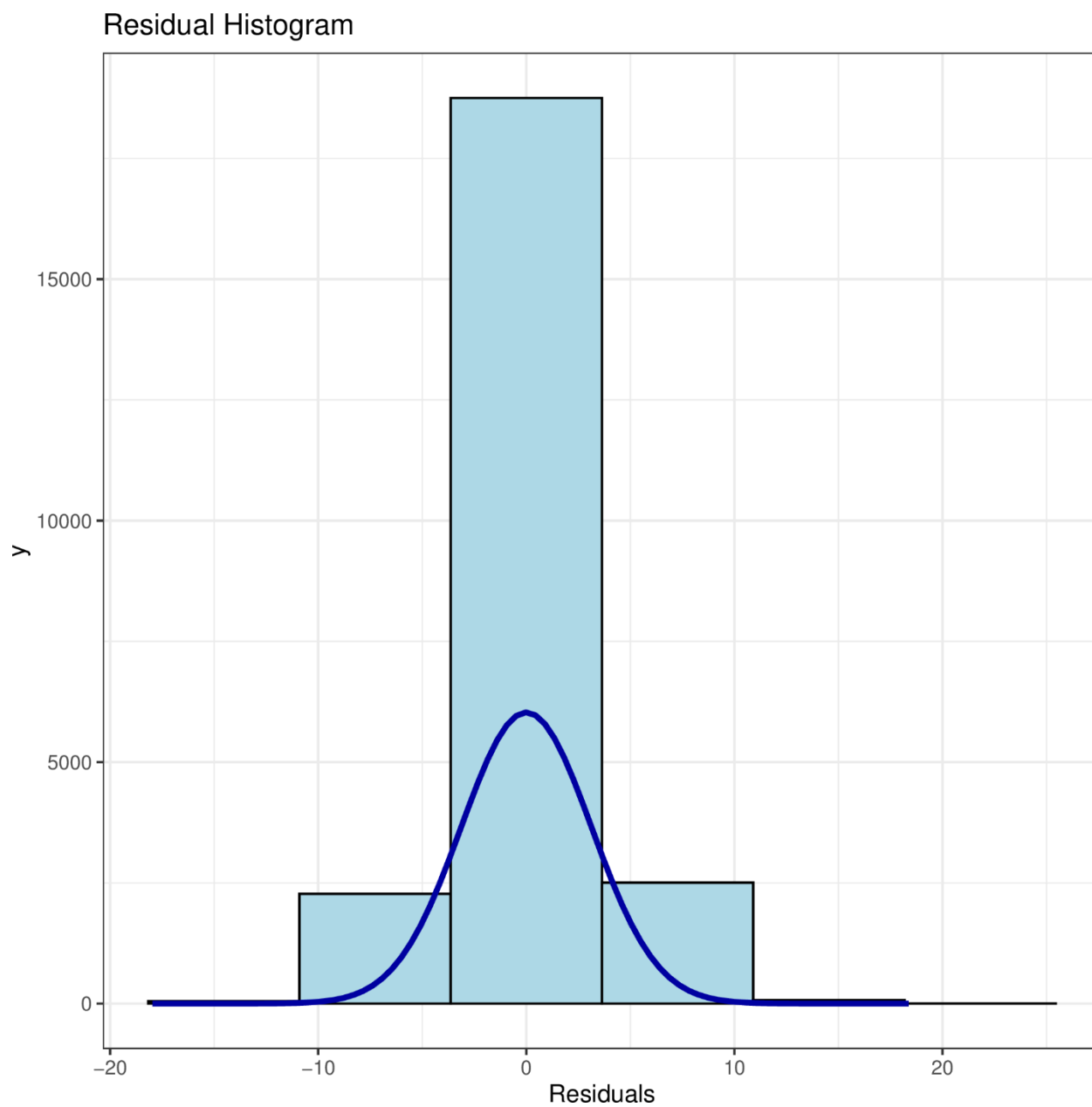
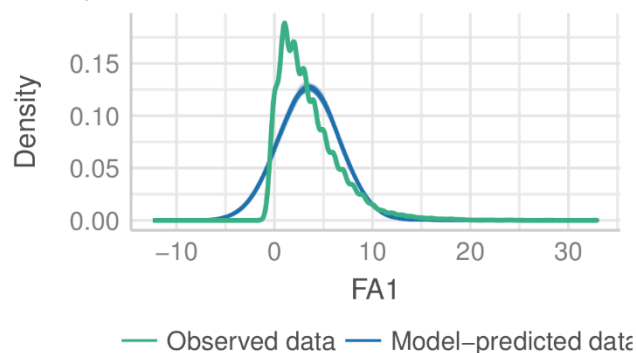


Figure S111. Histogram of residuals for baseline linear model of DA5 response variable, from SVR data set

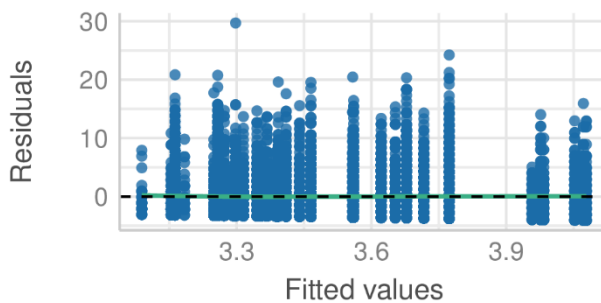
Posterior Predictive Check

Model-predicted lines should resemble observed data



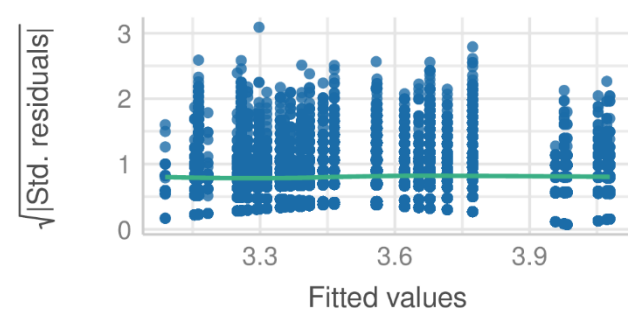
Linearity

Reference line should be flat and horizontal



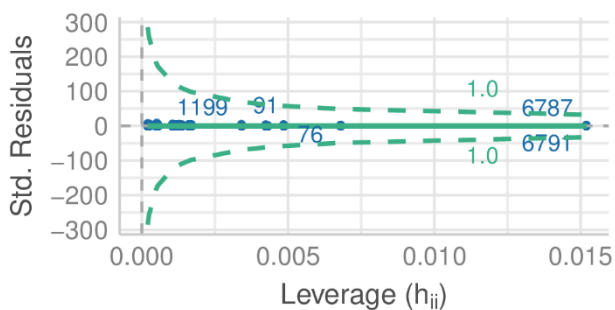
Homogeneity of Variance

Reference line should be flat and horizontal



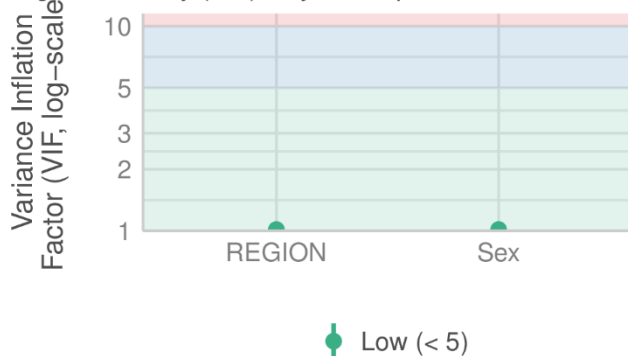
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

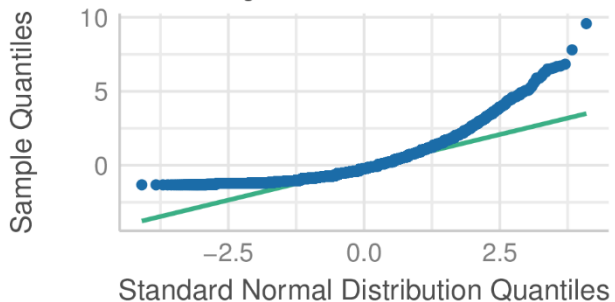


Figure S112. Linear model assumptions check for baseline linear model of FA1 response variable, from SVR data set

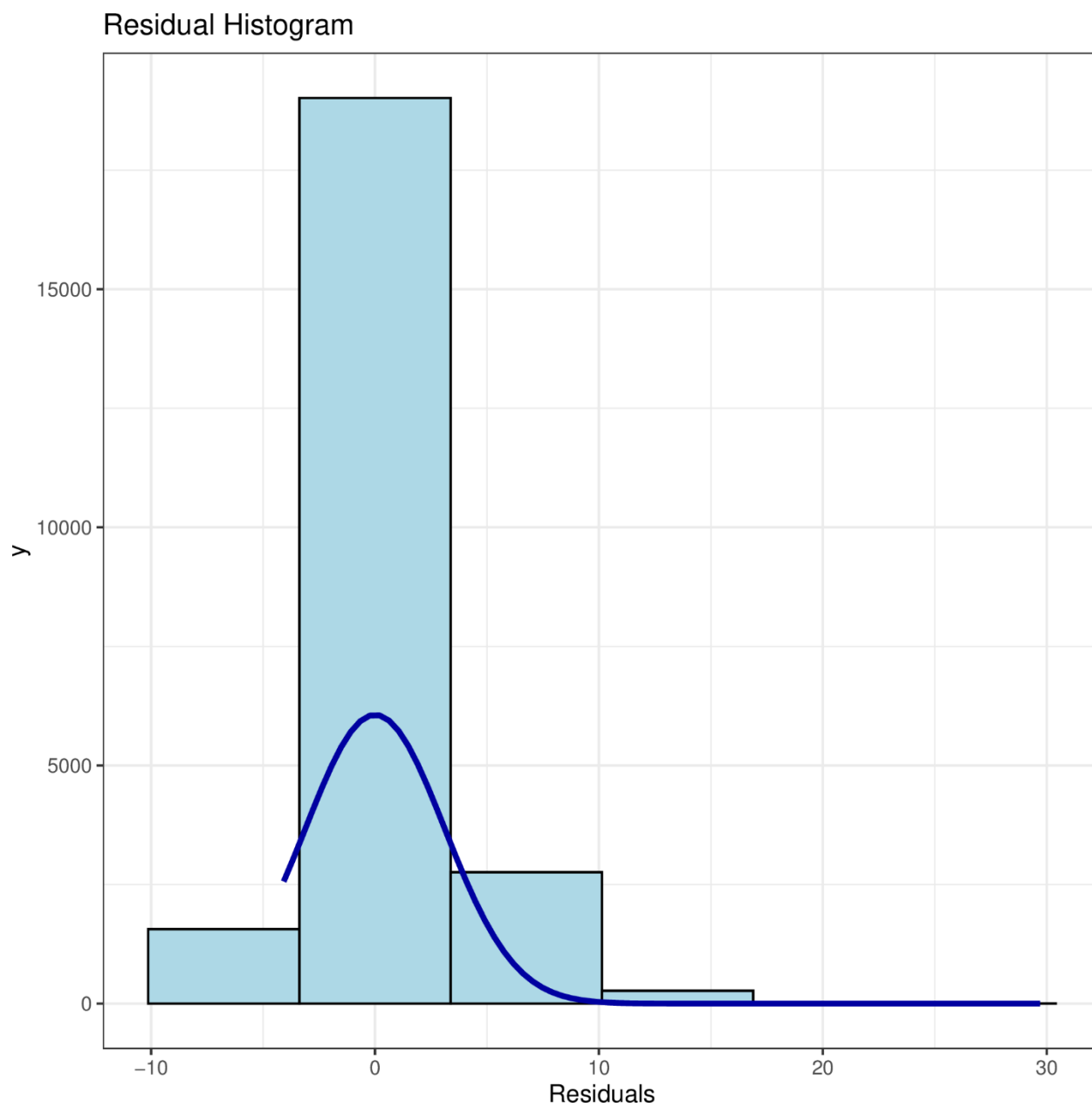
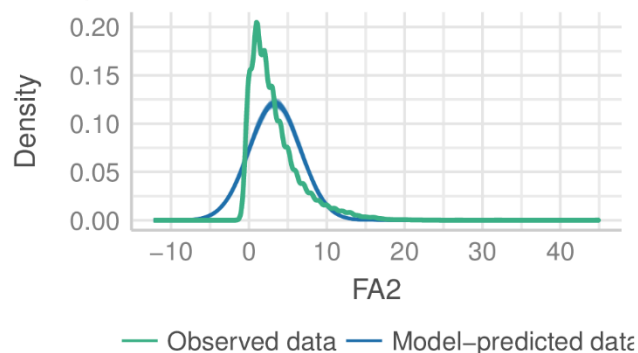


Figure S113. Histogram of residuals for baseline linear model of FA1 response variable, from SVR data set

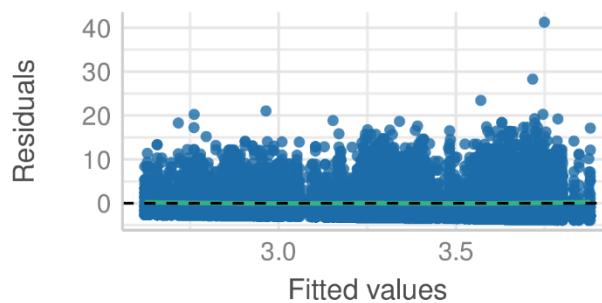
Posterior Predictive Check

Model-predicted lines should resemble observed data



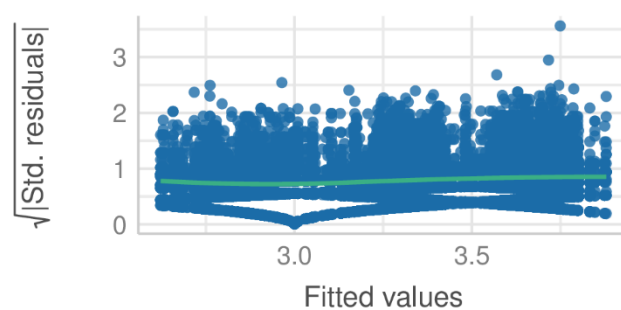
Linearity

Reference line should be flat and horizontal



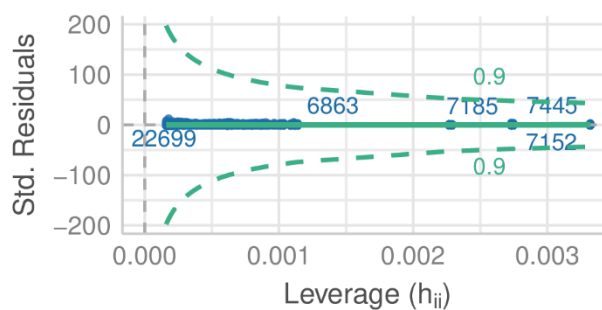
Homogeneity of Variance

Reference line should be flat and horizontal



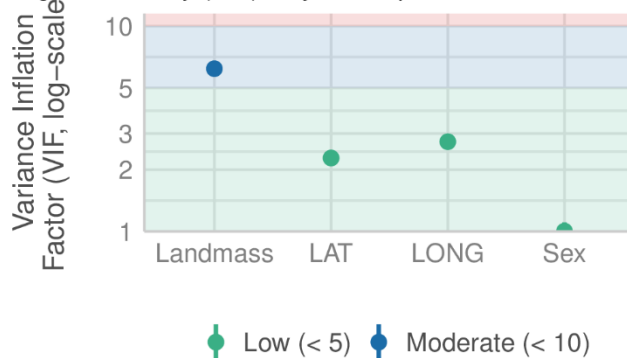
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

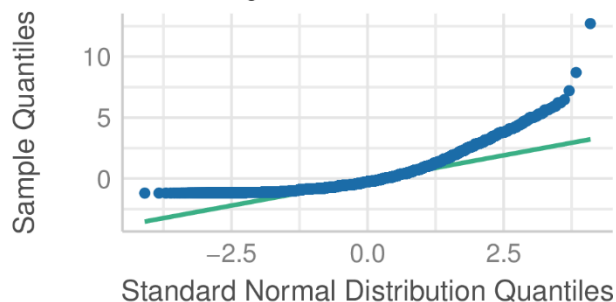


Figure S114. Linear model assumptions check for baseline linear model of FA2 response variable, from SVR data set

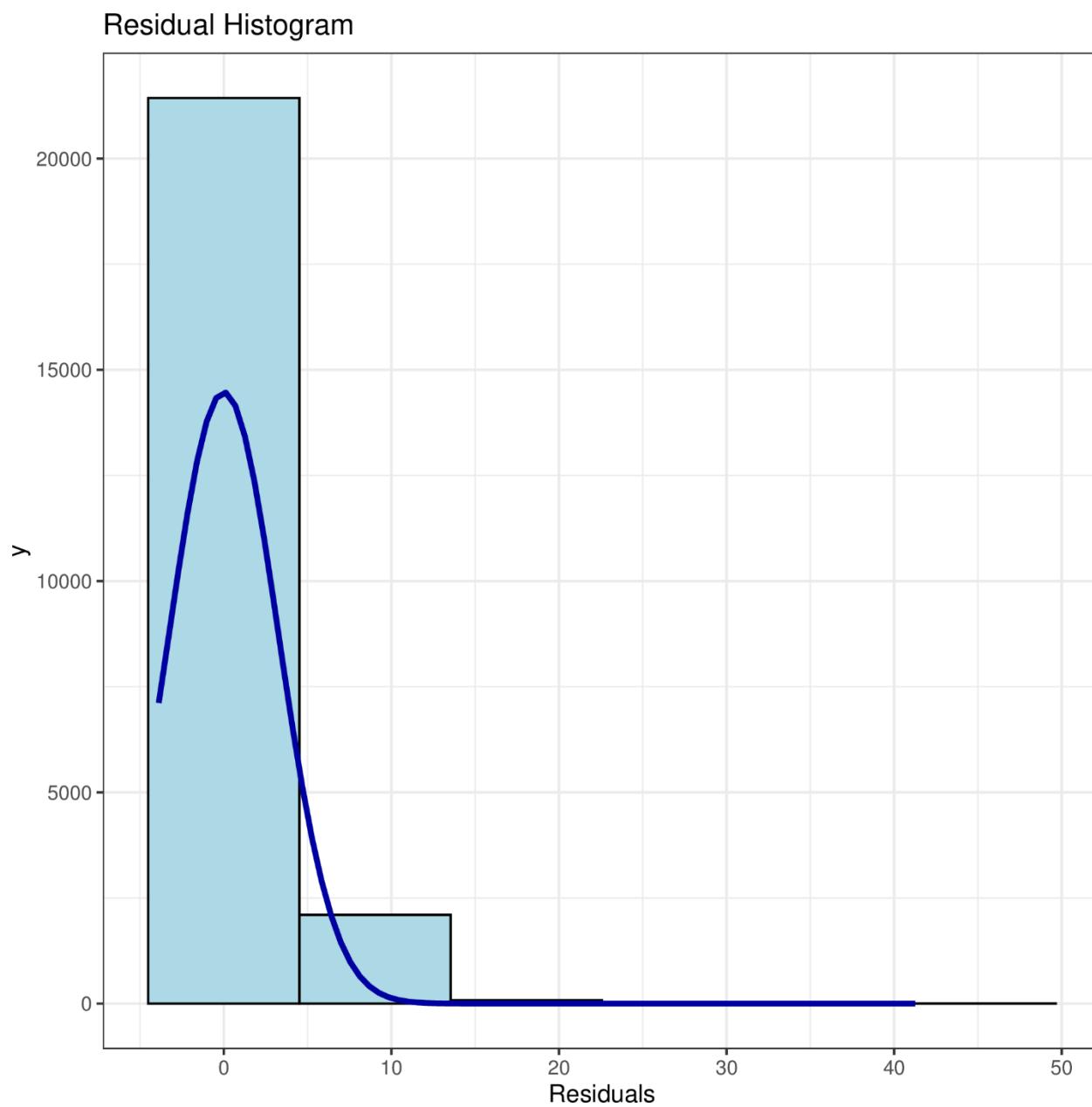
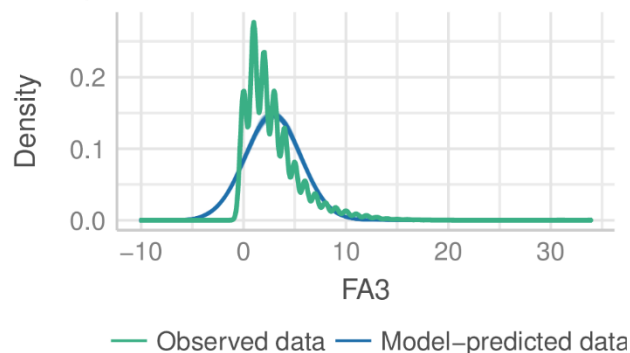


Figure S115. Histogram of residuals for baseline linear model of FA2 response variable, from SVR data set

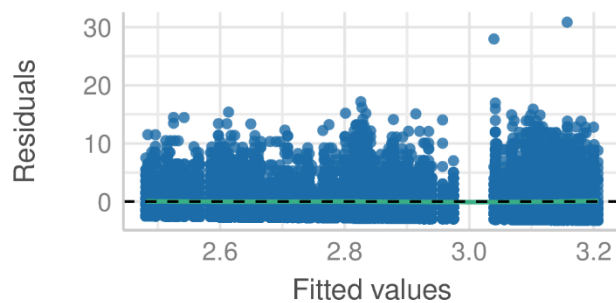
Posterior Predictive Check

Model-predicted lines should resemble observed data



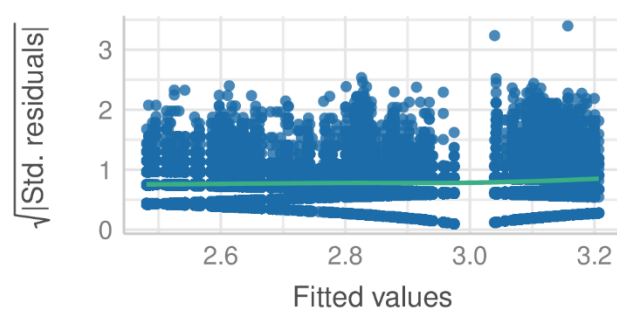
Linearity

Reference line should be flat and horizontal



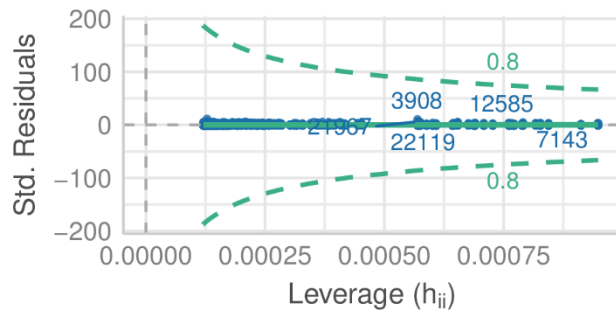
Homogeneity of Variance

Reference line should be flat and horizontal



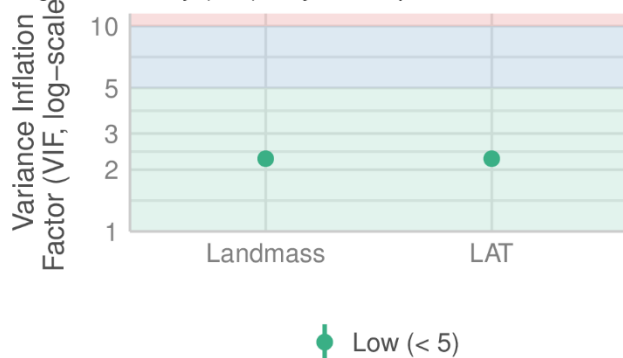
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

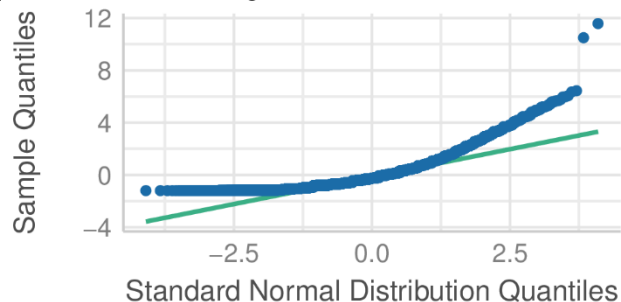


Figure S116. Linear model assumptions check for baseline linear model of FA3 response variable, from SVR data set

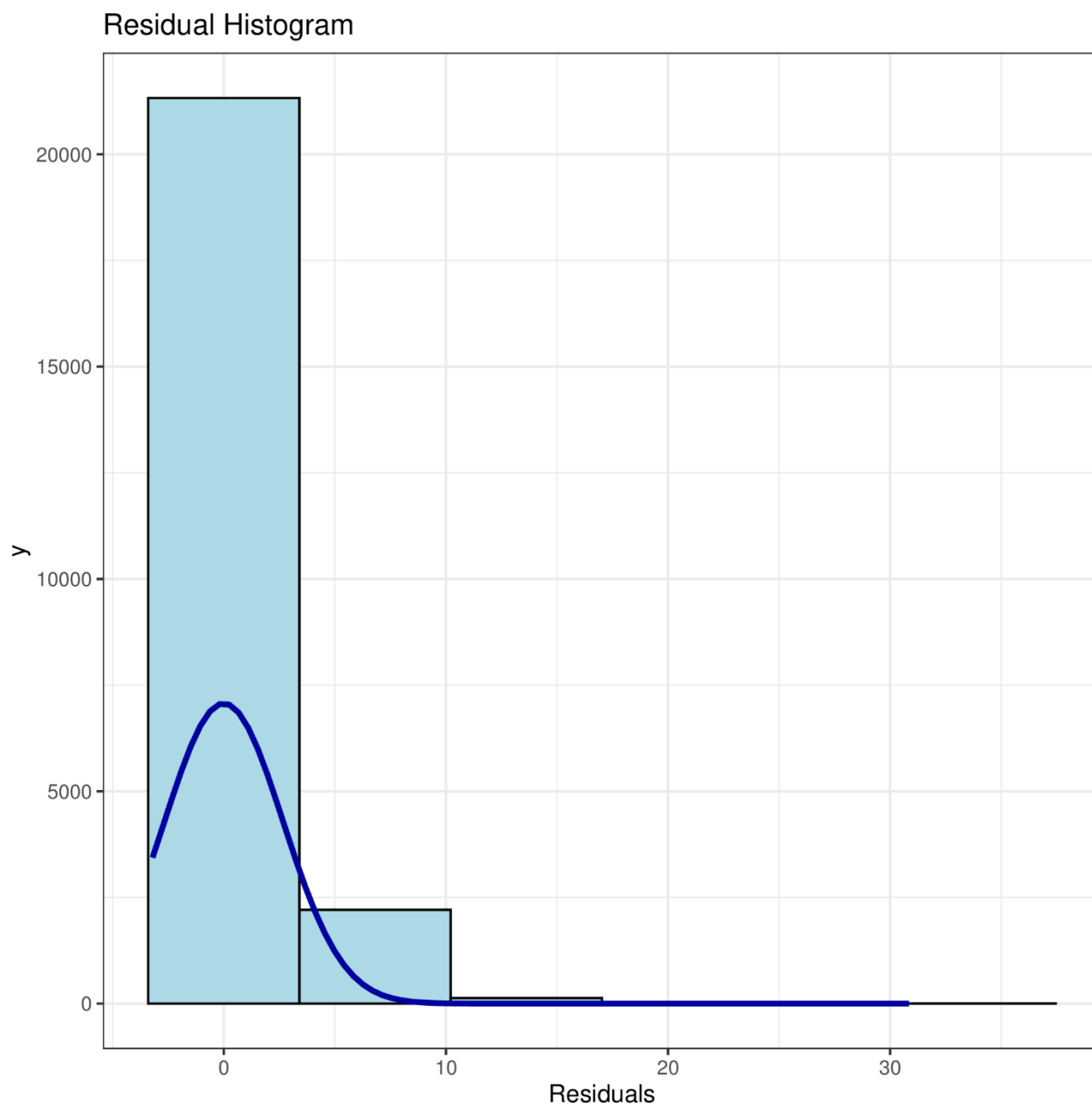
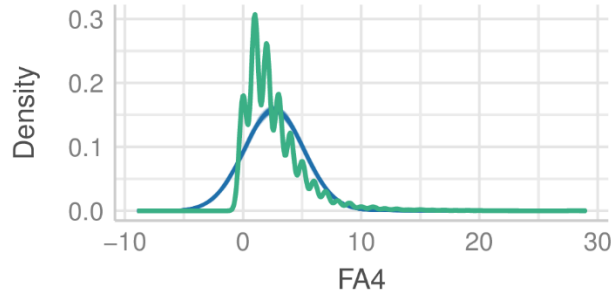


Figure S117. Histogram of residuals for baseline linear model of FA3 response variable, from SVR data set

Posterior Predictive Check

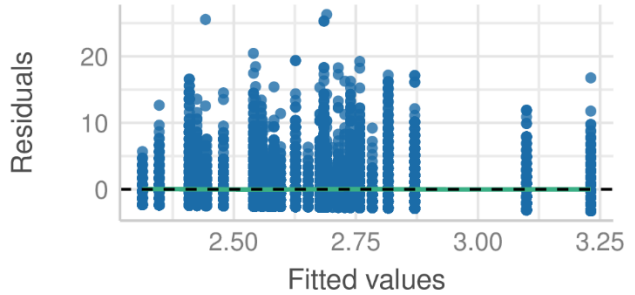
Model-predicted lines should resemble observed data



— Observed data — Model-predicted data

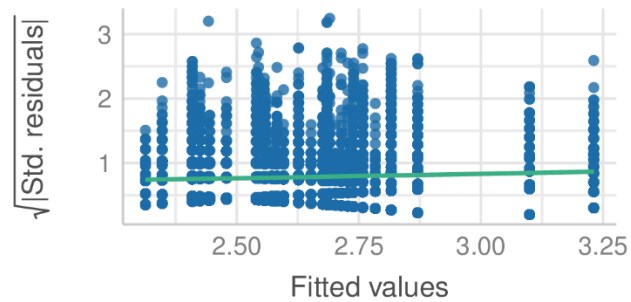
Linearity

Reference line should be flat and horizontal



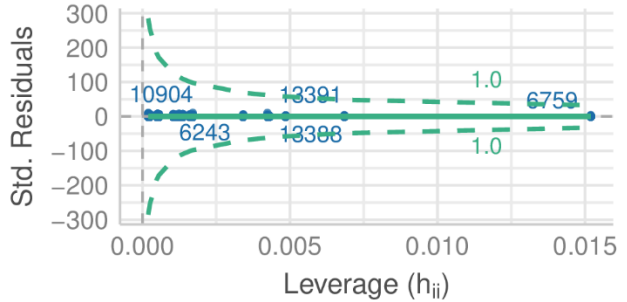
Homogeneity of Variance

Reference line should be flat and horizontal



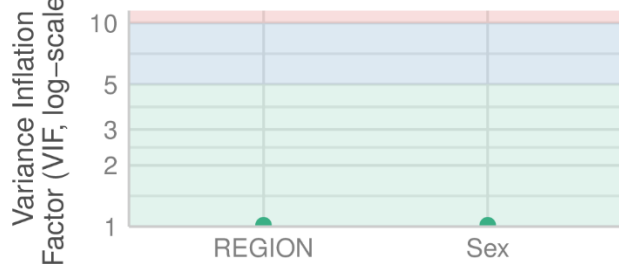
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5)

Normality of Residuals

Dots should fall along the line

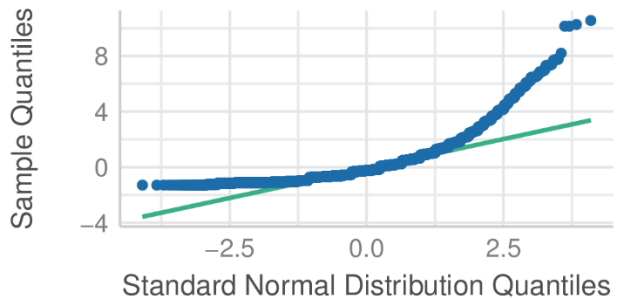


Figure S118. Linear model assumptions check for baseline linear model of FA4 response variable, from SVR data set

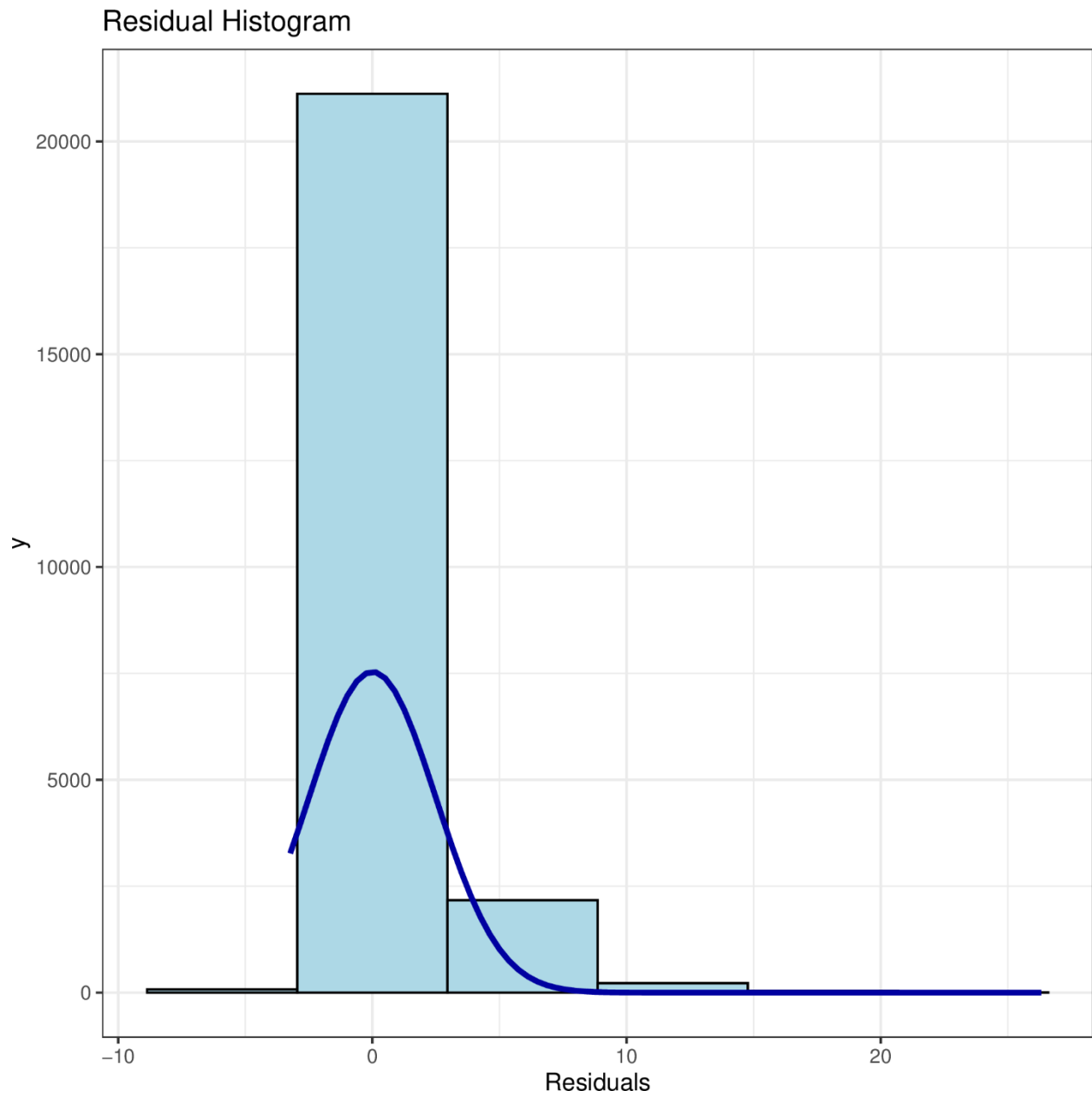
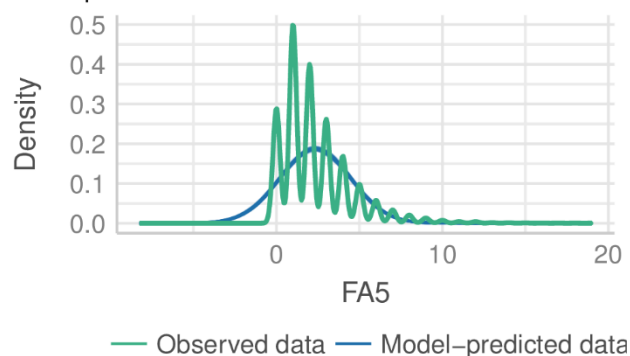


Figure S119. Histogram of residuals for baseline linear model of FA4 response variable, from SVR data set

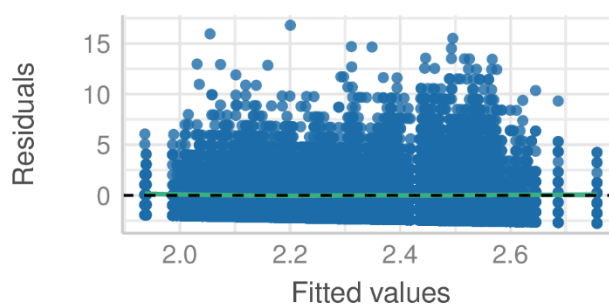
Posterior Predictive Check

Model-predicted lines should resemble observed data



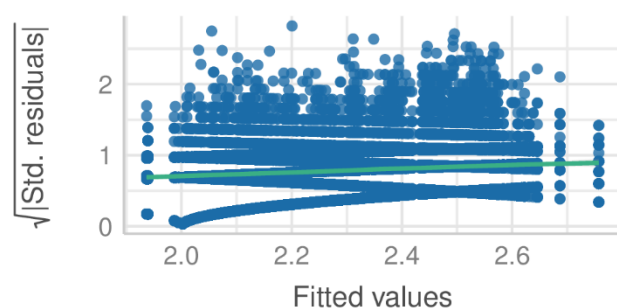
Linearity

Reference line should be flat and horizontal



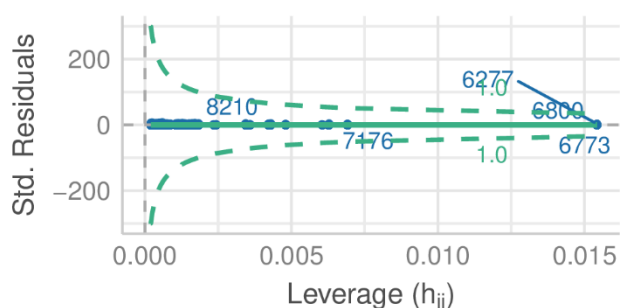
Homogeneity of Variance

Reference line should be flat and horizontal



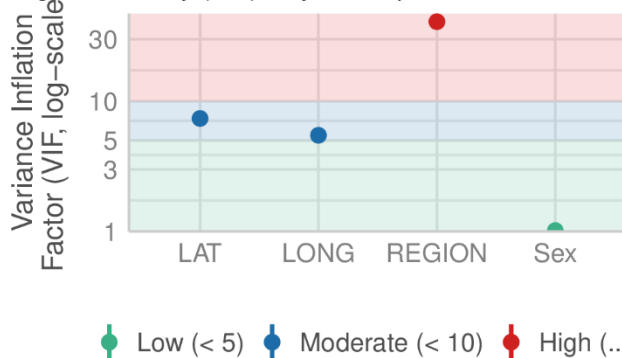
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

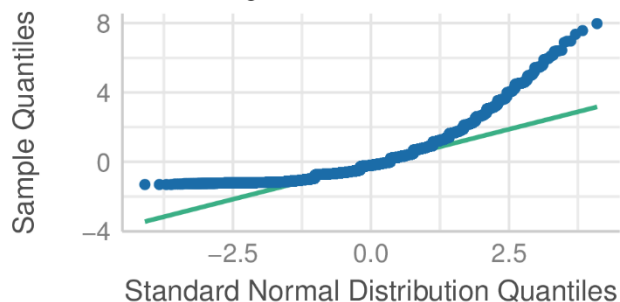


Figure S120. Linear model assumptions check for baseline linear model of FA5 response variable, from SVR data set

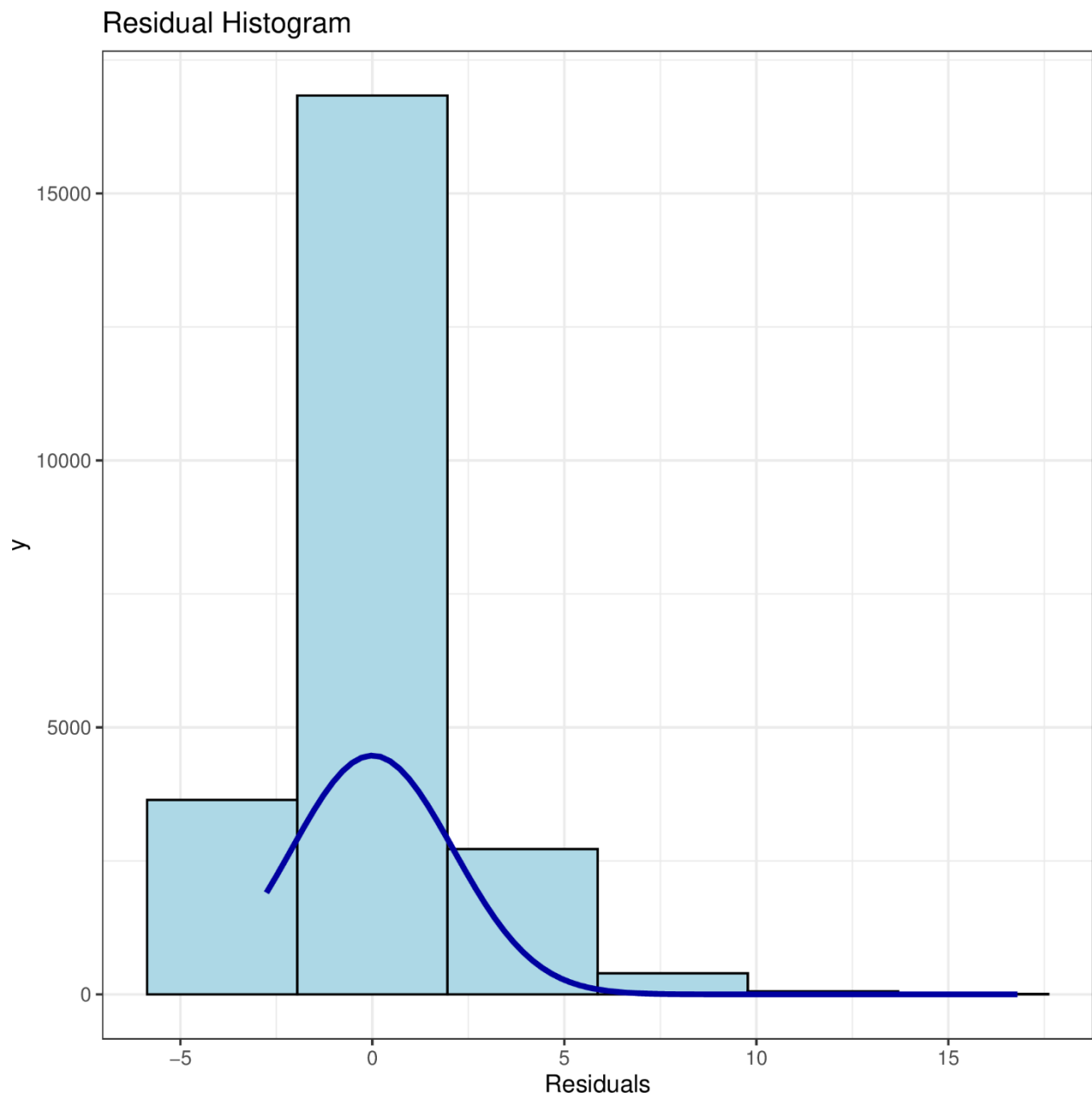
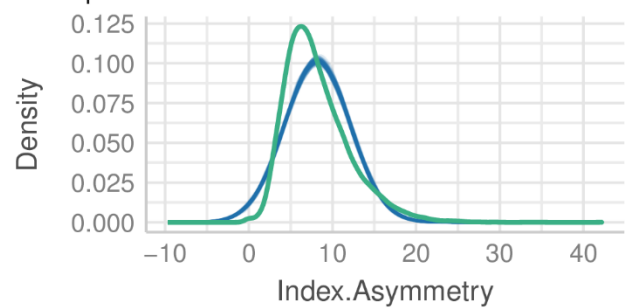


Figure S121. Histogram of residuals for baseline linear model of FA5 response variable, from SVR data set

Posterior Predictive Check

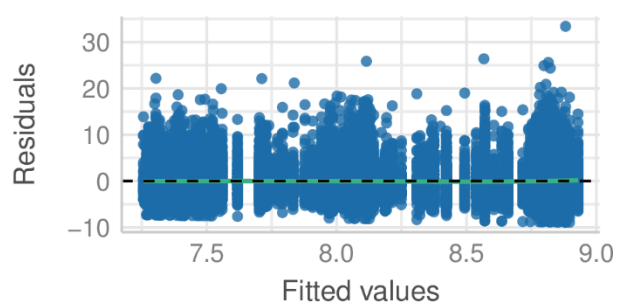
Model-predicted lines should resemble observed data



— Observed data — Model-predicted data

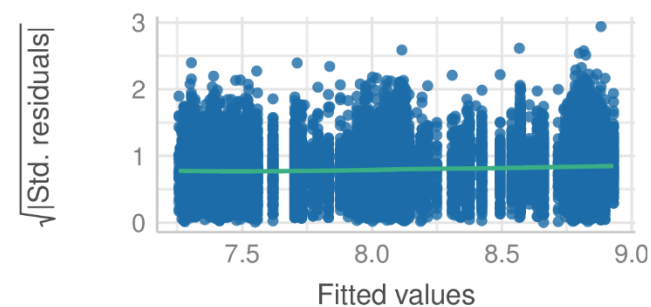
Linearity

Reference line should be flat and horizontal



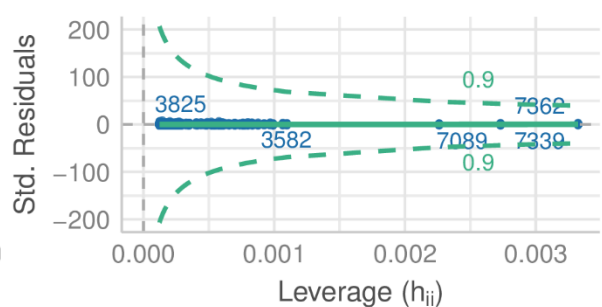
Homogeneity of Variance

Reference line should be flat and horizontal



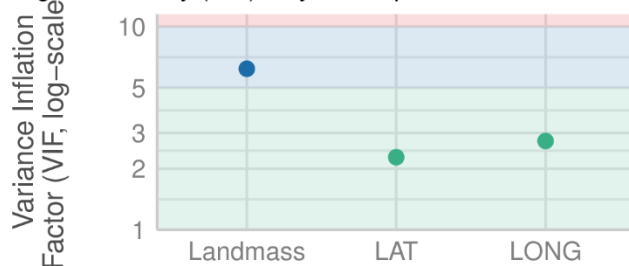
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



● Low (< 5) ● Moderate (< 10)

Normality of Residuals

Dots should fall along the line

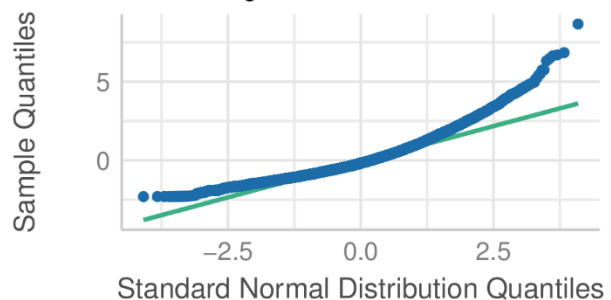


Figure S122. Linear model assumptions check for baseline linear model of Index.Asymmetry response variable, from SVR data set

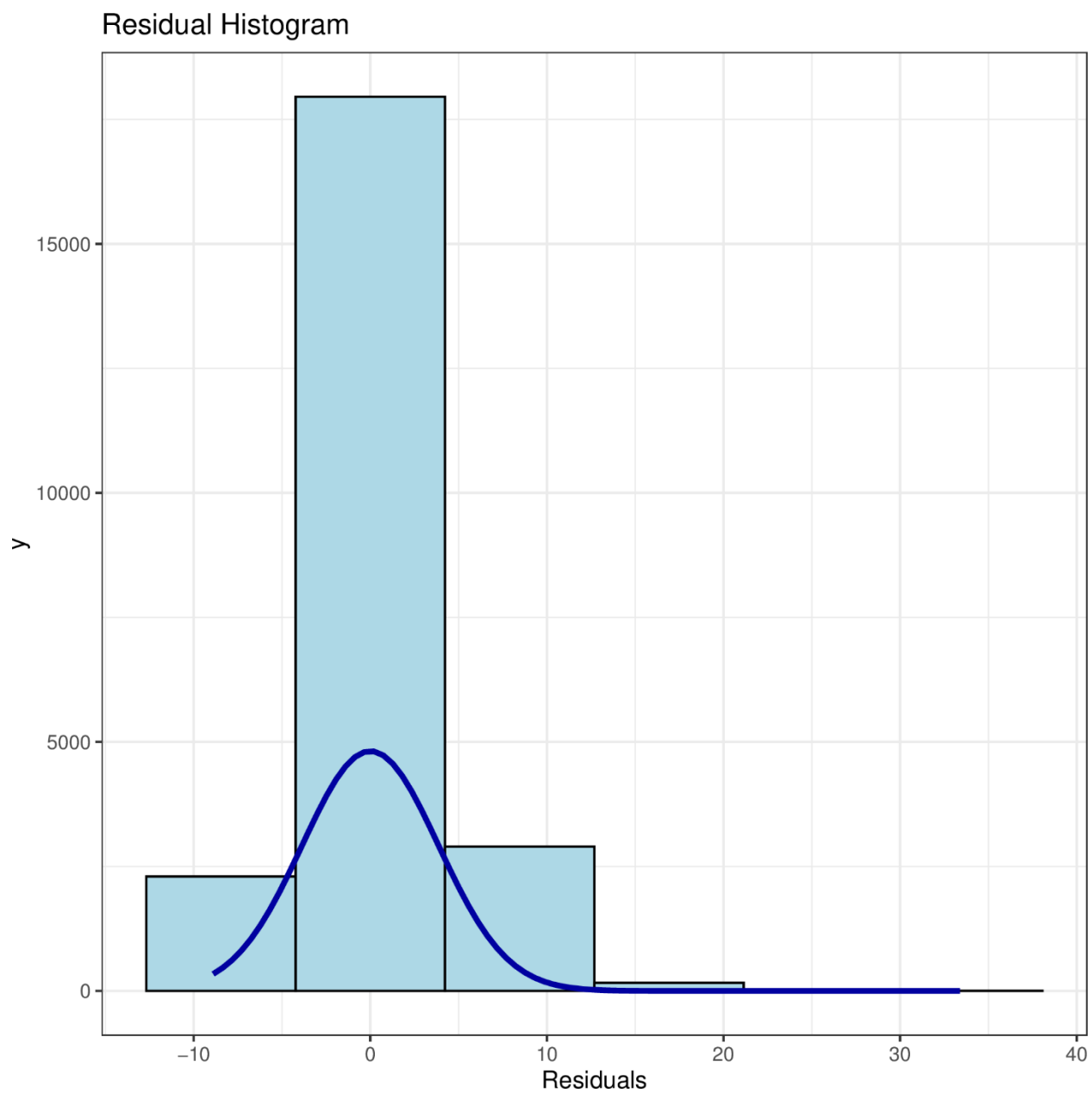
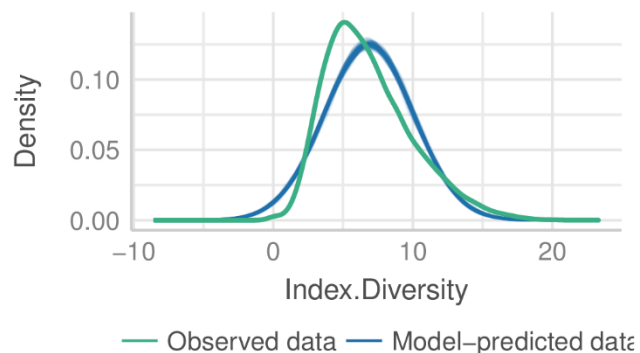


Figure S123. Histogram of residuals for baseline linear model of Index.Asymmetry response variable, from SVR data set

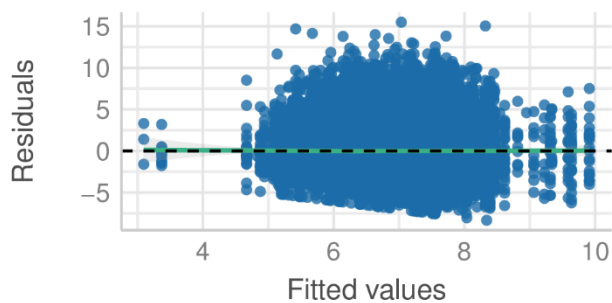
Posterior Predictive Check

Model-predicted lines should resemble observed data



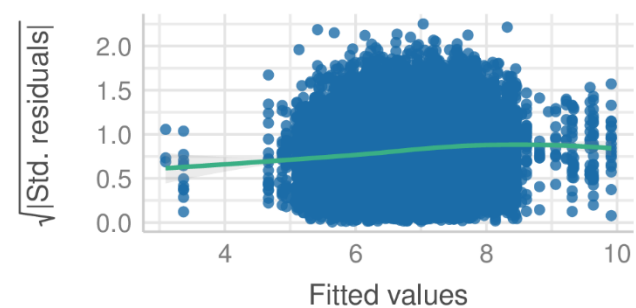
Linearity

Reference line should be flat and horizontal



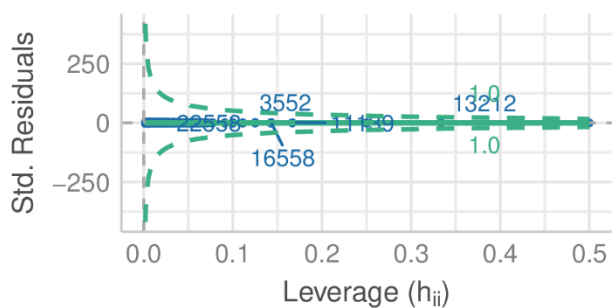
Homogeneity of Variance

Reference line should be flat and horizontal



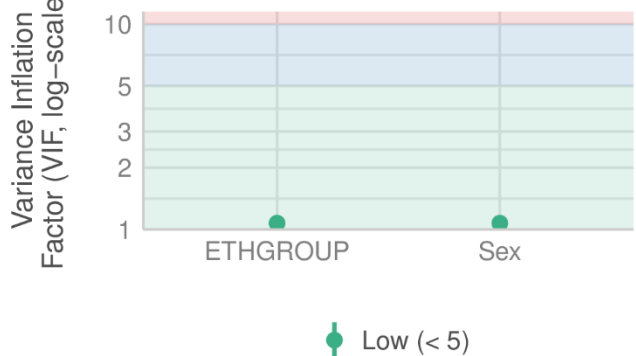
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

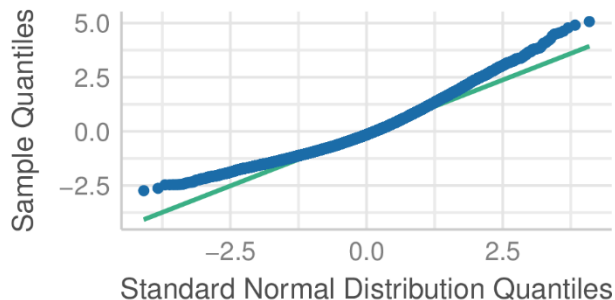


Figure S124. Linear model assumptions check for baseline linear model of Index.Diversity response variable, from SVR data set

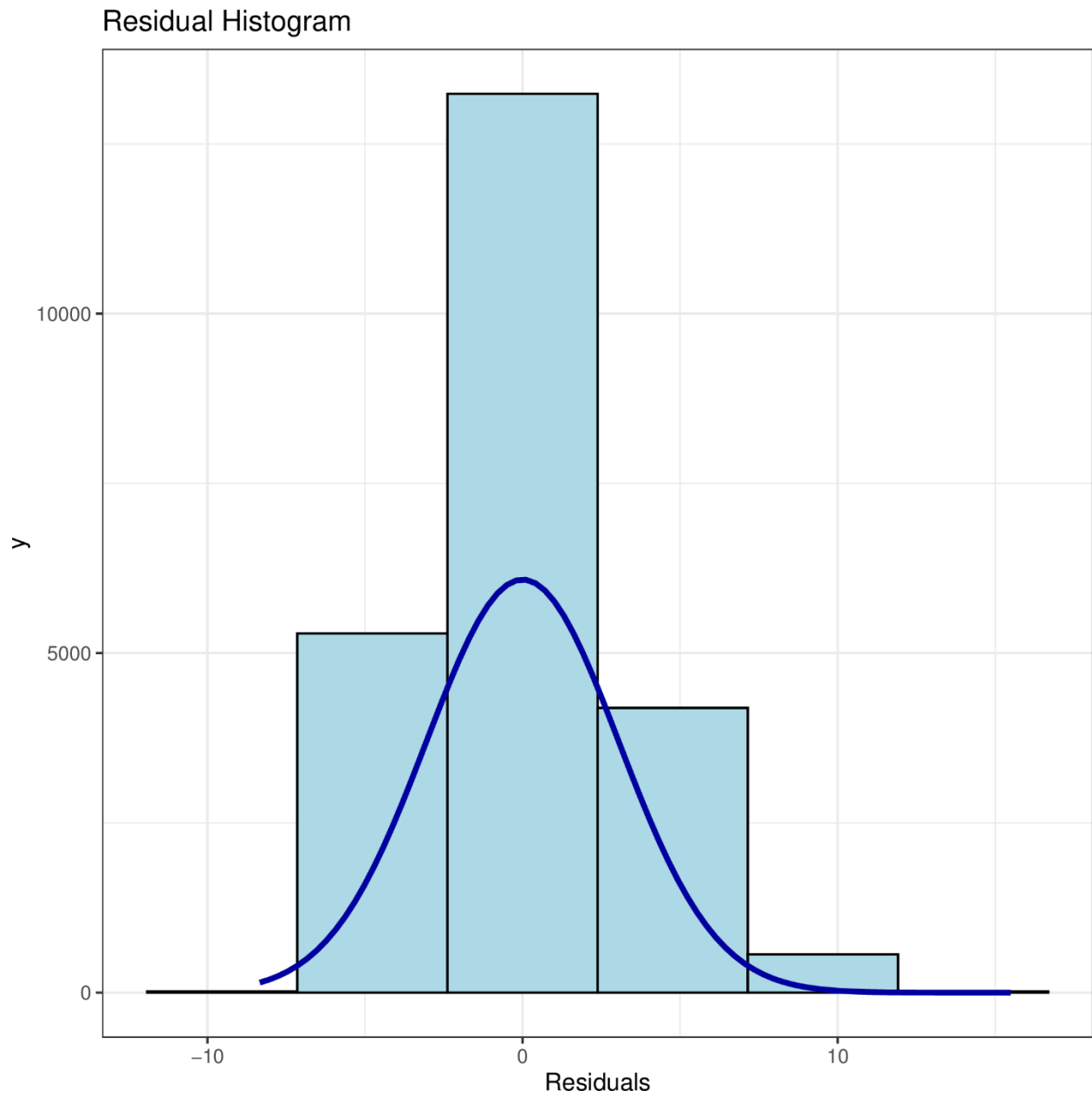
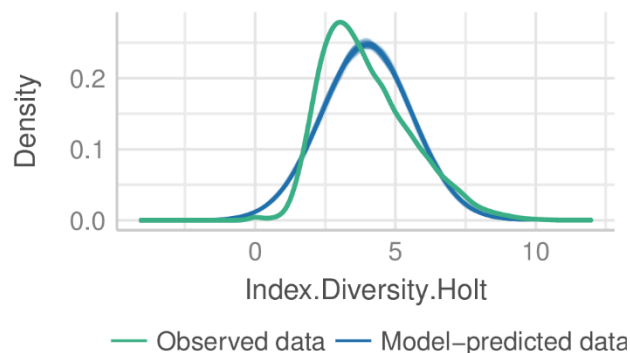


Figure S125. Histogram of residuals for baseline linear model of Index.Diversity response variable, from SVR data set

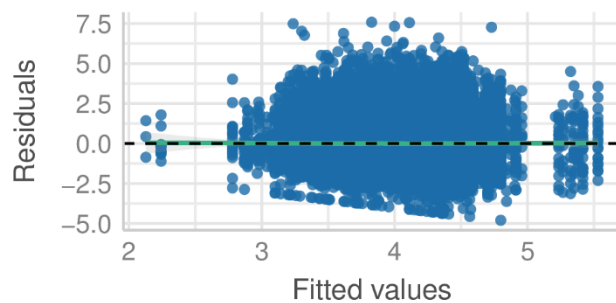
Posterior Predictive Check

Model-predicted lines should resemble observed data



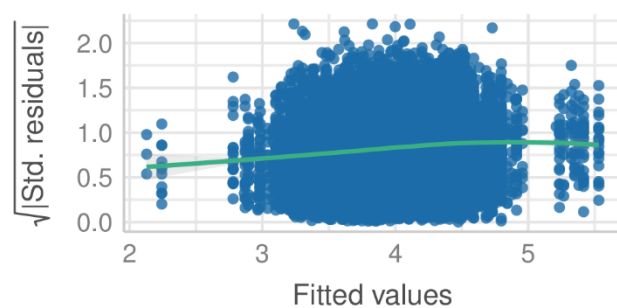
Linearity

Reference line should be flat and horizontal



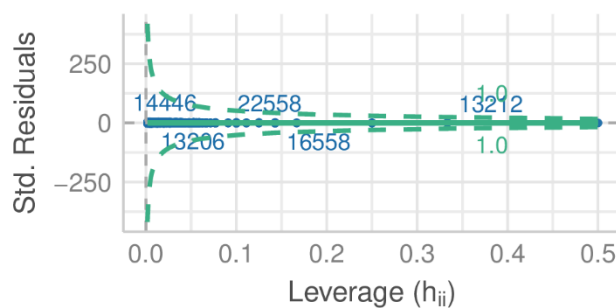
Homogeneity of Variance

Reference line should be flat and horizontal



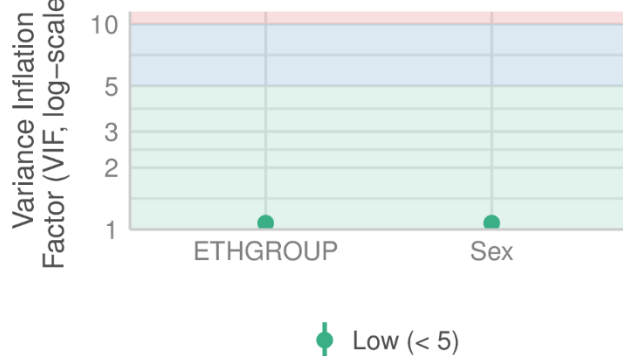
Influential Observations

Points should be inside the contour lines



Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line

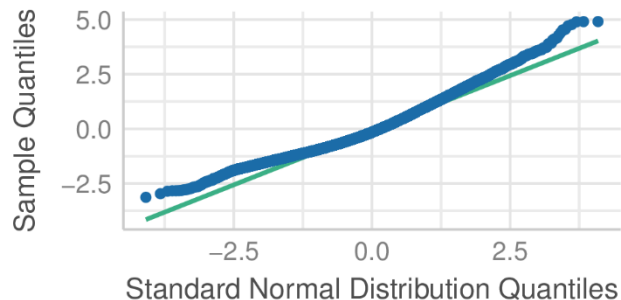


Figure S126. Linear model assumptions check for baseline linear model of Index.Diversity.Holt response variable, from SVR data set

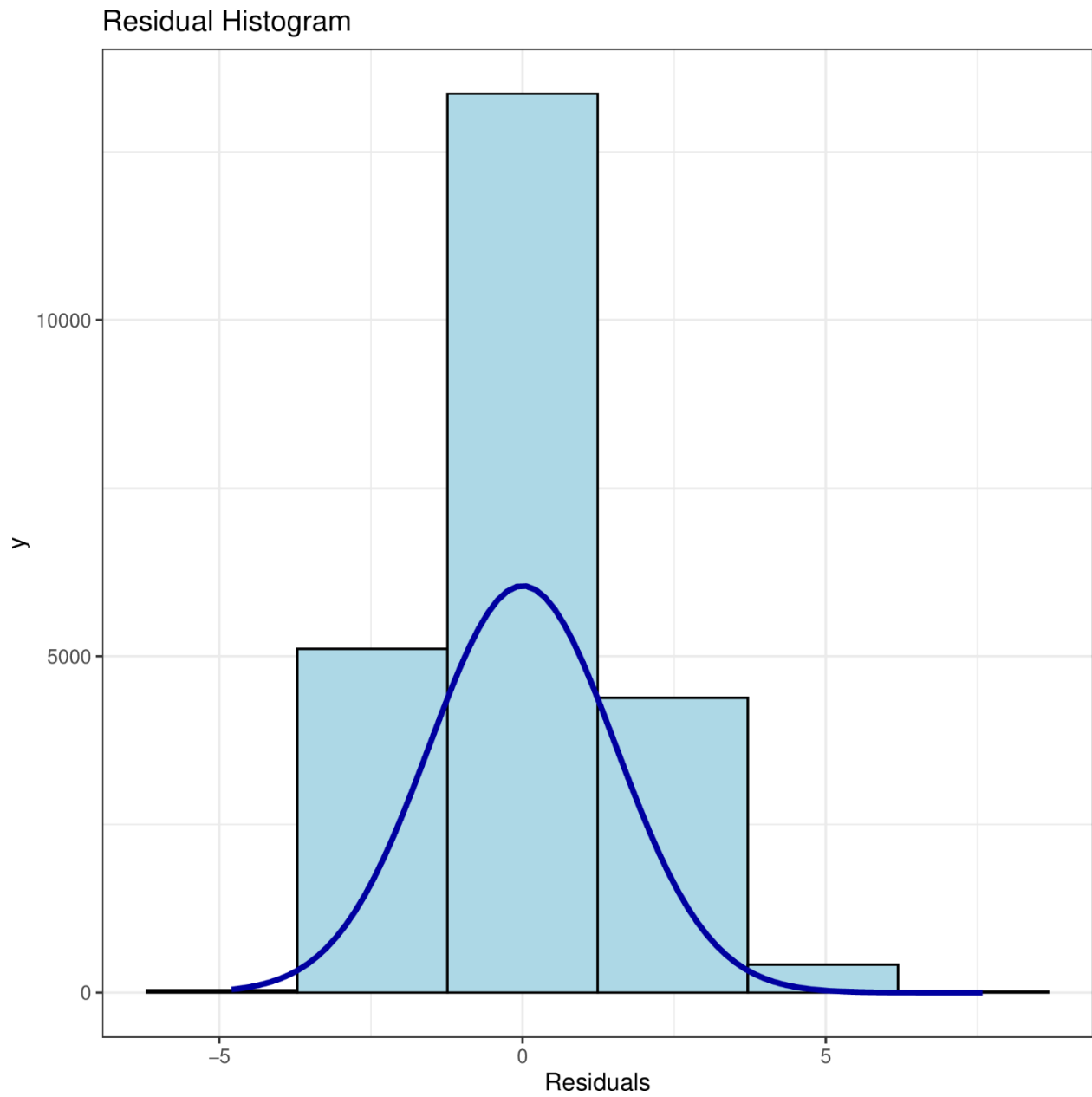


Figure S127. Histogram of residuals for baseline linear model of Index.Diversity.Holt response variable, from SVR data set

Appendix C.2.4 Supplementary figures for Support vector regression (SVR)

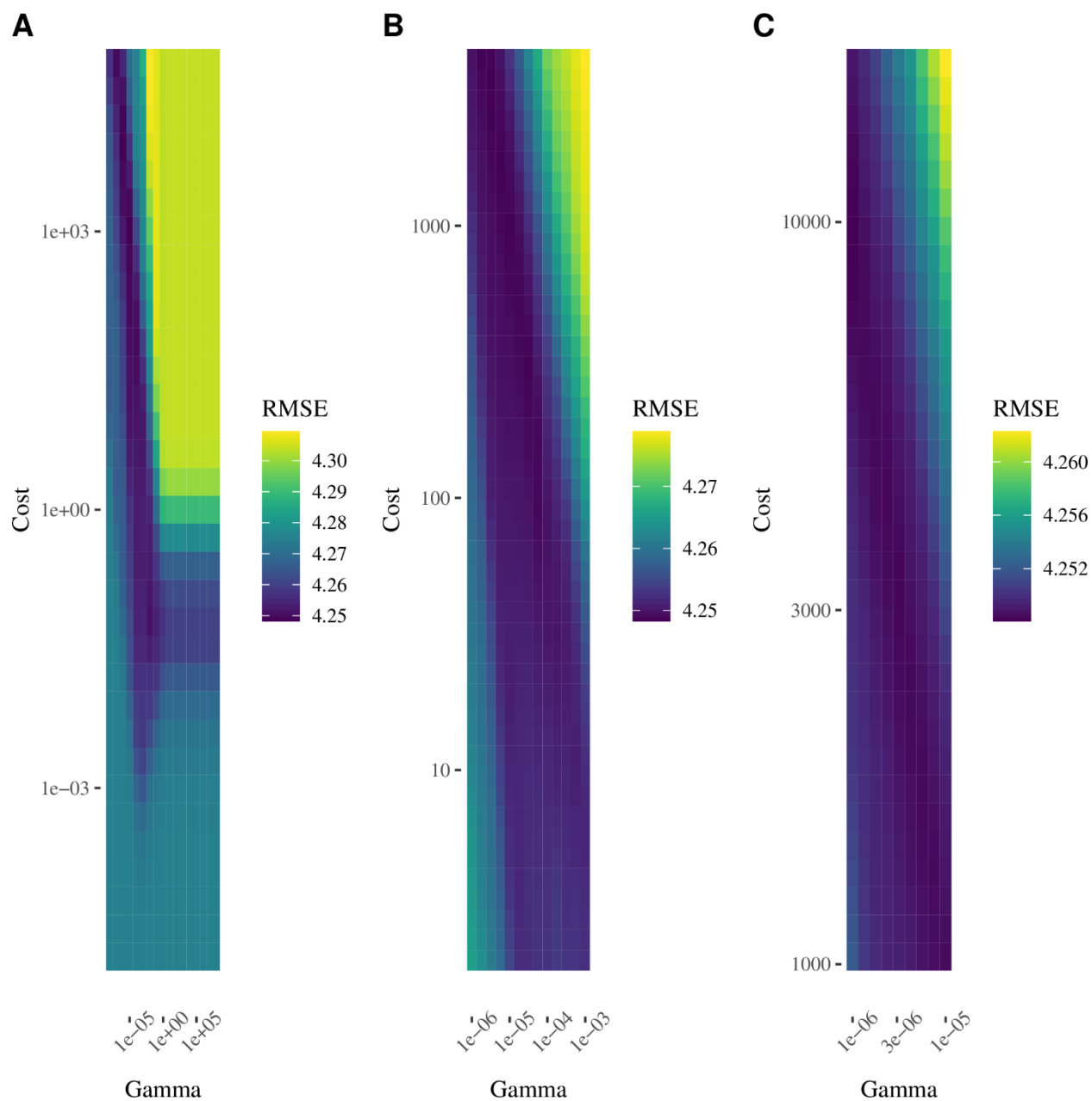


Figure S128. SVR DA1 grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

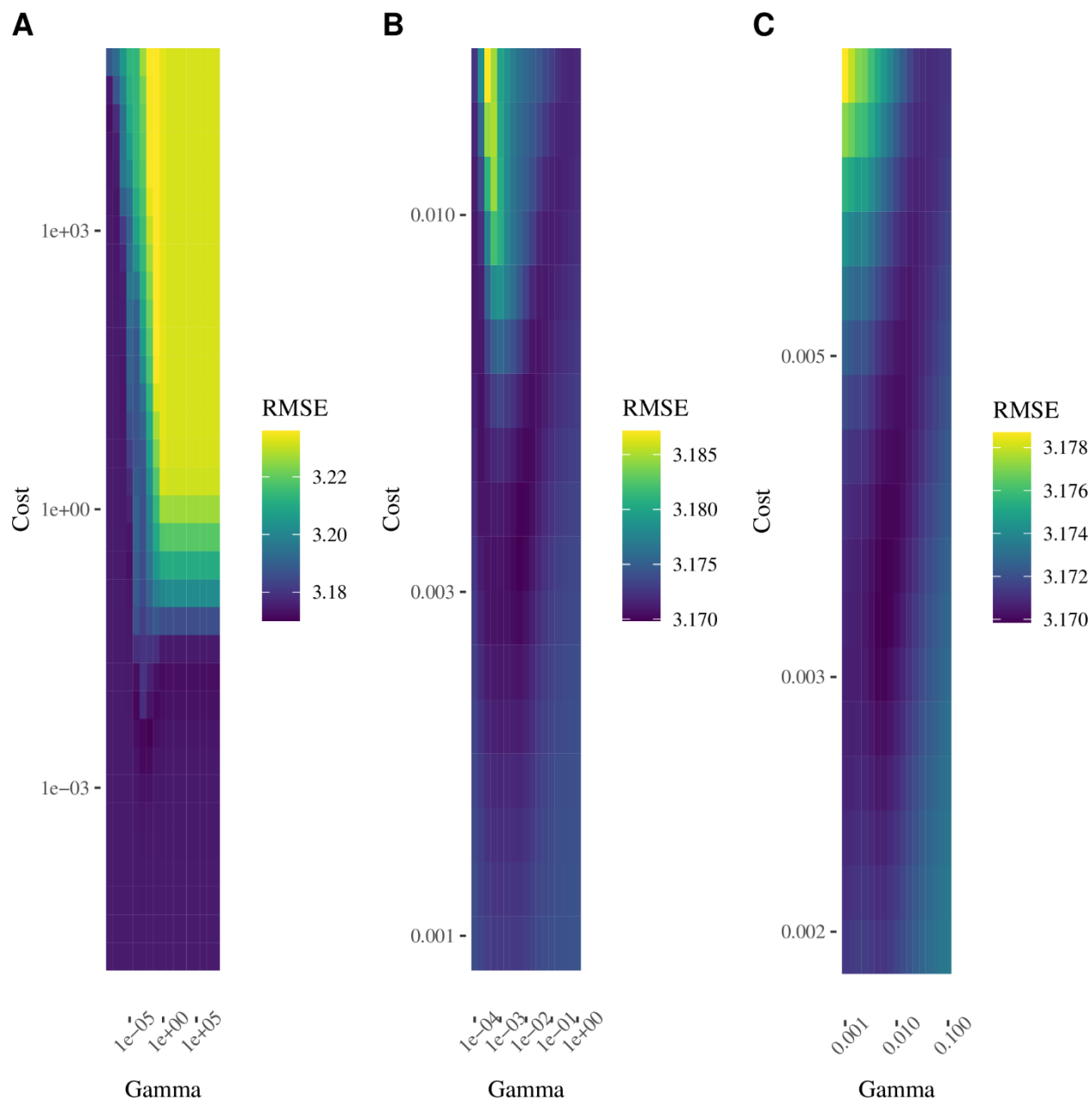


Figure S129. SVR FA1 grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

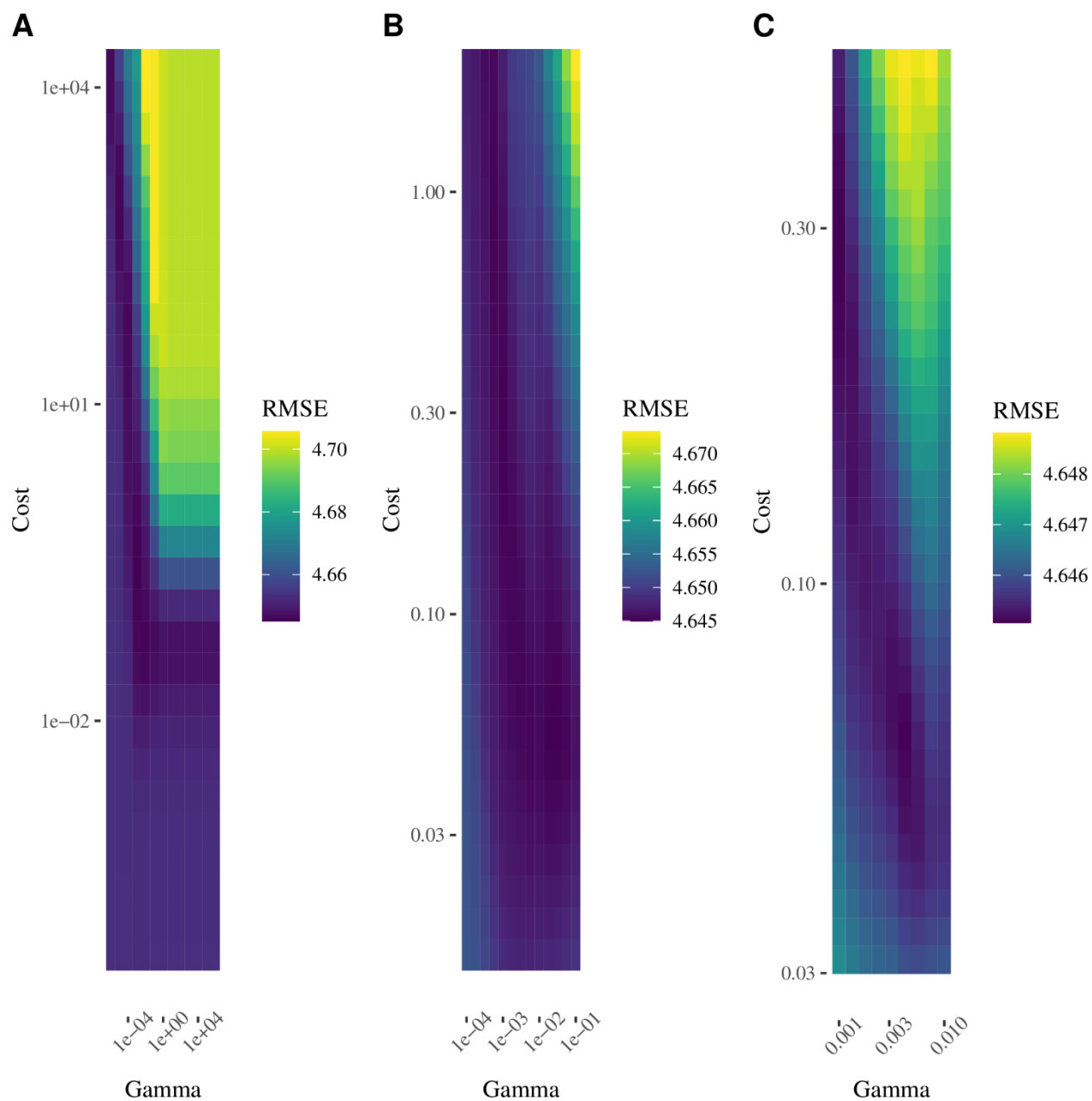


Figure S130. SVR DA2 grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

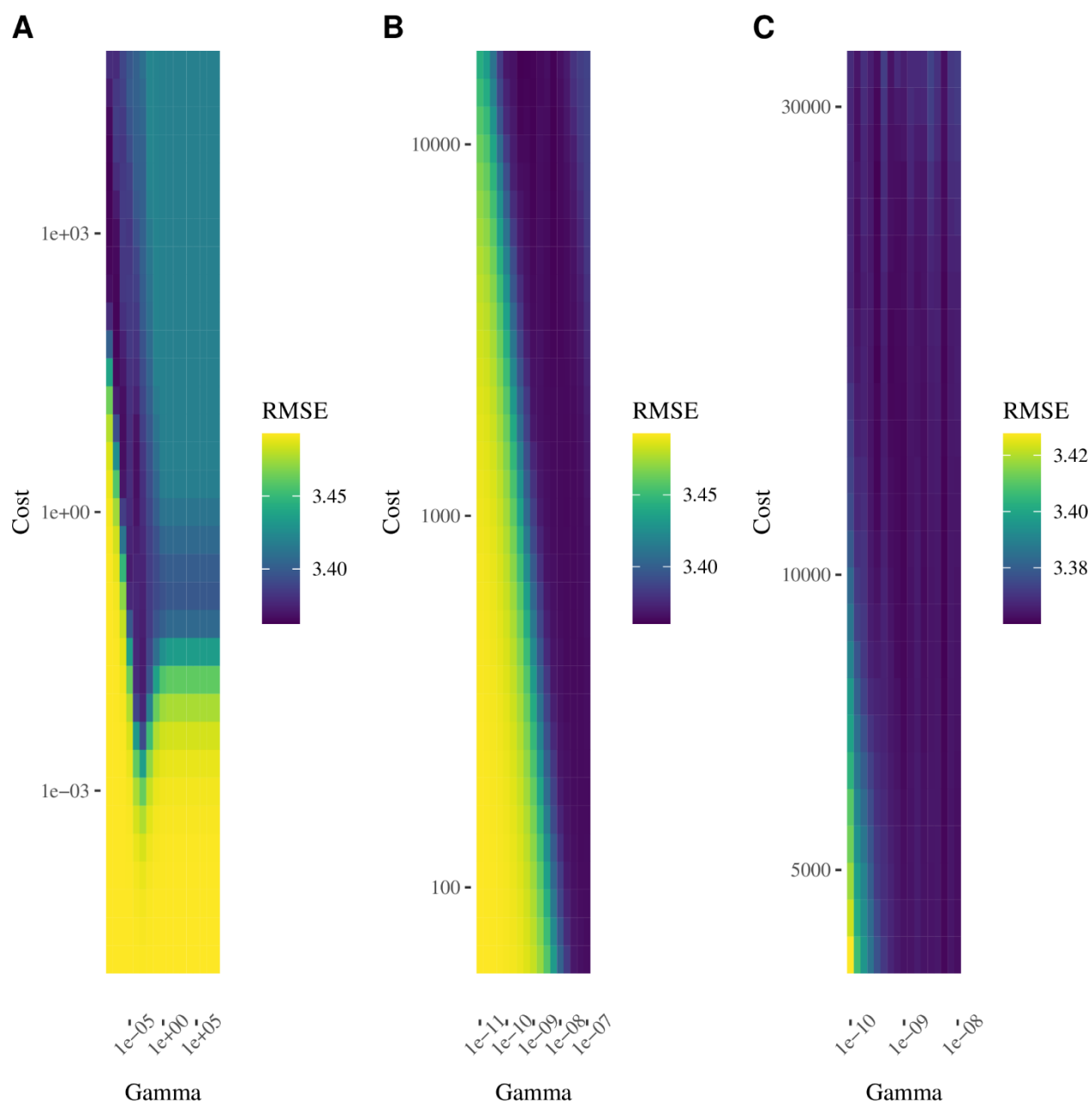


Figure S131. SVR FA2 grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

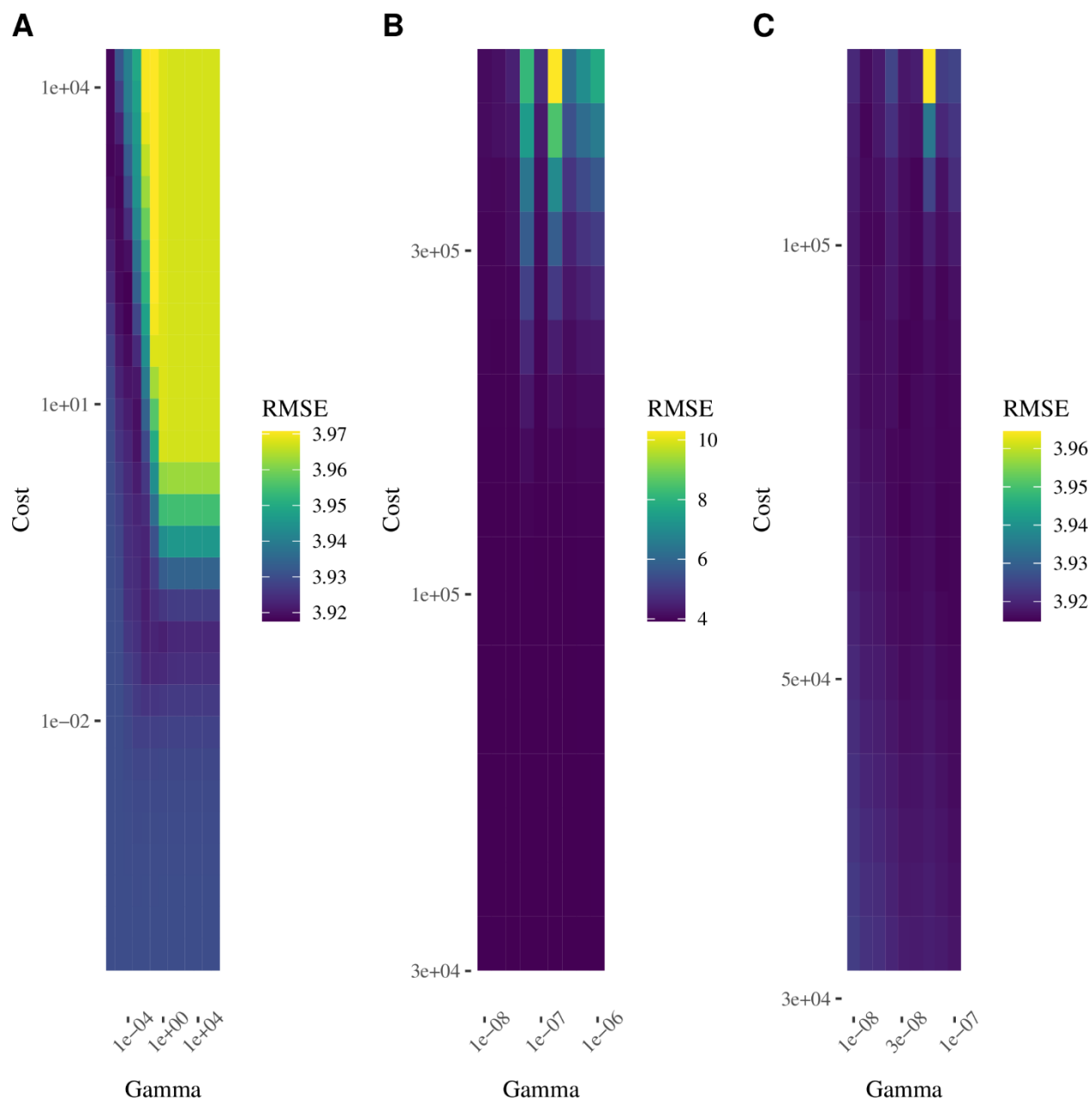


Figure S132. SVR DA3 grid searches gamma vs cost. (A) coarse grid search and (B) fine grid search

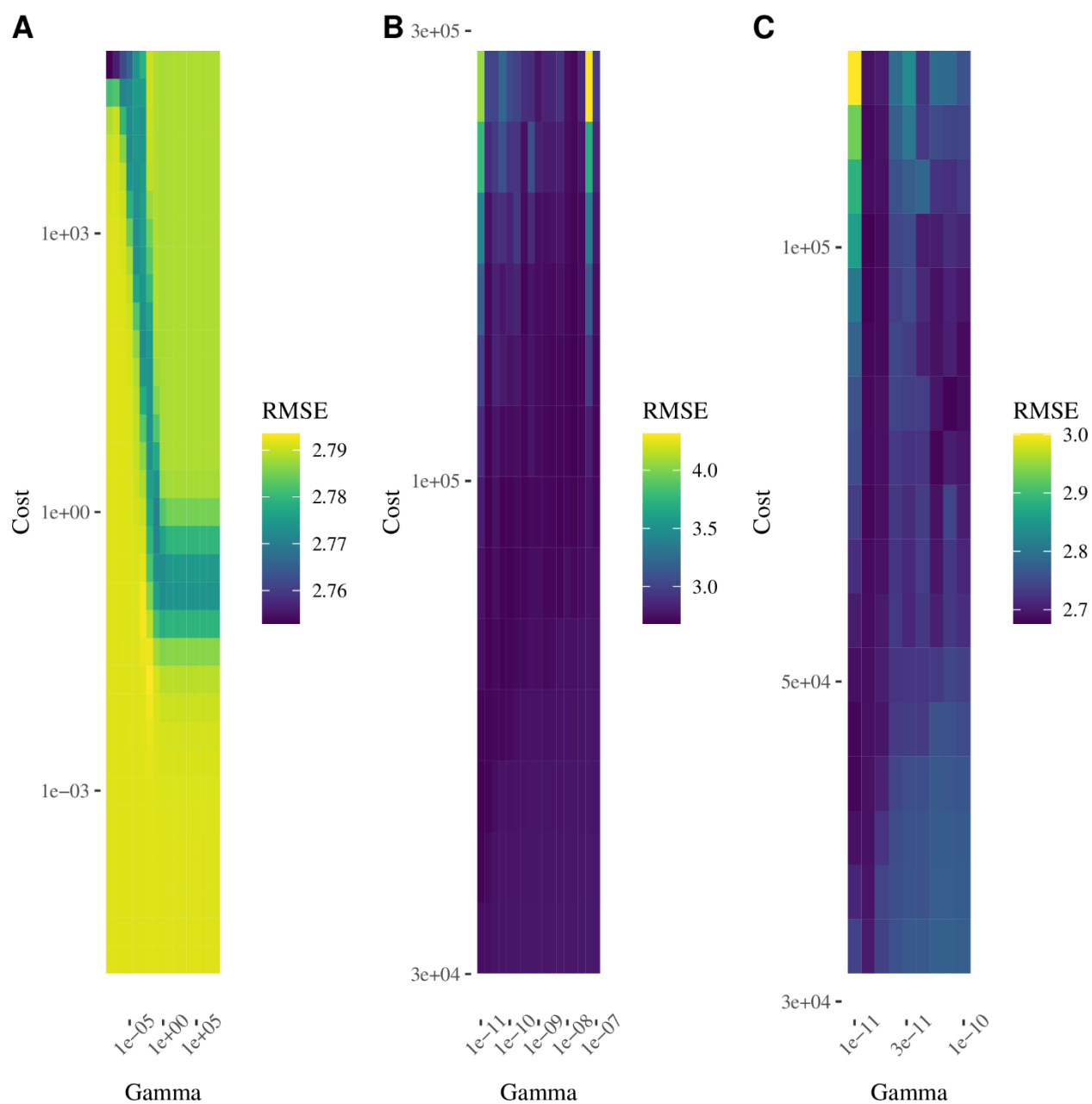


Figure S133. SVR FA3 grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

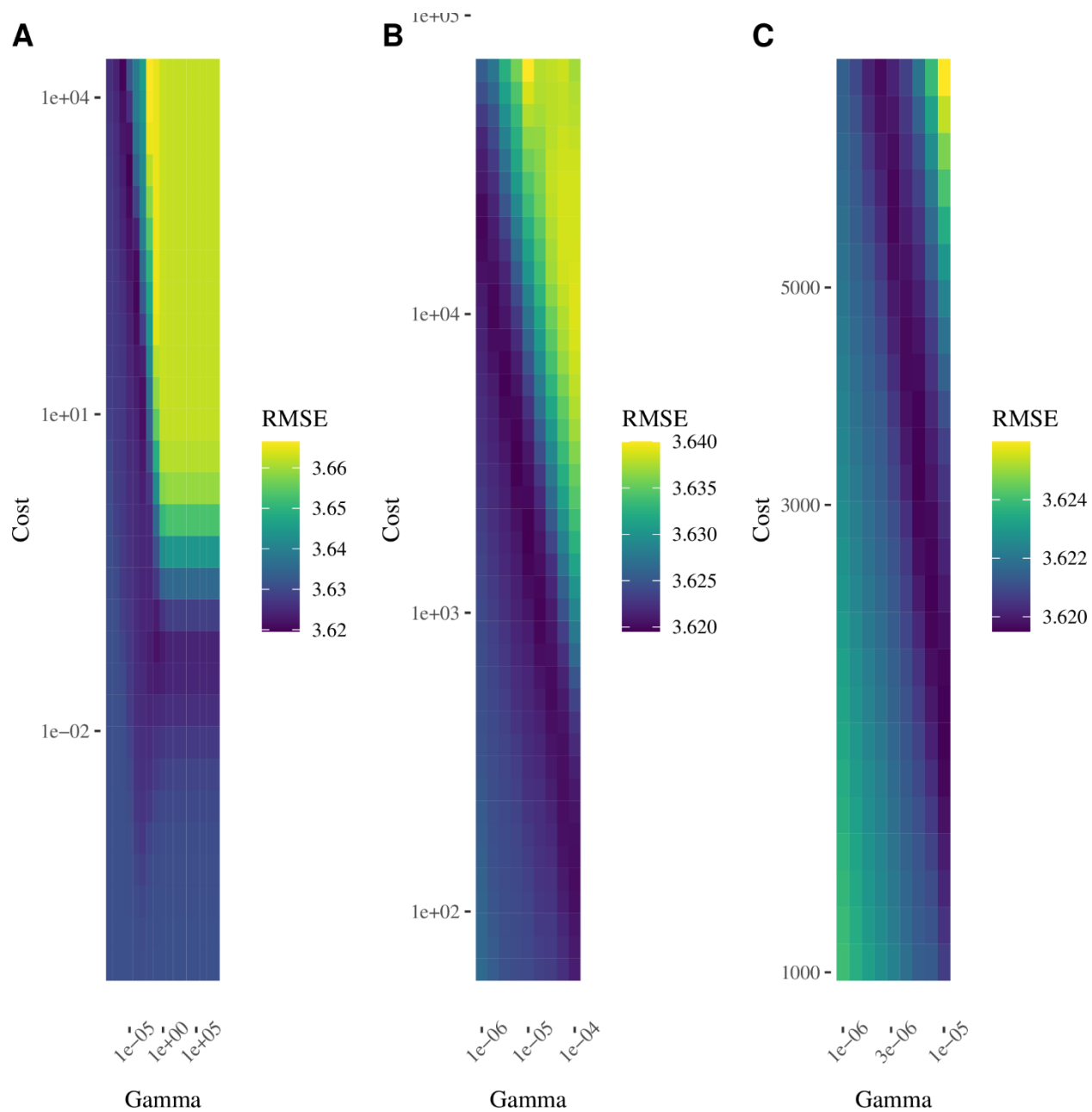


Figure S134. SVR DA4 grid searches gamma vs cost. (A) coarse grid search and (B) fine grid search

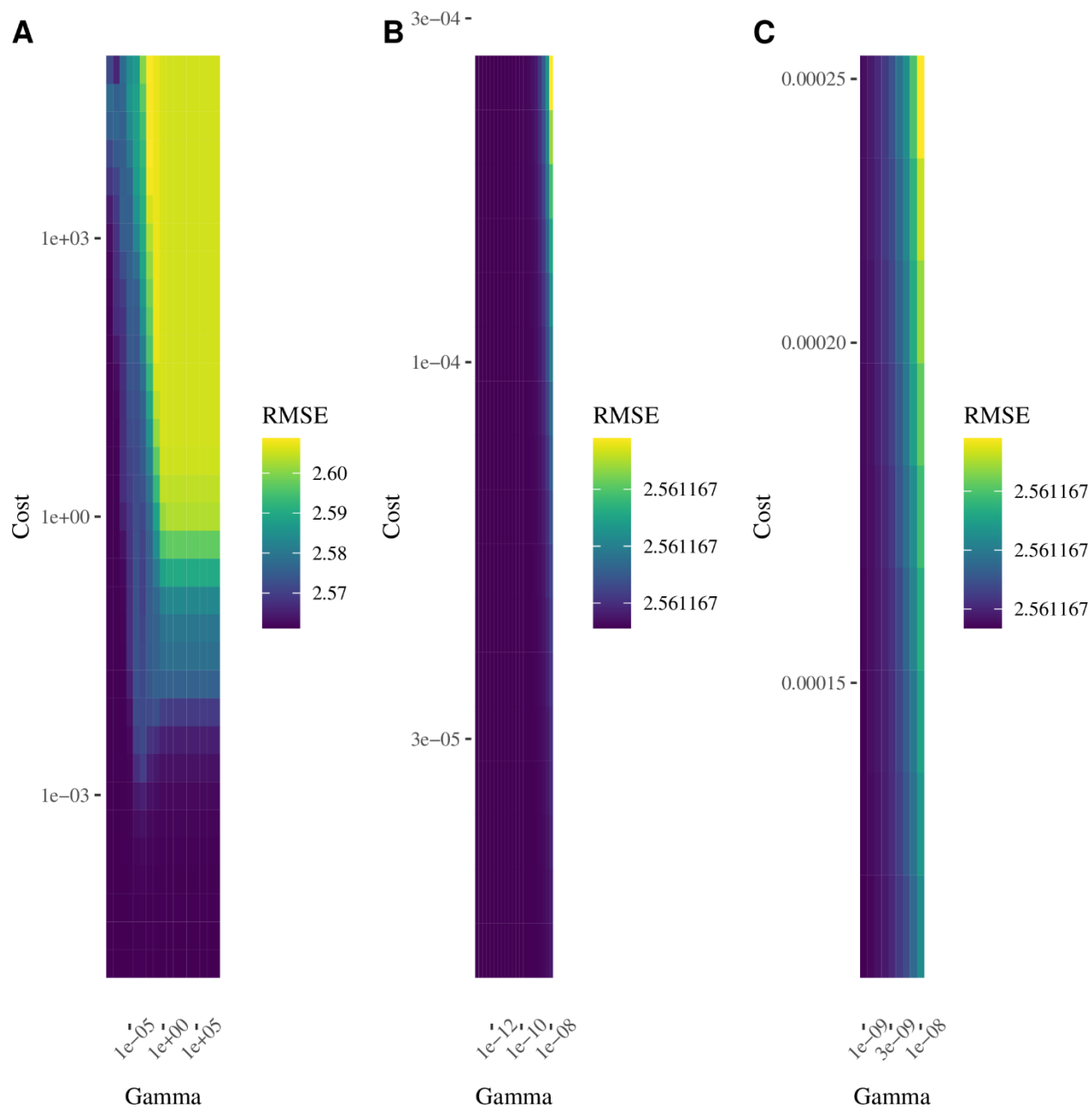


Figure S135. SVR FA4 grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

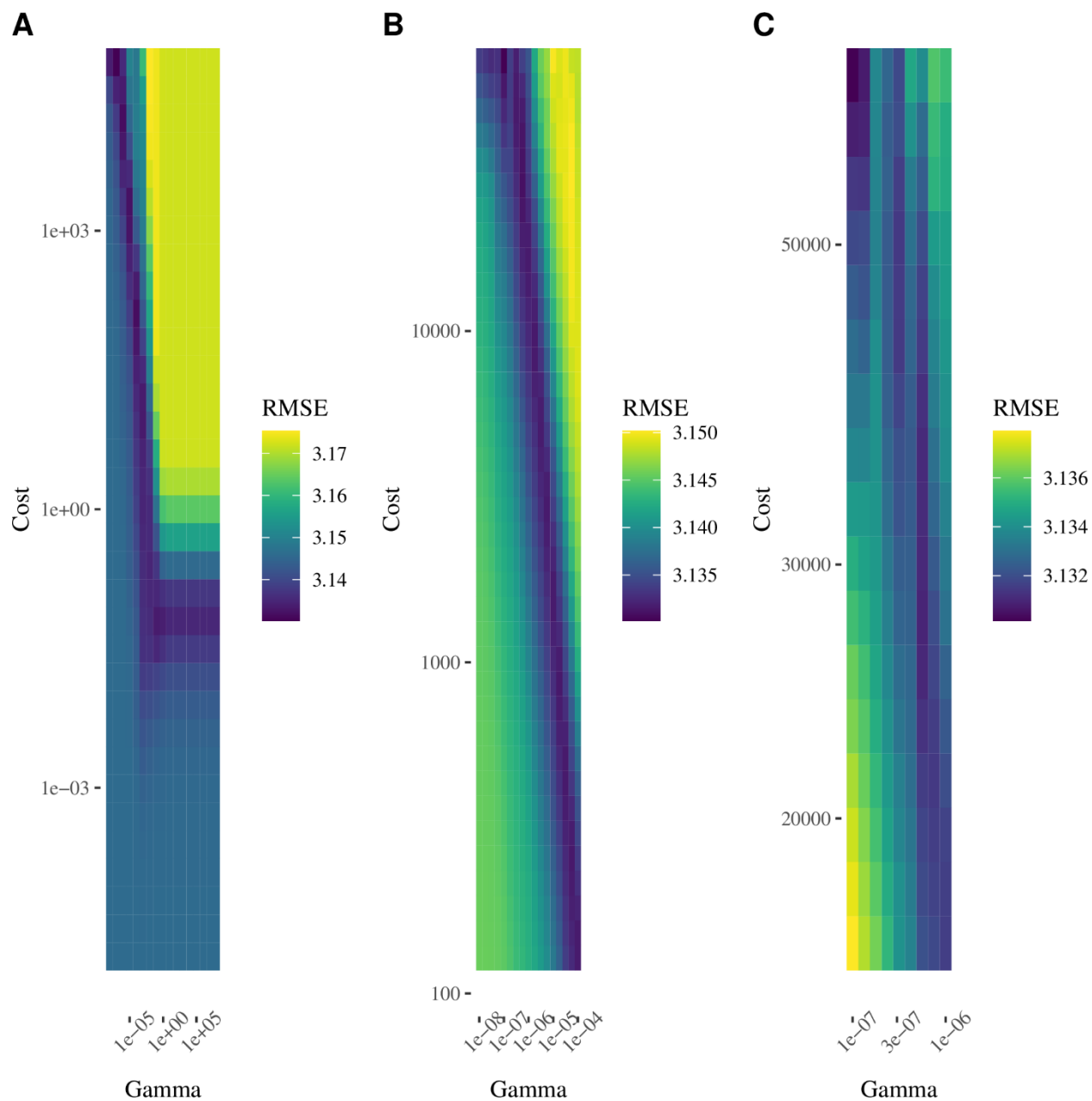


Figure S136. SVR DA5 grid searches gamma vs cost. (A) coarse grid search and (B) fine grid search

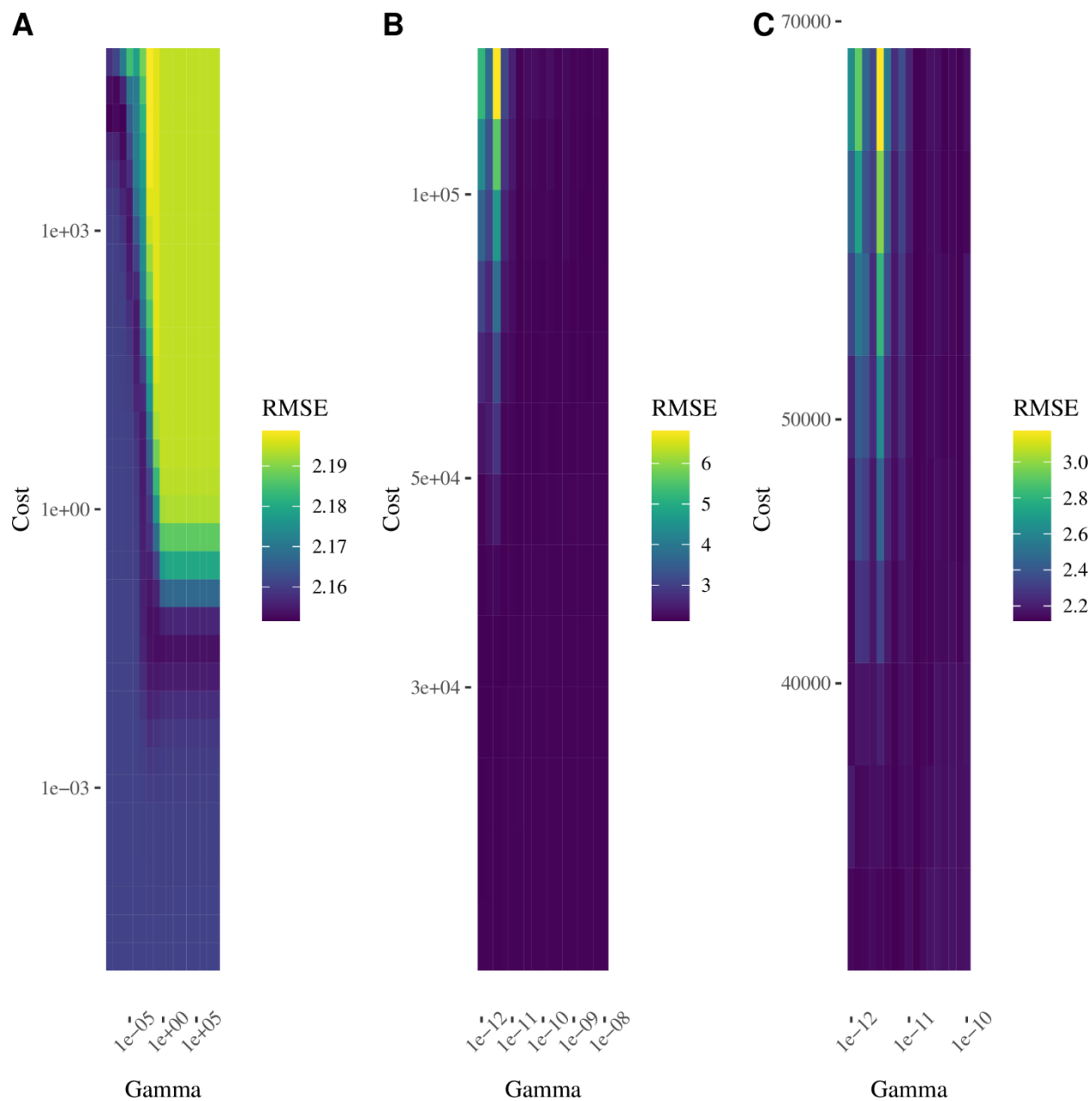


Figure S137. SVR FA5 grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

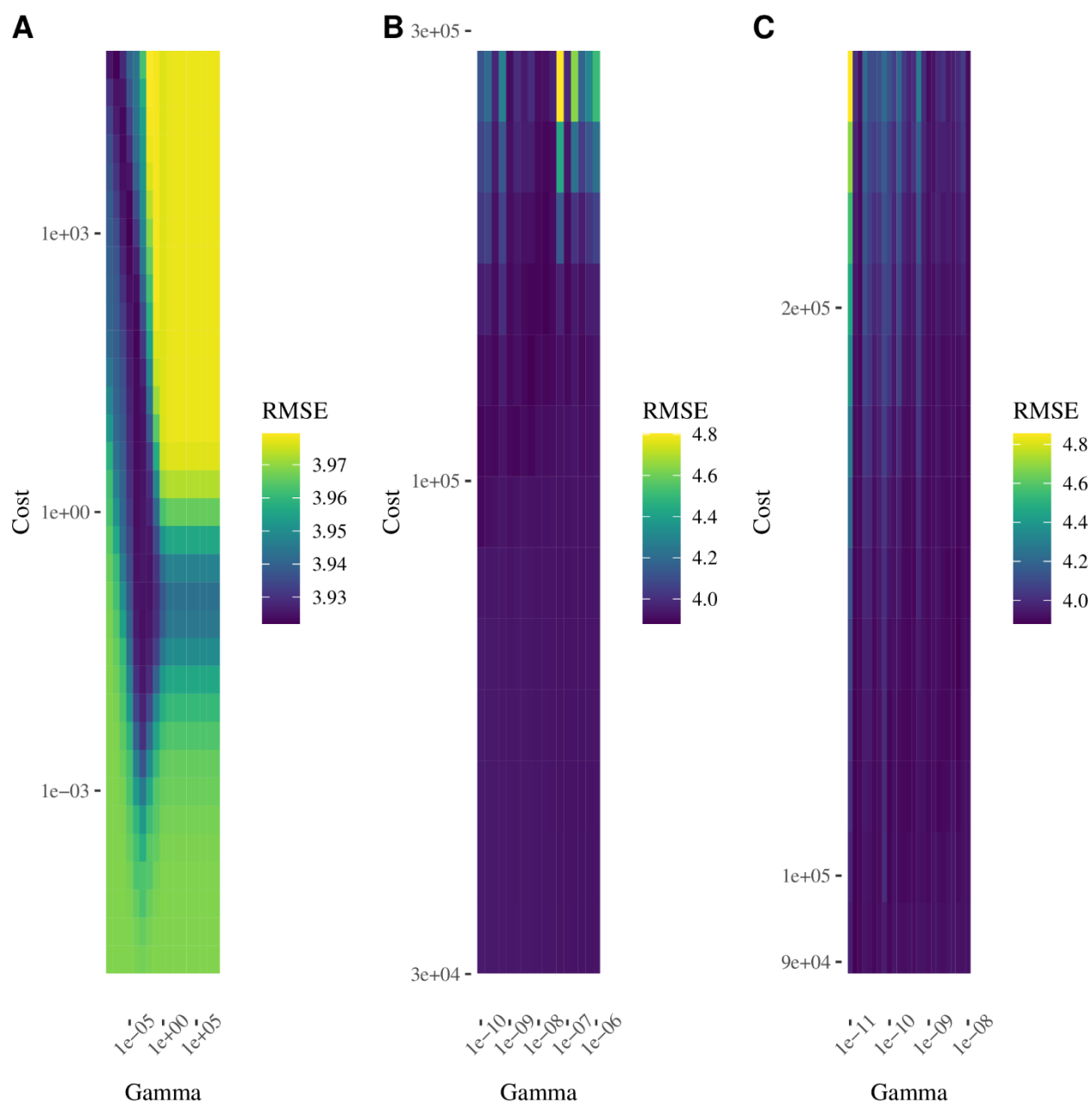


Figure S138. SVR Index. Asymmetry grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

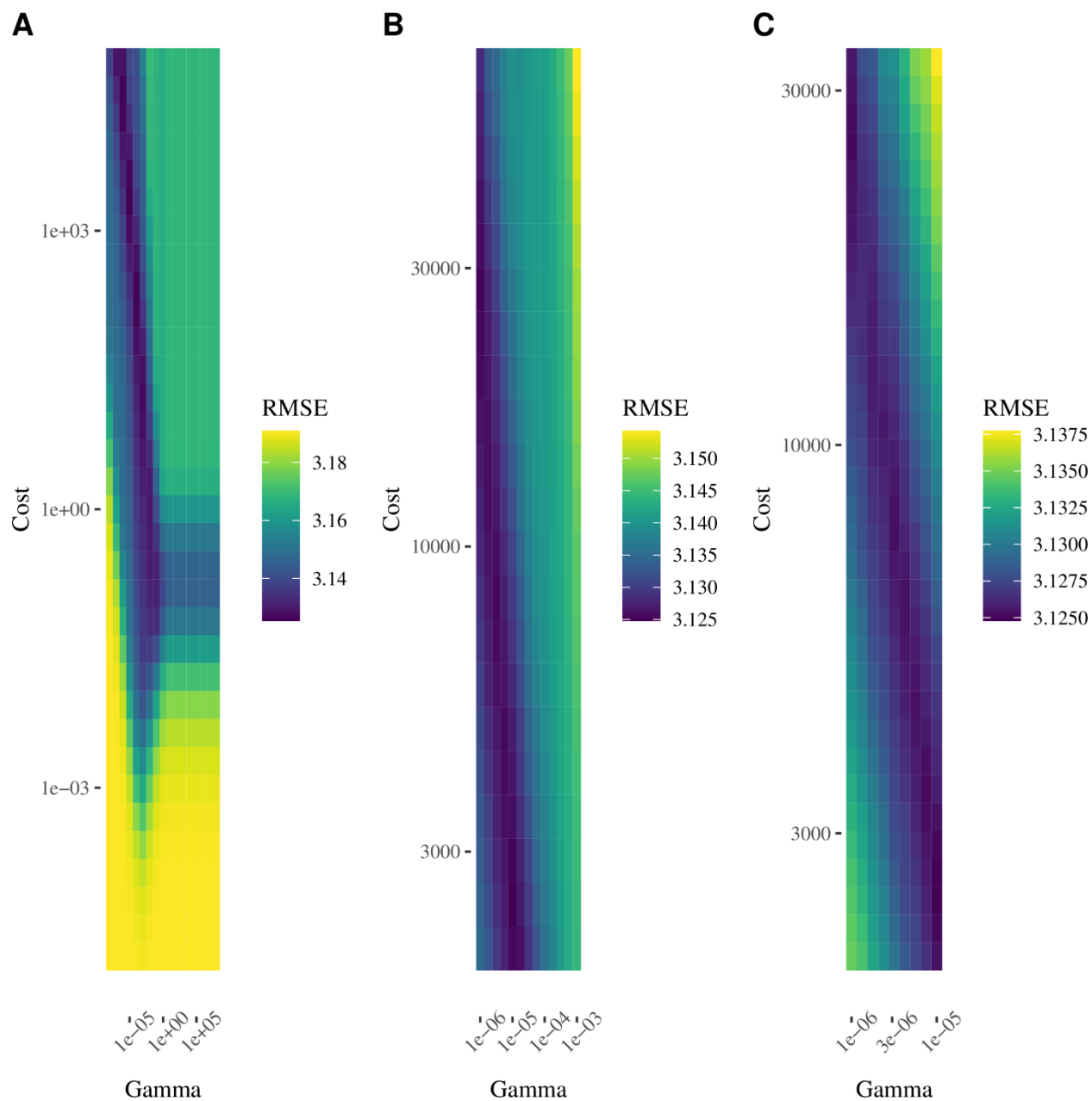


Figure S139. SVR Index.Diversity grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

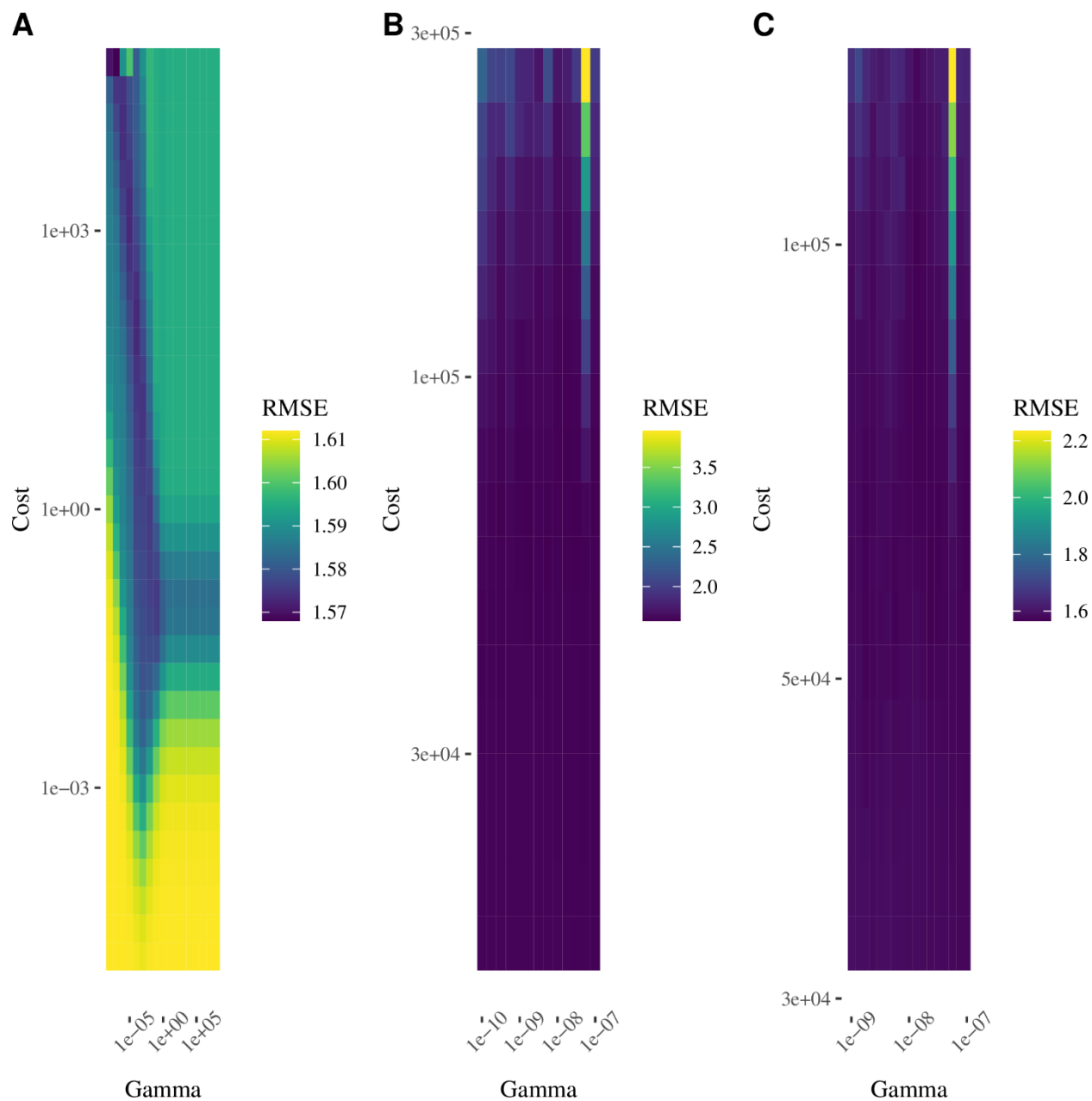


Figure S140. SVR Index.Diversity.Holt grid searches gamma vs cost. (A) coarse grid search, (B) medium grid search, and (C) fine grid search

Table S13. Change in SVR model performance through grid searching for Directional Asymmetry and Fluctuating Asymmetry variables for each digit, and indices of asymmetry and diversity. The coarse, medium, and fine grid search performance metrics were calculated by comparing the model to validation data. The final model performance metrics were calculated by comparing the model to unseen test data. MAE = mean absolute error; MAPE = mean absolute percentage error; RMSE = root mean squared error.

Response variable	Grid search	Hyperparameter range	MAE	RMSE	MAPE
Directional Asymmetry, digit 1	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	3.210	4.263	2.158
	Medium, by 0.25	cost: $2^1 : 2^{12}$ γ : $10^{-6} : 10^{-3}$	3.210	4.263	2.209
	Fine, by 0.125	cost: $2^{10} : 2^{14}$ γ : $10^{-6} : 10^{-5}$	3.210	4.263	2.182
	Final model	cost: 9742 γ : $1 \text{ e-}6$	3.183	4.209	2.143
Directional Asymmetry, digit 2	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-14} : 2^{14}$ γ : $10^{-6} : 10^6$	3.283	4.578	28.190
	Medium, by 0.25	cost: $2^{-6} : 2^1$ γ : $10^{-4} : 10^{-1}$	3.285	4.576	26.217
	Fine, by 0.125	cost: $2^{-5} : 2^{-1}$ γ : $10^{-3} : 10^{-2}$	3.283	4.578	470.776
	Final model	cost: 0.29 γ : 0.001	3.352	4.702	32.971

Table S13 Continued

Response variable	Grid search	Hyperparameter range	MAE	RMSE	MAPE
Directional Asymmetry, digit 3	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-14} : 2^{14}$ γ : $10^{-6} : 10^6$	2.880	3.922	30.497
	Medium, by 0.25	cost: $2^{15} : 2^{19}$ γ : $10^{-8} : 10^{-6}$	2.882	3.919	53.941
	Fine, by 0.125	cost: $2^{15} : 2^{17}$ γ : $10^{-8} : 10^{-7}$	2.882	3.919	53.941
	Final model	cost: 77936 γ : 3.16 e-8	2.924	3.997	31.396
Directional Asymmetry, digit 4	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-14} : 2^{14}$ γ : $10^{-8} : 10^8$	2.661	3.598	61.886
	Medium, by 0.25	cost: $2^6 : 2^{16}$ γ : $10^{-6} : 10^{-4}$	2.660	3.598	69.173
	Fine, by 0.125	cost: $2^{10} : 2^{13}$ γ : $10^{-6} : 10^{-5}$	2.660	3.598	95.847
	Final model	cost: 3158 γ : 5.6 e-6	2.706	3.711	70.291
Directional Asymmetry, digit 5	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	2.357	3.115	30.635
	Medium, by 0.25	cost: $2^7 : 2^{16}$ γ : $10^{-8} : 10^{-4}$	2.357	3.115	30.635
	Fine, by 0.125	cost: $2^{14} : 2^{16}$ γ : $10^{-7} : 10^{-6}$	2.357	3.115	30.635

Table S13 Continued

Response variable	Grid search	Hyperparameter range	MAE	RMSE	MAPE
	Final model	cost: 65536 γ : 1 e-7	2.378	3.157	86.351
Fluctuating Asymmetry, digit 1	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	2.272	3.184	0.784
	Medium, by 0.25	cost: $2^{-10} : 2^6$ γ : $10^{-4} : 10^0$	2.272	3.184	0.784
	Fine, by 0.125	cost: $2^{-9} : 2^{-7}$ γ : $10^{-3} : 10^{-1}$	2.272	3.184	0.784
	Final model	cost: 0.0035 γ : 0.0059	2.261	3.143	0.781
	Initial model	cost: 1 γ : 0.14286			
Fluctuating Asymmetry, digit 2	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	2.284	3.295	0.922
	Medium, by 0.25	cost: $2^6 : 2^{14}$ γ : $10^{-11} : 10^{-7}$	2.285	3.290	0.914
	Fine, by 0.125	cost: $2^{12} : 2^{15}$ γ : $10^{-10} : 10^{-8}$	2.285	3.292	0.916
	Final model	cost: 20826 γ : 3.16 e-10	2.336	3.419	0.933
	Initial model	cost: 1 γ : 0.14286			
Fluctuating Asymmetry, digit 3	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	1.934	2.757	0.847
	Medium, by 0.25	cost: $2^{15} : 2^{18}$ γ : $10^{-11} : 10^{-7}$	2.104	2.733	0.662
	Final model				

Table S13 Continued

Response variable	Grid search	Hyperparameter range	MAE	RMSE	MAPE
	Fine, by 0.125	cost: $2^{15} : 2^{17}$ γ : $10^{-11} : 10^{-10}$	2.104	2.733	0.662
	Final model	cost: 96876 γ : 1.33 e-11	2.041	2.761	0.726
Fluctuating Asymmetry, digit 4	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	1.714	2.507	0.816
	Medium, by 0.25	cost: $2^{-16} : 2^{-12}$ γ : $10^{-13} : 10^{-8}$	1.714	2.507	0.816
	Fine, by 0.125	cost: $2^{-13} : 2^{-12}$ γ : $10^{-9} : 10^{-8}$	1.714	2.507	0.816
	Final model	cost: 0.000244 γ : 1 e-8	1.749	2.632	0.833
	Initial model	cost: 1 γ : 0.14286			
Fluctuating Asymmetry, digit 5	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	1.490	2.122	0.768
	Medium, by 0.25	cost: $2^{14} : 2^{17}$ γ : $10^{-12} : 10^{-8}$	1.591	2.099	0.663
	Fine, by 0.125	cost: $2^{15} : 2^{16}$ γ : $10^{-12} : 10^{-10}$	1.591	2.099	0.663
	Final model	cost: 32768 γ : 5.6 e-12	1.620	2.132	0.678
	Initial model	cost: 1 γ : 0.14286			
Index.Asymmetry	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	2.877	3.836	0.384

Table S13 Continued

Response variable	Grid search	Hyperparameter range	MAE	RMSE	MAPE
	Medium, by 0.25	cost: $2^{15} : 2^{18}$ γ : $10^{-10} : 10^{-6}$	2.981	3.810	0.358
	Fine, by 0.125	cost: $2^{16.5} : 2^{18}$ γ : $10^{-11} : 10^{-8}$	2.981	3.810	0.358
	Final model	cost: 131072 γ : $5.6 \text{ e-}9$	2.996	3.959	0.384
Index.Diversity	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	2.402	3.096	0.377
	Medium, by 0.25	cost: $2^{11} : 2^{16}$ γ : $10^{-6} : 10^{-3}$	2.402	3.096	0.377
	Fine, by 0.125	cost: $2^{11} : 2^{15}$ γ : $10^{-6} : 10^{-5}$	2.402	3.096	0.377
	Final model	cost: $2^{11} : 2^{15}$ γ : $10^{-6} : 10^{-5}$	2.451	3.156	0.384
Index.Diversity.Holt	Initial model	cost: 1 γ : 0.14286			
	Coarse, by 1	cost: $2^{-16} : 2^{16}$ γ : $10^{-8} : 10^8$	1.243	1.563	0.322
	Medium, by 0.25	cost: $2^{14} : 2^{18}$ γ : $10^{-10} : 10^{-7}$	1.230	1.542	0.313
	Fine, by 0.125	cost: $2^{15} : 2^{17}$ γ : $10^{-9} : 10^{-7}$	1.237	1.544	0.312
	Final model	cost: 60097 γ : $2.37 \text{ e-}8$	1.245	1.582	0.327

VITA

Kathleen Ann Hauther lived in multiple states in her childhood and somehow ended up attending high school in Middle Tennessee. She attended the University of Tennessee, Knoxville initially for microbiology and then became enchanted by the introduction to biological anthropology course and then even more excited by the fragmentary human osteology course. In 2012 she received her BA in Anthropology with research interests of combining microbiology and anthropology to estimate the postmortem interval from postmortem human gut microbial communities. Upon publication of this research, she endeavored to continue study in this growing field.

Kathleen returned to the University of Tennessee in 2014 to pursue a doctoral degree. Since then she has worked with the Forensic Anthropology Center in donor correspondence, Facility maintenance and donor photos, and has established standard operating guidelines to train graduate and undergraduate students in these roles. She has served as a teaching assistant for Human Osteology, Introduction to Forensic Anthropology, and Anthropology Field Recovery for surface skeletal recoveries. Additionally, participation in instruction of short courses hosted by the Forensic Anthropology Center for various academic and law enforcement groups as well as participation in community outreach has kindled her interest in science communication and teaching, in general.

During this time, Kathleen oversaw and facilitated the field sampling at the University of Tennessee for an external, NIJ-funded grant to estimate the postmortem interval from bone using a metabolomics and lipidomics approach. She was then invited to oversee and facilitate the laboratory analysis at Lincoln Memorial University. Since then, Kathleen has been primary instructor for the Introduction to Biological Anthropology course, the Human Osteology course, and she hopes to continue with her research, teaching, and outreach.

Her research interests include genotype-phenotype interactions, postmortem interval estimation, the human microbiome, osteology, and biological profile estimation through the lens of human variation.