

Applications of comparative genomics and data science to agricultural and clinical research

A Dissertation Presented for the
Doctor of Philosophy
Degree

The University of Tennessee, Knoxville

Katrina Anne Schlum

May 2021

© by Katrina Anne Schlum, 2021
All Rights Reserved.

*To all my mentors (Bruce Leader, Robert W. Leach, Michael R. Leuze, Barry B. Bruce
and Scott J. Emrich), who saw potential in me and invested in my success.*

Acknowledgments

I would like to express my deepest gratitude to my mentor Dr. Scott J. Emrich for his supervision, guidance and support throughout my research. My skillset and scientific view of bioinformatics has expanded because of him. I appreciate my dissertation committee members, Dr. Barry B. Bruce, Dr. Juan Luis Juat-Fuentes and Dr. Dave W. Ussery for their helpful discussions and critiques.

Abstract

The advent of inexpensive, high-throughput whole genome sequencing (WGS) technologies has led to the generation of thousands of related genomes, even from a single study. Large-scale genome analysis has resulted in hypothesis-generating approaches in the fields of clinical, human and agriculture genomics. Additionally, population-level genomic sampling has resulted in a decrease in false positives in genotype-phenotype associations and an increase in understanding of the basis of disease, antibiotic and pesticide resistance. Deeper understanding of migration, genetic divergence and evolution has also been made possible due to WGS. This research applies comparative genomics, population genomics and data science approaches to whole genome sequence data at the individual gene and genome scale to identify phylogenetic lineages, species representative sets of genomes and potential phenotype associations.

The first chapter is focused on generating representative set of bacteria genomes per a species and determining phylogenetic lineages within each species based on data-driven metrics rather than data-informed manually curated lineages. This goal was accomplished by using Mash for comparing genomes based on whole genome sequence distances, unsupervised clustering metrics for determining lineages and unsupervised learning methods for classification. I first applied these methods to over 15,000 *Escherichia coli* genomes collected from NCBI RefSeq. Second, I applied these methods to the Cyanobacteria species, *Microcystis aeruginosa*, collected from NCBI RefSeq. In chapter two, we report the results of surveying the whole genome diversity of the fall armyworm (*Spodoptera frugiperda* (J.E. Smith)) and further the understanding of the populations of fall armyworms within the United States and in Western Hemisphere. This research was accomplished by aligning, calling variants and analyzing genomic diversity based on hierarchical clustering of genomic

Mash distances, PCA and comparing nucleotide diversity and population differentiation metrics. In the final chapter of this dissertation I presented a straightforward sequence alignment, variant calling and functional annotation pipeline to identify potential resistant *ABCC2* alleles to CryF1. This approach provides a benchmark for future targeted sequencing resistant gene methodologies and the use case for gene surveillance for identifying pesticide resistant alleles.

Table of Contents

1	Introduction	1
1.1	Biology in the era of whole genome sequencing	1
1.1.1	History of bacteria genomics	2
1.1.2	History of population genomics	3
1.1.3	History of insect genomics	4
1.1.4	Fall armyworm	5
1.2	Research Contribution	7
1.2.1	Identifying phylogenetic lineages	7
1.2.2	Comparing large number of whole genomes	9
1.2.3	Identification of causal mutations sequencing-based methods	11
1.2.4	Summary	12
2	Phylogroups and beyond: Representing whole genome diversity and selecting representative set of bacteria genomes	16
2.1	Abstract	17
2.2	Introduction	17
2.3	Methods	21
2.4	Results	25
2.4.1	<i>E. coli</i> whole genome diversity	25
2.4.2	Diversity in relation to existing phylogroups	26
2.4.3	Selecting representative sets of <i>E. coli</i> genomes	27
2.4.4	Effectiveness of our representative sets of <i>E. coli</i> genomes in capturing genomic diversity	29

2.4.5	Diversity across the Cyanobacteria Phylum	30
2.4.6	Selecting representative sets for cyanobacteria genomes	31
2.4.7	Effectiveness of our representative set of <i>Microcystis aeruginosa</i> genomes in capturing genomic diversity	31
	Appendix: Tables and Figures	36
3	Whole genome comparisons reveal panmixia among fall armyworm (<i>Spodoptera frugiperda</i>) from diverse locations	48
3.1	Abstract	49
3.2	Introduction	50
3.3	Methods	52
3.3.1	Samples and strain typing	52
3.3.2	DNA extraction, library preparation, and sequencing	53
3.3.3	Filtering and mapping	53
3.3.4	Phylogenomics	54
3.3.5	SNP diversity	54
3.3.6	Mitochondrial genome assembly	55
3.3.7	Structure	55
3.4	Results	56
3.4.1	Host strain typing	56
3.4.2	Whole genome diversity, nucleotide diversity and differentiation . . .	56
3.4.3	Gene flow between overwintering populations in the USA	58
3.4.4	Laboratory vs. field diversity	59
3.4.5	Diversity across the mitochondrial genome	59
3.5	Discussion	60
3.6	Conclusions	64
	Appendix: Tables and Figures	65
4	Targeted genotyping reveals common resistance alleles to Cry1F corn in fall armyworm from Puerto Rico and Brazil	75
4.1	Introduction	76

4.2	Materials and Methods	77
4.2.1	Sample Collection and genomic DNA extraction	77
4.2.2	TaqMan genotyping for <i>SfABCC2mut1</i>	78
4.2.3	Targeted DNA amplification and sequencing	79
4.2.4	Genotype assignment for SfABCC2mut1 for targeted-sequencing k-mer based approach	79
4.2.5	Alignment, variant calling and variant analysis	79
4.3	Results	81
4.3.1	Genotyping results using Taqman	81
4.3.2	Genotyping results for the 2bp INDEL using targeted sequencing . .	81
4.3.3	Variant analysis pipeline and genotyping	82
4.3.4	ABCC2 sequence variation	83
4.4	Discussion	84
4.5	Conclusions	86
	Appendix: Tables and Figures	87
5	Conclusions	95
	Bibliography	97
	Vita	119

List of Tables

2.1	Optimal clusters for 15,564 <i>E. coli</i> genomes across four different methods	37
2.2	Optimal cluster counts across six different Cyanobacteria species on three different methods across two different metrics	37
2.3	Phylogroup distribution across two different methods applied on 15,501 <i>E.coli</i> genomes	37
2.4	Evaluation metrics based on the number of genomes and summary statistics of Mash distances in full and reduced genome sets	38
3.1	List of <i>Spodoptera frugiperda</i> samples sequenced and used in this work. Shown is information on the location where collected (or facility for laboratory reference strains), Cry1F resistance phenotype (u for unknown, s for susceptible, r for resistant), collection (F for field-collected, L for lab-reared), host strain based on Tpi183 marker (corn, rice, hybrid or unknown), and the stage used for genomic DNA purification.	66
3.2	Pairwise Fst values across USA populations surveyed (51 samples total). FL = Florida, PR = Puerto Rico, MD = Maryland, MN = Minnesota, SC = South Carolina, TN = Tennessee, TX = Texas, MS = Mississippi, NC = North Carolina. NA was used when the samples overlap in the defined populations.	67
3.3	Average Mash distances based on country of origin based on all 55 samples.	67

4.1	Concordance rate between taqman assay vs. K-mer based genotype results for the <i>SfABCC2mut1</i> GC insertion represented by the allele r. The fraction of samples with the same genotype predicted are listed with the percentage in parentheses.	88
4.2	Non-synonymous mutations identified by SnpEFF across 41 <i>S. frugiperda</i> samples within 1,331 ABCC2 bi-allelic variants.	89

List of Figures

1.1	Number of bacterial genomes sequenced each year and submitted to NCBI. Source Genbank bacteria file assembly_summary.txt downloaded on October 15, 2020.	14
1.2	Conceptual framework of population genomics approaches. (copied from [140])	15
2.1	Overall workflow for selecting a representative set of genomes. The three unsupervised methods used were hierarchical clustering, k-means clustering on raw and PC's from PcoA and a custom random centroid approach. The four methods used for determing the number of clusters including purity, completeness, within sum of squared distances and silhoutte index. The software packages used are in bold.	39
2.2	Intraphylogroup Mash distance vs. frequency of occurrences across 15,649 <i>E. coli</i> genomes in panel A and 329 representative set of <i>E. coli</i> genomes in panel B.	40
2.3	Frequency of lineages represented across each representative set of <i>M. aeruginosa</i> genomes where random set of 18 genomes represented in A.), random centroid set of genomes in B.) and PcoA + k-means set in C.). . . .	41
2.4	PCoA + k-means clustering on 15,501 <i>E. coli</i> genomes	41
2.5	MLST phylogroup label versus closest genome-inferred label using Mash on 15,649 <i>E. coli</i> genomes in panel A, a reduced set of 329 genomes for random centroid method in B, a reduced set of 120 genomes for hierarchical clustering method in panel C, and 105 representative genome set from k-means clustering in D. The color represents the normalized count of genomes classified.	42

2.6	Phylogroup counts of genomes within each <i>E. coli</i> representative genome set generated by the random centroid approach (B.), HC (C.) and the PcoA+k-means clustering method (D.) compared against the ECOR set (A.)	43
2.7	Gene diversity profiles generated by Roary [127] across two previous representative sets [5, 135] and our three more inclusive representative sets. .	44
2.8	Intraspecies Mash distance across six cyanobacteria species	45
2.9	Mash distance tree of 120 <i>Microcystis aeruginosa</i> (A.) and Mash distance tree of 18 representative <i>Microcystis aeruginosa</i> genomes (B.) with branches colored by genome size range in Mbp as seen in key	46
2.10	Gene diversity profiles generated by Roary [127] for <i>Microcystis aeruginosa</i> (Non-reduced set) set, randomly selected set of genomes, and reduced representative set generated in this research by Hierarchical clustering, PcoA + k-means clustering, where numbers in parenthesis indicate genome count .	47
3.1	Principal component analysis (PCA) plot of 51 <i>Spodoptera frugiperda</i> samples (3 samples removed due to missing rate $\geq 50\%$) based on 2,762,958 SNPs. Country of origin and host strain as determined by the nuclear Tpi marker are visualized by color and shape, respectively, as shown in the legend. . . .	68
3.2	Dissimilarity heatmap of Mash distances across all 55 <i>Spodoptera frugiperda</i> samples, colored by country of origin as indicated. Two outgroups (<i>Spodoptera litura</i> and <i>Spodoptera exigua</i>) were added. Samples were named as described in Materials and Methods.	69
3.3	Complete clustering on Mash distances across 55 samples. Description of data: Mash distance based dendrogram on all 55 samples colored by host strain and rooted with <i>Spodoptera litura</i>	70
3.4	Nucleotide diversity box plots between 14 C-strain Florida (FL) vs. 2 R-strain Texas (TX) samples.	71
3.5	Mitochondrial haplotype network of all 54 <i>S. frugiperda</i> mitochondrial chromosome level assembled individuals, colored by country of origin.	72

3.6	UPGMA distance-based tree on biallelic variant SNPs in 54 mitochondrial chromosome level assembled individuals rooted with <i>Spodoptera litura</i> as an outgroup.	73
3.7	Amplification of a fragment of the wsp gene (610 bp) from <i>S. frugiperda</i> moths. A positive control (<i>Wolbachia pipiens</i>) was also included (+C). Samples shown ArgXXr2 (lane 1), PueSIu1 (lane 2), USAFLu8 (lane 3), USAFLu10 (lane 4), USAFLu12 (lane 5), USAMDu2 (lane 6), USATXu5 (lane 7) and <i>W. pipiens</i> (lane 8). Sequenced amplicons matched with 99.8% identity to <i>Wolbachia</i> endosymbiont of <i>Nasonia vitripennis</i> seqvar1 Nvit outer surface protein (wsp) gene (GenBank accession number DQ380865).	74
4.1	Pipeline of overall workflow from raw Fastq targeted sequencing reads to population genetics and phylo-genetic analyses based on bi-allelic SNP variant calling.	89
4.2	Schematic representation (not drawn to scale) of SfABCC2 two dimensional topology on the cell membrane and location of detected nonsynonymous mutations in sequenced <i>S. frugiperda</i> samples from Puerto Rico. Distinct colors as shown in the figure, indicate the number of individuals carrying each mutation nonsense mutations are shown in a bold box. Amino acid numbers are indicated, the ABCC2mut1 allele is represented as S748*. TD = transmembrane domain	90
4.3	PCA projection plot of 1,333 bi-allelic variants found within ABCC2 of 41 <i>S. frugiperda</i> samples and colored by 2 GC insertion known to be associated with resistance (<i>SfABCC2mut1</i>) in panel A and PCA colored by sample location collection in panel B. The <i>SfABCC2mut1</i> was determined by k-mer based genotyping approach as described in the Methods.	91
4.4	Manhattan plot of SNPs associated with PC2 on filtered bi-allelic dataset (1,333).	92

4.5	Genetic structure inferred by FastStructure on 82 haplotypes representing 41 samples at k =4 where samples in pink text indicate the <i>SfABCC2mut1</i> is present in the sample haplotype represented.	93
4.6	PCA projection plot of 1,332 bi-allelic variants found within ABCC2 of 41 <i>S. frugiperda</i> samples with allele associated with 2 bp GC insertion removed. The PCA projection is colored by the samples who have the 2 bp GC insertion known to be associated with resistance (<i>SfABCC2mut1</i>) in panel A and PCA colored by sample location collection in panel B. The <i>SfABCC2mut1</i> was determined by k-mer based genotyping approach as described in the Methods.	94

Chapter 1

Introduction

1.1 Biology in the era of whole genome sequencing

Next-generation sequencing (NGS) has enabled high throughput nucleotide sequencing at a low cost and population-scale. NGS improved upon Sanger sequencing by removing the requirement of bacterial cloning while enabling massively parallel sequencing reactions [130]. With these major improvements, scientists are now able to sequence whole genomes, the complete set of genetic information found in an organism or cell, at low cost and much shorter time periods for data generation. As of October 2020, there are currently 1,432,874,252 whole genome sequences (WGS) available in the National Center for Biotechnology Information (NCBI) GenBank [123]. This number of publicly available WGS is three thousand times more data compared to almost two decades ago when only 434,224 sequences were available in October 2002. The data in GenBank has grown at an exponential rate and has doubled roughly every 18 months [12]. With this increase in throughput, a single study can now contain between hundreds to thousands of samples [42], e.g., the 1,000 genomes project, the Human Microbiome Project and retrospective epidemiology studies [102].

The rapid growth of WGS data has revolutionized modern biology. For example, *de novo* sequencing with assembly can generate a new genome from pretty much any organism of interest [155]. Inexpensive sequencing has also enabled cataloging and comparisons of genetic variation of many organisms including single nucleotide polymorphisms (SNPs), structural variations, and copy number variations [130]. RNA-Seq can identify RNA splicing sites,

gene fusion, mutations and differential gene expression [58] while epigenome sequencing can identify DNA methylation sites and protein-DNA interactions. Finally, NGS has been used to study microbial ecology based on sequencing environmental samples [130].

This dissertation will address the second application above of NGS: measuring genetic variation of an organism at the whole genome and gene level in relation to an existing reference genome and quantification of SNPs in the insect pest fall armyworm (*Spodoptera frugiperda*) (Chapter 3 and 4) as well as genetic distance and gene family counts in pathogenic bacteria *E. coli* and the cyanobacteria species *Microcystis aeruginosa* (Chapter 2). Genomics research has shifted to genome analysis [81], the contribution of this shift impact the research focus presented in this thesis for both bacterial and eukaryotic data.

1.1.1 History of bacteria genomics

The first bacteria genome was sequenced in 1995 (*Haemophilus influenzae*). There has been a 500-fold increase in the number of bacteria genomes sequenced from 1999 to 2010 and 7-fold increase over the last 5 years as seen in Figure 1.1. With this increase in the number of genomes sequenced, now multiple genomes within the same species can be compared. From multiple genome comparisons, most bacterial genomes have been found to have an open system with constant gene loss and acquisition. Such plasticity of the bacteria genome is maintained by foreign DNA and mobile genetic elements including transposons, bacteriophages, and plasmids [98]. On the other hand, some species of bacteria appear to have closed genomes (i.e. *Bacillus anthracis*) [98]. The total number of genes found in a single species is often referred to as the “pangenome.” The open or closed property is dependent on the species’s capacity to acquire exogenous DNA, machinery to use the DNA, and have a large amount of rRNA to translate the DNA to RNA [144]. Additionally, bacteria pangenome sizes have been shown to depend on the lifestyle and niche of the bacteria in question. For example, sympatric species, i.e., species that live in the same location but are evolved from an ancestral species, typically have large and open pangenomes with high rates of horizontal gene transfer and ribosomal operons [144].

The set of genes found in $\geq 95\%$ of strains considered is referred to as the core genome. The core genome primarily consists of housekeeping genes involved in DNA replication,

transcription, translation, and cell envelope maintenance [128]. In contrast, the remaining genes tend to mix laterally [127]. This lateral gene transfer (LGT) can result in new trait combinations in strains (and also species). It has been speculated that each accessory genome is a reservoir of functionality [149]. LGT is important for maintaining the genetic integrity of a bacteria species. Unlike eukaryotes, prokaryotes reproduce asexually from ancestral cells and accumulate mutations that have selective advantages. LGT therefore provides an alternative mechanism for prokaryotes to adapt and advantageous genes with such advantages to spread [149].

1.1.2 History of population genomics

Historically, only one individual per species was sequenced and it represented the species as the “reference” genome. Phylogenetic analyses were primarily based on core genes in bacteria (see above) and either mitochondrial or nuclear genes in eukaryotic species (e.g., [161]). Now with time and cost constraints removed because of fast and inexpensive NGS, large-scale comparison of DNA sequences across populations can be performed to increase our knowledge on the roles of evolutionary processes (i.e. mutation, genetic drift, gene flow, natural selection, speciation and admixture) across and within genomes and populations. [130].

Before NGS, microsatellites were used to infer population structure in order to answer fundamental and applied research questions. After NGS, thousands of single nucleotide polymorphisms (SNPs) can now be used to answer the same questions instead. For example, consider the conceptual framework for population genomics established in 2010 by Hohenlohe et. al [140]. This conceptual framework is outlined in Fig. 1.2 where individuals from different phenotypes, populations or environments are sampled followed by genotyping. Next, outlier loci and genomic regions associated with studied trait or environmental niche are identified. Finally, evolutionary or demographic parameters can be modeled without the outlier loci and the cause of outlier loci can be further tested (i.e. selection, evolutionary inferences) [140]. With increased power from larger amounts of data, we are now able to identify DNA loci related to fitness, adaptation, and important variation, e.g., population persistence and growth [15].

Accuracy has also increased due to innovative software, statistical approaches, and models to identify meaningful results from NGS datasets such as the identification of candidate adaptive loci by detecting sweeps, local adaptation and/or speciation, and adaptive introgression and associated loci [15]. Population genomics can also address the association with environmental variation (landscape genomics) and the detection of signatures of polygenic adaptation.

1.1.3 History of insect genomics

During the last decade, the field of insect population genetics has similarly shifted from analyzing a couple of genes/microsatellites to more of a population genomics approach [88]. For example, initiatives such as the USDA-led 'ik5' initiative [69] have championed insect genome sequencing and more contemporary insect genome projects have involved both genome assembly and large-scale resequencing [88, 87]. As of May 2019, there are 401 insect genomes assembled at the National Center for Biotechnology Information (NCBI). The order Diptera (true flies) have the most publicly available assembled sequenced insect genomes to date [171]. This bias in genome sequencing of true flies is a consequence of the importance of *Drosophila melanogaster* as a model organism and the role of mosquitoes in transmission of vector-borne diseases. Because of fly genomics and comparisons within and between different populations, landmark research in gene regulation, repair, development, neurobiology, sex determination, insecticide resistance, to name only a few areas, has been performed [171].

Despite these clear advances, the genome sequencing of insect pests is lagging even though these genomes will be highly important for alternative pest control practices [21, 18] – and similar to mosquito genomes' role in improved control of malaria. The first agricultural pest was sequenced in 2008, *Tribolium castaneum* (red flour beetle), and as of May 2019, a little over two dozen agricultural pests have been sequenced. Advances in new methods for integrated pest management (IPM) are now possible with functional, population and comparative genomic research. For example, the characterization of the molecular diversity of an insect population can generate gene-targeted alternative interventions such as RNA-interference-based pest control. Genome-wide association studies (GWAS) have also been

used to identify mechanisms behind insecticide resistance and host adaptation [168, 10, 41, 121].

Despite the rapid development in the field of insect genomics, there are some main challenges that remain. It is still difficult to generate for most insects a high-quality genome assembly using inexpensive Illumina HiSeq read data. Further, with draft assemblies, there are errors in sequencing and assembly that can lead to inaccurate genome alignments [161]. The first challenge, lack of high-quality agricultural genomes, can be addressed by newer long read sequencing, HiC scaffolding and other scaffolding methods. For example, BioNano physical maps have been used to generate chromosome-level assemblies for *T. castaneum* (red flour beetle) and *Cydia pomonella* (codling moth) [62, 164]. The second challenge of in-depth analysis of genomic data—but also an opportunity tackled in this thesis—can be addressed by exploring non-traditional data driven methods for identifying patterns in populations of insect genomes of interest, i.e., associating diversity (SNPs) with pesticide resistance in field-obtained samples.

1.1.4 Fall armyworm

The order Lepidoptera has the second most assembled insect genomes at almost 100 in NCBI [88]. Within the order Lepidoptera, there is the *genus Spodoptera* with 25 species currently described. One of these species of great agricultural interest is the fall armyworm, *Spodoptera frugiperda*, which was first described as *Phaleana frugiperda* in 1797 [61]. *Spodoptera frugiperda* is an economically important species in United States, South America and Africa because it is such a crop damaging pest [8]. In Brazil, it is estimated to cause over 400 million in annual losses [92]. *S. frugiperda* can feed on over 80 plants including maize, rice, sorghum, sugarcane, cabbage, beet, cotton with preference for pasture grasses [111].

The fall armyworm has been predominately controlled in the Western Hemisphere by synthetic insecticides and genetically engineered (GE) crops producing *Bacillus thuringiensis* (Bt) insecticidal proteins. This reliance on insecticides has resulted in insecticide resistance in many regions including Puerto Rico, several Mexican states and in the United States [57]. Currently, the fall armyworm is resistant to at least 29 insecticidal active ingredients in the Americas [57]. Although there are currently three identified mutations related to Cry1F

resistance reported in the fall armyworm, it is unclear if the resistance alleles occurred naturally or are the result of selection pressure to pest control interventions. As such the genetic diversity, migration patterns, and reproduction preferences of this agricultural pest are also important to stop the spread of resistant alleles to less developed areas of the world.

There are two host strains used to describe *S. frugiperda*, the corn strain (C-strain) that is known to preferably feed on maize and sorghum and the rice strain (R-strain) known to preferably feed on rice and turfgrass. The two host strains are not only differentiated by host preference but also allelic differences found in the mitochondria gene *COI* and sex linked gene *Tpi*. The two host strains have been further differentiated by ratios of allelic differences within the *COI* gene. These two haplotypes are called the Florida haplotype profile representing the Florida and Caribbean region and the Texas haplotype profile representing the Texas through Central America to Argentina region. These two haplotype profiles have different migration paths. The Florida haplotype migrates north upward to Canada staying east of the Appalachian mountains every summer while the Texas haplotype stays west of the Appalachian mountains when they travel north up to Canada in the summer[111].

The first two fall armyworms sequenced and made publicly available originated from Guadeloupe and Florida, USA [52]. With the largest study to date on fall armyworm genome comparison of United States of America (USA) samples consisting of only two states in the USA (Florida and Mississippi) [159], there is a gap in our knowledge of our understanding of fall armyworm genomics from individuals originating from the USA. This research helps overcome this limitation in knowledge by surveying over eight states in the USA (Texas, Tennessee, North Carolina, South Carolina, Maryland, Minnesota, Mississippi, and Florida). Specifically, our research presented in Chapter 2 surveys 55 fall armyworms collected from the Western and Africa at the whole genome level to evaluate if genome wide level differences are observed between corn and rice host strains. This research will help further the understanding of the global migration of the fall armyworm and how host strain may be impacting reproductive isolation, while also tying these patterns to Cry1F resistance.

1.2 Research Contribution

1.2.1 Identifying phylogenetic lineages

Historically, multilocus sequence typing (MLST) has been extensively applied to bacteria genomes. MLST involves comparing sequences of at least 7 house keeping genes and constructing lineages based on phylogeny constructions [95]. In *E. coli*, there are five main MLSTs recognized as A, B1, B2, D, and E [85]. As well as minor MLSTs as C, F, G and cryptic (clade I-V) [160]. These MLST groups have been verified to match phylogenetic signal via correlation of clades in phylogenetic tree based on alignment of core genes to MLST labels [76]. Although the majority of these MLST groupings correlate with whole genome single nucleotide polymorphisms (SNP) and core genome based trees, there are exceptions observed in *E. coli*. Specifically, admixture has been observed between A and B1 and between A, B and D [172]. Furthermore, recent research has identified 14 phylogenetic lineages in *E. coli* [5] based on an analysis of 100,000 *E. coli* genomes which is five more lineages in *E. coli* than previous MLST based analysis has shown. The current MLST schemas do not account for admixture and/or new mutations in housekeeping genes that may result in a bacteria genome where no MLST can be assigned. Our research in Chapter 2 overcomes this limitation of MLST by demonstrating the use case of applying four different clustering methods (hierarchical clustering , K-means clustering , Principal component of analysis (PcoA) combined with k-means clustering, and a custom random centroid method) to genome by genome distance matrices in order MLST type *E. coli* genomes. With clustering, we are able to identify structure in the data based on a similarity criterion and each group/cluster has maximal homogeneity within the cluster and maximal heterogeneity between clusters. We show that hierarchical clustering recovers the five main phylogenetic lineages of *E. coli* while k-means or the custom random centroid approach resulted in more fine grained phylogenetic groups.

Furthermore, there are many bacteria species such as *Microcystis aeruginosa* where MLST schemas are not as established as *E. coli* and/or not established yet thus there is constant need to continue updating databases for MLST schemas per bacteria species and increase the number of phylogenetic lineages within a MLST schema if the current MLST

schema does not cover the entire diversity of bacteria species studied to date. This continuous need to establish and/or update MLST schemas is unsustainable as the number of bacteria genomes increases. Another challenge with identifying phylogenetic lineages within a bacteria species is the uneven rates of sexual exchange within and between phylogenetic lineages. As the number of bacteria genomes analyzed increased, it has become clear that there were differences in not only genetic content but also rate of sexual exchange of DNA where some bacteria such as *Thermotoga maritima* showed HGT between Archaea and Bacteria while *E. coli* showed high rates of HGT within the *E. coli* species. This difference in sexual exchange of DNA (recombination) results in differences in ancestral history of a bacteria species. Some bacteria lineages within a species show evidence for no sexual exchange of DNA between or within different species while some have bacteria lineages show high rates of recombination.

We present one solution in Chapter 2 to overcome the dependency on establishing a MLST schema based on house keeping gene allelic differences. Our solution as presented above is to apply different clustering methods to a genome by genome distance matrix which represents the entire set of bacteria genomes sequenced to date. Based on the highest silhouette index, the number of phylogenetic lineages within the dataset (clusters) can be determined based on how similar clusters are to its own cluster compared to other clusters. This data driven metric enables the number phylogenetic lineages to be defined by a non-bias approach rather than bias human determined approach. To demonstrate this use case, we applied four different clustering methods to the cyanobacteria species, *Microcystis aeruginosa*, since no MLST schema is established for this species and identified at least 7 lineages. The advantages of applying clustering to genome by genome distances generated by Mash include: 1. computational efficiency, 2. ability to classify and account for "admixture" genomes, 3. remove the need to continue updating a MLST schema as more genomes are sequenced and 4. phylogenetic lineages are determined by data rather than human determination. There are also limitations of this work including no "optimum" way to determine the correct number of clusters as this is an "open" problem [80] and obtaining a different count of the optimal number of clusters depending on clustering algorithm and/or clustering metric applied.

With this clustering solution to identifying phylogenetic lineages we present in this thesis, researchers are now able to identify relationships between the presence of toxic genes and phylogenetic lineages without the need to establish a MLST schema and/or align core genes/whole genomes, generate distance metric between bacteria species and perform tree building in order to generate a phylogenetic tree to identify phylogenetic lineages. Since both bacteria species *E. coli* and *Microcystis aeruginosa* are a threat to human health [? 31] with toxic (pathogenic) and non-toxic (commensal) strains present within each species, this research presented in Chapter 2 provides a framework for clinical researchers to use when analyzing relationships between phylogenetic lineages and toxic genes.

1.2.2 Comparing large number of whole genomes

Since Charles Darwin’s time, it has been important and beneficial to compare species to each other and understand ancestral relationships. Comparing genomes involves alignments comparing individual/groups of DNA or amino acid sequences, and these methods assume every sequence can be categorized as conserved/similar or non-conserved/dissimilar [174].

One of the first sequence similarity methods was FASTP. The number of alignment-based methods has continued to grow from protein-sequence aligners to whole-genome sequence aligners which include progressive Mauve, BLASTZ, TBA. Alignment free methods have also been around since roughly the same time with the first method comparing word counts of coding sequences in different eukaryotic genes [143]. Alignment-free methods have expanded to Maximal Unique Match (mer) (MUMmer) and FAsANI [39, 71]. FAsANI compare genomes based on percentage of similarity of coding regions between two genomes (Average Nucleotide Identity (ANI)). Another popular alignment-free approach is Mash which generates a sketch of random sets of k-mers between two genomes and estimates genomic distance based on Jaccard distance between both hash sets of k-mers [125].

Current comparative genomics approaches focus on shared features and therefore may omit important genomic determinants that could better characterize and discriminate sub-populations and/or phenotypes, especially in bacterial genomes. For example, the accessory genome has been shown to be important for phenotypes such as antibiotic resistance, adaptation to specific environments and colonization of different hosts [42]. Given the

importance of the accessory genome in pathogenicity including antibiotic resistant genes and virulent genes, K-mer alignment-free based methodologies could be useful since they estimate the proportion of shared features (i.e. conserved genes). Work presented in this thesis applies K-mer alignment-free methods to both understand of the ancestral phylogenetic lineages (Chapter 2) as well as host strain differentiation (Chapter 3). Specifically, we show that with alignment-free methods, there is little to no differentiation of the fall armyworm at the whole genome level based on host strain and the fall armyworm populations originating from the USA have no population structure and are panmictic. Furthermore, we show how k-mer alignment-free methods (Mash) can be used to generate genome by genome distance matrices in bacteria and used as input to clustering methods in order to identify phylogenetic lineages. This large scale application to bacteria genome comparison demonstrates how k-mer based comparison provides enough resolution for identifying phylogenetic lineages compared to the traditional SNP based approach on conserved genes. Future research can now use k-mer alignment-free methods combined with clustering algorithms to identify phylogenetic lineages as demonstrated in our work (Chapter 2).

As the number of genomes sequenced within a species increases, it is not feasible to perform large number of genome comparisons for every new analysis. One solution we present in this thesis (Chapter 2) to large genome comparisons is to generate a representative set of genomes per bacteria species which represents the maximum diversity of the species and minimizes the number of genomes within the representative genome set. A representative set of genomes enables quick and efficient comparisons to assign a newly sequenced genome a phylogenetic lineage. Representative sets of genomes have been predominately generated via hierarchical clustering or Gaussian mixture model (GMM) approach of genome distance metrics (genomic similarity scores (GSS)[101], BlastP score ratios (BSRs)[141] or average nucleotide identity (ANI)). Yet, all these prior methods result in a different number of representative genome set. Thus, it is not clear what method results in a representative genome set that captures the known diversity of the species of interest. Our research in Chapter 2 presents a comprehensive comparison of different representative genome sets generated for *E. coli* and *Microcystis aeruginosa* based on taking the centroid of the clusters generated by four different clustering algorithms as query and database as the

cluster and running a new open sourceware software package called referenceseeker to select representative genomes per cluster per bacteria species. Referenceseeker selects genomes from the database provided which are closest to the genome supplied by the query based on Mash distances and conserved gene score [148]. We show that PcoA + k-means clustering generates the least diverse representative genome set while hierarchical clustering and custom random centroid method generated the most diverse representative genome sets.

Additionally, a representative set of genomes is beneficial in genome wide association studies (GWAS) where a complete set of genomes are scanned for genetic markers that are associated with a particular disease/phenotype. Because of the advent of WGS and focus on microbes, microbial genome-wide association studies (mGWAS) are now emerging as an exciting field of research with adaptation from human GWAS methods [145]. The goal of mGWAS is to help understand how variations in microbial genomes affect host or pathogen phenotypes such as antibiotic resistance, virulence and host specificity. Because performing a mGWAS on multiple clonal isolates can reduce the amount of power to detect variants, reducing the extent of clones has a substantial impact. Furthermore, the strong population structure that is maintained in some species of bacteria (i.e. *E. coli*) makes it difficult to identify features related to phenotype and not associated strains (lineage effects) [84]. To circumvent the possibility of identifying lineage differences rather than phenotype differences, lineage effects have been incorporated in association models as covariate and have improved accuracy [84]. We anticipate our representative set method could have similar impact on future mGWAS studies.

1.2.3 Identification of causal mutations sequencing-based methods

Genome wide association studies, like the mGWAS studies discussed above, are geared towards alleles that are single-gene and homozygous variants because dominant and/or semi-dominant alleles are more challenging to identify in WGS data [154]. One approach to overcome this limitation is to boost power by sequencing more overall samples of interest at a similar cost using targeted sequencing, which reduces sequencing cost, time, provides deeper coverage, can detect and quantify low-frequency variants [53]. Targeted sequencing aids in rapid diagnosis of many genetic diseases and important phenotypes (i.e. pesticide resistance)

in the field of entomology. Previous research has shown that monitoring via targeted sequencing can result in identification of new mutations due to reducing false negatives and positives that have been observed in prior sequencing approaches of pharmacogenes in human blood samples [56]. Additionally, standards in filtering targeted sequencing variant calls have enabled higher precision and accuracy of identifying novel mutations [151] but uneven/low sequence coverage, absent data and ambiguous alleles still persist in targeted sequencing data.

We demonstrate the power of targeted sequencing in Chapter 4 of this thesis where we targeted sequenced and analyzed 45 fall armyworm *ABCC2* samples. With this approach, we increased allele depth and sequence coverage of *ABCC2* alleles compared to the coverage we had observed in the whole genome sequencing samples from the dataset analyzed in Chapter 3. Furthermore, targeted sequence of *ABCC2* gene resulted in identification of known and discovery of candidate resistance alleles to Cry1F in the fall armyworm. This approach also resulted in identifying mutations potentially associated with resistance to Cry1F in Brazil among samples from Puerto Rico. This provides the foundational research for future work in monitoring the presence of Puerto Rican resistant alleles to Cry1F in Brazil and vice versa.

1.2.4 Summary

The research presented in this thesis overcomes the limitations of comparing large number of genomes via the software tool Mash (Chapter 2 and 3) and generating representative sets of bacteria genomes which maximize the genome diversity of species and minimizes the number of genomes compared via unsupervised clustering algorithms (Chapter 2). Our research also overcomes the limitation of generating, updating and maintaining MLST schemas by providing the application of clustering genome by genome distance matrices and determining the number of MLSTs (phylogenetic lineages) by using the clustering metric, Silhouette index, to determine the number of clusters (phylogenetic lineages) within a bacteria species (Chapter 2). Finally, this research demonstrates the advantages of monitoring resistant alleles in fall armyworm populations via targeted sequencing of known resistant genes (*ABCC2*) which results in identification of known resistant alleles, potentially novel resistant alleles and ability to show evidence for gene flow of resistant alleles from different populations

of fall armyworms. Without the power of whole genome and targeted sequencing, k-mer alignment free methods (Mash), and clustering algorithms, the research presented here would not be possible. This thesis demonstrates that applied comparative genomics research is dependent on technology improvements in genome/gene sequencing and development of data science algorithms and software.

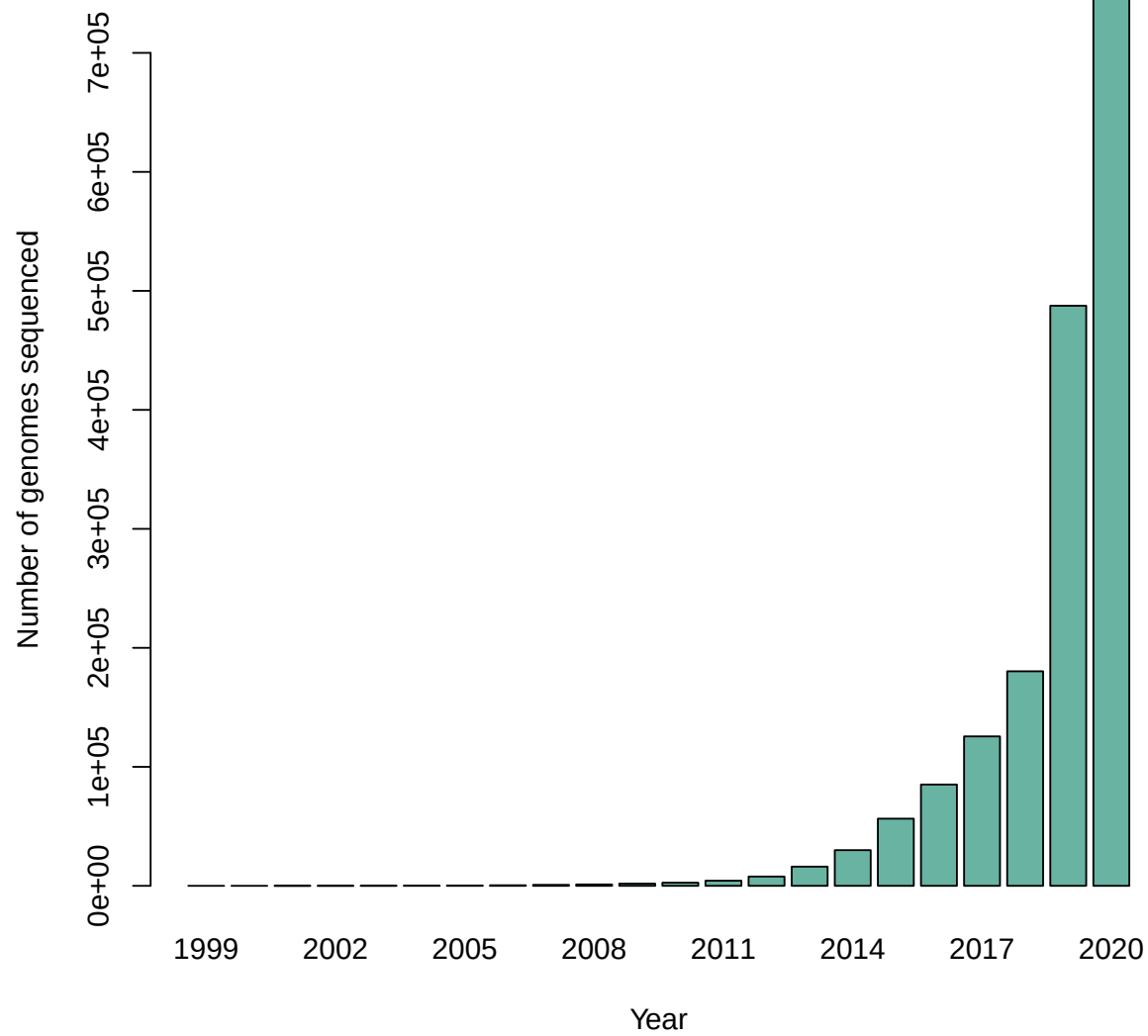


Figure 1.1: Number of bacterial genomes sequenced each year and submitted to NCBI. Source Genbank bacteria file assembly_summary.txt downloaded on October 15, 2020.

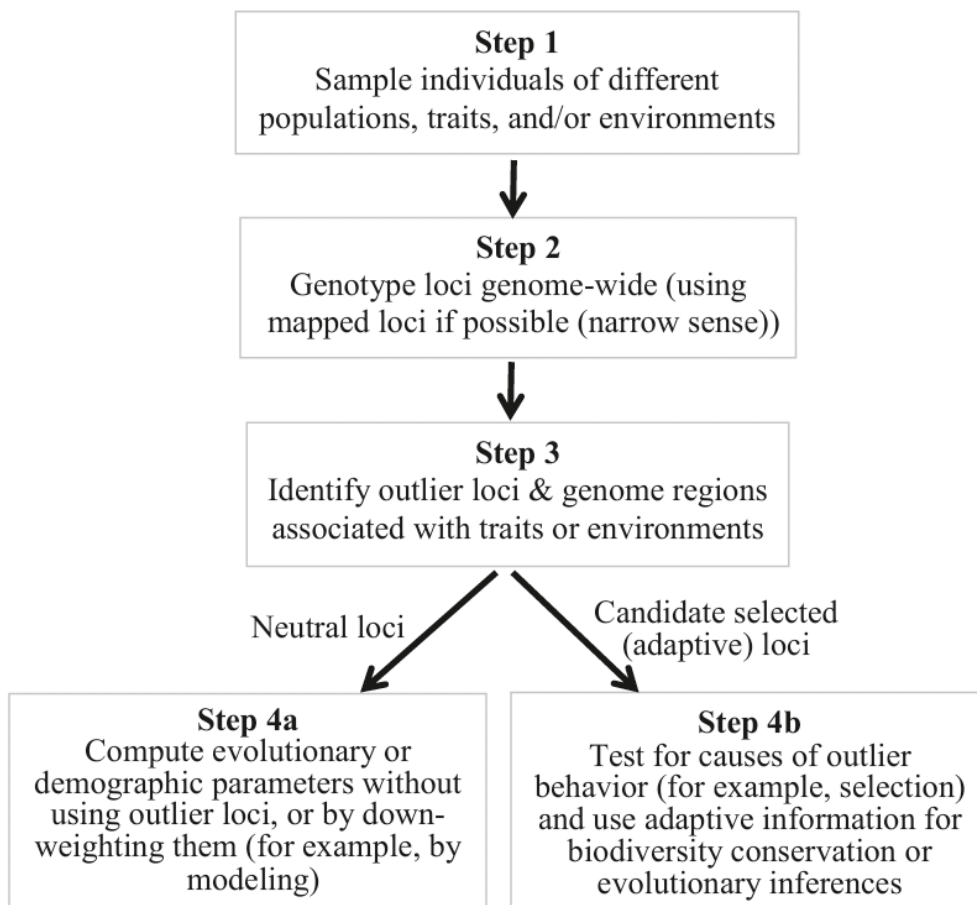


Figure 1.2: Conceptual framework of population genomics approaches. (copied from [140])

Chapter 2

Phylogroups and beyond:

Representing whole genome diversity
and selecting representative set of
bacteria genomes

This following chapter is a revised version of work in submission for publication: Under submission, Schlum, K.A., Emrich, S.J., Ussery, D.W.

My primary contributions to this paper include: (i) pipeline (ii) bionformatics data analyses, (iii) results interpretation (iv) generated results (v) writing of manuscript.

2.1 Abstract

The advent of inexpensive whole genome sequencing has led to the generation of thousands of related genomes. The best studied and most sequenced bacteria species to date is *Escherichia coli* (*E. coli*), and there are now over 17,900 assembled *E. coli* genomes publicly available in NCBI GenBank. Because there is a strong public health interest in sequencing genomes of pathogenic bacteria and genomes from disease outbreaks, a large number of clonal or nearly clonal genomes are sequenced that can bias comparative genomics analyses and genome-wide association studies. We applied four different data-driven approaches (k-means, Principal Coordinate Analysis (PcoA) + k-means, hierarchical clustering, and custom random centroid method) to identify "phylogenetic" groups and representative genome sets from each lineage group identified within a species. We found that across all methods, based on four different evaluation metrics, the optimal number of clusters (phylogenetic groups) is largely dependent on the metric and clustering method applied. We generate species representative genome sets for two different bacterial species: *E. coli* and *Microcystis aeruginosa*. The framework presented here results in an unbiased subset of genomes that is smaller in size compared to the original genome set, captures the most genome diversity of original genome set compared to other subsets, and has low redundancy among the representative genomes. The species representative genome sets generated in this work are potentially better suited for important analyses such as associating genotypes/genomic background to specific phenotypes.

2.2 Introduction

E. coli is a gram-negative anaerobe found both inside animals and in the broader environment [160]. The genome of *E. coli* is extremely plastic; only about half of the typical number of

genes per genome ($\sim 3,000$) are found in most other strains of this species and its “pan-genome”, i.e. the genes observed across the species, is at least tens of thousands of genes [156]. Currently, there are five largely accepted phylogroups – or phylogenetic lineages – in *E. coli* based on allelic differences found across seven house keeping genes. These main distinctions, which are most commonly called Multilocus sequence types (MLSTs), are A, B1, B2, D, and E [85]. There have also been four minor phylogroups reported: C, F, G and cryptic (clade I-V) [160]. Members of the cryptic lineages (clade I-V) are largely understudied, but are still clearly different on a genomic level from the current major and minor characterized *E. coli* groups [162].

Because the historical MLST lineages correlate well with more recent whole genome SNP and core gene SNP based trees [76], *in silico* genome-based “typing” is now performed using multiple approaches. For example, clereonTyping [11] uses a database of 83 *E. coli* genomes to represent seven of the characterized MLSTs in combination with genome sketching (Mash). A phylotype is determined based on the phylogroup of the closest genome in the database to the query genome using mash. On a test set of 334 *E. coli* genomes, the predicted phylogroup had a 99.4% concordance rate with its house keeping gene-based assignment [11]. Although GenBank accessions are not provided for this clereonTyping representative set, the majority of genomes were obtained from a prior *E. coli* reference set (called ECOR) and the overall distribution per phylogroup can be inferred from a dendrogram in their accompanying paper: 9 A’s, 7 B1’s, 14 B2’s, 7 C’s, 6 *Escherichia* clade I’s, 6 *Escherichia* encompassing clade II to V, 7 D’s, 7 E’s, 6 *Escherichia albertii*, 4 *Escherichia fergusonii* and 9 F’s. The original ECOR set was established in 1984 by Ochman and Selander and remains the most widely used representative set of *E. coli* genomes to date [135]. The initial collection of *E. coli* genomes in this ECOR set was collected from thousands of archived isolates from around the world. Although this ECOR set has largely scaled well, the cryptic group (Clade I-V) is not well represented because the original ECOR collection did not include samples *ex vivo* [163]. Additionally, the membership of phylogroups A, B1, and B2 in this collection is likely over-represented, especially with respect to the potential phylogroups of C, F, and G that were not included in early isolate collections [166, 135]. Further, members of phylogroup E have been shown to cluster inconsistently [172]. For these reasons there is still a need to

define scalable, repeatable, and balanced representative sets of *E. coli* genomes that evenly represent genomic diversity and can include known cryptic lineages.

Representative sets of genomes have been predominately generated using clustering methods, specifically hierarchical clustering, in part because these methods have found associations between pathogens and clinically-relevant traits such as host and phenotypes of interest including virulence and antimicrobial resistance [83]. Clustering is also important for identifying functions of unknown genes, disease profiles, regulatory mechanisms, and to visualize population structure (e.g., Principal Components Analysis (PcoA)). To generate representative sets of genomes, hierarchical clustering of different distance metrics has been performed including genomic similarity scores (GSS)[101], BlastP score ratios (BSRs)[141] or average nucleotide identity (ANI). One example of hierarchical clustering was applied to reduce 249 *Acinetobacter baumannii* genomes to 100 genomes by manually selecting cluster representatives based on BSRs[23]. Similarly, a set of 4,600 *E. coli* genomes was reduced to as few as 315 representative genomes using GGRasP based on an ANI-based distance matrix[30]. Most of these earlier hierarchical clustering methods require either a distance cutoff or a pre-determined number of clusters. In contrast to the manual approach of [141], GGRasP can also automatically select clusters using a Gaussian mixture model (GMM) approach in which distances between closely related genomes and more diverse genomes are maximized. Similarly, popPUNK computes both core and accessory Mash distances and then clusters using GMMs. Based on a prior popPUNK analysis of 10,158 *E. coli* genomes, 11 total clusters [63] and 50 lineages (18 B1's, 13 B2's, 4 A's, 5 D's, 4 F's, 3 E's, 1 C, and 2 of the Shigellas, *S. sonnei* and *S. flexneri*) were defined. Despite these prior efforts in generating representative *E. coli* lineages and representative genome sets, it is still not clear how many lineages truly exist within this species and what representative genome set captures the known diversity of this species.

While prior clustering methods have been somewhat successful, this approach in general can be sensitive to outliers, does not always scale well for large datasets, and downstream results can be impacted by non-repeatable factors such as the order of data. Most importantly, a rigid structure between individuals or groups may not accurately represent genome data if the evolution of the species/individuals being studied is not tree-like, e.g.,

genomes with large amounts of horizontal gene transfer or the existence of a single branch (phylogroup), which will have too few representatives generated [54]. This lack of rigid structure has been shown in *E. coli* with the major lineage (A) as well as admixture between both A and B1 and between A, B and D [172]. More recently, research based on on current *in silico*-derived phylogenetic labels and hierarchical clustering of Mash distances suggests there could be as many as 14 *E. coli* groups: 3 D's, 2 B2's, 1 E2(O157), E1, F, G, A, C, B1, and two *Shigella* clades: Shig1 and Shig2 [5]. It is unclear if the existing clustering frameworks, which try to achieve the highest concordance to a fixed number of groups (eight in [5]), can deal with any admixture, which is not always known prior and incorporated into established phylogroups. The framework presented below does not assume a rigid tree-like structure in the data and therefore may perform better for identifying *E. coli* lineages with admixture [172].

Furthermore, there are numerous bacteria where phylogenetic lineage schemas are even more unclear than *E. coli*. For example, the largest collection of Cyanobacteria at the phylum level studied to date identified at least 8 lineages based on a maximum likelihood and neighbor joining (NJ) tree of 32 conserved genes in 140 samples [64]. Within the the phylum of Cyanobacteria, there are over 2,698 described species and 6,280 predicted species [103]. Additionally, there exists a huge range in the genome size within the Cyanobacteria phylum at 1.6×10^6 bp to 13.2×10^6 bp [78]. There are some Cyanobacteria such as the toxic cyanobacterium *Microcystis aeruginosa* where MLST schemas have been proposed based on seven housekeeping genes but allelic profiles for schema are not publicly available. A NJ phylogenetic tree of the concatenated housekeeping genes across 164 strains of *M. aeruginosa* identified 5 main phylogenetic lineages [158]. Identifying phylogenetic lineages within cyanobacteria species without a publicly available MLST would increase our understanding of Cyanobacteria evolutionary history even further.

We propose to apply an alternative approach, namely unsupervised clustering, to uncover phylogenetic lineages. Unsupervised clustering can identify admixed lineages that do not conform to a rigid tree structure as well as test how robust existing phylogenetic lineages/frameworks are as more whole genome sequences are generated. Here, we apply two data-driven, unsupervised learning alternative methods and compare them to traditional

hierarchical clustering for generating genome representative sets. These methods are k-means and a custom random centroid approach. This “bake off” of established hierarchical clustering vs. less traditional methods addresses the need for a high quality, non-redundant set of bacteria genomes for a species, but also shows advantages and limitations of these methods by comprehensively comparing the resulting genome clusters, gene diversity profiles across representative set and concordance between existing phylogenetic lineages and the clusters defined by our data driven methods. High quality, non-redundant sets of genomes are also essential, especially as we continue to associate genotypes, genomic features, phylogenetical lineages, and/or genomic backgrounds with specific phenotypes (i.e. virulence) over tens of thousands of sequenced genomes.

2.3 Methods

Genome Datasets

Assembled genomes were downloaded from NCBI Refseq on February 11, 2020 and filtered by keyword *Escherichia coli*. To select high quality genomes, dBBQs was used to filter the initial 18,832 *E. coli* Refseq genomes to 15,649 with a total quality score cutoff of 0.8 or greater [165]. This quality score incorporates sequence quality, rRNA, tRNA and essential genes and this specific cut-off was selected because it is the average score for the majority of genomes per [82]. Phylotypes were determined for the *E. coli* dataset using Ez-cleremont version 0.6.2 downloaded from GitHub as of September 4, 2020 [167]. For clustering analyses, we generated clusters with (15,549) and without “U/cryptic” and cryptic genomes (15,501) included. Genome distances were computed via Mash [70, 4]. This method is based on MinHash sketching, a technique first invented to detect duplicate web pages for Alta Vista [17]. Briefly, MinHash sketching uses a fixed vector of random hash values per sample to estimate the genomes’ k-mer based Jaccard similarity (distance). Sketching has also been applied to unassembled data (see [4] for *E. coli* comparisons using Mash and different sequencing platform data), and therefore can also be applied on raw bacterial isolate sequences. Here each complete assembled genome was represented as a sketch with 10,000

random hash functions (21 k-mers), with the parameters being the same as those previously determined for *E. coli* in [70]. Mash distances were then computed for the 15,649 high quality Refseq *E. coli* genomes.

Additionally, assembled Cyanobacteria genomes were downloaded from NCBI RefSeq on November 12, 2020 with keyword Cyanobacteria and all phages/viruses were removed. Based on the dBBQ quality score cut-off of 0.8, no Cyanobacteria genomes were filtered. Because no characterized Cyanobacteria has a known genome size less than 2 Mb, likely partial genomes were filtered using this size criterion. Because the final Cyanobacteria dataset (1631 genomes $\geq$ 2MB out of the initial set of 2942 genomes) represents multiple species, we generated additional likely species subsets based on the suggested Mash species distance cut-off of 0.05 [4], which generated 695 total subgroups (species). When random sets of genome were selected from the entire set of genomes within each genome datasets, we used the Numpy Random function [59].

Applying unsupervised learning methods to select representative genome set

To reduce redundancy and represent each phylogroup equally because there is an inherent bias in the distribution of sequenced phylogroup members, we generated clusters of genomes based on four different methods: 1. hierarchical clustering; 2. k-means clustering; 3. PcoA + k-means clustering; and 4. custom random centroid approach. All methods except the custom random centroid approach were standard approaches. A genome by genome matrix of Mash distances was provided as input to all methods. Hierarchical clustering was then performed with complete linkage in R via fastcluster. Although four different hierarchical clustering methods (UPGMA, Ward's, complete, average) were applied to the Mash distance matrix, the *hclust* complete function in R seems to be the best and is used throughout (data not shown). Clusters were visualized in R as dendrograms using *dendextend* [46] and *plot*. The dendrogram generated from hierarchical clustering was then cut into k groups via the *cutree* function in *dendextend* [47]. PcoA and k-means clustering was performed with default parameters in Scikit-learn [19] and visualized via matplotlib [68]. The custom

random centroid approach first involved setting an arbitrary random genome as the seed and the centroid of the cluster, generating clusters based on an intra-cluster distance cutoff where each genome was placed in the cluster if it was within a given distance or less to the randomly chosen seed genome. Clusters were continually generated until all genomes were either assigned to a cluster or remained as a singleton. For the random centroid approach applied to the *E. coli* datasets, the arbitrarily chosen Mash distance cutoffs ranged from 0.005 to 0.045 with an interval of 0.01 since 0.005 was the minimum Mash distance in this dataset and 0.045 was the maximum Mash distance. Likewise, we generated a Mash distance cut-off range for the cyanobacteria species, *Microcystis aeruginosa*, where the first value in range was the minimum Mash distance in the dataset and last value in range was the maximum distance in the range set. Because the range of Mash distances was greater for *M. aeruginosa* genomes as compared to the *E. coli* genome set, we determined the Mash distance interval by the numpy linspace function where evenly spaced distances are generated based on parameters of the start, end value and the quantity of numbers generated between the start and end value. We selected the start value as the minimum value within the mash distances and the stop value as the maximum value within the mash distances with the number of samples generated set to 10.

To determine the optimum Mash distance cutoff(s) for *E. coli* we used purity, completeness and the silhouette index to survey the quality of the clusters generated from the random centroid approach. The “true” label of the cluster was assigned by majority rule based on the five major ez-cleremont phylogroups. For k-means clustering, the optimal number of clusters, k, was surveyed based on both the silhouette index and within-cluster-sum-of-squares (WCSS). The silhouette index is the measure of how similar objects are within the cluster compared to other clusters. The larger the silhouette index, the more well-matched the clusters are to its own cluster and poorly matched to its neighboring clusters. The WCSS is calculated by the sum of the squares of the distances of each data point in all clusters to their respective centroid. As the accepted standard for WCSS, we choose the last point after the elbow in the graph as the optimal number of clusters where the distance no longer changes as the number of clusters increases. Since both the silhouette index and WCSS can result in a different “optimal” number for k, here we choose the larger value of k. For

hierarchical clustering, we also used the silhouette index to determine k . Since there are no established phylogenetic lineages in the cyanobacteria species, *M. aeruginosa*, we were only able to use the silhouette index and WCSS to determine k for all of the methods.

We selected the representative genome from each cluster based on a custom Python script. This representative was defined as the genome in the cluster with the smallest mash distance sum relative to all other cluster members. This selects the most “central” genome in the cluster (smallest average) instead of selecting a “random” genome from a cluster. To evenly represent the membership of each of the non-singleton clusters (23), we then used the most central genome in each cluster as the seed in the referenceseecker program (<https://github.com/oschwengers/referenceseecker>) [148] with 15 as the maximum number of candidate reference genomes. This resulted in a reduction of the original 15,649 high quality *E. coli* genomes, which includes “U/cryptic” members, down to 329 genomes. The final number of representative genomes was smaller than expected using referenceseecker (23 x 15) because some genomes were selected more than once; when this happened we kept only one instance. The process of selecting the most central genome in each cluster and running referenceseecker to select representatives was repeated for the clusters generated by PcoA + k -means and hierarchical clustering. Finally, we ran referenceseecker with the same parameters as the *E. coli* set for the clusters generated for *M. aeruginosa* except the database was defined as the set of 1,631 Cyanobacteria genomes and candidate reference genomes was set to 5 to account for the lower overall number of genomes within this species.

For the representative sets generated for both *E. coli* and *M. aeruginosa*, genes were annotated via Prodigal version 2.6.3. We ran Roary to generate gene presence/absence matrices based on the prodigal gff output files with default parameters [127]. Additionally, we annotated and generated gene presence/absence matrices for the ECOR set and the 14 genome set from [5] for comparison. The core genome (genes found in 95% of strains analyzed) was also aligned via Roary (using MAFFT) and a tree was generated using FastTree [137]. The overall workflow of the entire framework is outlined in Figure 1.

Determining phylogroup specific features across the representative *E. coli* genome set

Based on the three different representative sets of genomes generated for *E. coli* (random centroid, PcoA +k-means, and HC), phylogroup specific genes were defined by Non-negative Matrix Factorization (NMF; [36]). NMF has been previously used to cluster metagenomic samples [22, 73] and, more recently, to cluster and refine marker genes for single cell RNAseq [6]. We applied the NMF R package [48] to the gene presence/absence matrices of the representative *E. coli* genome sets. The number of clusters initialized for the NMF runs was determined based on the silhouette index. NMF results were visualized in R using heatmap [79]. The NMF predict function was used to determine the features defining each cluster (basis component). The probability of each feature defining the basis component is defined as the relative contributions of the feature to the dominant basis component. We defined arbitrarily defined features with probability > 0.8 as cluster defining genes. The basis component was assigned a phylogroup label based on majority rules for number of genomes with the same ez-cleremont *in silcio* phylogroup label. If there was a tie between the counts of phylotypes within each cluster, we defined the cluster as “cryptic,” consistent with prior atypical genomes in *E. coli*, and did not report its phylogroup specific genes.

2.4 Results

2.4.1 *E. coli* whole genome diversity

The first dataset used is a filtered set of 15,649 *E. coli* higher quality genomes (see Methods). To determine the current genome distance distribution within this dataset, we first survey the frequency distribution of the Mash distances between each genome pair within each phylogroup (Figure 2). The distribution of the Mash distances does not show a clear single peak for most phylogroups. This suggests these phylogroups may be more diverse relative to groups A and B1, at least with respect to what has been previously sequenced. We note that cryptic, F’s and E’s are the most diverse with a Mash distance range of 0.001 to 0.04 and happen to also be the most recent phylogroups uncovered in *E. coli*. The least diverse

phylogroup is B1. The average genome Mash distance within the full 15,649 set is 0.024 with minimum Mash distance of 0 and maximum Mash distance of 0.045 (Supplementary Table 3).

2.4.2 Diversity in relation to existing phylogroups

We first attempted to assign *E. coli* genome representatives by combining Principal Component of Analysis (PCoA) with k-means clustering (Figure 3). From this analysis it is clear that the existing labels may not all be accurate, e.g., at least two D genomes grouping with Es and a handful of F genomes grouping with B2. Further, A's overlap with B1's and C's in the PCoA plot and there is no clear k-means centroid (X in the figure) for G/D and F (Figure 3). In conclusion, this projection of the genome by genome distances into a 2-D space shows that a lot of the diversity within known phylogroups is not clearly unique.

To see if phylogroups are assigned better using pairwise genome-based comparisons, as previously done by us [147] and others [11, 5], we identify the closest genome in different datasets according to Mash distance and compare this label to its MLST phylogroup designation in Figure 4. Based on the full collection of 15,649 *E. coli* genomes, the majority of the phylotyped genomes closest genomes belong to the same phylotype with 54 mismatches between U/"U/cryptic"/cryptic and the 8 main phylogenetic lineages. The number of mismatches between the phylotyped genomes and the phylotype of the closest genome decreases when using the representative sets as seen in Figure 4.B-D. Specifically, there are 11 mismatches in the random centroid method (Fig. 4B), 8 mismatches in the hierarchical clustering representative set (Fig. 4C), and only one mismatch in the k-means representative set (Fig. 4D). The reduced number of mismatches between the representative set and the full *E. coli* genome set demonstrates that the genomes represented in the reduced representative set are less "heterogeneous" and therefore overall less ambiguous for use in assigning phylogroups. Although, we still observe discordance between MLST phylogroup designation and the MLST phylogroup of the closest genome thus suggesting there is still ambiguity between D, F, E, G and B2 across all representative sets except k-means clustering.

We observe at least a 98.7% concordance, which is similar to the rate observed by others [11].

2.4.3 Selecting representative sets of *E. coli* genomes

Although pairwise Mash-based distances appeared to work well for the use case of assigning a specific phylotype to most *E. coli* genomes, we want to also comprehensively consider alternative approaches for representing diversity in *E. coli*. So we apply three different unsupervised learning methods to select representative clusters from our *E. coli* 15,549 genome dataset and compare the phylogroup distribution within each cluster’s centroid as well as the representative genome sets generated by running referenceseecker on the genome of the cluster’s centroid. For comparison to the most widely used representative set of *E. coli* genomes (ECOR set), we plotted the phylogroup counts of the genomes within this set. As Fig. 5.A shows, phylogroup D and E have low representation and no cryptic genomes are present. Additionally, the phylogroups A, B1 and B2 have the highest representation while the minor phylogroups (C-G) are the least represented.

Using the hierarchical clustering method, which is closest to the pairwise assignments visualized in Figure 3, we determined 8 clusters based on the silhouette index. After the most centroid genome per each HC cluster was selected, we determined representative genomes for the cluster’s centroid genome via referenceseecker as described in the methods. Although HC was the only method considered where the optimal number of clusters matched the same count for the previously established 8 phylogenetic *E. coli* lineages, the actual phylogroup of the 8 cluster’s centroid only covered five phylogroups (B1, B2, D, 2 G’s, 3 F’s). With these eight centroids as the seed reference to generate representative set of genomes for the cluster via referenceseecker as described in Methods, the phylogroup representation increased to 7 phylogroups. These 7 phylogroup were A, B1, B2, cryptic, D, E, F, G (Fig. 5.C.). The least representative phylogroup was A at 2 genomes and the most represented was D.

The second method we applied to the genome by genome distance matrix was k-means clustering. We identified 5 to 6 clusters based on both sum of squared distances and the silhouette index. Of the 6 clusters generated, only one is mostly pure (phylogroup E) while

the remaining clusters are mixed including two clusters with A, B1, and C and another three clusters with B2, G and D (data not shown). In contrast, when we apply k-means clustering to the first 10 components of PCoA, we generate two pure B2 clusters and another pure cluster for phylogroup E, but also one “cryptic” cluster containing phylogroups G, F and D and two separate admixed clusters containing A, B1 and C labeled genomes(Figure 2). Based on the most central genome of each of these PcoA+k-means generated clusters with running referenceseeker on each centroid, we identified three pure phylogroup A clusters, one phylogroup B1 cluster, one phylogroup E cluster, and one admixture cluster of D and G phylogroups. The five main phylogroups are represented within this Pcoa + k-means clustering representative set as seen in Fig. 5.D. Furthermore, this representative set resulted in the most even counts across each phylogroup represented where A and B2 had the largest representation and G was the least represented.

The final method we considered is the random centroid clustering approach. As described in our Methods section, this approach enables “random” clusters to be generated based on random selection of a cluster centroid and then adding genomes to this cluster based on other genome distances to each randomly selected genome. This approach allowed us to generate a data-driven representative set of genome clusters with indirect guidance from existing phylogroup labels. In contrast to the prior two methods, the random centroid approach is sensitive to the “U/cryptic” and “cryptic” genomes: the optimal number of clusters increased when these unclear genomes are included. The optimal number of clusters selected by this method is 30, which represents 23 phylogroup clusters and 7 remaining “singletons.” The final phylogroup count of the random centroid representative set was the most comprehensive compared to the prior two methods (Fig. 5.B). The five main phylogroups are represented (A, B1, B2, D, E) and the most represented phylogroup is D, as we also observe in PcoA + k-means clustering, consistent with the two subgroups known to exist within phylogroup D.

To rule out the possible outlier effect of the “U/cryptic” genomes to the clustering methods applied, we evaluated the effect of removing these genomes prior to clustering. We hypothesized that removing genomes with phylogroup of “U/cryptic” and/or cryptic would minimize the effect of anomalous genomes and reduce the number of singleton genome clusters. Based on hierarchical and k-means clustering, the optimal number of clusters

predicted is the same when "U/cryptic" and "cryptic" genomes were included or not included in the analysis.

2.4.4 Effectiveness of our representative sets of *E. coli* genomes in capturing genomic diversity

Across all methods, the optimal number of clusters is method-dependent and could be as many as 30 (Table 1). Therefore, we next evaluated how well the phylogroup of each cluster's "centroid" compared to the eight established phylogroups in *E. coli* (see supplemental Table 2). Although none of our methods resulted in at least one phylogroup per each cluster's centroid, our random centroid method generated the highest number of phylogroup representation at the centroid level (Supp Table 3) and the representative genome set from these cluster's centroid combined with referenceseeker represented 8 phylogroups excluding C. In contrast, the centroids of clusters generated by HC excluded A, C and E(supplementary table 3) but included these genomes in HC representative genome set (Figure 5.C). Finally, the phylogroup of the centroids from PcoA combined with k-means (data not shown) encompassed five phylogroups (2 A's, 2 B2's, E, G, B1). The representative genome set from the 7 centroids covered the least number of phylogroups of only six phylogroups excluding C, cryptic and F(Fig 5.D).

One phylogroup-agnostic assessment alternative is to consider relative genomic features. If our representative sets are biased in any way, the core gene set would be higher or lower than expected according to previous research [147]. Relevant gene categories were defined by Roary [127] where the total genes (pan-genome) is the set of genes found in any strain, core genes are defined by the genes found in $\geq 95\%$ of the strains considered, shell genes are defined as being found $< 95\%$ of these strains but also in $\geq 15\%$ of the same strains, and finally the cloud genes found in $< 15\%$ of the genomes considered. The largest core genome was 3,033 genes for the ECOR set and the smallest core genome was 2,157 genes for the random centroid representative set (see Figure 5), which is consistent with a prior core gene estimate of $\sim 3,188$ using 2,000 *E. coli* genomes [81] and roughly 2,613 core genes found in a total of 10,667 *E. coli* genomes [5]. As evidence that the ECOR set is relatively less diverse

than other representative sets, note that its estimated core genome is larger than Abram et. al [5] even though there are 57 more genomes in the ECOR set. The number of genomes selected can also impact the total genome size. *E. coli* is an example of this with an open pan-genome [81]. For our representative sets, the pan-genome (total gene set) increases in all genome sets as the number of genomes increases except for the k-means representative set. The k-means representative set of 105 genomes has a pan-genome size of 19,381 while the ECOR set has 33 fewer genomes but still a slightly larger pan-genome size at 19,968. Based on the smaller pan-genome size generated by k-means clustering, the k-means generated representative set is more core and less diverse than the other sets generated by the HC and random centroid approaches.

2.4.5 Diversity across the Cyanobacteria Phylum

To begin selecting representative sets of Cyanobacteria, we first generated likely cyanobacteria species datasets using a 0.05 Mash distance cut-off (see Methods). Because the majority of the 695 species are singletons, we select six putative species clusters that have at least 20 genomes for further analysis. These six species are *Microcystis wesenbergii* (52 genomes), *Microcystis aeruginosa* (120 genomes), *Fischerella thermalis* (43 genomes), *Synechococcus lacustris* (26 genomes), *Cyanobacteria bacterium* (24 genomes) and *Cylindrospermopsis raciborskii* (21 genomes). The whole genome diversity for each species is surveyed based on Mash distances as was done earlier for *E. coli* (Figure 7). The most diverse species is *Microcystis aeruginosa* with the largest Mash distance range of 3.6e-05 to 0.06 between pairs. This mash distance range is 0.02 larger than analyzed *E. coli* genomes. The least diverse is *Cyanobacteria bacterium* with Mash distance range of 0.004 to 0.0100832. The average genome Mash distance of *M. aeruginosa* is larger than the *E. coli* genome Mash range at 0.04186884 despite the fact that the genome set of *M. aeruginosa* had 16,000 fewer genomes than the *E. coli* genomes considered (Supplementary Table 3). The minimum Mash distance within *M. aeruginosa* is 3.57545E-05 and maximum Mash distance is 0.0501626.

2.4.6 Selecting representative sets for cyanobacteria genomes

We apply the same unsupervised clustering methods (PcoA + k-means clustering, hierarchical clustering, and random centroid approach) to one of our cyanobacteria species sets, *Microcystis aeruginosa*. The number of clusters for each method is determined by WCSS and silhouette indices as shown in Supplemental Table 1. As observed in *E. coli*, there is little consistency with respect to the inferred optimal number of clusters. Even though *M. aeruginosa* is a more diverse species than *E. coli*, the optimal number of clusters was slightly smaller with a range from 7 to 21. Because all metrics and methods result in a different number of groups (k) across the four unsupervised learning methods, we select the lowest number of groups for consistency for each method. For pcoA + k-means, we selected k = 7 based on elbow method of within-sum-of-squared. Additionally, we selected k=21 for the random centroid approach and k=18 for hierarchical clustering based on the largest silhouette index.

2.4.7 Effectiveness of our representative set of *Microcystis aeruginosa* genomes in capturing genomic diversity

Because to the best of our knowledge there are no phylogenetic lineages established in *M. aeruginosa*, we visualized the phylogenetic lineages in the full genome set and reduced set via a dendrogram (Supp Figure 2). In the full genome set, based on the silhouette index, we cut the tree to identify 18 lineages. We used these lineages identified by HC as the lineages in *M. aeruginosa* and compared how many of these lineages are represented across each representative set of genomes generated (Figure 8). The random centroid method has all the HC lineages represented at 18 while PcoA + k-means representative genome set has 6 lineages represented and the random set of genomes has 9 lineages. The genomes generated from the CH the most number of lineages represented at 18 Within the reduced representative set of genomes generated from PcoA + k-means method, 6 of these lineages are also represented.

Moreover, the full genome set has genome sizes ranging from 2 to 6 Mpb. The three primary genome sizes within this range are all represented in our reduced *M. aeruginosa*

representative set. Based on HC, we identify eighteen subclusters. Within these subclusters, six centroid genomes had a different taxonomy name to *M. aeruginosa* and included two *Microcystis wesenbergii*, one *Microcystis panniformis*, two *Microcystis flos-aquae*, and a *Microcystis sp. LEGE*. One advantage of our clustering method for generating representative clusters/“lineages” within a set of genomes is we are not dependent on the taxonomy name present on the database record. These six genomes all have an ANI value inconsistent with its declared species name as indicated in NCBI [28].

We also generate gene diversity profiles across the three representative sets of genomes we generated for *Microcystis aeruginosa* compared to a random set of 18 *Microcystis aeruginosa* genomes and the non-reduced genome set to evaluate the effectiveness in maximizing gene diversity in each set. We observe a largest total gene count of 39,277 (Figure 8). The number of core genes is largest in the HC representative set at 1,883 genes. Compared to the other sets, the HC representative shares the most number of core genes with the other sets generated. For another comparison, we generated a randomly selected set of genomes equal to the number genomes in the PcoA + k-means set. As expected, the number of total genes was smallest in the random set (30,282 genes in total) but this set also had the lowest number of lineages represented (6) and the low gene diversity is expected since there was low representation of genomes per each lineage.

Discussion

One major objective of generating representative sets of genomes is preserving as much of the diversity in a species with the least amount of selected genomes. As this basically can be viewed as the “set cover” problem in computer science, which has been shown to be NP-complete [77] and therefore computationally intractable to obtain the best set, here we compare gene diversity profiles as well as phylogroup counts, if available (i.e. *E. coli*), across each representative set generated using a different method. We find that the total gene count can be used as an indicator for how comprehensive a genome representative set is. Further, our research addresses the question of how many phylogroups/phylogenetic lineages exist within a species via data-driven cluster metrics including purity, completeness, WCSS,

and silhouette index. Our comparison of different cluster metrics across four unsupervised learning methods shows that the number of *E. coli* phylogroups identified in the 15,549 genomes analyzed is dependent on the unsupervised method used and ranges from 6 to 14. Each method is likely identifying distinct lineages but the actual number varies based on how fine grained it was partitioned. For example, recovering the “primary” five *E. coli* phylogenetic lineages at the coarsest grain is achieved by hierarchical clustering on *E. coli* data. In contrast, if one wants to identify additional sub-clusters, then k-means or the custom random centroid approach are more appropriate.

For all pre-established multilocus sequence types (MLSTs), except U/cryptic, we find that 0.02 is the best Mash distance cutoff for identifying phylogroups on these *E. coli* genome data based on the highest purity and completeness of clusters identified by the custom random centroid approach. We note that a prior publication found 0.0115 as the best Mash distance cutoff for separating phylogroups in a more network-based approach [93]. Our Mash distance cutoff may be higher because our downloaded set of complete genomes potentially contains plasmid sequences, which can generate slightly increased distance estimates for EIEC/*Shigella* and other *Shigella* toxin-ETEC *E. coli*.

The selection of the representative sets in this study may also be biased by unequal distribution of phylogroups in public databases. To address this possibility, we randomly down sampled phylogroup A, B2, and B1 into smaller phylogroup sets of 1,350 genomes each while the other phylogroups were left as is. The overall down sampled genome set contains 7,901 genomes, and we re-ran our random centroid method based on Mash distance cutoffs from 0.005 to 0.045. This analysis results in a similar number of clusters as compared to the original non-down sampled data (results not shown). Thus, at least the random custom centroid method presented here does not appear to be affected by any potential the bias in over represented phylogroups. That said, our research also shows there are additional lineages (diversity) that is not accounted for in current MLST schemes. Our approach is therefore a more extensible framework that can capture diversity even within a well-studied bacterial species such as *E. coli*.

Given the six different cyanobacteria species identified based on the 0.05 Mash distance cut-off, we focus our efforts on one species, *Microcystis aeruginosa*, since this species had

largest diversity and the most Cyanobacteria genomes at 120. Within *M. aeruginosa*, we determine that 0.03 is the best Mash distance cut-off for identifying lineages. For the remaining five species, we ran the same clustering methods except on the centroids of the clusters due to low overall representation of the species (21 – 52 genomes each) and the numbers of clusters are shown in Supplementary Table 2. Interestingly, when we ran our framework on *Synechococcus lacustris*, *Cylindrospermopsis raciborskii* Cr2010 and *Cyanobacteria bacterium* QS-6-48-18, the representative sets include almost all of the original genomes considered. A centroid genome approach for these species may actually decrease representation of diversity, and for this reason we believe it is best to only generate a representative set if there are at least 100 genomes available. A further limitation of representative sets is the potential removal of outlier genomes, which may contain atypical gene diversity relative to other sequenced members. For this reason this framework provides the option of including any such “cryptic” genome (or others) if desired.

Conclusion

We present the first unbiased comparison of four different unsupervised learning methods for generating representative sets of genomes. Our research shows that the prior established *E. coli* phylogroups do not exactly correspond to the data-driven clusters identified. Rather, the data-driven clusters show ambiguity and even possible large-scale gene sharing between A, B1 and C as well as between cryptic, G, F and D. Furthermore, we found that the set generated by combining the centroids found in our random centroid approach with referenceseeker generated the most comprehensive *E. coli* representative genome set. Although phylogroup C was not represented in this set, it was also absent from all other sets including [5]. We also capture at least seven phylogenetic lineages within *M. aeruginosa*, which is larger than the 5 previously established lineages [158]. We also show that the relative sizes of core, cloud, shell and pan-genomes can be used to compare multiple competing representative sets and visualize the relative genomic diversity captured by each one. More diverse representative sets may help improve the prediction accuracy in machine learning approaches, as seen in a prior study on predicting pathotypes in *E. coli*[94]. Finally, a deeper characterization of

genomic diversity within a bacterial species is important for understanding these species including *E. coli* and *Microcystis aeruginosa*.

Appendix: Tables and Figures

Table 2.1: Optimal clusters for 15,564 *E. coli* genomes across four different methods

Method	Method chosen to select k	k (optimal number of clusters)	Distance cutoff for groups
HC	silhouette index	8	NA
random centroid approach	purity	30	0.02
random centroid approach	completeness	30	0.02
random centroid approach	silhouette index	6	0.02
k-means	sum of squared distances	5	NA
k-means	silhouette index	6	NA
PCoA + k-means	sum of squared distances	5	NA
PCoA + k-means	silhouette index	6/7	NA

Table 2.2: Optimal cluster counts across six different Cyanobacteria species on three different methods across two different metrics

Species (number of genomes)	Method	Method chosen to select k	k (optimal number of clusters)
<i>Synechococcus lacustris</i> (26)	PcoA + k-means	WCSS	6
<i>Synechococcus lacustris</i> (26)	PcoA + k-means	silhouette index	3/4
<i>Synechococcus lacustris</i> (26)	random centroid	silhouette index	2
<i>Synechococcus lacustris</i> (26)	HC	silhouette index	2
<i>Cylindrospermopsis raciborskii</i> Cr2010 (21)	PcoA + k-means	WCSS	5/6
<i>Cylindrospermopsis raciborskii</i> Cr2010 (21)	PcoA + k-means	silhouette index	3/5
<i>Cylindrospermopsis raciborskii</i> Cr2010 (21)	random centroid	silhouette index	2
<i>Cylindrospermopsis raciborskii</i> Cr2010 (21)	HC	silhouette index	2
<i>Cyanobacteria bacterium</i> QS_6.48.18 (24)	PcoA + k-means	WCSS	4/11
<i>Cyanobacteria bacterium</i> QS_6.48.18 (24)	PcoA + k-means	silhouette index	3
<i>Cyanobacteria bacterium</i> QS_6.48.18 (24)	random centroid	silhouette index	2
<i>Cyanobacteria bacterium</i> QS_6.48.18 (24)	HC	silhouette index	4
<i>Fischerella thermalis</i> WC441 (43)	PcoA + k-means	WCSS	7
<i>Fischerella thermalis</i> WC441 (43)	PcoA + k-means	silhouette index	5
<i>Fischerella thermalis</i> WC441 (43)	random centroid	silhouette index	10
<i>Fischerella thermalis</i> WC441 (43)	HC	silhouette index	7
<i>Microcystis aeruginosa</i> DA14 (120)	PcoA + k-means	WCSS	7
<i>Microcystis aeruginosa</i> DA14 (120)	PcoA + k-means	silhouette index	13
<i>Microcystis aeruginosa</i> DA14 (120)	random centroid	silhouette index	21
<i>Microcystis aeruginosa</i> DA14 (120)	HC	silhouette index	18/19
<i>Microcystis wessenbergii</i> Mw_QC.S.20081001.S30 (52)	PcoA + k-means	WCSS	6
<i>Microcystis wessenbergii</i> Mw_QC.S.20081001.S30 (52)	PcoA + k-means	silhouette index	3/7
<i>Microcystis wessenbergii</i> Mw_QC.S.20081001.S30 (52)	random centroid	silhouette index	5
<i>Microcystis wessenbergii</i> Mw_QC.S.20081001.S30 (52)	HC	silhouette index	4

Table 2.3: Phylogroup distribution across two different methods applied on 15,501 *E. coli* genomes

	HC (k=8)	random centroid approach (k=30)	random centroid approach (k=14)
A	0	1	2
B1	1	6	1
B2	1	3	2
C	0	0	0
D	1	12	0
E	0	5	1
F	3	2	0
G	2	1	1

Table 2.4: Evaluation metrics based on the number of genomes and summary statistics of Mash distances in full and reduced genome sets

	<i>E. coli</i>	<i>Microcystis aeruginosa</i>
Number of genomes in full set	15641	120
Number of genomes in reduced set	329	18
Number of genome reduced (%)	2.1	15
Method for reduction	random centroid	pcoA + k-means
Max mash diversity full set	0.04486	0.0501626
Max mash diversity reduced set	0.0425102	0.0563751
Min mash diversity full set	0	3.57545E-05
Min mash diversity reduced set	0	0.0277952
Median diversity full set	0.024471	0.0437092
Median diversity reduced set	0.0272156	0.0444822
Average diversity full set	0.023751697	0.04186884
Average diversity reduced set	0.026017122	0.043298981

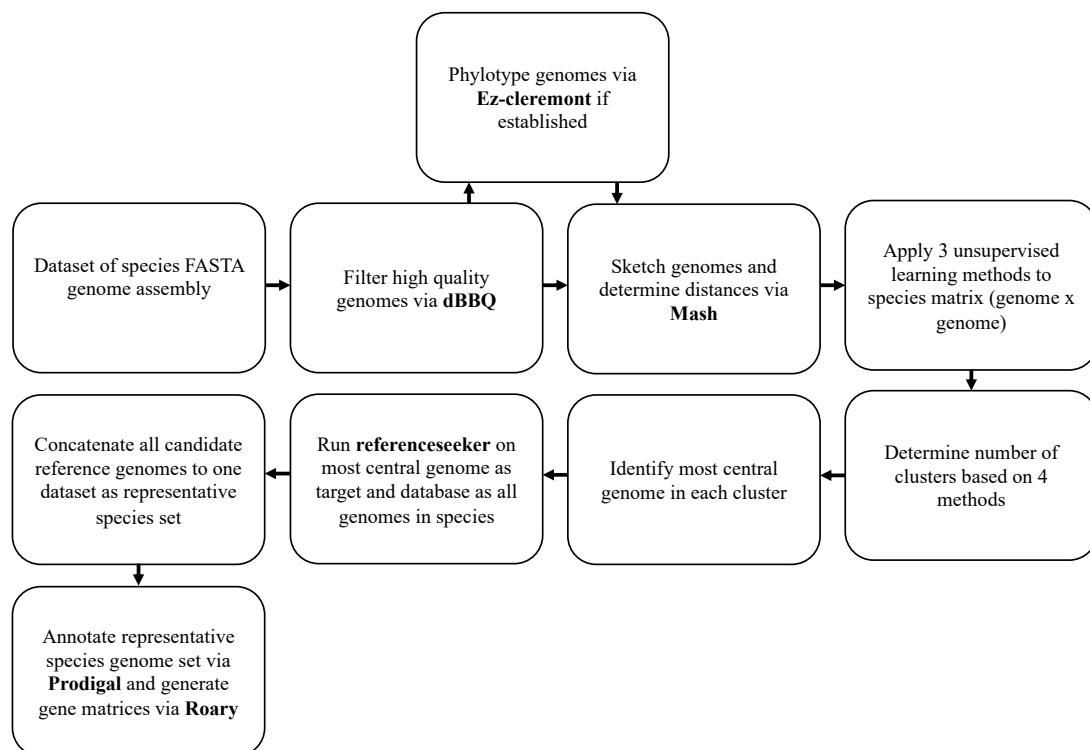


Figure 2.1: Overall workflow for selecting a representative set of genomes. The three unsupervised methods used were hierarchical clustering, k-means clustering on raw and PC's from PcoA and a custom random centroid approach. The four methods used for determining the number of clusters including purity, completeness, within sum of squared distances and silhouette index. The software packages used are in bold.



Figure 2.2: Intra-phylogroup Mash distance vs. frequency of occurrences across 15,649 *E. coli* genomes in panel A and 329 representative set of *E. coli* genomes in panel B.

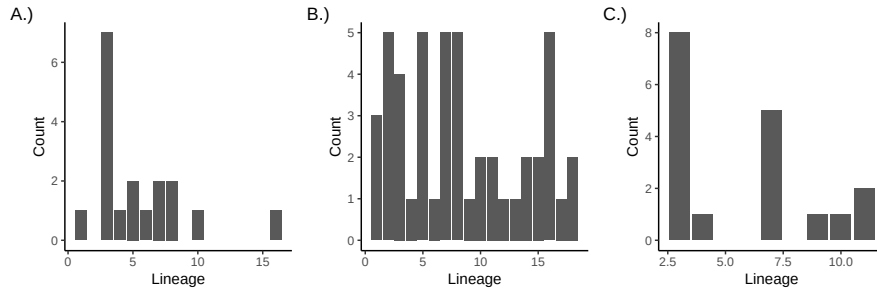


Figure 2.3: Frequency of lineages represented across each representative set of *M. aeruginosa* genomes where random set of 18 genomes represented in A.), random centroid set of genomes in B.) and PcoA + k-means set in C.).

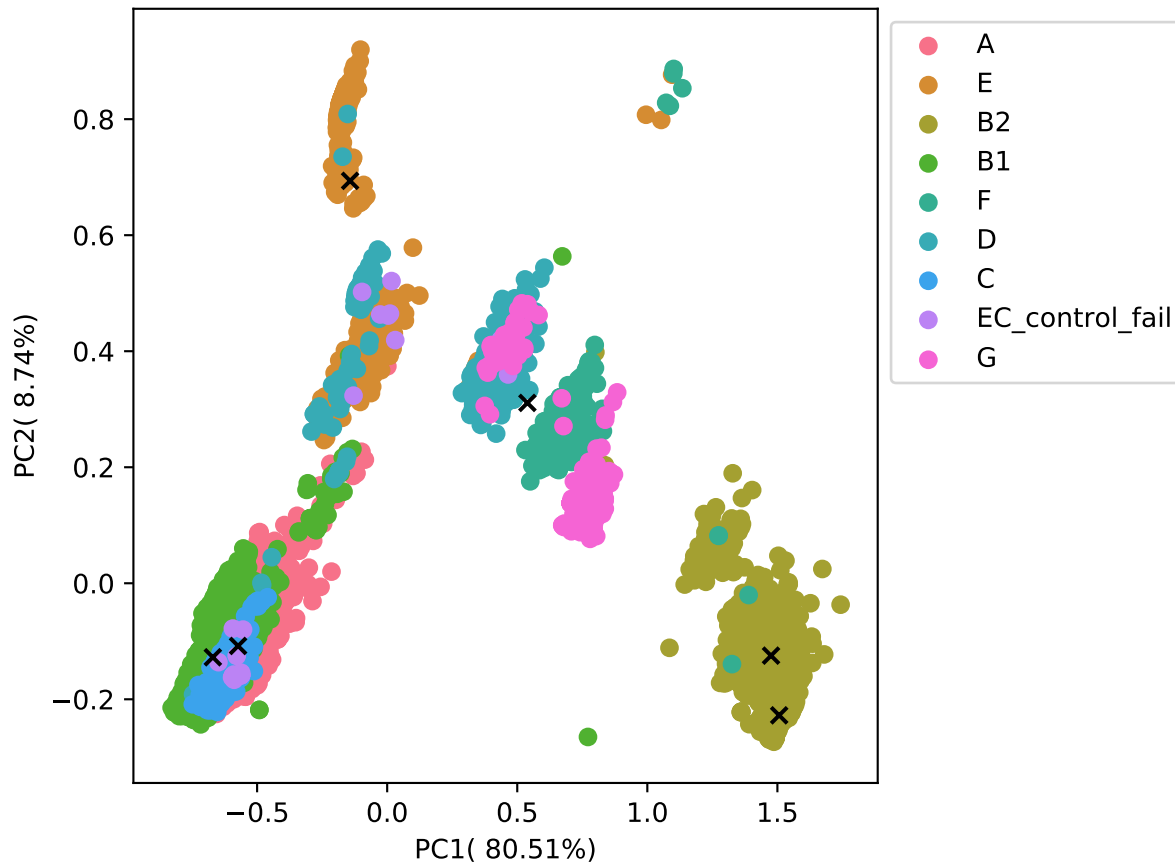


Figure 2.4: PCoA + k-means clustering on 15,501 *E. coli* genomes

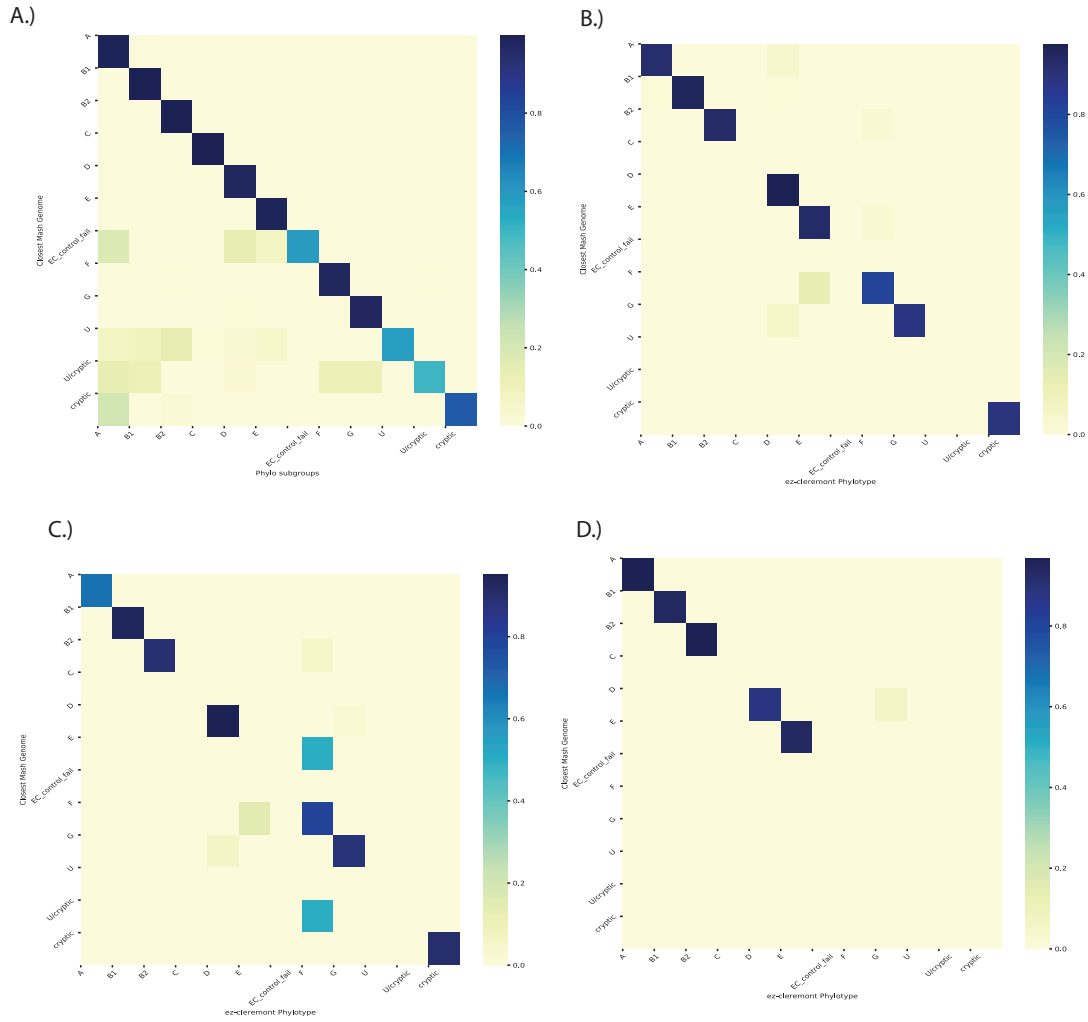


Figure 2.5: MLST phylogroup label versus closest genome-inferred label using Mash on 15,649 *E. coli* genomes in panel A, a reduced set of 329 genomes for random centroid method in B, a reduced set of 120 genomes for hierarchical clustering method in panel C, and 105 representative genome set from k-means clustering in D. The color represents the normalized count of genomes classified.

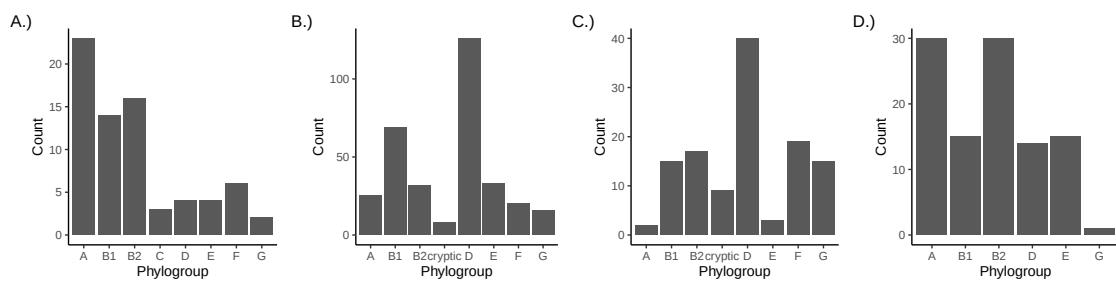


Figure 2.6: Phylogroup counts of genomes within each *E. coli* representative genome set generated by the random centroid approach (B.), HC (C.) and the PcoA+k-means clustering method (D.) compared against the ECOR set (A.)

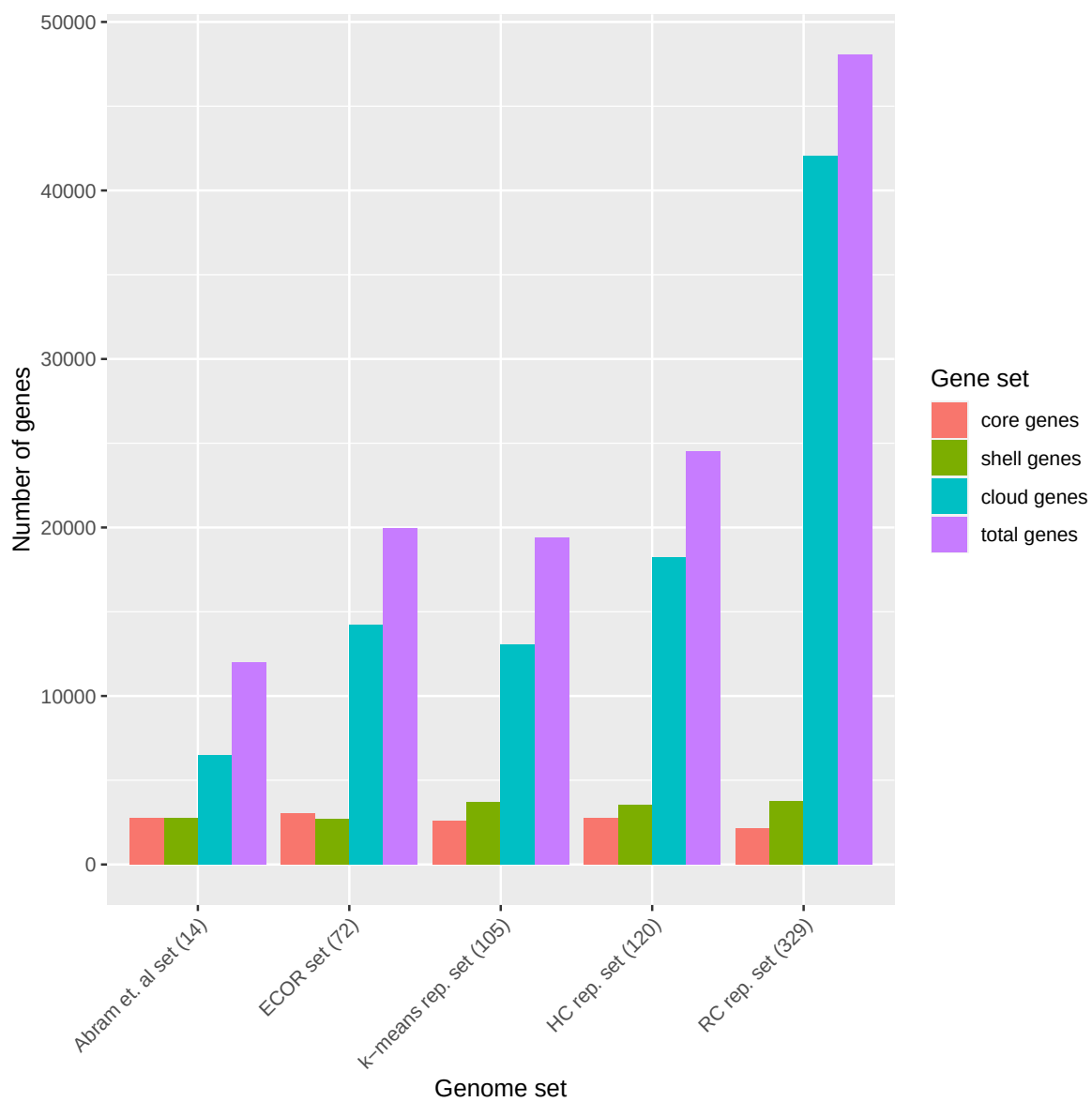


Figure 2.7: Gene diversity profiles generated by Roary [127] across two previous representative sets [5, 135] and our three more inclusive representative sets.

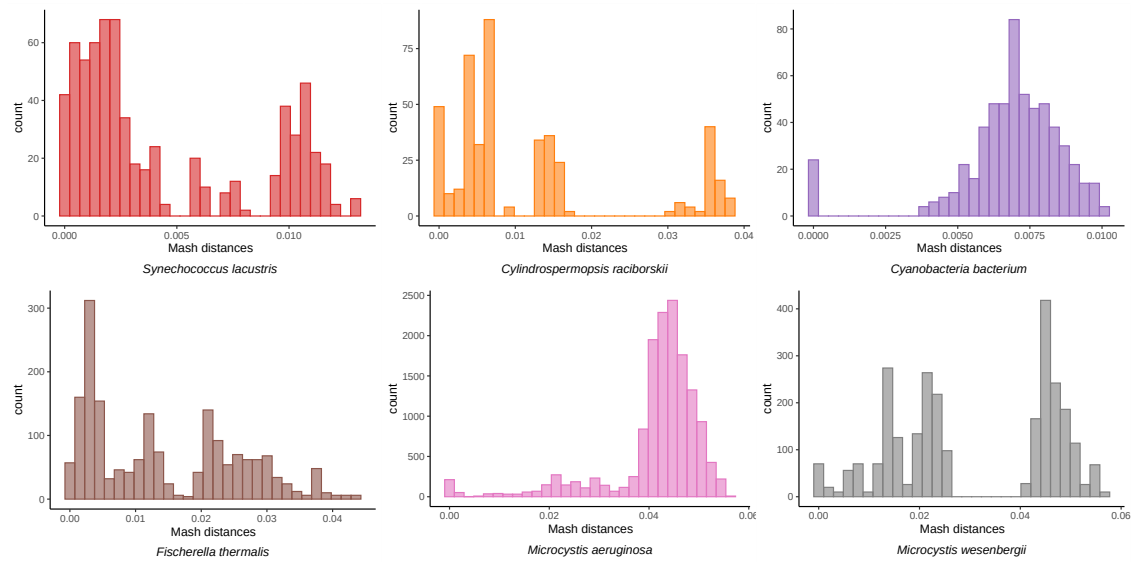


Figure 2.8: Intraspecies Mash distance across six cyanobacteria species

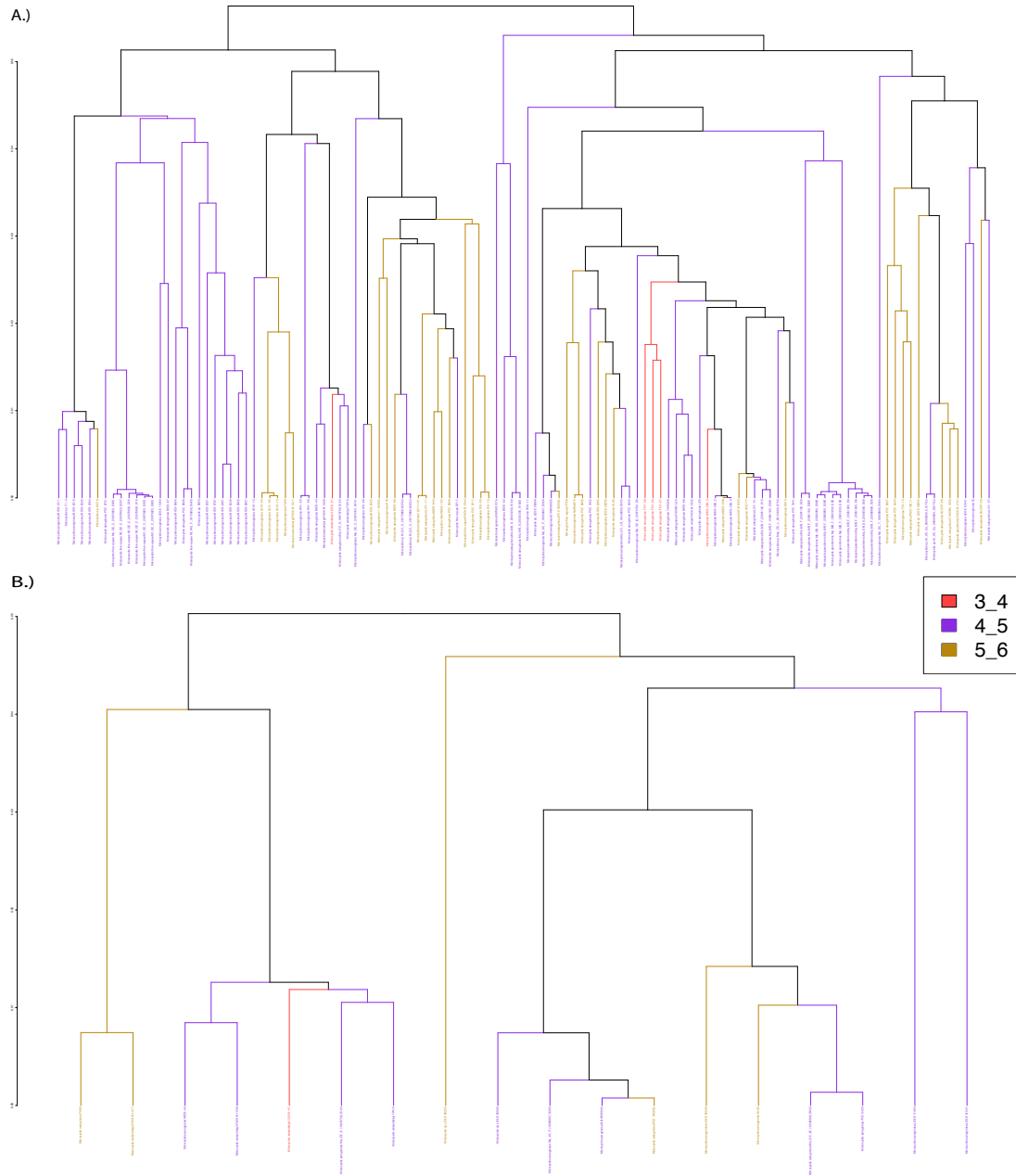


Figure 2.9: Mash distance tree of 120 *Microcystis aeruginosa* (A.) and Mash distance tree of 18 representative *Microcystis aeruginosa* genomes (B.) with branches colored by genome size range in Mbp as seen in key

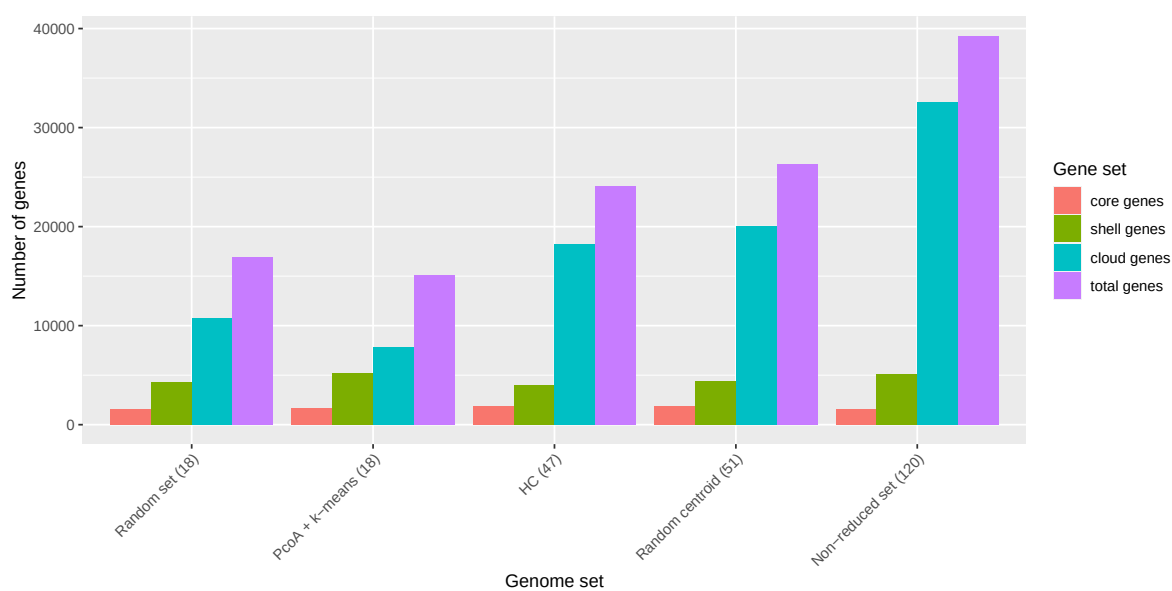


Figure 2.10: Gene diversity profiles generated by Roary [127] for *Microcystis aeruginosa* (Non-reduced set) set, randomly selected set of genomes, and reduced representative set generated in this research by Hierarchical clustering, PcoA + k-means clustering, where numbers in parenthesis indicate genome count

Chapter 3

Whole genome comparisons reveal
panmixia among fall armyworm
(*Spodoptera frugiperda*) from diverse
locations

A version of this chapter was originally published by Schlum et al. in BMC Genomics: Schlum, K.A., Lamour, K., de Bortoli, C.P. et al. Whole genome comparisons reveal panmixia among fall armyworm (*Spodoptera frugiperda*) from diverse locations. BMC Genomics 22, 179 (2021). <https://doi.org/10.1186/s12864-021-07492-7>

My primary contributions to this paper include: (i) bioinformatics data analyses, (ii) results interpretation (iii) generated results (iv) writing of manuscript.

3.1 Abstract

The fall armyworm (*Spodoptera frugiperda* (J.E. Smith)) is a highly polyphagous agricultural pest with long-distance migratory behavior threatening food security worldwide. This pest has a host range of >80 plant species, but two host strains are recognized based on their association with corn (C-strain) or rice and smaller grasses (R-strain). The population genomics of the United States (USA) fall armyworm remains poorly characterized to date despite its agricultural threat. In this study, the population structure and genetic diversity in 55 *S. frugiperda* samples from Argentina, Brazil, Kenya, Puerto Rico and USA were surveyed to further our understanding of whole genome nuclear diversity. Comparisons at the genomic level suggest panmixia in this population, other than a minor reduction in gene flow between the two overwintering populations in the continental USA that also corresponded to genetically distinct host strains. Two maternal lines were detected from analysis of mitochondrial genomes. We found members from the Eastern Hemisphere interspersed within both continental USA overwintering subpopulations, suggesting multiple individuals were likely introduced to Africa. Our research is the largest diverse collection of United States *S. frugiperda* whole genome sequences characterized to date, covering eight continental states and a USA territory (Puerto Rico). The genomic resources presented provide foundational information to understand gene flow at the whole genome level among *S. frugiperda* populations. Based on the genomic similarities found between host strains and laboratory vs. field samples, our findings validate the experimental use of laboratory strains and the host strain differentiation based on mitochondria and sex-linked genetic markers

extends to minor genome wide differences with some exceptions showing mixture between host strains is likely occurring in field populations.

3.2 Introduction

Larvae of the fall armyworm, *Spodoptera frugiperda* (J.E. Smith, 1797) (Lepidoptera: Noctuidae), are highly polyphagous agricultural pest affecting key food and fiber staples such as corn (*Zea mays* L.), cotton (*Gossypium* spp. L.), sorghum (*Sorghum bicolor* L.), rice (*Oryza sativa* L.), and vegetable crops [100]. Damage by *S. frugiperda* in its native subtropical range in the Americas resulted in 15-100% yield loss, depending on the level of infestation [65]. In most of the continental USA this pest does not diapause or survive cold winters. Overwintering populations in southern Texas and Florida migrate northward yearly over several generations to populate northern regions in USA and Canada [169]. Effective control of *S. frugiperda* has been provided mainly by genetically modified corn and cotton producing insecticidal Cry and Vip3A proteins from the bacterium *Bacillus thuringiensis* (Bt). However, resistance to Cry1F, Cry1Ab and Cry1A.105 insecticidal proteins quickly developed in Puerto Rico and the continental USA (Florida and North Carolina) [157, 66], as well as Brazil [43] and Argentina [24]. More recently, the economic importance of this pest has further increased with its introduction in sub-Saharan Africa [51], subsequent spread to India and Southeastern Asia [109], and more recently Australia [139]. An estimate using data from twelve African countries indicates that yield losses resulting from *S. frugiperda* injury could be 21-53% of their annual corn production, which equals to a loss of US\$2.5-\$6.2 billion in losses [37]. While yet to be detected in molecular screens [110], resistance alleles to Bt toxins could have been carried by invasive *S. frugiperda* to the Eastern hemisphere.

Populations of *S. frugiperda* are composed of sympatric mixtures of two genetically differentiated strains based on host preference, a “rice” (R) strain feeding on rice, millet and smaller grasses, and a “corn” (C) strain feeding preferentially on corn and sorghum [132, 134]. There is evidence supporting that this differentiation involves reproductive incompatibility [133] and differential susceptibility to xenobiotics [60]. Signatures of positive selection for genes involved in chemoreception, detoxification, and digestion were also detected in

whole-genome comparisons of nine C- and R-strain individuals [52]. Being morphologically indistinguishable, C and R individuals are discriminated using genetic markers located on the mitochondrial COI (cytochrome oxidase subunit I) and sex-linked (Z chromosome) Tpi (triosephosphate isomerase) genes [86, 113, 112, 104]. However, host preference is not absolute and discrepancies among haplotype markers have been reported. For example, the Tpi marker agrees with assortative mating and host assessments [105] in describing *S. frugiperda* samples from corn in the Eastern Hemisphere as C-strain, yet a predominant *COI-R* strain marker is obtained in these collections [109]. These conflicting results may reflect interstrain hybridization [106] or be driven by maternally inherited symbionts that skew the distribution and diversity of certain haplotypes [32]. Additional factors influencing these conflicting haplotyping results may include incomplete lineage sorting and plant host behavioral plasticity. Nucleotides found at COI host strain marker sites have been commonly grouped into “haplotypes” with relative proportions associated to geographic origin [119], enabling differentiation of “Texas” and “Florida” overwintering *S. frugiperda* populations in the USA. Migratory studies with these haplotypes determined that the “Texas” population is distributed throughout South, Central and North America [107, 116], while the “Florida” population locates to Florida and the Caribbean but migrates through the eastern USA seaboard to reach Canada [115, 117]. Haplotype profiling of *S. frugiperda* from Africa, India and Southeastern Asia supports introductions in Africa of individuals from the “Florida” population and subsequent spread to Asia [109, 110].

In this work, we collected and sequenced 55 genomic DNA (gDNA) samples of *S. frugiperda* from three continents, with an emphasis on C-strain individuals from North and South America. Genome-wide comparisons allowed testing for gene flow between geographically distant *S. frugiperda* populations and comparing genetic diversity between field and laboratory-reared *S. frugiperda*. Based on F_{st} values there is limited genetic differentiation between Texas vs. Florida populations and samples from these locations display similar genetic nucleotide diversity. Host strain differentiation was supported by nuclear and mitochondrial genome differences. Comparison of mitochondrial genomes detected two clusters where the clusters correlate by and large with host strain and resistance to the Cry1F toxin. The exceptions to this correlation suggest that C-strain individuals

based on nuclear genome have likely mated with R-strain individuals. We also observed no detectable reduction in genetic diversity in well-established *S. frugiperda* laboratory-reared colonies, which supports using laboratory strains of this pest as a model for field populations.

3.3 Methods

3.3.1 Samples and strain typing

Details of the 55 laboratory and field-collected *S. frugiperda* samples analyzed are found in Table 3.1. Adults (moths) were captured using sex pheromone baited traps [97] and collected near corn planting sites in order to optimize trap capture efficiency of C-strain males. Laboratory-reared samples were obtained from rearing facilities. Susceptibility to Cry1F toxin from *B. thuringiensis* for these laboratory samples was tested in bioassays presented elsewhere [17, 126, 146, 169]. Larval samples were collected at field and laboratory rearing locations and used when available. The collected specimens were identified as *S. frugiperda* by morphology features [74] and stored at -20 °C until required for analysis. We followed a sample naming protocol that included the first three letters representing the country of origin (Bra for Brazil, Arg for Argentina, USA for United States of America, Ken for Kenya, and Pue for Puerto Rico), the next two letters representing the first two letters of the state/province of origin (SP for Sao Paulo, TX for Texas, etc. XX when unknown), and the third letter representing the Cry1Fa susceptibility phenotype when known (“r” for resistant, “s” for susceptible and “u” for unknown). A number was used when necessary to differentiate samples with the same geographic origin and Cry1F susceptibility phenotype. Host strain (C versus R) was determined based on sequence identity at specific marker positions in reference mtCoI1164 and Tpi183 sequences, as described elsewhere [110]. The mitochondrial COI1287 marker identified all samples as C-strain, so the nuclear Tpi183 marker was used as a more reliable marker of host strain [120]. The mitochondrial COI1164 marker was not used because it did not fall under any of the pre-defined COI haplotypes [120] for any of the samples.

3.3.2 DNA extraction, library preparation, and sequencing

Genomic DNA was isolated from individual carcasses of fifth instar *S. frugiperda* larvae after dissecting the gut tissue or from legs or heads of adults using the Pure Link Genomic DNA mini kit (Invitrogen), following manufacturer’s protocols, and then quantified using a Nanodrop spectrophotometer (Thermo Scientific). Extracted DNA was then sheared randomly to between 250-500 bp using a Covaris M220 focused ultrasonicator (Woburn, MA), according to manufacturer’s instructions. The fragmented DNA was then ligated with dual-indices using a KAPA Hyper prep PCR-free library kit (Roche) according to the manufacturer’s directions. The ligated fragments were quantified using quantitative PCR and a KAPA Library quantification kit (Roche), and then submitted for sequencing on an Illumina HiSeqX device running a 2 x 150 bp paired-end configuration (Admera Health, NJ). Raw paired-end sequence reads for each sample (55) are available for download in the NCBI Sequence Read Archive (SRA) under SRR12044614-SRR12044668 with associated metadata available under NCBI BioProject id PRJNA640063. The raw data was processed to remove low quality reads using the CLC Genomics Workbench v9.5.2 (Qiagen) trim function with default parameters.

3.3.3 Filtering and mapping

Raw reads were quality-trimmed at both ends, filtered for adapter sequences and error corrected for known Illumina artifacts and PhiX sequencing control using BBduk (<https://jgi.doe.gov/data-and-tools/bbtools/>) with options trimq = 15 and filter = 23. Further quality control and confirmation that adapter and non-relevant (contaminants, primer artifacts) sequences were removed from filtered reads were performed using FastQC [7]. Given the prevalence of C-strain among our samples and similar number of variants called when using both *S. frugiperda* corn and rice genomes [122], we mapped the remaining reads to the *S. frugiperda* corn reference genome (v3.1) downloaded from https://bipaa.genouest.org/sp/spodoptera_frugiperda_pub/ (assembled length of 312 Mb across 29,949 scaffolds [86]) using a Burrows-Wheeler aligner algorithm (bwa-mem) [90]. Variants were then called using the SAMtools mpileup utility [89], resulting in detection of

126,977,977 single nucleotide polymorphisms (SNPs) and indels. The unfiltered dataset contained 120,398,863 SNPs. Three samples (USATXu1, BraSPR2 and BraSPR1) were removed from SNP analysis because they had missing variant call rate greater than 50%, resulting in 52 samples being used for SNP PCA-based analyses. For nucleotide diversity analyses, SNPs were filtered using VCFtools [35] based on both minor allele frequency being less than or equal to 0.05 and presence in at least 50% of the 52 samples surveyed. This filtering further reduced SNP number to 204,784.

3.3.4 Phylogenomics

The Mash distance method was used to measure pairwise dissimilarity of genomic sequences [125]. One key advantage of sketching-based approaches is that they neither require de novo genome assembly nor a pre-existing reference genome to identify related individuals. The Mash method estimates sequence similarity via the Jaccard similarity coefficient. All *S. frugiperda* individuals were sketched with Mash [125] using a 10,000 sketch size, k-mer size of 21 and minimum copy of each k-mer equal to 2, per previous recommendations of representing genomes with minimal computational costs [125]. All sample sketches were screened against the Mash Refseq database and other Spodoptera species (*Spodoptera exigua*, GCA_011316535.1; and *Spodoptera litura*, GCA_002706865.2) as outgroups.

3.3.5 SNP diversity

Weighted Fst values as indicators of total genetic variance in a subpopulation were calculated using VCFtools [35] by the Weir and Cockerham estimator on 1 kbp, 3 kbp and 5 kbp windows. The variation in window size was used to confirm that there was little to no change in Fst when the window was varied. For Fst calculations using the mitochondrial genome, each base pair position was compared. Nuclear genome (π) values were calculated using 1 kbp windows when comparing corn vs. rice host strain, and 3 kbp when comparing Fst values for overwintering FL/TX vs. USA laboratory samples. We also used 3 kbp windows for Fst comparisons across the nuclear genome with populations defined by mitochondria haplotype network groupings, Mash genome defined “haplotypes”, and host strain haplotypes. For all

nucleotide diversity analyses, a total of 3,558,854 SNPs were used to assess diversity using VCFtools [35].

3.3.6 Mitochondrial genome assembly

Mitochondrial sequences were extracted from the whole genome sequence reads using NOVOplasty [40] with the mitochondrial partial COI sequence MH932092.1 as the seed. A continuous complete mitochondrial chromosome of around 15 kbp was generated for all 55 samples, except for BraBar5, which was excluded from further analysis. The whole mitochondrial FASTA sequences were assembled for each sample and then aligned via bwa-mem (<https://github.com/lh3/bwa>) with default options. Mitochondrial variants were called using SAMtools mpileup based on the *S. frugiperda* corn reference genome at LepidoDB (<https://bipaa.genouest.org/sp/spodoptera.frugiperda-pub/>). The vcf files were compressed using SAMtools bgzip, indexed using SAMtools tabix and merged using SAMtools BCFtools merge [89]. Missing genotypes were replaced with sequence at the mitochondrial reference genome from the *S. frugiperda* corn genome at LepidoDB. The vcf files were then filtered by removing 16 multi-allelic sites and 92 uninformative sites due to being found in 2 samples or less via the informloci function in the adegenet R package [75], resulting in 298 sites remaining. These remaining mitochondrial sites were used to build a haplotype network using the R package pegas haploNet function [129] with Euclidean distance and an infinite site model used for building the network.

3.3.7 Structure

PLINK --make-bed [25] was used to generate the input file to FastStructure [138], a variation of Structure made for larger SNP datasets. FastStructure is a generative model-based approach based on Hardy-Weinberg equilibrium assumptions between alleles and linkage disequilibrium between genotyped loci. The FastStructure script structure.py was used to determine k, the number of assumed populations or genetic groups that share a subset of allele frequencies from $k = 2$ to 10. Then $k = 2$ was chosen by the FastStructure's choosek.py script for model complexity that maximizes marginal likelihood, and $k = 1$ for model components

to explain structure. Additionally, PCA eigenvalues and eigenvectors were generated using PLINK 2.0 [25, 34] and visualized in R using ggplot2 [170].

3.4 Results

3.4.1 Host strain typing

Of the 55 *S. frugiperda* samples used in this study, almost half (26, 47.3%) originated from field collections representing eight states in the USA. The average mapping rate, defined as the proportion of sample sequencing reads mapping to the reference genome assembly, for all samples was 88.2% with a range of between 60.62% to 96.82% range per SAMtools flagstat [89]. Identification of the C- and R- host strains for each individual was performed using nuclear Tpi and mitochondrial COI genetic markers described elsewhere [116]. Given that sample collections were from or around cornfields, we expected to observe a majority of C-strain samples. The COI1164 marker identified all samples as C-strain. In contrast, typing based on the nucleotides at the TPi183 site resulted in 42 samples identified as C-strain individuals (C/C), 8 as R-strain individuals (T/T), 3 as interstrain hybrids (C/T) and 2 that were undetermined due to low sequencing coverage at this locus. All eight identified R-strain samples were from the USA (3 from Texas, 2 from Maryland and Tennessee each, and 1 from Puerto Rico).

3.4.2 Whole genome diversity, nucleotide diversity and differentiation

The FastStructure tool was used first to determine possible population structure among the samples. The number of groups (k) was analyzed for $k = 1$ to $k = 10$, but all samples clustered as one group even when the number of groups was increased. This lack of clear population structure was further confirmed using principal component analysis (PCA) on a filtered set of 2,762,958 nuclear SNPs (see Methods). PC1 through PC10 was plotted and no structure was found based on country of origin, host strain or susceptibility to the Cry1F insecticidal protein from Bt (data not shown for PC2 to PC10). We noted that PCA using

the first two principal components uncovered weak substructure within the USA and Brazil samples, where the majority of samples clustered in a larger cluster and two subclusters formed. The first subcluster was comprised of Brazil, Florida, Texas and Tennessee samples (BraMGr1, BraMGr3, BraBAr1, USAFLu1, USAFlu3, USAFLu6, USAFLu10, USAFLu12, USATxu6, USATNu1). The second subcluster is formed by five samples, three from Florida (USAFLu9, USAFLu11, and USAFlr2) and two from Brazil (BraMGr2, BraBAr2) (Fig. 2.1). Since the three clusters do not correlate with country or city of origin, the structure observed may be an artifact of atypical migration paths due to humans where the samples that are more similar cluster together due to migration via human movement from common origin.

Given the overall lack of clear population structure observed based on origin, we next applied genome sketching [125] and traditional distance-based hierarchical clustering as a third strategy to analyze these samples. The hierarchical cluster dendrogram defined by the genome-wide distance estimates based on the fraction of shared k-mers between each pair of genomes was cut at height 0.05 as it resulted in at least two primary clusters (Texas vs. non-Texas) among the samples analyzed (Fig. 2.2).

This cut height generated 8 clusters, 6 containing less than 3 samples and were merged to the major cluster, while the remaining 10 samples were left in the remaining 2 minor cluster. The first cluster was comprised of samples from the USA (USATXu2-5 from Texas, USATNu1-2 from Tennessee, USAMDu1-2 from Maryland, USAFlu2 from Florida and PueSlu1 from Puerto Rico), while the second cluster included individuals from all countries sampled with the majority of samples from USA representing Florida/eastern seaboard. The three samples from Puerto Rico were interspersed across both clusters and always clustered with Florida samples, yet genomic distances between clusters were small (≈ 0.05 Mash distance). Similar to the PCA clusters, these two clusters did not correlate with host strain designation or geographic origin, with the first cluster including both R- and C-strain samples and the second cluster containing both R- and C-strain samples as well as three interstrain hybrids (Fig. 2.3).

Mash distance is a proxy for genome pairwise distance, with 0 representing identical genomes (all k-mers that comprise the sketches are present in both sequences). The Mash

distances across all countries surveyed were averaged to be ≈ 0.05 (Table 2, range of average distance was 0.043 to 0.045), probably reflecting recent population expansion. Moreover, these Mash distances support that R and C-strains are members of the same species, as the estimated genetic variation is comparable to at least 95% average nucleotide identity (ANI), which is the ANI cutoff for eukaryotic species [125].

3.4.3 Gene flow between overwintering populations in the USA

Previous work based on the ratios of four pre-determined COI haplotypes suggested that the overwintering Florida *S. frugiperda* population is mostly reproductively isolated from the overwintering population in Texas, with genetic exchange mostly occurring at the north and south ends of the Appalachian mountain chains [115]. As mentioned above, we observed that C-strain samples from Texas and Florida/eastern seaboard did not separate based on hierarchical clustering of genome-wide distances, and Florida samples were found across both groups (Fig. 2.2). Further, when comparing the two R-strain samples from Texas and the fourteen C-strain samples from Florida/eastern seaboard, we detected a low pairwise weighted F_{st} value (0.0111) based on 3,558,854 nuclear sites. This very low F_{st} value suggests no population separation at the nuclear level, supporting recent population separation or ongoing gene flow between Texas and Florida overwintering populations, independently of host strain. Based on the same Texas R-strain and Florida C-strain groupings, mean diversity was determined. Mean diversity for the field samples from Texas R-strain and Florida C-strain samples was estimated at 0.0631 and 0.0556, respectively (Fig. 2.4). Next, we determined the F_{st} value based on 406 mitochondrial SNPs found across all samples in these groups. The mitochondrial F_{st} for Texas R-strain and Florida C-strain samples was estimated at 0.7961. This F_{st} estimate being close to 1 suggests genetic differentiation between the R and C strains is the strongest at the mitochondrial level, possibly due to two main ancestral lineages. Although we see a stronger difference between geographic regions than between R and C strains in this study, the estimated mitochondrial to nuclear F_{st} ratio of 46.8 is consistent with previously reported sex-based differences in dispersal [142] and/or mating preference differences between strains [44]. Even so, our data suggest hybridization

occurs and some C-strain nuclear genomes have a clearly R-strain associated mitochondrial genome (see Fig. 2.6).

3.4.4 Laboratory vs. field diversity

Laboratory-reared strains of *S. frugiperda* are commonly used for diverse physiological, toxicological and genetic studies, yet their genetic resemblance to field populations has not been investigated. To investigate the differentiation level between two well established laboratory strains (>20 years) originating from the same locations in the USA (Mississippi, MS) and all field USA populations surveyed, we estimated a weighted F_{st} (fixation index). As shown in Table 1, there was moderate differentiation between the MS laboratory strains and field USA field populations, with an F_{st} range of 0.0056 to 0.1109. Moderate differentiation was also observed when comparing the MS laboratory strains with other more recently established (<10 years) laboratory populations from North Carolina (NC) and Florida (FL), with an F_{st} range of 0.0024 to 0.0874. The samples least differentiated ($F_{st} = 0.0631$) from the MS laboratory populations were collected from TX/MN locations, while the most distinct ($F_{st} = 0.1263$) were the samples from Maryland (MD). Overall, these moderate F_{st} values [50] suggest small differences at the allelic level between laboratory and field-collected groups. Genetic diversity in the two laboratory MS colonies was slightly different but comparable to diversity in field USA populations ($F_{st} = 0.0981$ on average) and to either of the two overwintering populations ($F_{st} = 0.0594$). Additionally, we surveyed overall genomic diversity using 3 kbp windows and estimated mean diversity for the same two MS laboratory strains compared to all field and laboratory *S. frugiperda* samples surveyed. Mean nuclear diversity was 0.0011 for Benzon (USAMSs1) and SIMRU (USAMSs2) laboratory strains (n=2), compared to 0.0015 for field Texas samples (n=7), and 0.0015 for field Florida/eastern USA samples (n=14).

3.4.5 Diversity across the mitochondrial genome

We investigated the maternal lineage among all but one of the *S. frugiperda* samples (due to lack of a complete mitochondrial genome in sample BraBAr4) by comparing genetic

diversity of whole mitochondrial (mt) genomes. Using a total of 390 SNPs, we generated a haplotype network as described in Materials and Methods (Fig. 2.5). Consistent with a large amount of standing variation in the species, the mitochondrial sequences appeared unique except for 22 samples forming ten haplogroups (A to J) with more than one sample (Fig. 2.4). Two of these haplogroups contained only samples from USA, haplogroup A formed by samples of the same laboratory *S. frugiperda* strain (USAMSS1 and USAMSS2), and haplogroup B containing USAFLU5 and USAFLU9. Clusters with samples only from Brazil included haplogroups C, D and E containing samples from Cry1F-resistant strains from Bahia (BraBAR2, BraBAR3 and BRABAR4), Sao Paulo (BraSPR1 and BraSPR2), and Minas Gerais (BraMGR1 and BraMGR2), respectively. Samples from Kenya clustered in haplogroup F (KenXXU3 and USAMAu2) and G (KenXXU1 and KenXXU2). Non-unique haplotype samples from Puerto Rico clustered in haplogroups H (PueGUR1 and USAFLU3) and I (PueLAU1 and USAFLU7) with samples from Florida. Haplogroup J contained two samples (BraMGR4 and BraMGR5) from the Cry1F-resistant strain from Minas Gerais also represented in haplogroup E, and one sample from USA (USAFLU6). There were two main “haplotype” clusters that separated in the network, with average genetic difference between the two clusters of 29.53, supporting evidence for two main maternal lineages within the samples surveyed. Using these two haplotype clusters as labels, we mapped the mt haplotype clusters to the two whole genome sequence (WGS) “haplotype” clusters found in Mash distance-based tree against a mitochondrial variant UPGMA distance-based tree (Fig. 2.6). In accordance with the previously estimated F_{st} , the mt and WGS clusters did not correlate completely. However, all samples in one of the mt haplotypes clustered in one of the WGS haplotypes (“one” in purple in Fig. 2.6), while the other mt haplotype contained a mix of WGS haplotypes. While not all samples were phenotyped, all samples classified as resistant to the Cry1F toxin clustered in the mt haplotype group with majority of corn strain samples.

3.5 Discussion

We report on the genome-level comparison of 55 *S. frugiperda* samples from diverse locations, mostly from the native range in the Americas but also including samples from an African

location (Kenya) where the insect is a devastating exotic pest. While using the C-strain *S. frugiperda* genome for alignments could have underestimated diversity in R-strain individuals, this approach is supported by previous surveys using both corn and rice genomes with minimal differences found based on the number of variants called and Fst values [122].

The population structure of the 55 genomic *S. frugiperda* samples was surveyed using parametric and non-parametric methods. Based on FastStructure (parametric), no population structure was found either within samples from the Eastern Hemisphere or between the Eastern and Western Hemisphere samples. Since methods based on parametric estimation of allele frequencies methods are sensitive to sample size, one possibility is that the lack of structure detected is a reflection of the relatively small number of R-strain samples (only 8 individuals and potentially 2 hybrids) analyzed. A more even distribution would address potential limited sampling of some locations surveyed, especially samples representing the Eastern Hemisphere (e.g. Africa). However, results from a previous study of genetic *S. frugiperda* diversity already supported a large interbreeding population in the Western Hemisphere[29]. Both PCA and FastStrucutre analyses found little to no population structure based on the nuclear genome. Furthermore, the Mash distance-based hierarchical clusters showed weak correlation between nuclear genome and host strain. In contrast, the Hamming Distance tree based on the mitochondrial genome detected two subpopulations within the samples surveyed, which were associated with host strain and the southern Texas and Florida subpopulations previously described based on four mitochondrial COI haplotypes [114]. Further, we identified individuals that appeared to have a C-strain nuclear genome with an R-strain associated mitochondria (maternal lineage). This discrepancy is consistent with conflicting host strain assignments obtained using specific genetic markers between R-strain and interstrain hybrids, as previously reported [106]. The lack of consistent clustering based on host strain at the nuclear genome level and the weak correlation between mitochondrial haplotype and nuclear genome groupings suggest that either there is some level of hybridization between host strains or that the current locus markers may not be sufficient to discriminate host strains. Although previous research found the groupings based on a neighbor joining tree of nuclear genome SNPs to correlate with host strain (9 corn and 9 rice *S. frugiperda* samples from Mississippi) [52], that finding could be affected by limited

sample size. Low amount of genetic variation across the whole nuclear genome between the C- and R-strains is suggestive of genome differentiation after the ecological differentiation of host strains and that reproductive and behavioral barriers are isolating C- and R-strains [55]. In testing this hypothesis, future research should involve analyses of a larger sample size of C- and R-strain individuals from Texas and Florida. Our analysis included samples collected from locations (Maryland & Tennessee) close to predicted hybridization zones for TX and FL overwintering *S. frugiperda* populations [115], and as expected these associated with both geographic regions. There were also three exceptions to the geographic separation: two Texas samples that were found in the predominately Florida represented group and one Florida sample was found in the predominately Texas population. This observation suggests that the reproductive isolation between Texas and Florida/Eastern seaboard is not absolute and mating may be occurring in overlapping locations, as previously suggested [115]. Future work should include increased sampling from potential hybrid zones (Georgia/New York) and Texas. The Mash distance-based clusters suggested that Puerto Rico samples are genetically closer to Florida, since they were found in sub-clusters with Florida-based samples. However, due to low representation and small genomic distances between each cluster, our data is not able to show conclusive support for migration models suggesting significant exchanges between *S. frugiperda* from Puerto Rico and Florida [118]. Mitochondrial F_{st} estimates obtained between C- and R- host strains (0.7961) were highly similar to the F_{st} value (0.938) from a prior analysis of 9 corn and 9 rice sympatric samples from Mississippi (USA) [52]. Moreover, the nuclear DNA F_{st} found in our study comparing all corn versus rice strains (0.0173) is also similar to the previous study's estimate (0.019) [52]. The two mitochondrial haplotypes found in our analysis could be explained by selective mating or some form of reproductive incompatibility. Despite the fact that susceptibility to Cry1F was not known for all samples, we observed that all Cry1F-resistant strains clustered to one of the mitochondrial haplotypes, despite their geographic origin. Although speculative given the comparatively low number of confirmed Cry1F resistant samples, this grouping may reflect partitioning of resistance with the haplotype group only containing C-strain samples. Further work including Cry1F-resistant R-strain samples would be necessary to test for associations between mitochondrial haplotype and resistance to Cry1F. One

possibility to explain the mitochondrial haplotypes detected could be that they are based on reproductive incompatibility from infection by parasites such as Wolbachia [33]. However, we only detected short fragments (≤ 100 bp) matching to the RefSeq Wolbachia genomes on NCBI as of May 9, 2020 using BLAST (data not shown). This observation is in agreement with leg and head tissues being mostly used for isolating genetic material. In testing for Wolbachia infection, we extracted genomic DNA from a limited number of available abdominal samples and performed sequencing of PCR amplicons using primers amplifying the outer surface protein precursor (wsp) Wolbachia gene. Amplicons of the same size (Fig. 2.7) and containing exactly the same sequence (Fig. 2.8), were detected in all samples tested independently of mitochondrial haplotype. These observations support no influence of Wolbachia infection on reproductive isolation of clustered populations. However, due to lack of detailed data on Wolbachia strains infecting *S. frugiperda*, a more detailed characterization would be needed to conclusively rule out the possibility of incompatibility. Further, lack of clear correspondence between genome-based and mitochondria-based haplotypes suggest ongoing hybridization between C- and R-strains, consistent with detection of two potential hybrids among our 54 samples. Consistent lack of concordance between the nuclear and mitochondrial F_{st} values suggests that biological mitochondria-nuclear interactions are not maintained, further supporting that the samples analyzed are part of a panmictic population with little to no population structure. Previous reports providing comparisons of genetic diversity in laboratory-reared and wild populations of the lepidopteran *Plutella xylostella* [136] Tephritid fruit flies [50], and locusts [13] support loss of allele number and heterozygosity during adaptation to mass rearing. In contrast, in Drosophilid fruit flies [153] and in long-established laboratory lines of *Anastrepha fraterculus* (Tephritidae) [131] high genetic variability remained. We compared diversity in two long-established (≥ 20 years) *S. frugiperda* laboratory lines. Both laboratory lines, Benzon (USAMSs1) and SIMRU (USAMSs2), originated from field collections in Mississippi and they were the least differentiated from the TX/MN field collections. Interestingly, more recently established laboratory populations from NC and FL [66] presented lower F_{st} values compared to field samples. This moderate differentiation between laboratory and field samples could be indicative of good culture management involving maintenance of large populations and/or regular introgression of

field individuals. Another possibility could be that some of the field samples compared display low diversity, yet this is unlikely given samples from Central and South America (for Texas population) and the Caribbean (for Florida populations) based on current *S. frugiperda* migratory models [107]. It is important to consider that despite the lack of overall differentiation between laboratory and field *S. frugiperda*, differentiation probably occurs during adaptation to mass rearing, resulting in selection of specific alleles that may not be representative of field populations. Thus, while our observations suggest laboratory strains contain genetic diversity similar to field populations, laboratory strains may have specific allele frequencies that may not be representative of field populations.

3.6 Conclusions

Overall, this study provides a diverse characterization of *S. frugiperda* whole and mitochondrial genome diversity. Findings from the study support lack of clear *S. frugiperda* population structure among the locations sampled, with only mitochondrial genomes indicating two different maternal lines mostly separating host strains. The genomic resources generated allow further exploration of gene flow and how it may affect management of *S. frugiperda* as an expanding global superpest using laboratory-reared and field-collected individuals.

Appendix: Tables and Figures

Table 3.1: List of *Spodoptera frugiperda* samples sequenced and used in this work. Shown is information on the location where collected (or facility for laboratory reference strains), Cry1F resistance phenotype (u for unknown, s for susceptible, r for resistant), collection (F for field-collected, L for lab-reared), host strain based on Tpi183 marker (corn, rice, hybrid or unknown), and the stage used for genomic DNA purification.

Sample Name	Country	City/County/Facility	Municipality/State originally collected	Cry1F	Collection	Host strain	Stage
ArgBAu1	Argentina	Buenos Aires	Buenos Aires	u	F	Corn	Moth
ArgJsl	Argentina	Estacion Experimental Agroindustrial Obispo Colombres (EEAOC, Tucuman)	Humahuaca (Jujuy)	s	L	Corn	Moth
ArgSFu1	Argentina	San Justo	Santa Fe	u	F	Corn	Moth
ArgXXr1	Argentina	Overo Pozo	Tucuman	r	L	Corn	Moth
ArgXXr2	Argentina	La Cocha	Tucuman	r	L	Corn	Moth
BraBAr1	Brazil	Luís Eduardo Magalhães	Bahia	r	L	Corn	Larva
BraBAr2	Brazil	Luís Eduardo Magalhães	Bahia	r	L	Corn	Larva
BraBAr3	Brazil	Luís Eduardo Magalhães	Bahia	r	L	Corn	Larva
BraBAr4	Brazil	Luís Eduardo Magalhães	Bahia	r	L	Corn	Larva
BraBAr5	Brazil	Luís Eduardo Magalhães	Bahia	r	L	Corn	Larva
BraMGr1	Brazil	Viçosa	Minas Gerais	r	L	Corn	Larva
BraMGr2	Brazil	Viçosa	Minas Gerais	r	L	Corn	Larva
BraMGr3	Brazil	Viçosa	Minas Gerais	r	L	Corn	Larva
BraMGr4	Brazil	Viçosa	Minas Gerais	r	L	Corn	Larva
BraMGr5	Brazil	Viçosa	Minas Gerais	r	L	Corn	Larva
BraSPr1	Brazil	Casa Branca	Sao Paulo	r	L	Corn	Larva
BraSPr2	Brazil	Casa Branca	Sao Paulo	r	L	Unknown	Larva
BraSPr3	Brazil	Casa Branca	Sao Paulo	r	L	Corn	Larva
KenXXu1	Kenya	International Centre of Insect Physiology and Ecology (ICIPE, Nairobi)	Siaya and Homa Bay counties	u	L	Corn	Moth
KenXXu2	Kenya	International Centre of Insect Physiology and Ecology (ICIPE, Nairobi)	Siaya and Homa Bay counties	u	L	Corn	Moth
KenXXu3	Kenya	International Centre of Insect Physiology and Ecology (ICIPE, Nairobi)	Siaya and Homa Bay counties	u	L	Corn	Moth
PueGUr1	Puerto Rico		Guayama	r	F	Corn	Larva
PueLAu1	Puerto Rico		Lajas	u	F	Corn	Moth
PueSLu1	Puerto Rico		Santa Isabel	u	F	Rice	Moth
USAF1r1	USA	Collier County	Florida	r	L	Corn	Larva
USAF1r2	USA	Collier County	Florida	r	L	Corn	Larva
USAF1u1	USA	Belle Glade (Palm Beach County)	Florida	u	F	Corn	Moth
USAF1u10	USA	Hague (Alachua County)	Florida	u	F	Corn	Moth
USAF1u11	USA	Hague (Alachua County)	Florida	u	F	Corn	Moth
USAF1u12	USA	Hague (Alachua County)	Florida	u	F	Corn	Moth
USAF1u2	USA	Belle Glade (Palm Beach County)	Florida	u	F	Rice	Moth
USAF1u3	USA	Miami (Miami-Dade County)	Florida	u	F	Corn	Moth
USAF1u4	USA	Miami (Miami-Dade County)	Florida	u	F	Corn	Moth
USAF1u5	USA	Miami (Miami-Dade County)	Florida	u	F	Hybrid	Moth
USAF1u6	USA	Miami (Miami-Dade County)	Florida	u	F	Corn	Moth
USAF1u7	USA	Miami (Miami-Dade County)	Florida	u	F	Corn	Moth
USAF1u8	USA	Hague (Alachua County)	Florida	u	F	Corn	Moth
USAF1u9	USA	Hague (Alachua County)	Florida	u	F	Corn	Moth
USAMDu1	USA	Jarrettsville (Harford County)	Maryland	u	F	Rice	Moth
USAMDu2	USA	Jarrettsville (Harford County)	Maryland	u	F	Rice	Moth
USAMNu1	USA	Rosemount (Dakota County)	Minnesota	u	F	Corn	Moth
USAMNu2	USA	Rosemount (Dakota County)	Minnesota	u	F	Corn	Moth
USAMSe1	USA	Benzon Research Inc (Carlisle, PA)	Mississippi	s	L	Corn	Moth
USAMSe2	USA	USDA-ARS Southern Insect Management Research Unit (SIMRU, Stoneville)	Mississippi	s	L	Corn	Moth
USANCr1	USA	Hyde County	North Carolina	r	L	Corn	Moth
USASCu1	USA	Charleston (Charleston County)	South Carolina	u	F	Corn	Moth
USASCu2	USA	Charleston (Charleston County)	South Carolina	u	F	Corn	Moth
USATNu1	USA	Crossville (Cumberland County)	Tennessee	u	F	Rice	Moth
USATNu2	USA	Crossville (Cumberland County)	Tennessee	u	F	Rice	Moth
USATXu1	USA	Corpus Christi (Nueces County)	Texas	u	F	Unknown	Moth
USATXu2	USA	College Station (Brazos County)	Texas	u	F	Rice	Moth
USATXu3	USA	College Station (Brazos County)	Texas	u	F	Rice	Moth
USATXu4	USA	Corpus Christi (Nueces County)	Texas	u	F	Hybrid	Moth
USATXu5	USA	Lubbock (Lubbock County)	Texas	u	F	Hybrid	Moth
USATXu6	USA	Lubbock (Lubbock County)	Texas	u	F	Corn	Moth

Table 3.2: Pairwise Fst values across USA populations surveyed (51 samples total). FL = Florida, PR = Puerto Rico, MD = Maryland, MN = Minnesota, SC = South Carolina, TN = Tennessee, TX = Texas, MS = Mississippi, NC = North Carolina. NA was used when the samples overlap in the defined populations.

	FL field	PR field	MD field	MN field	SC field	TN field	TX/MN field	TX field	FL lab	MS lab	NC lab
FL field	-										
PR field	0.0022	-									
MD field	0.001	0.0058	-								
MN field	0.003	0.0026	0.0257	-							
SC field	0.0004	0.002	0.0247	0.0015	-						
TN field	0.0096	0.0057	-0.001	0.0253	0.0236	-					
TX/MN field	0.0043	0.0012	-0.0026	NA	0.0066	-0.0026	-				
TX field	0.0075	0.0044	-0.0032	0.0159	0.0147	-0.0033	NA	-			
FL lab	0.0189	0.0314	0.0657	0.046	0.046	0.0655	0.0251	0.0361	-		
MS lab	0.031	0.0546	0.0961	0.0878	0.087	0.0947	0.0364	0.0485	0.0131	-	
NC lab	0.0052	-0.0032	0.0399	0.0074	0.0067	0.0389	0.0101	0.0184	0.1302	0.1302	-

Table 3.3: Average Mash distances based on country of origin based on all 55 samples.

Country of origin	Argentina	Brazil	Kenya	Puerto Rico	USA
Argentina		0.043	0.044	0.044	0.044
Brazil	0.043		0.043	0.045	0.045
Kenya	0.044	0.043		0.043	0.043
Puerto Rico	0.044	0.045	0.043		0.045
USA	0.044	0.045	0.043	0.045	

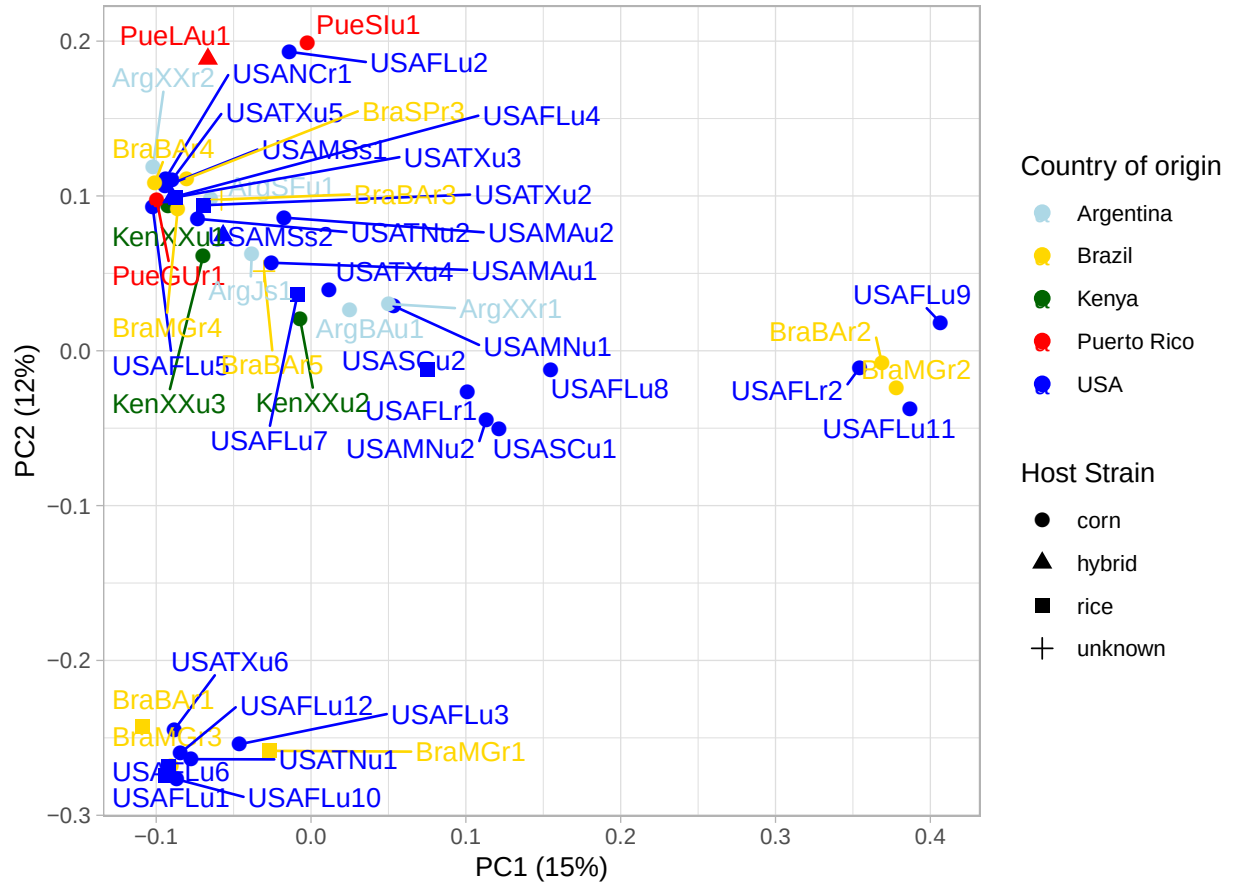


Figure 3.1: Principal component analysis (PCA) plot of 51 *Spodoptera frugiperda* samples (3 samples removed due to missing rate $\geq 50\%$) based on 2,762,958 SNPs. Country of origin and host strain as determined by the nuclear Tpi marker are visualized by color and shape, respectively, as shown in the legend.

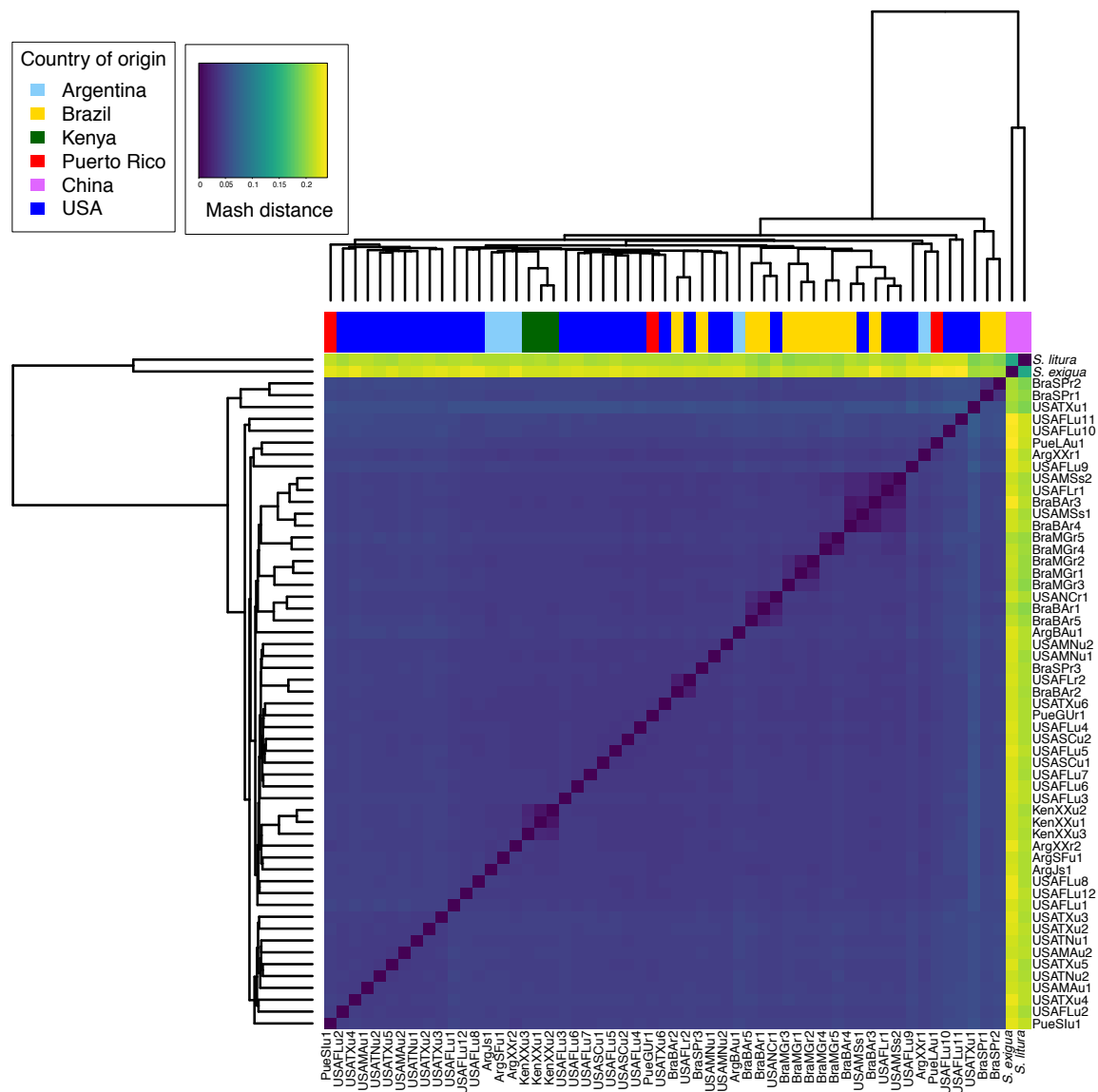


Figure 3.2: Dissimilarity heatmap of Mash distances across all 55 *Spodoptera frugiperda* samples, colored by country of origin as indicated. Two outgroups (*Spodoptera litura* and *Spodoptera exigua*) were added. Samples were named as described in Materials and Methods.

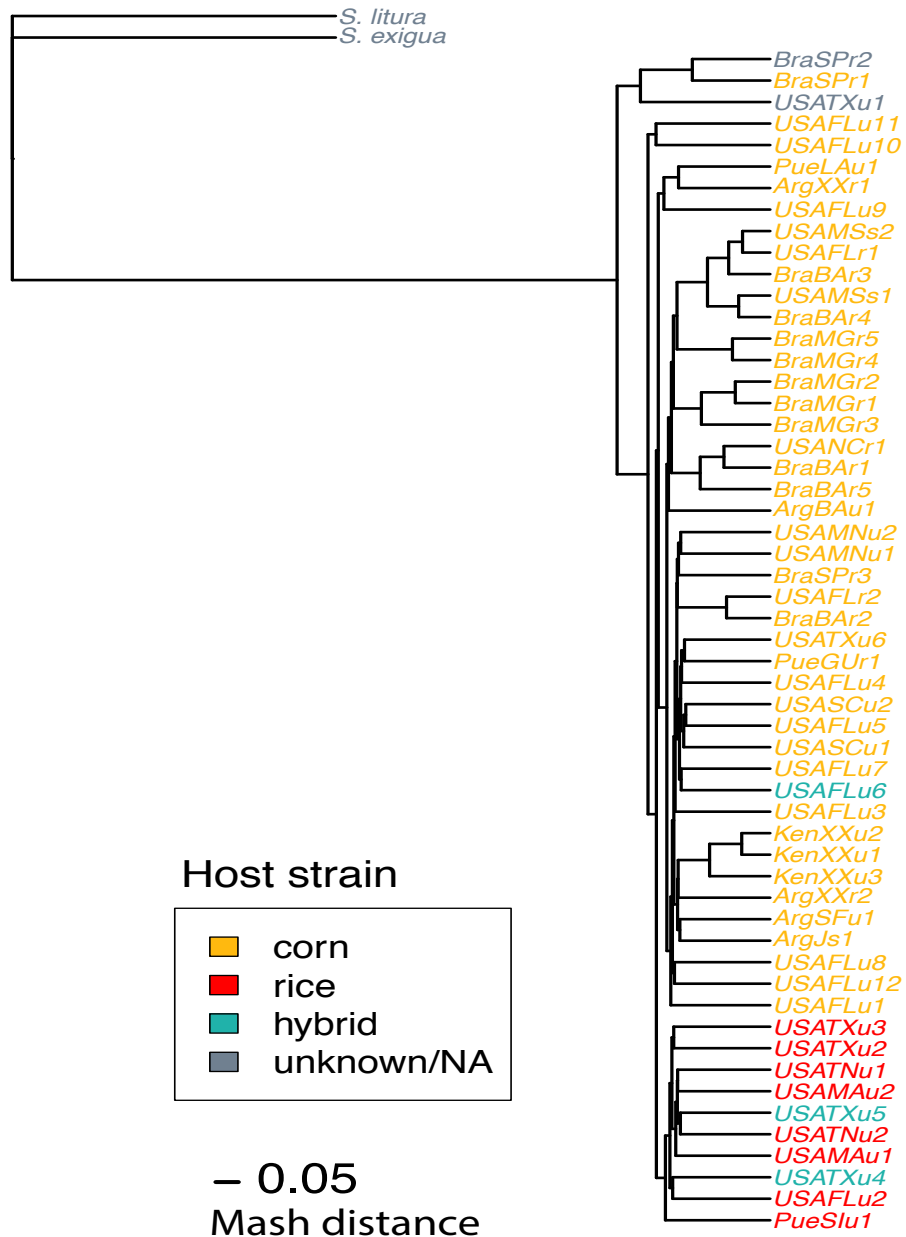


Figure 3.3: Complete clustering on Mash distances across 55 samples. Description of data: Mash distance based dendrogram on all 55 samples colored by host strain and rooted with *Spodoptera litura*.

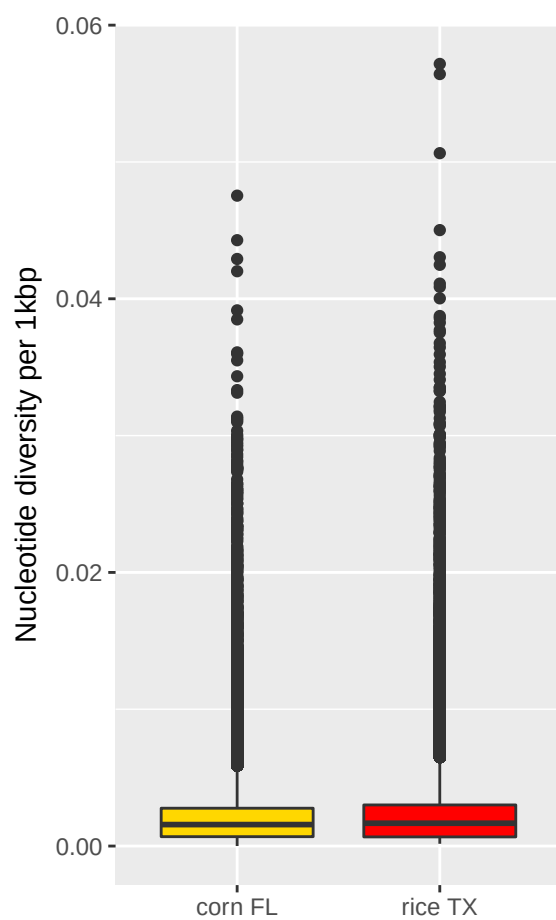


Figure 3.4: Nucleotide diversity box plots between 14 C-strain Florida (FL) vs. 2 R-strain Texas (TX) samples.

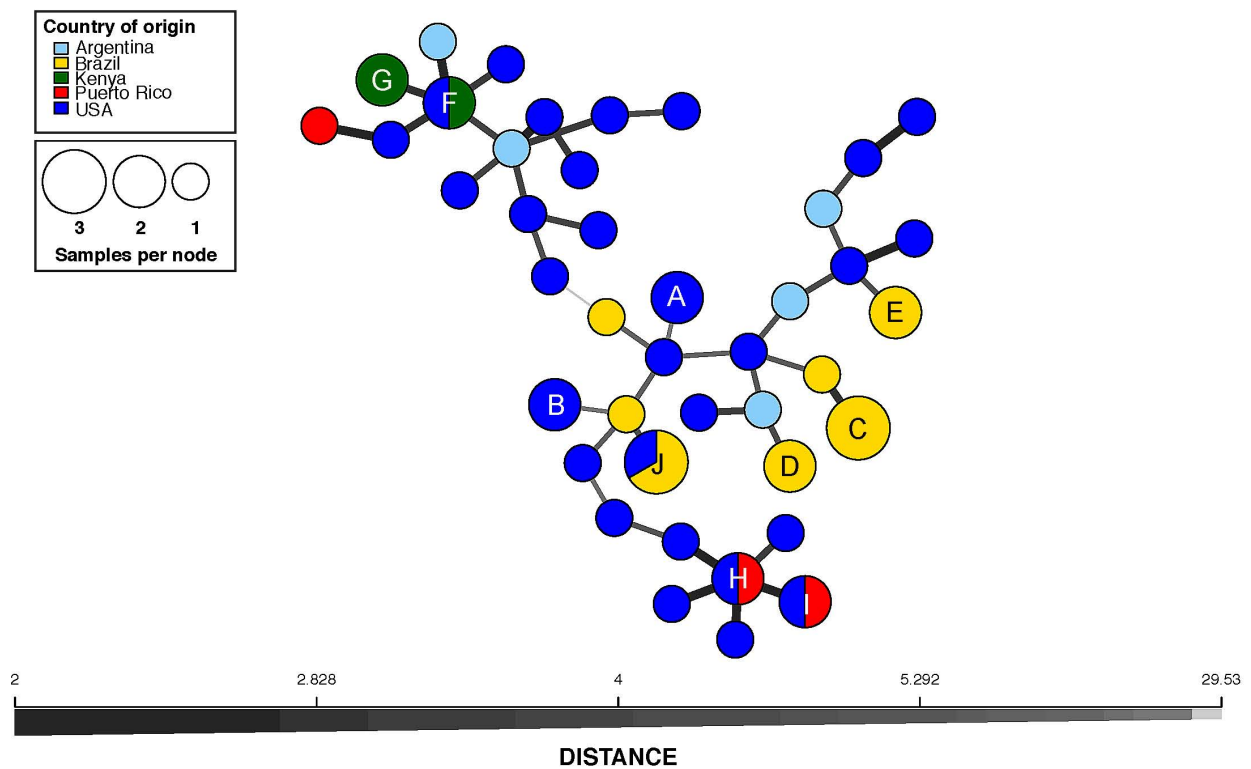


Figure 3.5: Mitochondrial haplotype network of all 54 *S. frugiperda* mitochondrial chromosome level assembled individuals, colored by country of origin.

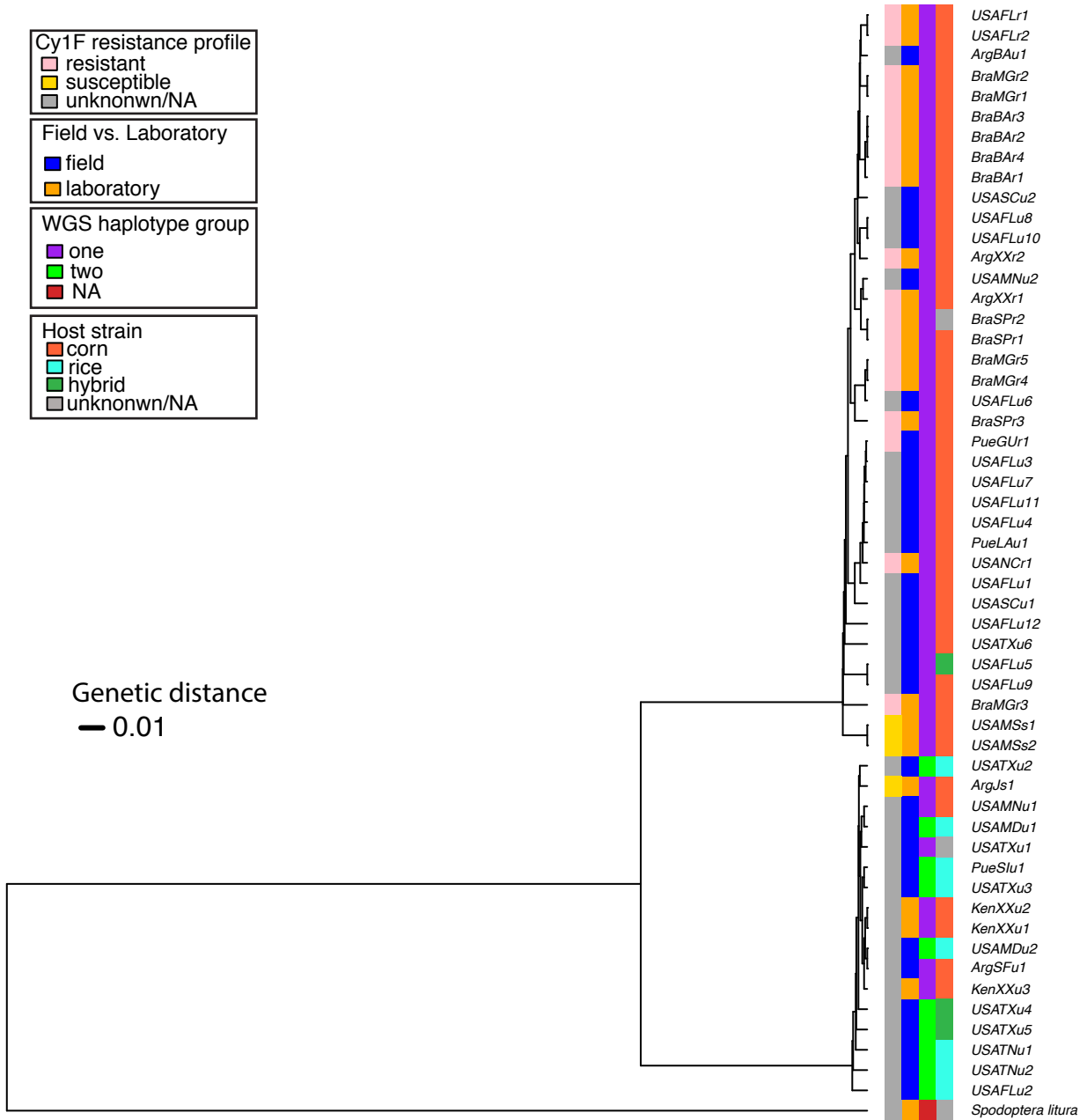


Figure 3.6: UPGMA distance-based tree on biallelic variant SNPs in 54 mitochondrial chromosome level assembled individuals rooted with *Spodoptera litura* as an outgroup.

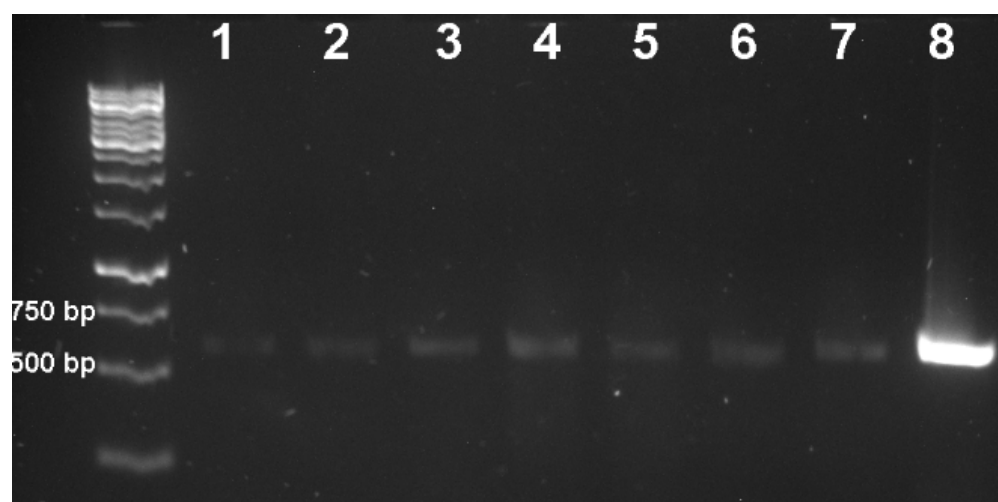


Figure 3.7: Amplification of a fragment of the wsp gene (610 bp) from *S. frugiperda* moths. A positive control (*Wolbachia pipiens*) was also included (+C). Samples shown ArgXXr2 (lane 1), PueSIu1 (lane 2), USAFLu8 (lane 3), USAFLu10 (lane 4), USAFLu12 (lane 5), USAMDu2 (lane 6), USATXu5 (lane 7) and *W. pipiens* (lane 8). Sequenced amplicons matched with 99.8% identity to *Wolbachia* endosymbiont of *Nasonia vitripennis* seqvar1 Nvit outer surface protein (wsp) gene (GenBank accession number DQ380865).

Chapter 4

Targeted genotyping reveals common resistance alleles to Cry1F corn in fall armyworm from Puerto Rico and Brazil

This following chapter is a revised version of work in submission for publication: Under submission, Schlum, K.A., Lamour, K., Tandy, P., Emrich, S.J., Placidi de Bortoli, C., Rao, T., Viteri Dillon, D., Linares Ramirez, A.M., and Jurat-Fuentes, J.L.

My primary contributions to this paper include: (i) pipeline (ii) bioinformatics data analyses, (iii) results interpretation (iv) generated results (v) writing of manuscript.

4.1 Introduction

The fall armyworm (*Spodoptera frugiperda* (J.E. Smith)) is a global invasive pest affecting numerous food and fiber staple crops, although the highest damage is observed in corn (*Zea mays* L.) [100]. In North America, *S. frugiperda* moths migrate a long distance northward over several generations from overwintering populations in southern Texas and Florida, reaching Canada in the summer months [169]. Substantial gene flow occurs between Caribbean and Florida *S. frugiperda* populations, but not with moths from South America [108]. In addition to pesticides, effective control of *S. frugiperda* has been provided by genetically modified corn and cotton producing insecticidal Cry proteins from the bacterium *Bacillus thuringiensis* (Bt). However, resistance to Cry1F, Cry1Ab and Cry1A.105 has been reported for *S. frugiperda* populations from Puerto Rico and the continental US (Florida and North Carolina) [157, 67], Brazil [152] and Argentina [24]. Resistance to these Cry proteins is recessive and genetically linked to mutations in an ABC transporter superfamily C2 gene (SfABCC2) [110, 45, 16]. In Puerto Rico two SfABCC2 resistance alleles have been reported, a nonsense 2 bp insertion (*SfABCC2mut1* allele) [110] and an insertion near the start of the fourth exon associated with aberrant splicing [45]. In contrast, resistance in Brazil is linked to missense mutations primarily localized to the longest extracellular loop in SfABCC2, suggesting a region of the fourth loop may contain a Cry protein binding domain [16].

Identification of resistance alleles advances the development of DNA-based genotyping tools which are more sensitive and less laborious and costly compared to traditional resistance screening using F1 or F2 screenings. With F1/F2 screenings, resistance allele frequencies are estimated based on mating a field collected individual whose genotype is unknown with

an individual whose phenotype is known to be resistant (RR). The screening will result in the progeny (F1) having the genotype RS if the field collected individual's genotype was susceptible (SS) or the ratio of 1RR:1RS if the field-collected parent contains one resistant allele [150]. In fact, successful SfABCC2 genotyping efforts were described using TaqMan probes [110, 45] and pyrosequencing [73] in populations from Puerto Rico, USA and Brazil. Results from these tests support high resistance allele frequencies in areas where practical resistance has been reported, further associating SfABCC2 alleles with field-relevant resistance. Unexpectedly, based on predicted migratory models [107], the *SfABCC2mut1* allele from Puerto Rico was not found at migratory Caribbean destinations or in Florida populations [110, 45], suggesting limited spread of resistance in that case. Lack of significant fitness costs associated with *SfABCC2mut1* [72] does not appear to explain the observed distribution limited to Puerto Rico. The goal of this project was to test the use of multiplexed targeted SfABCC2 sequencing to genotype field collected *S. frugiperda* as a potentially more sensitive method to monitor resistance. Here we focus our analysis on SfABCC2 alleles of *S. frugiperda* samples from Puerto Rico. Analysis of targeted SfABCC2 sequences detected the known *SfABCC2mut1* allele in high frequency (34% of samples) and identified additional eighteen novel candidate resistance alleles. Importantly, our analysis documents, for the first time, the presence of SfABCC2 resistance alleles previously described in *S. frugiperda* from Brazil in Puerto Rican fall armyworm populations. The pipeline presented here shows the power of combining targeted sequencing with population genomics to perform surveillance and field monitoring of previously resistant and candidate resistant alleles.

4.2 Materials and Methods

4.2.1 Sample Collection and genomic DNA extraction

To optimize capture efficiency of C-strain males, adult moths were captured at three locations in Puerto Rico. To optimize capture efficiency of C-strain males, adult moths were captured at three locations in Puerto Rico (Lajas, Salinas and Santa Isabel) using sex pheromone baited traps [97] at sites near corn plantings. Details of the 41 *S. frugiperda*

individuals are in Supplementary Table 1. Captured moths were confirmed as *S. frugiperda* by visual inspection [74] and kept refrigerated until shipped to the University of Tennessee. Samples were stored at -20°C until processed for purification of genomic DNA. Samples were named and numbered according to trapping location as PRL (Lajas), PRS (Salinas) and PRSI (Santa Isabel). A sample from the previously characterized Cry1F-resistant 456LSD4 strain, which also originated from collections in Puerto Rico [3, 110], was also included. Genomic DNA was isolated from moth legs using the Pure Link Genomic DNA mini kit (Invitrogen), following manufacturer’s protocols, and then quantified using a Nanodrop spectrophotometer (Thermo Scientific).

4.2.2 TaqMan genotyping for *SfABCC2mut1*

Detection of the *SfABCC2mut1* allele using Taqman probes was as previously described [110]. Briefly, 10µl (final volume) reactions included 10–20 ng of gDNA as template, a VIC-labeled probe specific to the *SfABCC2mut1* allele (5’ AA-GCACATCGCCCACTT 3’), a FAM-labeled probe specific to the wild type *SfABCC2* allele (5’ CCAAGCACATCCCACTT 3’), and forward (5’ TGGAGGCCGAAGAGAGACA 3’) and reverse (5’ AGGAGTTGACT-GACTTCATGTACCT 3’) primers. Plates (Micro Amp Fast optical 96 well reaction plate, Applied Biosystems) were loaded in the Quant studio 6 Real Time PCR instrument (Applied Biosystems) and amplified as follows: pre read stage at 60°C for 30 seconds, hold stage at 95°C for 10 minutes, PCR stage at 95°C for 15 seconds and 60°C for 1 minute for 40 cycles, post read stage at 60°C for 30 seconds. The fluorescence in each well was measured in the post read stage of the PCR, and the software generated an allelic discrimination plot based on the post amplification intensity of the fluorescent probes. Frequency of the *SfABCC2mut1* allele was determined using the Hardy-Weinberg equation with the formula: $F = (2 \times \text{ObsAa} + \text{Obsaa}) / [2 \times (\text{ObsAA} + \text{ObsAa} + \text{Obsaa})]$; where “F” is the frequency of the “a” allele (*SfABCC2mut1*) and “Obs” the observed frequency of each of the three possible genotypes.

4.2.3 Targeted DNA amplification and sequencing

Using BLASTn searches the SfABCC2 gene was found to be split between scaffolds 7154 and 11087 from the *S. frugiperda* v3.1 genome from LepidoDB [14]. Both scaffolds were combined and the resulting SfABCC2 gene sequence was used as template to design TILING primers using PCRTiler [49] at default settings allowing 50 bp overlap on each end and producing 180-200 bp final amplicon sizes. The two approximately 1Kbp intronic regions at the beginning of the gene (Scaffold 11087) were not targeted. All primers (238) were mixed into a single multiplex (119 overlapping targets) (Supplementary Table 1) and PCR performed using the Hi-Plex targeted sequencing strategy, as previously described [124]. The resulting amplicons were pooled and sequenced on a HiSeqX device running a 2x150 bp configuration and merged to produce single, high quality reads using PEAR [173]. Demultiplexing used the sample-specific barcode sequences that were incorporated during the Hi-Plex PCR reaction.

4.2.4 Genotype assignment for SfABCC2mut1 for targeted-sequencing k-mer based approach

4.2.5 Alignment, variant calling and variant analysis

The complete pipeline used for trimming, alignment and variant calling and analysis is summarized in Figure 1. Raw reads were quality-trimmed at both ends and filtered for adapter sequences using bbdduk [20]. The quality of samples was checked before and after trimming/processing to confirm adapter and non-relevant (e.g., contaminants, primer artifacts) sequences were removed using FastQC [7]. The overall ABCC2 gene interval was determined by aligning a publicly available SfABCC2 cDNA from a Cry1F-susceptible *S. frugiperda* isolate (GenBank accession numbers KY489760.1) to the *S. frugiperda* corn reference genome v6.0 [52] via the BLAST server hosted on the LepidoDB website. Reads were aligned using a Burrows-Wheeler aligner algorithm (bwa-mem) to the latest version of SfABCC2 with default options [90].

The mapping rate for all samples was determined via SAMtools flagstat [91]. Only samples with a mapping rate of 80% or higher were kept, which led to the removal of only

three samples from further analyses. We determined the count of DNA sequence reads across the ABCC2 gene via the software bam-readcount The McDonnell Genome Institute. We note that that two regions (~ 138 -1518 and ~ 7195 -7858) within the ABCC2 gene were not sequenced due to the primers being deigned based on the old assembly (version 3.0). Further, the second region was the former inter-scaffold gap, and is currently filled with ‘Ns’, precluding alignment of any reads for this region (e.g, WGS reads from [3]).

We performed alignment cleanup by sorting the bam file via SAMtools [91], deduplicated via Picard Tools [2] and then added read groups to each bam file. Variants were called via Genome Analysis Tool Kit (GATK) Haplotypecaller and default parameters were set based on the GATK Best Practice pipeline [96]. The initial variant set was 2,006 in number, all of which with a quality score greater than 40. After considering different filters based on depth and minor allele frequencies, the final filter applied was depth of coverage ≥ 8 , each individual allele depth ≥ 2 , alternate allele frequency $\geq 5\%$, and the allele must be called in at least 25% of the samples (> 10). This filter produced 1,861 variants (1,428 SNPs and 433 indels), over 92% of the unfiltered set. We further filtered this variant set to bi-allelic variants only using GATK SelectVariants function [96], which resulted in 1,333 bi-allelic variants that were used in all variant based analysis. This final set included 1,143 SNPs and 190 indels.

The intron/exon boundaries were determined by aligning the previously mentioned cDNA KY489760.1 to the full ABCC2 gene via Exonerate with est2genome model [25]. To confirm the accuracy of the intron/exon boundaries, we additionally aligned three other cDNA sequences (MG387043.1, MN399979.1 and KY646296.1). Across all four alignments of the ABCC2 cDNA to the full the ABCC2 gene, the same intron/exon boundaries were identified in the generic feature files (GFFs) generated by Exonerate. Functional annotation was performed with SNPEff [27], and SnpSift was used to filter annotations to a tabular format [26]. The final filtered set of 1,333 variants contained 17 missense and 1 nonsense mutations in total, which was only two fewer than the unfiltered variant set. One nonsense mutation was filtered out due to alternate allele frequency and read depth, and two missense mutations due to alternate allele frequency.

We also computed PCA projections via PLINK [34], and visualized them in ggplot2 [170]. Given the highly polymorphic nature of these samples, we also determined haplotypes for PCA and generating neighbor joining trees using the ShapeIT software [38]. Since, no recombination maps are available for *S. frugiperda*, the recombination rate in ShapeIT was set to 0.000023 based on previous research for *Spodoptera litura* [1].

Admixture model-based clustering was performed using FastStructure [138] in order to identify genetic structure within the ABCC2 gene. FastStructure is an clustering algorithm for inferring population structure from SNP data based on Bayesian framework. The FastStructure script structure.py was used to determine k, the number of assumed populations or genetic groups that share a subset of allele frequencies from $k = 2$ to 10. Then $k = 3$ was chosen by the FastStructure’s choosek.py script for model complexity that maximizes marginal likelihood, and $k = 4$ for model components to explain structure in the data. The final output of Structure was visualized using the distruct.py at $k = 4$.

4.3 Results

4.3.1 Genotyping results using Taqman

Out of the 41 tested *S. frugiperda* samples, 18 carried the *SfABCC2mut1* allele with nine being homozygotes. Overall, these detections suggested a frequency of 0.44 for *SfABCC2mut1* in the analyzed samples, similar to previous estimations for the frequency of this allele in Puerto Rico [9]. The relative proportion of individuals carrying the *SfABCC2mut1* allele at each location was higher at the Lajas (53% carriers, 33% homozygotes, 20% heterozygotes) compared to Salinas (41% carriers, 8% homozygotes, 33% heterozygotes) and Santa Isabel (21% homozygotes) locations.

4.3.2 Genotyping results for the 2bp INDEL using targeted sequencing

The multiplexed TILING primers amplified > 95% of the targeted regions of the ABCC2 gene. The Hi-Plex strategy does not rely on equilibration of primer efficiencies (or any kind of

primer optimization) and as such, some targets will amplify more efficiently than others (or fail completely) and thus receive more sequencing (or none at all). To accommodate these expected variations in template concentration, the samples were sequenced to high depth (> 100X per target amplicon). Of the 41 samples genotyped via targeted sequencing, 14 contained the *SfABCC2mut1* GC insertion with 8 scored as heterozygous and 6 as homozygous. There was a 74% (31/42) concordance rate between the Taqman assay based *SfABCC2mut1* genotype and the k-mer based genotype (Table 1). There were five samples identified as susceptible by the k-mer based genotyping and 3 of them were identified as heterozygotes based on Taqman assay and the other two samples as homozygote resistant. Furthermore, we identified four samples as heterozygous per the k-mer based genotype and one homozygote susceptible and three resistant samples per Taqman. Finally, there was 2 samples identified as resistant per k-mer based method and homozygote susceptible per Taqman.

4.3.3 Variant analysis pipeline and genotyping

Functional annotation of variants detected using the bioinformatics pipeline shown in Fig. 1 detected *SfABCC2mut1* (2 bp GC insertion) as a frameshift synonymous mutation. There were 18 samples in total where the *SfABCC2mut1* was potentially detected using GATK variant calling. Within these 18 samples, 12 samples were heterozygous for the mutation and 6 samples were homozygous for the mutation. Although both the k-mer method and GATK agree on the same six homozygotes, GATK identified four samples as potentially heterozygous for the *SfABCC2mut1* allele because its alternative allele threshold was much lower (≥ 3). If we apply a similar more stringent cutoff ($\geq 10X$ coverage and $> 25\%$ alternate allele frequency), GATK has a 100% concordance with the k-mer based genotyping based approach. In addition to identifying the known PRL1 allele (*SfABCC2mut1*) known to be associated with Cry1F resistance, we also detected eighteen additional nonsynonymous mutations in *SfABCC2* in the filtered dataset. The majority (17, 94%) were missense mutations with a predicted moderate effect. The most abundant missense mutation (37 individuals, 90% of samples) was M488I, which is located within the intracellular region connecting transmembrane domain 1 to the first ATP binding domain (Fig. 2). The second

most abundant (26 individuals, 63%) mutation (E1195D) was found in the second ATP binding cassette. In contrast, only one missense mutation with low frequency in this Puerto Rican collection (1 heterozygous individual) was found in the fourth extracellular loop, a region previously hypothesized to be crucial for Cry1F toxicity [16]. Only one nonsense mutation (E700*) was predicted that would result in a truncated SfABCC2 form. The mutation was detected in individual PRS_6 and is located in the first ATP binding cassette (E700*).

4.3.4 ABCC2 sequence variation

Although we do not have the Cry1F resistant phenotypes verified for these samples, we surveyed the existing sequence diversity of the ABCC2 in relation to *SfABCC2mut1* genotype. We found that first two components differentiated the SS (two susceptible alleles for *SfABCC2mut1*) versus the Sr (one susceptible and one resistant allele for the *SfABCC2mut1* genotype) and rr (two resistance alleles for the *SfABCC2mut1* genotype) (Fig. 2.A). There are two clusters based on PC1 (the two blue SS clusters) and two clusters differentiated based on PC2 (Sr and rr). The population of *SfABCC2mut1* rr genotype showed the most compact cluster while the *SfABCC2mut1* SS genotype cluster showed the least compact cluster suggesting there is more sequence variation within the SS genotyped samples compared to the rr samples. The higher sequence variation within the SS genotyped samples suggests there is a large amount of variation which exists naturally in the *ABCC2* gene and supports the alternative hypothesis that *ABCC2* resistant alleles exist naturally in the population and are selected for.

To determine if the clusters differentiated based on PCs grouped based on sample location, we plotted PC1-PC10 of the ABCC2 bi-allelic variants with data points colored by location (data shown for PC1 vs. PC2 in Figure 4.3.B). No clusters differentiated by PC1-PC10 correlated with sample location. We observed samples from Santa Isabel, Puerto Rico were found across all four clusters in PC1 versus PC2 (Figure 4.3.B) while samples from Salinas and Lajas were predominately found in two of the clusters differentiated by PC1 and PC2. We wanted to test if we can perform logistic regression association test on PC1 vs. PC2 since the susceptible for *SfABCC2mut1* associated with PC1 while PC2 associated

with heterozygote and homozygote *SfABCC2mut1*. As Figure 4.4 shows, there were eight SNPs with significant $-\log_{10}$ p-value > 5 . These SNPs associated with SNP at position 3874, 3936, 3944, 4085, 4097, 5485, 5788, and 13206 which are all SNPs not predicted to have any functional effect change by SNPeff and/or located where previously resistant Cry1F alleles are located (7034, 9085,912) in ABCC2 6.0.

Based on the PCA, it was not clear if the clusters generated are associating with population structure or the resistant allele (*SfABCC2mut1*). To determine the source of the PCA clusters, we re-ran the PCA analysis without the *SfABCC2mut1* variant and observed the same three clusters differentiated based on PC1 versus PC2 (Figure 4.6). To further elucidate the ABCC2 genetic structure of the populations in relation to the *SfABCC2mut1* locus, we ran FastStructure on the haplified ABCC2 dataset (1,333 variants). Based on the FastStructure plot (Figure 4.5), we observed four populations and the *SfABCC2mut1* allele always corresponded to one population (purple color). Furthermore, we observed admixture occurring between the purple population (haplotypes with the *SfABCC2mut1* allele) and populations where the *SfABCC2mut1* allele is not present. The frequency of the *SfABCC2mut1* allele was highest in Lajas, Puerto Rico and indicated the *SfABCC2mut1* allele likely originated from this location and spread to other fall armyworms in Puerto Rico as seen in the Structure plot. We observed moderate levels of admixture between the fall armyworms studied across all three Puerto Rican populations. Based on the population represented by the color purple which represents the *SfABCC2mut1* associated populations, the *SfABCC2mut1* has spread to other populations indicated by the samples which have at least a quarter of purple color in the barplot and has potential to spread to other populations.

4.4 Discussion

To determine if the missense and nonsense mutations we identified matched the same missense mutations found in Brazil ABCC2 samples, we compared the mutations found in our Puerto Rico samples versus the mutations found in Brazil samples. We found that the GY Brazil mutation occurred in a region where we predicted a missense mutation

at Q788P. The remaining positions where there was Brazilian mutations which included P799K/R, G1088D and the variable region between 783-789 did not align to any of the missense or nonsense mutations found in our Puerto Rico samples. We did find that the Brazil mutation, G1088D, was predicted in two of our samples (PRL_168 and PRL_164) as CODON_CHANGE_PLUS_CODON_DELETION with amino acid change DGY784D. Although these two samples are heterozygote for this Brazilian mutation with alternate allele depth of 3, one of these samples also had the *SfABCC2mut1* allele as heterozygote (PRL_168) . Future research with higher coverage of this region is required to verify this mutation exists in these two Puerto Rican samples.

We note that a similar group surveyed partially resistant Brazilian populations and did not find the 2 known Puerto Rico mutations in the Brazil samples [45] thus suggesting there are multiple ways for Cry1F resistance to occur independently in *S. frugiperda*. Although 2 bp insertion was identified in a subset of our samples, we found the second mutation in all our samples as G at position 3420 in our ABCC2 gene which was aligned to MG387070.1, one of the resistant samples identified in [45] with SNP of G in resistant and A in susceptible. Based on the Faststructure and PCA based analysis, no selective sweep was observed. The most parsimonious explanation based on our data suggests the *SfABCC2mut1* allele existed in the Puerto Rico population before Bt corn was introduced and was selected for after Bt corn was introduced. The Faststructure plot further shows that the genomic region around *SfABCC2mut1* is subject to recombination and occurring frequently within this Puerto Rican Population, more between Santa Isabel and Lajas. Furthermore, our research shows there is no barrier to gene flow for the *SfABCC2mut1* with showing introgression between highly likely resistant genomes with the *SfABCC2mut1* and genomes without the *SfABCC2mut1* allele. With the *S. frugiperda* collected from Lajas, Puerto Rico having the highest number of *SfABCC2mut1* alleles, the genomic background of these *S. frugiperda* may be more likely to have the *SfABCC2mut1* allele.

4.5 Conclusions

As previous research has shown, there are at least three mechanisms for Cry1F resistance identified with two mutations identified in Puerto Rico and two mutations identified in natural populations of Brazil fall armyworms. We verified one of mutations (the *SfABCC2mut1* allele) in at least 14 of our samples surveyed across three different methods (33%) and found evidence for the Brazilian Cry1F resistant associated mutation (GY deletion) in two of the fall armyworm samples we studied from Puerto Rico (PRL_164) which did not have the *SfABCC2mut1* allele. Our research shows there are additional nonsense and missense mutations which exist in this population of Puerto Rican samples which may be linked to Cry1F resistance and the advantage of monitoring resistant alleles via targeted sequencing since previously identified mutations and novel mutations can be identified compared to Taqman sequencing where only previously identified mutations can be verified. Furthermore, we show that there is admixture occurring between Puerto Rican populations which have the *SfABCC2mut1* allele and populations that do not have the *SfABCC2mut1* allele indicating there is barrier for gene flow for Cry1F resistant alleles and potential to spread to other non-Cry1F resistant populations.

Appendix: Tables and Figures

Table 4.1: Concordance rate between taqman assay vs. K-mer based genotype results for the *SfABCC2mut1* GC insertion represented by the allele r. The fraction of samples with the same genotype predicted are listed with the percentage in parentheses.

		K-mer based genotype		
		SS	Sr	rr
Taqman based genotype	SS	22/42 (52%)	1/42 (2%)	2/42 (5%)
	Sr	3/42 (7%)	4/42 (9%)	0/42 (0%)
	rr	2/42 (5%)	3/42 (7%)	5/42 (12%)

Table 4.2: Non-synonymous mutations identified by SnpEFF across 41 *S. frugiperda* samples within 1,331 ABCC2 bi-allelic variants.

Position	Reference allele	Alt allele	Effect	Functional Class	DNA change	Amino acid change	Number of samples affected
31	T	G	NONSYNONYMOUS CODING	MISSENSE	Tca/Gca	S11A	26
41	C	T	NONSYNONYMOUS CODING	MISSENSE	aCg/aTg	T14M	4
47	T	C	NONSYNONYMOUS CODING	MISSENSE	gTg/gCg	V16A	22
3972	A	G	NONSYNONYMOUS CODING	MISSENSE	Atc/Gtc	I295V	1
4077	C	T	NONSYNONYMOUS CODING	MISSENSE	Ctc/Ttc	L330F	2
4357	T	G	NONSYNONYMOUS CODING	MISSENSE	Ttt/Gtt	F398V	1
4540	T	C	NONSYNONYMOUS CODING	MISSENSE	gTg/gCg	V414A	9
4643	G	T	NONSYNONYMOUS CODING	MISSENSE	atG/atT	M448I	37
6020	A	G	NONSYNONYMOUS CODING	MISSENSE	cAa/cGa	Q627R	1
7077	A	G,*	NONSYNONYMOUS CODING	MISSENSE	Atg/Gtg	M754V	1
7120	C	G	NONSYNONYMOUS CODING	MISSENSE	gCt/gGt	A768G	3
9098	A	C	NONSYNONYMOUS CODING	MISSENSE	cAa/cCa	Q788P	1
9335	C	T	NONSYNONYMOUS CODING	MISSENSE	gCc/gTc	A822V	3
11708	C	T	NONSYNONYMOUS CODING+SPLICE _S TE _R REGION	MISSENSE	cCg/cTg	P1079L	1
11726	G	A	NONSYNONYMOUS CODING	MISSENSE	gGa/gAa	G1085E	10
12507	G	A	NONSYNONYMOUS CODING	MISSENSE	Gaa/Aaa	E1195K	2
12509	A	T	NONSYNONYMOUS CODING	MISSENSE	gaA/gaT	E1195D	28
12748	G	A	NONSYNONYMOUS CODING	MISSENSE	aGt/aAt	S1209N	10
13285	T	A	NONSYNONYMOUS CODING	MISSENSE	Tcc/Acc	S1310T	1
1637	C	T	STOP GAINED	NONSENSE	Caa/Taa	Q102*	1
6915	G	T	STOP GAINED+SPLICE SITE REGION	NONSENSE	Gaa/Taa	E700*	1

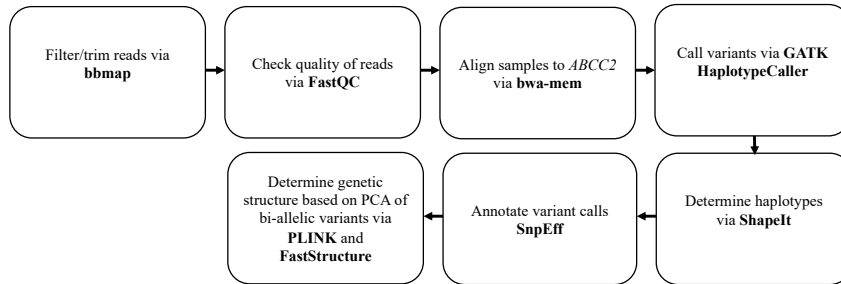


Figure 4.1: Pipeline of overall workflow from raw Fastq targeted sequencing reads to population genetics and phylo-genetic analyses based on bi-allelic SNP variant calling.

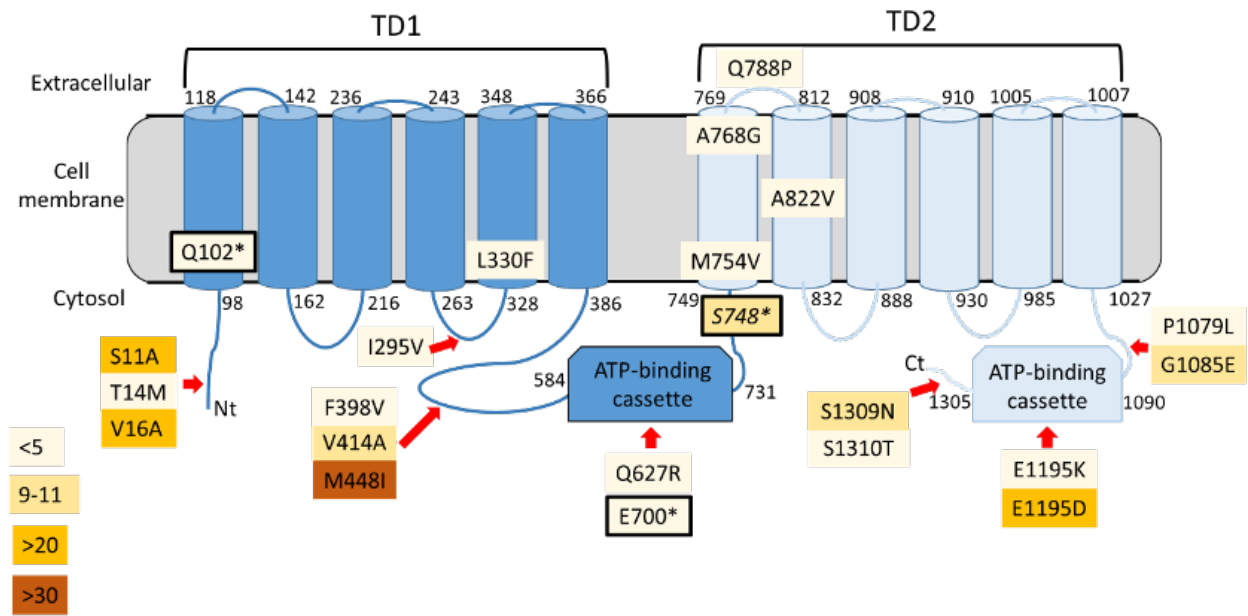


Figure 4.2: Schematic representation (not drawn to scale) of SfABCC2 two dimensional topology on the cell membrane and location of detected nonsynonymous mutations in sequenced *S. frugiperda* samples from Puerto Rico. Distinct colors as shown in the figure, indicate the number of individuals carrying each mutation nonsense mutations are shown in a bold box. Amino acid numbers are indicated, the ABCC2mut1 allele is represented as S748*. TD = transmembrane domain

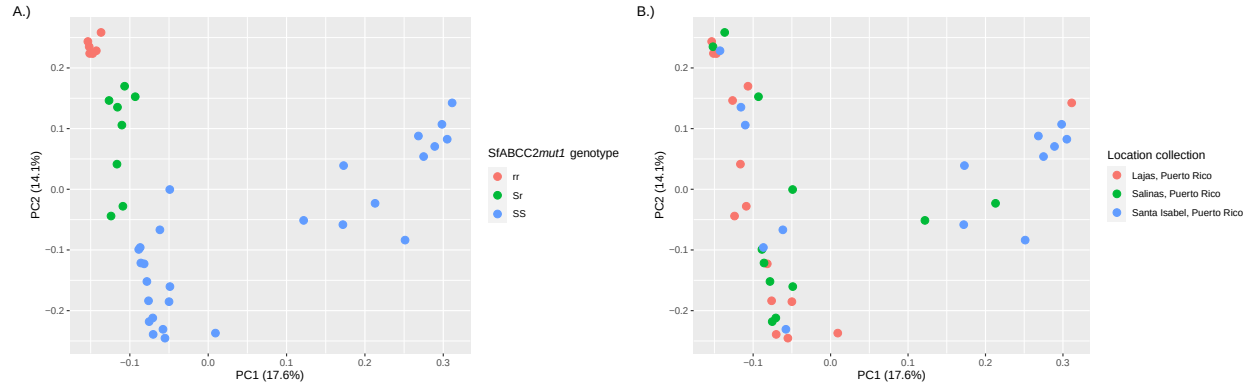


Figure 4.3: PCA projection plot of 1,333 bi-allelic variants found within ABCC2 of 41 *S. frugiperda* samples and colored by 2 GC insertion known to be associated with resistance (*SfABCC2mut1*) in panel A and PCA colored by sample location collection in panel B. The *SfABCC2mut1* was determined by k-mer based genotyping approach as described in the Methods.

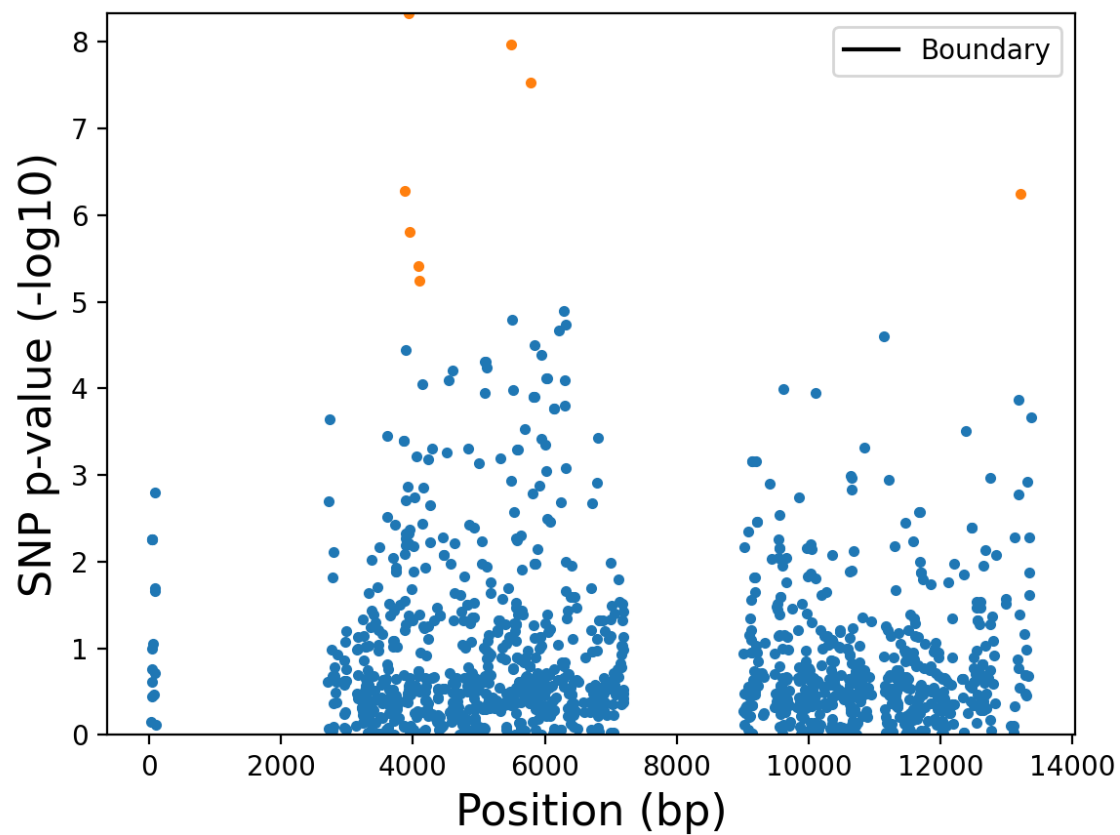


Figure 4.4: Manhattan plot of SNPs associated with PC2 on filtered bi-allelic dataset (1,333).

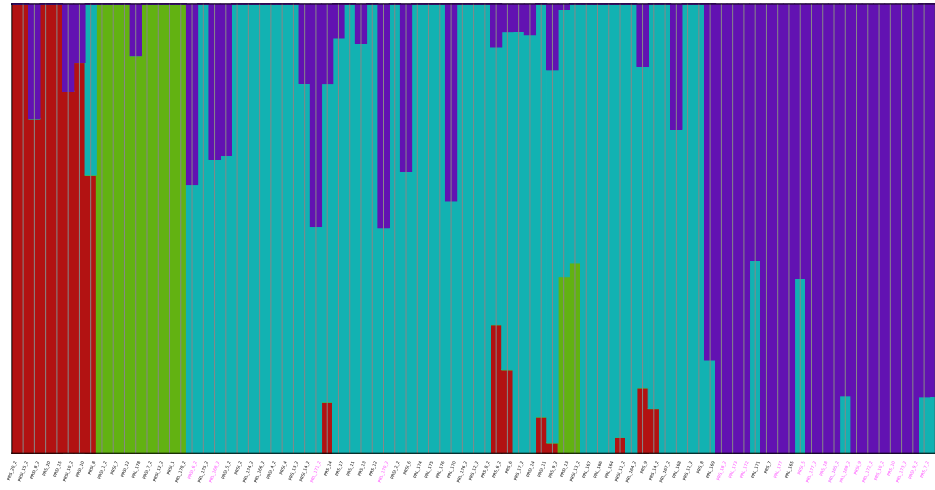


Figure 4.5: Genetic structure inferred by FastStructure on 82 haplotypes representing 41 samples at $k=4$ where samples in pink text indicate the *SfABCC2mut1* is present in the sample haplotype represented.

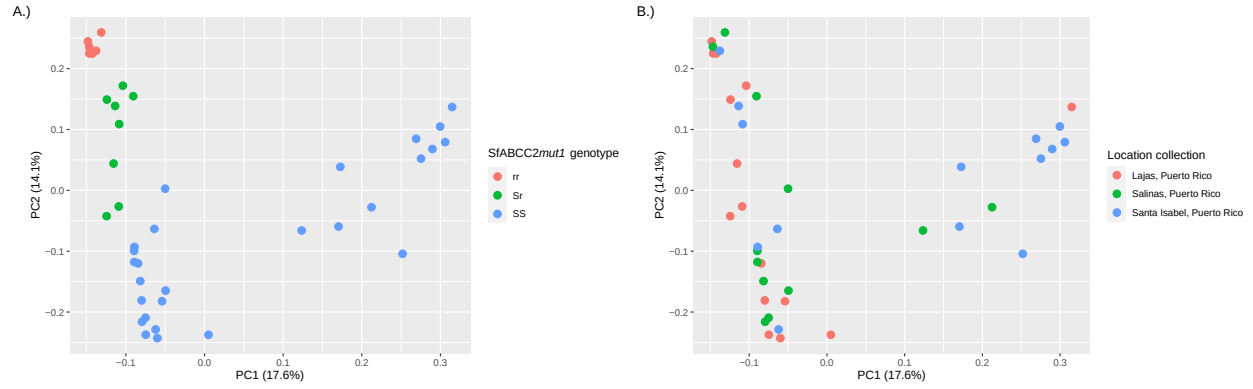


Figure 4.6: PCA projection plot of 1,332 bi-allelic variants found within *ABCC2* of 41 *S. frugiperda* samples with allele associated with 2 bp GC insertion removed. The PCA projection is colored by the samples who have the 2 bp GC insertion known to be associated with resistance (*SfABCC2mut1*) in panel A and PCA colored by sample location collection in panel B. The *SfABCC2mut1* was determined by k-mer based genotyping approach as described in the Methods.

Chapter 5

Conclusions

While WGS and targeted sequencing are both powerful approaches to identifying potential casual mutation alleles related to a disease or phenotype of interest, only nonsense, missense codon changes are identified. To comprehensively identify altered transcript splicing and gene expression levels, RNASeq based mapping and candidate gene identification of mutations would extend the current limitations of WGS and targeted sequencing mapped based approaches to identify candidate mutations. The power of RNASeq has been exemplified in zebrafish as reliable approach to harness the power of identifying candidate genes via forward genetics [99].

Based on the unsupervised methods applied for selecting the number of phylogenetic lineages and representative genome sets based on the centroid of the phylogenetic lineage/cluster, this pipeline would be useful for future researchers to run in user friendly environment. Future research would include implementing the three unsupervised methods, centroid selection and referenceseeker software into a dedicated software program which automates the entire pipeline. This would standardize the process for selecting clusters/phylogenetic lineages within a bacteria genome species and representative genome subsets from a collection of bacteria WGS.

For surveying potential alleles for pesticide resistance as we performed in chapter 4, potential mutations could be prioritized by comparing the frequency of the mutations and focusing on the most frequent mutations identified by SnpEff. Additionally, comparing conserved positions across the *ABCC2* gene in different domains of life as well as Bombyx species would allow one to prioritize mutations that are conserved across different species.

Bibliography

- [1] (2017). Genomic adaptation to polyphagy and insecticides in a major east asian noctuid pest. *Nature ecology evolution*, 1(11):1747–1756. [81](#)
- [2] (2018). Picard toolkit. <http://broadinstitute.github.io/picard/>. [80](#)
- [3] (2021). Whole genome comparisons reveal panmixia among fall armyworm (spodoptera frugiperda) from diverse locations. *BMC genomics*, 22(1):179–179. [78](#), [80](#)
- [4] Abdul-Rauf, M. and Ellar, D. J. (1999). Toxicity and receptor binding properties of a bacillus thuringiensis cryic toxin active against both lepidoptera and diptera. *Journal of invertebrate pathology.*, 73(1):52–8. [21](#), [22](#)
- [5] Abram, K., Udaondo, Z., Bleker, C., Wanchai, V., Wassenaar, T. M., Robeson, 2nd, M. S., and Ussery, D. W. (2021). Mash-based analyses of escherichia coli genomes reveal 14 distinct phylogroups. *Communications biology*, 4(1):117–117. [xiii](#), [7](#), [20](#), [24](#), [26](#), [29](#), [30](#), [34](#), [44](#)
- [6] Aeevermann, B. D., Novotny, M., Bakken, T., Miller, J. A., Diehl, A. D., Osumi-Sutherland, D., Lasken, R. S., Lein, E. S., and Scheuermann, R. H. (2018). Cell type discovery using single-cell transcriptomics: implications for ontological representation. *Human Molecular Genetics*, 27(R1):R40–R47. [25](#)
- [7] Andrews, S. (2018). Fastqc. <https://qubeshub.org/resources/fastqc>. [53](#), [79](#)
- [8] Assefa, F. and Ayalew, D. (2019). Status and control measures of fall armyworm (spodoptera frugiperda) infestations in maize fields in ethiopia: A review. *Cogent food agriculture*, 5(1):1641902–. [5](#)
- [9] Banerjee, R., Hasler, J., Meagher, R., Nagoshi, R., Hietala, L., Huang, F., Narva, K., and Jurat-Fuentes, J. L. (2017). Mechanism and dna-based detection of field-evolved resistance to transgenic bt corn in fall armyworm (spodoptera frugiperda). *Scientific reports*, 7(1):10877–10. [81](#)
- [10] Battlay, P., Schmidt, J. M., Fournier-Level, A., and Robin, C. (2016). *G3 Genes—Genomes—Genetics*, 6(8):2573–2581. [5](#)

- [11] Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E., and Clermont, O. (2018). Clermonttyping: an easy-to-use and accurate in silico method for escherichia genus strain phylotyping. *Microbial Genomics*, 4(7). 18, 26, 27
- [12] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2008). GenBank. *Nucleic Acids Research*, 37(suppl₁) : D26 – –D31. 1
- [13] Berthier, K., Chapuis, M., Simpson, S., Ferenz, H., Habib Kane, C., Kang, L., Lange, A., Ott, S., Babah Ebbe, M., Rodenburg, K., Rogers, S., Torto, B., Vanden Broeck, J., Loon, van, J., and Sword, G. (2010). Laboratory populations as a resource for understanding the relationship between genotypes and phenotypes: A global case study in locusts. *Advances in insect physiology*, 39:1–37. 63
- [14] BIPAA (2014). Spodoptera frugiperda. https://bipaa.genouest.org/sp/spodoptera_frugiperda_pub/). 79
- [15] Black IV, W. C., Baer, C. F., Antolin, M. F., and DuTeau, N. M. (2001). Population genomics: Genome-wide sampling of insect populations. *Annual review of entomology*, 46(1):441–469. 3, 4
- [16] Boaventura, D., Ulrich, J., Lueke, B., Bolzan, A., Okuma, D., Gutbrod, O., Geibel, S., Zeng, Q., Dourado, P. M., Martinelli, S., Flagel, L., Head, G., and Nauen, R. (2020). Molecular characterization of cry1f resistance in fall armyworm, spodoptera frugiperda from brazil. *Insect Biochem Mol Biol*, 116:103280. 76, 83
- [17] Broder, A. Z., Charikar, M., Frieze, A. M., and Mitzenmacher, M. (2000). Min-wise independent permutations. *Journal of Computer and System Sciences*, 60(3):630–659. 21, 52
- [18] Buchman, A., Marshall, J. M., Ostrovski, D., Yang, T., and Akbari, O. S. (2018). Synthetically engineered medea gene drive system in the worldwide crop pest drosophila suzukii. *Proceedings of the National Academy of Sciences*, 115(18):4725–4730. 4
- [19] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B.,

- and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122. [22](#)
- [20] Bushnell, B., Rood, J., and Singer, E. (2017). Bbmerge – accurate paired shotgun read merging via overlap. *PLoS ONE*, 12(10). [79](#)
- [21] Cagliari, D., Smagghe, G., Zotti, M., and Taning, C. N. T. (2020). Rnai and crispr/cas9 as functional genomics tools in the neotropical stink bug, euschistus heros. *Insects*, 11(12). [4](#)
- [22] Cai, Y., Gu, H., and Kenney, T. (2017). Learning microbial community structures with supervised and unsupervised non-negative matrix factorization. *Microbiome*, 5(1). [25](#)
- [23] Chan, A. P., Sutton, G., DePew, J., Krishnakumar, R., Choi, Y., Huang, X.-Z., Beck, E., Harkins, D. M., Kim, M., Lesho, E. P., Nikolich, M. P., and Fouts, D. E. (2015). A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of acinetobacter baumannii. *GenomeBiology.com*, 16(1):143–143. [19](#)
- [24] Chandrasena, D. I., Signorini, A. M., Abratti, G., Storer, N. P., Olaciregui, M. L., Alves, A. P., and Pilcher, C. D. (2018). Characterization of field-evolved resistance to bacillus thuringiensis-derived cry1f delta-endotoxin in spodoptera frugiperda populations from argentina. *Pest Management Science*, 74(3):746–754. [50](#), [76](#)
- [25] Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1). [55](#), [56](#)
- [26] Cingolani, P., Patel, V., Coon, M., Nguyen, T., Land, S., Ruden, D., and Lu, X. (2012a). Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Frontiers in Genetics*, 3. [80](#)
- [27] Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S., Lu, X., and Ruden, D. (2012b). A program for annotating and predicting the effects of single nucleotide

polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92. [80](#)

- [28] Ciufu, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., Brover, S., Schoch, C. L., Kimchi, A., and DiCuccio, M. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the ncbi. *International Journal of Systematic and Evolutionary Microbiology*, 68(7):2386–2392. [32](#)
- [29] Clark, P. L., Molina-Ochoa, J., Martinelli, S., Skoda, S. R., Isenhour, D. J., Lee, D. J., Krumm, J. T., and Foster, J. E. (2007). Population variation of the fall armyworm, *spodoptera frugiperda*, in the western hemisphere. *Journal of Insect Science*, 7(5):1–10. [61](#)
- [30] Clarke, T. H., Brinkac, L. M., Sutton, G., and Fouts, D. E. (2018). GGRaSP: a R-package for selecting representative genomes using Gaussian mixture models. *Bioinformatics*, 34(17):3032–3034. [19](#)
- [31] CODD, G. A., BELL, S. G., KAYA, K., WARD, C. J., BEATTIE, K. A., and METCALF, J. S. (1999). Cyanobacterial toxins, exposure routes and human health. *European journal of phycology*, 34(4):405–415. [9](#)
- [32] Cook James, M. and Delgado Ana, M. (2009). Effects of a sex-ratio distorting endosymbiont on mtdna variation in a global insect pest. *BMC evolutionary biology*, 9(1):49. [51](#)
- [33] Correa, C. C. and Ballard, J. W. O. (2016). Wolbachia associations with insects: winning or losing against a master manipulator. *Frontiers in Ecology and Evolution*, 3. [63](#)
- [34] Cristopher, C. (2019). Plink 2.0. www.cog-genomics.org/plink/2.0/. [56](#), [81](#)
- [35] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158. [54](#), [55](#)

- [36] Daniel, D. L. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788. [25](#)
- [37] Day, R., Abrahams, P., Bateman, M., Beale, T., Clottey, V., Cock, M., Colmenarez, Y., Corniani, N., Early, R., Godwin, J., Gomez, J., Moreno, P., Murphy, S., Oppong-Mensah, B., Phiri, N., Pratt, C., Silvestri, S., and Witt, A. (2017). Fall armyworm: Impacts and implications for africa. *Outlooks on Pest Management*, 28(5):196–201. [50](#)
- [38] Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181. [81](#)
- [39] Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., and Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic acids research*, 27(11):2369–2376. [9](#)
- [40] Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). Novoplasty: de novo assembly of organelle genomes from whole genome data. *Nucleic acids research*, 45(4):e18. [55](#)
- [41] Doellman, M. M., Ragland, G. J., Hood, G. R., Meyers, P. J., Egan, S. P., Powell, T. H. Q., Lazorchak, P., Glover, M. M., Tait, C., Schuler, H., Hahn, D. A., Berlocher, S. H., Smith, J. J., Nosil, P., and Feder, J. L. (2018). Genomic differentiation during speciation-with-gene-flow: Comparing geographic and host-related variation in divergent life history adaptation in *rhagoletis pomonella*. *Genes*, 9(5). [5](#)
- [42] Déraspe, M., Raymond, F., Boisvert, S., Culley, A., Roy, P. H., Laviolette, F., and Corbeil, J. (2017). Phenetic comparison of prokaryotic genomes using k-mers. *Molecular biology and evolution*, 34(10):2716–2729. [1](#), [9](#)
- [43] Farias, J. R., Andow, D. A., Horikoshi, R. J., Sorgatto, R. J., Fresia, P., dos Santos, A. C., and Omoto, C. (2014). Field-evolved resistance to cry1f maize by *spodoptera frugiperda* (lepidoptera: Noctuidae) in brazil. *Crop Protection*, 64:150–158. [50](#)
- [44] Finley, L. W. S. and Haigis, M. C. (2009). The coordination of nuclear and mitochondrial communication during aging and calorie restriction. *Ageing Research Reviews*, 8(3):173–188. [58](#)

- [45] Flagel, L., Lee, Y. W., Wanjugi, H., Swarup, S., Brown, A., Wang, J., Kraft, E., Greenplate, J., Simmons, J., Adams, N., Wang, Y., Martinelli, S., Haas, J. A., Gowda, A., and Head, G. (2018). Mutational disruption of the *abcc2* gene in fall armyworm, *spodoptera frugiperda*, confers resistance to the *cry1fa* and *cry1a.105* insecticidal proteins. *Scientific Reports*, 8(1):7255. [76](#), [77](#), [85](#)
- [46] Galili, T. (2015a). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, 31(22):3718–3720. [22](#)
- [47] Galili, T. (2015b). dendextend: an r package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*. [22](#)
- [48] Gaujoux, R. and Seoighe, C. (2010). A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11(1):367–367. [25](#)
- [49] Gervais, A. L., Marques, M., and Gaudreau, L. (2010). Pcrtiler: automated design of tiled and specific pcr primer pairs. *Nucleic acids research*, 38(Web Server issue):W308–W312. [79](#)
- [50] Gilchrist, A. S., Cameron, E. C., Sved, J. A., and Meats, A. W. (2012). Genetic consequences of domestication and mass rearing of pest fruit fly *bactrocera tryoni* (diptera: Tephritidae). *Journal of Economic Entomology*, 105(3):1051–1056. [63](#)
- [51] Goergen, G., Kumar, P. L., Sankung, S. B., Togola, A., and Tamò, M. (2016). First report of outbreaks of the fall armyworm *spodoptera frugiperda* (j e smith) (lepidoptera, noctuidae), a new alien invasive pest in west and central africa. *PLoS ONE*, 11(10). [50](#)
- [52] Gouin, A., Bretaudeau, A., Nam, K., Gimenez, S., Aury, J.-M., Duvic, B., Hilliou, F., Durand, N., Montagné, N., Darboux, I., Kuwar, S., Chertemps, T., Siaussat, D., Bretschneider, A., Moné, Y., Ahn, S.-J., Hänniger, S., Grenet, A.-S., Neunemann, D., Maumus, F., Luyten, I., Labadie, K., Xu, W., Koutroumpa, F., Escoubas, J.-M., Llopis, A., Maïbèche-Coisne, M., Salasc, F., Tomar, A., Anderson, A., Khan, S., Dumas, P., Orsucci, M., Guy, J., Belser, C., Alberti, A., Noel, B., Couloux, A., Mercier, J., Nidelet, S., Dubois, E., Liu, N.-Y., Boulogne, I., Mirabeau, O., Le Goff, G., Gordon, K., Oakeshott, J., Consoli, F., Volkoff, A.-N., Fescemyer, H., Marden, J., Luthe, D., Herrero, S., Heckel, D., Wincker, P.,

- Kergoat, G., Amselem, J., Quesneville, H., Groot, A., Jacquín-Joly, E., Nègre, N., Lemaitre, C., Legeai, F., d’Alençon, E., and Fournier, P. (2017). Two genomes of highly polyphagous lepidopteran pests (*spodoptera frugiperda*, *noctuidae*) with different host-plant ranges. *Sci Rep*, 7(1):11816–11816. [6](#), [51](#), [61](#), [62](#), [79](#)
- [53] Grada, A. and Weinbrecht, K. (2013). Next-generation sequencing: Methodology and application. *Journal of investigative dermatology*, 133(8):1–4. [11](#)
- [54] Greenbaum, G., Rubin, A., Templeton, A. R., and Rosenberg, N. A. (2019). Network-based hierarchical population structure analysis for large genomic data sets. *Genome research*, 29(12):2020–2033. [20](#)
- [55] GROOT, A. T., MARR, M., HECKEL, D. G., and SCHÖFL, G. (2010). The roles and interactions of reproductive isolation mechanisms in fall armyworm (*lepidoptera*: *Noctuidae*) host strains. *Ecological Entomology*, 35(s1):105–118. [62](#)
- [56] Gulilat, M., Lamb, T., Teft, W. A., Wang, J., Dron, J. S., Robinson, J. F., Tirona, R. G., Hegele, R. A., Kim, R. B., and Schwarz, U. I. (2019). Targeted next generation sequencing as a tool for precision medicine. *BMC medical genomics*, 12(1):81–81. [12](#)
- [57] Gutiérrez-Moreno, R., Mota-Sanchez, D., Blanco, C. A., Whalon, M. E., Terán-Santofimio, H., Rodriguez-Maciel, J. C., and DiFonzo, C. (2018). Field-Evolved Resistance of the Fall Armyworm (*Lepidoptera*: *Noctuidae*) to Synthetic Insecticides in Puerto Rico and Mexico. *Journal of Economic Entomology*, 112(2):792–802. [5](#)
- [58] Han, Y.-X., Gao, S.-G., Muegge, K., Zhang, W., and Zhou, B. (2015). Advanced applications of rna sequencing and challenges. *Bioinformatics and biology insights*, 2015(S1):29–46. [2](#)
- [59] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R’io, J. F., Wiebe, M., Peterson, P., G’erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and

Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362. 22

[60] Hay-Roe, M. M., Meagher, R. L., and Nagoshi, R. N. (2011). Effects of cyanogenic plants on fitness in two host strains of the fall armyworm (spodoptera frugiperda). *Journal of Chemical Ecology*, 37(12):1314–22. 50

[61] Hazlett, D. L. (1979). The fall armyworm, spodoptera frugiperda (j.e. smith) a review f.b. peairs and j. l. saunders. <https://revistas.zamorano.edu/index.php/CEIBA/article/view/77>. 5

[62] Herndon, N., Shelton, J., Gerischer, L., Ioannidis, P., Ninova, M., Dönitz, J., Waterhouse, R. M., Liang, C., Damm, C., Siemanowski, J., Kitzmann, P., Ulrich, J., Dippel, S., Oberhofer, G., Hu, Y., Schwirz, J., Schacht, M., Lehmann, S., Montino, A., Posnien, N., Gurska, D., Horn, T., Seibert, J., Vargas Jentzsch, I. M., Panfilio, K. A., Li, J., Wimmer, E. A., Stappert, D., Roth, S., Schröder, R., Park, Y., Schoppmeier, M., Chung, H.-R., Klingler, M., Kittelmann, S., Friedrich, M., Chen, R., Altincicek, B., Vilcinskis, A., Zdobnov, E., Griffiths-Jones, S., Ronshaugen, M., Stanke, M., Brown, S. J., and Bucher, G. (2020). Enhanced genome assembly and a new official gene set for tribolium castaneum. *BMC genomics*, 21(1):47–47. 5

[63] Horesh, G., Blackwell, G. A., Tonkin-Hill, G., Corander, J., Heinz, E., and Thomson, N. R. (2021). A comprehensive and high-quality collection of escherichia coli genomes and their genes. 19

[64] Howard-Azzeh, M., Shamseer, L., Schellhorn, H. E., and Gupta, R. S. (2014). Phylogenetic analysis and molecular signatures defining a monophyletic clade of heterocystous cyanobacteria and identifying its closest relatives. *Photosynthesis research*, 122(2):171–185. 20

[65] Hruska, A. J. and Gould, F. (1997). Fall armyworm (lepidoptera: Noctuidae) and diatraea lineolata (lepidoptera: Pyralidae): Impact of larval population level and temporal occurrence on maize yield in nicaragua. *Journal of Economic Entomology*, 90(2):611–622. 50

- [66] Huang, F., Qureshi, J. A., Meagher, R. L., J., Reisig, D. D., Head, G. P., Andow, D. A., Ni, X., Kerns, D., Buntin, G. D., Niu, Y., Yang, F., and Dangal, V. (2014a). Cry1f resistance in fall armyworm *spodoptera frugiperda*: single gene versus pyramided bt maize. *PLoS One*, 9(11):e112958. 50, 63
- [67] Huang, F., Qureshi, J. A., Meagher, Jr., R. L., Reisig, D. D., Head, G. P., Andow, D. A., Ni, X., Kerns, D., Buntin, G. D., Niu, Y., Yang, F., and Dangal, V. (2014b). Cry1f resistance in fall armyworm *spodoptera frugiperda*: Single gene versus pyramided bt maize. *PLOS ONE*, 9(11):1–10. 76
- [68] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95. 22
- [69] i5K Consortium (2013). The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. *Journal of Heredity*, 104(5):595–600. 4
- [70] Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2017). High-throughput ANI analysis of 90k Prokaryotic genomes reveals clear species boundaries. *bioRxiv*. 21, 22
- [71] Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1). 9
- [72] Jakka, S. R., Knight, V. R., and Jurat-Fuentes, J. L. (2014). Fitness costs associated with field-evolved resistance to bt maize in *spodoptera frugiperda* (lepidoptera: Noctuidae). *Journal of Economic Entomology*, 107(1):342–51. 77
- [73] Jiang, X., Langille, M. G. I., Neches, R. Y., Elliot, M., Levin, S. A., Eisen, J. A., Weitz, J. S., and Dushoff, J. (2012). Functional biogeography of ocean microbes revealed through non-negative matrix factorization. *PLOS ONE*, 7(9). 25, 77
- [74] John L., C. (2017). Fall armyworm. http://entnemdept.ufl.edu/creatures/field/fall_armyworm.htm. 52, 78

- [75] Jombart, T. and Ahmed, I. (2011). Adegnet 1.3-1: new tools for the analysis of genome-wide snp data. *Bioinformatics (Oxford, England)*, 27(21):3070. [55](#)
- [76] Kaas, R. S., Friis, C., Ussery, D. W., and Aarestrup, F. M. (2012). Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC genomics*, 13:577. [7](#), [18](#)
- [77] Karp, R. M. (2009). Reducibility among combinatorial problems. In *50 Years of Integer Programming 1958-2008*, pages 219–241. Springer Berlin Heidelberg, Berlin, Heidelberg. [32](#)
- [78] Kauff, F. and Büdel, B. (2010). Phylogeny of cyanobacteria: An overview. In *Progress in Botany 72*, Progress in Botany, pages 209–224. Springer Berlin Heidelberg, Berlin, Heidelberg. [20](#)
- [79] Kolde, R. (2018). *pheatmap: Pretty Heatmaps*. R package version 1.0.10. [25](#)
- [80] Kothari, R. and Pitts, D. (1999). On finding the number of clusters. *Pattern Recognition Letters*, 20(4):405–416. [8](#)
- [81] Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S., and Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. *Functional integrative genomics*, 15(2):141–161. [2](#), [29](#), [30](#)
- [82] Land, M. L., Hyatt, D., Jun, S.-R., Kora, G. H., Hauser, L. J., Lukjancenko, O., and Ussery, D. W. (2014). Quality scores for 32,000 genomes. *Standards in genomic sciences*, 9(1):20–. [21](#)
- [83] Lees, J. A., Harris, S. R., Tonkin-Hill, G., Gladstone, R. A., Lo, S. W., Weiser, J. N., Corander, J., Bentley, S. D., and Croucher, N. J. (2019). Fast and flexible bacterial genomic epidemiology with poppunk. *Genome research*, 29(2):304–316. [19](#)
- [84] Lees, J. A., Mai, T. T., Galardini, M., Wheeler, N. E., Horsfield, S. T., Parkhill, J., and Corander, J. (2020). Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *mBio*, 11(4). [11](#)

- [85] Leimbach, A., Hacker, J., and Dobrindt, U. (2013). *E. coli as an All-Rounder: The Thin Line Between Commensalism and Pathogenicity*, pages 3–32. Springer Berlin Heidelberg, Berlin, Heidelberg. [7](#), [18](#)
- [86] Levy, H., Garcia-Maruniak, A., and Maruniak, J. (2002). Strain identification of *spodoptera frugiperda* (lepidoptera: Noctuidae) insects and cell line; pcr-rflp of cytochrome oxidase c subunit 1 gene. *Florida Entomologist*, 85(1):186. [51](#), [53](#)
- [87] Levy, S. E. and Myers, R. M. (2016). Advancements in next-generation sequencing. *Annual review of genomics and human genetics*, 17(1):95–115. [4](#)
- [88] Li, F., Zhao, X., Li, M., He, K., Huang, C., Zhou, Y., Li, Z., and Walters, J. R. (2019). Insect genomes: progress and challenges. *Insect Molecular Biology*, 28(6):739–758. [4](#), [5](#)
- [89] Li, H. (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993. [53](#), [55](#), [56](#)
- [90] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. [53](#), [79](#)
- [91] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079. [79](#), [80](#)
- [92] Lima, M., Silva, P., Oliveira, O., Silva, K., and Freitas, F. (2010). Corn yield response to weed and fall armyworm controls. *Planta Daninha*, 28:103 – 111. [5](#)
- [93] Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M., Konstantinidis, K. T., and Doolittle, W. F. (2011). Genome sequencing of environmental *escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences - PNAS*, 108(17):7200–7205. [33](#)
- [94] Mageiros, L., Méric, G., Bayliss, S. C., Pensar, J., Pascoe, B., Mourkas, E., Calland, J. K., Yahara, K., Murray, S., Wilkinson, T. S., Williams, L. K., Hitchings, M. D., Porter,

- J., Kemmett, K., Feil, E. J., Jolley, K. A., Williams, N. J., Corander, J., and Sheppard, S. K. (2021). Genome evolution and the emergence of pathogenicity in avian *escherichia coli*. *Nature communications*, 12(1):765–765. [34](#)
- [95] MAIDEN, M. C. J. (2006). Multilocus sequence typing of bacteria. *Annual review of microbiology*, 60(1):561–588. [7](#)
- [96] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303. [80](#)
- [97] Meagher, R. L. (2001). Trapping fall armyworm (lepidoptera: Noctuidae) adults in traps baited with pheromone and a synthetic floral volatile compound. *The Florida Entomologist*, 84(2):288–292. [52](#), [77](#)
- [98] Medini, D., Donati, C., Tettelin, H., Massignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics Development*, 15(6):589–594. Genomes and evolution. [2](#)
- [99] Miller, A. C., Obholzer, N. D., Shah, A. N., Megason, S. G., and Moens, C. B. (2013). Rna-seq-based mapping and candidate identification of mutations from forward genetic screens. *Genome research*, 23(4):679–686. [96](#)
- [100] Montezano, D. G., Specht, A., Sosa-Gómez, D., Roque-Specht, V., Sousa-Silva, J., Paula-Moraes, S. V., Peterson, J., and Hunt, T. (2018). Host plants of *spodoptera frugiperda* (lepidoptera: Noctuidae) in the americas. *African Entomology*, 26(2):286–300. [50](#), [76](#)
- [101] Moreno-Hagelsieb, G., Wang, Z., Walsh, S., and ElSherbiny, A. (2013). Phylogenomic clustering for selecting non-redundant genomes for comparative genomics. *Bioinformatics*, 29(7):947–949. [10](#), [19](#)
- [102] Murphy, R., Palm, M., Mustonen, V., Warringer, J., Farewell, A., Parts, L., and Moradigaravand, D. (2021). Genomic epidemiology and evolution of *escherichia coli* in wild animals in mexico. *mSphere*, 6(1). [1](#)

- [103] Nabout, J. C., da Silva Rocha, B., Carneiro, F. M., and Sant’Anna, C. L. (2013). How many species of cyanobacteria are there? using a discovery curve to predict the species number. *Biodiversity and conservation*, 22(12):2907–2918. [20](#)
- [104] Nagoshi, R. N. (2010). The fall armyworm triose phosphate isomerase (tpi) gene as a marker of strain identity and interstrain mating. *Annals of the Entomological Society of America*, 103(2):283–292. [51](#)
- [105] Nagoshi, R. N. (2019a). Evidence that a major subpopulation of fall armyworm found in the western hemisphere is rare or absent in africa, which may limit the range of crops at risk of infestation.(research article). *PLoS ONE*, 14(4):e0208966. [51](#)
- [106] Nagoshi, R. N. (2019b). Evidence that a major subpopulation of fall armyworm found in the western hemisphere is rare or absent in africa, which may limit the range of crops at risk of infestation.(research article). *PLoS ONE*, 14(4):e0208966. [51](#), [61](#)
- [107] Nagoshi, R. N., Fleischer, S., Meagher, R. L., Hay-Roe, M., Khan, A., Murua, M. G., Silvie, P., Vergara, C., and Westbrook, J. (2017a). Fall armyworm migration across the lesser antilles and the potential for genetic exchanges between north and south american populations. *PLoS One*, 12(2):e0171743. [51](#), [64](#), [77](#)
- [108] Nagoshi, R. N., Fleischer, S., Meagher, R. L., Hay-Roe, M., Khan, A., Murúa, M. G., Silvie, P., Vergara, C., and Westbrook, J. (2017b). Fall armyworm migration across the lesser antilles and the potential for genetic exchanges between north and south american populations. *PloS one*, 12(2):e0171743. [76](#)
- [109] Nagoshi, R. N., Htain, N. N., Boughton, D., Zhang, L., Xiao, Y., Nagoshi, B. Y., and Mota-Sanchez, D. (2020). Southeastern asia fall armyworms are closely related to populations in africa and india, consistent with common origin and recent migration. *Scientific Reports*, 10(1). [50](#), [51](#)
- [110] Nagoshi, R. N., Koffi, D., Agboka, K., Tounou, K. A., Banerjee, R., Jurat-Fuentes, J. L., and Meagher, R. L. (2017c). Comparative molecular analyses of invasive fall armyworm in

- togo reveal strong similarities to populations from the eastern united states and the greater antilles. *PLoS One*, 12(7):e0181982. [50](#), [51](#), [52](#), [76](#), [77](#), [78](#)
- [111] Nagoshi, R. N. and Meagher, R. L. (2008). Review of fall armyworm (lepidoptera: Noctuidae) genetic complexity and migration. *Florida Entomologist*, 91(4):546–554. [5](#), [6](#)
- [112] Nagoshi, R. N. and Meagher, R. L. (2016). Using intron sequence comparisons in the triose-phosphate isomerase gene to study the divergence of the fall armyworm host strains. *Insect Mol Biol*, 25(3):324–37. [51](#)
- [113] Nagoshi, R. N., Meagher, R. L., Adamczyk, J. J., J., Braman, S. K., Brandenburg, R. L., and Nuessly, G. (2006). New restriction fragment length polymorphisms in the cytochrome oxidase i gene facilitate host strain identification of fall armyworm (lepidoptera: Noctuidae) populations in the southeastern united states. *Journal of Economic Entomology*, 99(3):671–7. [51](#)
- [114] Nagoshi, R. N., Meagher, R. L., Flanders, K., Gore, J., Jackson, R., Lopez, J., Armstrong, J. S., Buntin, G. D., Sansone, C., and Leonard, B. R. (2008). Using haplotypes to monitor the migration of fall armyworm (lepidoptera: Noctuidae) corn-strain populations from texas and florida. *Journal of Economic Entomology*, 101(3):742–9. [61](#)
- [115] Nagoshi, R. N., Meagher, R. L., and Hay-Roe, M. (2012). Inferring the annual migration patterns of fall armyworm (lepidoptera: Noctuidae) in the united states from mitochondrial haplotypes. *Ecol Evol*, 2(7):1458–67. [51](#), [58](#), [62](#)
- [116] Nagoshi, R. N., Nagoshi, B. Y., Canarte, E., Navarrete, B., Solorzano, R., and Garces-Carrera, S. (2019). Genetic characterization of fall armyworm (spodoptera frugiperda) in ecuador and comparisons with regional populations identify likely migratory relationships. *PLoS ONE*, 14(9):e0222332. [51](#), [56](#)
- [117] Nagoshi, R. N., Rosas-Garcia, N. M., Meagher, R. L., Fleischer, S. J., Westbrook, J. K., Sappington, T. W., Hay-Roe, M., Thomas, J. M., and Murua, G. M. (2015a). Haplotype profile comparisons between spodoptera frugiperda (lepidoptera: Noctuidae) populations

- from mexico with those from puerto rico, south america, and the united states and their implications to migratory behavior. *Journal of Economic Entomology*, 108(1):135–44. [51](#)
- [118] Nagoshi, R. N., Rosas-García, N. M., Meagher, R. L., Fleischer, S. J., Westbrook, J. K., Sappington, T. W., Hay-Roe, M., Thomas, J. M., and Murúa, G. M. (2015b). Haplotype profile comparisons between *spodoptera frugiperda* (lepidoptera: Noctuidae) populations from mexico with those from puerto rico, south america, and the united states and their implications to migratory behavior. *Journal of Economic Entomology*, 108(1):135–144. [62](#)
- [119] Nagoshi, R. N., Silvie, P., and Meagher, R. L. (2007a). Comparison of haplotype frequencies differentiate fall armyworm (lepidoptera: Noctuidae) corn-strain populations from florida and brazil. *Journal of Economic Entomology*, 100(3):954–61. [51](#)
- [120] Nagoshi, R. N., Silvie, P., Meagher, R. L., Lopez, J., and Machado, V. (2007b). Identification and comparison of fall armyworm (lepidoptera: Noctuidae) host strains in brazil, texas, and florida. *Annals of the Entomological Society of America*, 100(3):394–402. [52](#)
- [121] Nallu, S., Hill, J. A., Don, K., Sahagun, C., Zhang, W., Meslin, C., Snell-Rood, E., Clark, N. L., Morehouse, N. I., Bergelson, J., Wheat, C. W., and Kronforst, M. R. (2018). The molecular genetic basis of herbivory between butterflies and their host plants. *Nature ecology evolution*, 2(9):1418–1427. [5](#)
- [122] Nam, K., Nhim, S., Robin, S., Bretaudeau, A., Nègre, N., and d’Alençon, E. (2018). Divergent selection causes whole genome differentiation without physical linkage among the targets in *spodoptera frugiperda* (noctuidae). *bioRxiv*, page 452870. [53](#), [61](#)
- [123] National Library of Medicine (US), N. C. f. B. I. (2020). Genbank. <https://www.ncbi.nlm.nih.gov/genBank/>. [1](#)
- [124] Nguyen-Dumont, T., Teo, Z. L., Pope, B. J., Hammet, F., Mahmoodi, M., Tsimiklis, H., Sabbaghian, N., Tischkowitz, M., Foulkes, W. D., Giles, G. G., Hopper, J. L., Southey, M. C., and Park, D. J. (2013). Hi-plex for high-throughput mutation screening: application to the breast cancer susceptibility gene *palb2*. *BMC medical genomics*, 6(1):48–48. [79](#)

- [125] Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using minhash.(report). *Genome Biology (Online Edition)*, 17(1). [9](#), [54](#), [57](#), [58](#)
- [126] Oscar, F. S.-A., João, V. C. R., Thadeu, C. S., Clébson, S. T., Silverio, O. C., Raul, N. C. G., and Eliseu, J. G. P. (2015). Resistance to dual-gene bt maize in spodoptera frugiperda: selection, inheritance, and cross-resistance to other transgenic events. *Scientific Reports*, 5(1). [52](#)
- [127] Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A., and Parkhill, J. (2015a). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693. [xiii](#), [3](#), [24](#), [29](#), [44](#), [47](#)
- [128] Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., and Parkhill, J. (2015b). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics (Oxford, England)*, 31(22):3691–3693. [3](#)
- [129] Paradis, E. (2010). pegas: an r package for population genetics with an integrated-modular approach. *Bioinformatics*, 26(3):419–420. [55](#)
- [130] Park, S. T. and Kim, J. (2016). Trends in next-generation sequencing and a new era for whole genome sequencing. *International neurourology journal*, 20(Suppl 2):S76–83. [1](#), [2](#), [3](#)
- [131] Parreño, M. A., Scannapieco, A. C., Remis, M. I., Juri, M., Vera, M. T., Segura, D. F., Cladera, J. L., and Lanzavecchia, S. B. (2014). Dynamics of genetic variability in anastrepha fraterculus (diptera: Tephritidae) during adaptation to laboratory rearing conditions. *BMC genetics*, 15 Suppl 2(Suppl 2):S14. [63](#)
- [132] Pashley, D. (1989). *Host-associated differentiation in armyworms (Lepidoptera: Noctuidae): An allozyme and mitochondrial perspective*, pages 103–114. Clarendon Press, Oxford. [50](#)
- [133] Pashley, D. and Martin, J. (1987). Reproductive incompatibility between host strains of the fall armyworm (lepidoptera: Noctuidae). *Annals of the Entomological Society of America*, 80:731–733. [50](#)

- [134] Pashley, D. P. (1988). Quantitative genetics, development, and physiological adaptation in host strains of fall armyworm. *Evolution*, 42(1):93–102. 50
- [135] Patel, I. R., Gangiredla, J., Mammel, M. K., Lampel, K. A., Elkins, C. A., and Lacher, D. W. (2018). Draft genome sequences of the escherichia coli reference (ecor) collection. *Microbiology resource announcements*, 7(14). xiii, 18, 44
- [136] Perry, K. D., Keller, M. A., and Baxter, S. W. (2020). Genome-wide analysis of diamondback moth, *plutella xylostella* l., from brassica crops and wild host plants reveals no genetic structure in australia. *Scientific reports*, 10(1):12047–12047. 63
- [137] Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, 26(7):1641–1650. 24
- [138] Pritchard, J. K., Stephens, M., and Raj, A. (2014). faststructure: Variational inference of population structure in large snp data sets.(report)(author abstract). *Genetics*, 197(2):573–589. 55, 81
- [139] Queensland, B. (2020). Department of agriculture and fisheries, queensland government, australia. first mainland detection of fall armyworm. <https://www.business.qld.gov.au/industries/farms-fishing-forestry/agriculture/crop-growing/fall-armyworm/detections>. 50
- [140] Rajora, O. P. (2019). *Population Genomics : Concepts, Approaches and Applications*. Population Genomics. Springer, Cham, Switzerland. xii, 3, 15
- [141] Rasko, D. A., Myers, G. S. A., and Ravel, J. (2005). Visualization of comparative genomic analyses by blast score ratio. *BMC bioinformatics*, 6(1):2–2. 10, 19
- [142] Rašić, G., Schama, R., Powell, R., Maciel-de Freitas, R., Endersby-Harshman, N. M., Filipović, I., Sylvestre, G., Máspero, R. C., and Hoffmann, A. A. (2015). Contrasting genetic structure between mitochondrial and nuclear markers in the dengue fever mosquito from rio de janeiro: implications for vector control. *Evolutionary Applications*, 8(9):901–915. 58

- [143] Ren, J., Bai, X., Lu, Y. Y., Tang, K., Wang, Y., Reinert, G., and Sun, F. (2018). Alignment-free sequence analysis and applications. *Annual review of biomedical data science*, 1(1):93–114. [9](#)
- [144] Rouli, L., Merhej, V., Fournier, P.-E., and Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New microbes and new infections*, 7(C):72–85. [2](#)
- [145] San, J. E., Baichoo, S., Kanzi, A., Moosa, Y., Lessells, R., Fonseca, V., Mogaka, J., Power, R., and de Oliveira, T. (2019). Current affairs of microbial genome-wide association studies: Approaches, bottlenecks and analytical pitfalls. *Frontiers in microbiology*, 10:3119–3119. [11](#)
- [146] Santos-Amaya, O. F., Tavares, C. S., Rodrigues, J. V. C., Campos, S. O., Guedes, R. N. C., Alves, A. P., and Pereira, E. J. G. (2017). Fitness costs and stability of cry1fa resistance in brazilian populations of *spodoptera frugiperda*. *Pest Management Science*, 73(1):35–43. [52](#)
- [147] Schlum, K., Jun, S.-R., Udaondo, Z., Ussery, D. W., and Emrich, S. J. (2019). Improved bacteria population structure analysis on thousands of genomes using unsupervised methods. *bioRxiv*. [26](#), [29](#)
- [148] Schwengers, O., Hain, T., Chakraborty, T., and Goesmann, A. (2020). Referenceseecker: rapid determination of appropriate reference genomes. *Journal of Open Source Software*, 5(46):1994. [11](#), [24](#)
- [149] Segerman, B. (2012). The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Frontiers in cellular and infection microbiology*, 2:116–116. [3](#)
- [150] Siegfried, B. D., Rangasamy, M., Wang, H., Spencer, T., Haridas, C. V., Tenhuberg, B., Sumerford, D. V., and Storer, N. P. (2014). Estimating the frequency of cry1f resistance in field populations of the european corn borer (lepidoptera: Crambidae). *Pest management science*, 70(5):725–733. [77](#)

- [151] Sikkema-Raddatz, B., Johansson, L. F., Boer, E. N., Almomani, R., Boven, L. G., Berg, M. P., Spaendonck-Zwarts, K. Y., Tintelen, J. P., Sijmons, R. H., Jongbloed, J. D. H., and Sinke, R. J. (2013). Targeted next-generation sequencing can replace sanger sequencing in clinical diagnostics. *Human mutation*, 34(7):1035–1042. [12](#)
- [152] Silva, G. A., Picanco, M. C., Ferreira, L. R., Ferreira, D. O., Farias, E. S., Souza, T. C., Rodrigues-Silva, N., and Pereira, E. J. G. (2018). Yield losses in transgenic cry1ab and non-bt corn as assessed using a crop-life-table approach. *J Econ Entomol.* [76](#)
- [153] Simões, P., Pascual, M., Coelho, M. M., and Matos, M. (2010). Divergent evolution of molecular markers during laboratory adaptation in drosophila subobscura. *Genetica*, 138(9-10):999–1009. [63](#)
- [154] Smith, H. E., Fabritius, A. S., Jaramillo-Lambert, A., and Golden, A. (2016). Mapping challenging mutations by whole-genome sequencing. *G3: Genes, Genomes, Genetics*, 6(5):1297–1304. [11](#)
- [155] Sohn, J.-i. and Nam, J.-W. (2016). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, 19(1):23–40. [1](#)
- [156] Stoppe, N. d. C., Silva, J. S., Carlos, C., Sato, M. I. Z., Saraiva, A. M., Ottoboni, L. M. M., and Torres, T. T. (2017). Worldwide phylogenetic group patterns of escherichia coli from commensal human and wastewater treatment plant isolates. *Frontiers in Microbiology*, 8:2512. [18](#)
- [157] Storer, N. P., Babcock, J. M., Schlenz, M., Meade, T., Thompson, G. D., Bing, J. W., and Huckaba, R. M. (2010). Discovery and characterization of field resistance to bt maize: *Spodoptera frugiperda* (lepidoptera: Noctuidae) in puerto rico. *Journal of Economic Entomology*, 103(4):1031–1038. [50](#), [76](#)
- [158] Tanabe, Y., Kasai, F., and Watanabe, M. M. (2007). Multilocus sequence typing (mlst) reveals high genetic diversity and clonal population structure of the toxic cyanobacterium *microcystis aeruginosa*. *Microbiology*, 153(11):3695–3703. [20](#), [34](#)

- [159] Tay, W., Rane, R., Padovan, A., Walsh, T., Elfekih, S., Downes, S., Nam, K., d’Alençon, E., Zhang, J., Wu, Y., Nègre, N., Kunz, D., Kriticos, D., Czepak, C., Otim, M., and Gordon, K. (2020). Global faw population genomic signature supports complex introduction events across the old world. *bioRxiv*. 6
- [160] Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. (2010). The population genetics of commensal escherichia coli. *Nature reviews. Microbiology*, 8(3):207–217. 7, 17, 18
- [161] Triant, D. A., Cinel, S. D., and Kawahara, A. Y. (2018). Lepidoptera genomes: current knowledge, gaps and future directions. *Current opinion in insect science*, 25:99–105. 3, 5
- [162] Walk, S. T., Alm, E. W., Gordon, D. M., Ram, J. L., Toranzos, G. A., Tiedje, J. M., and Whittam, T. S. (2009a). Cryptic lineages of the genus escherichia. *Applied and Environmental Microbiology*, 75(20):6534–6544. 18
- [163] Walk, S. T., Alm, E. W., Gordon, D. M., Ram, J. L., Toranzos, G. A., Tiedje, J. M., and Whittam, T. S. (2009b). Cryptic lineages of the genus escherichia. *Applied and Environmental Microbiology*, 75(20):6534–6544. 18
- [164] Wan, F., Yin, C., Tang, R., Chen, M., Wu, Q., Huang, C., Qian, W., Rota-Stabelli, O., Yang, N., Wang, S., Wang, G., Zhang, G., Guo, J., Gu, L. A., Chen, L., Xing, L., Xi, Y., Liu, F., Lin, K., Guo, M., Liu, W., He, K., Tian, R., Jacquín-Joly, E., Franck, P., Siegwart, M., Ometto, L., Anfora, G., Blaxter, M., Meslin, C., Nguyen, P., Dalíková, M., Marec, F., Olivares, J., Maugin, S., Shen, J., Liu, J., Guo, J., Luo, J., Liu, B., Fan, W., Feng, L., Zhao, X., Peng, X., Wang, K., Liu, L., Zhan, H., Liu, W., Shi, G., Jiang, C., Jin, J., Xian, X., Lu, S., Ye, M., Li, M., Yang, M., Xiong, R., Walters, J. R., and Li, F. (2019). A chromosome-level genome assembly of cydia pomonella provides insights into chemical ecology and insecticide resistance. *Nature communications*, 10(1):4237–4237. 5
- [165] Wanchai, V., Patumcharoenpol, P., Nookaew, I., and Ussery, D. (2017). dbbqs: database of bacterial quality scores. *BMC bioinformatics*, 18(Suppl 14):483–153. 21
- [166] Waters, N. (2020 (accessed September 4, 2020)). *EzClermont: The E. coli Clermont PCR phylotyping tool*. 18

- [167] Waters, N. R., Abram, F., Brennan, F., Holmes, A., and Pritchard, L. (2020). Easy phylotyping of escherichia coli via the ezclermont web app and command-line tool. *Access Microbiology*, 2(9). [21](#)
- [168] Weetman, D., Wilding, C. S., Neafsey, D. E., Müller, P., Ochomo, E., Isaacs, A. T., Steen, K., Rippon, E. J., Morgan, J. C., Mawejje, H. D., Rigden, D. J., Okedi, L. M., and Donnelly, M. J. (2018). Candidate-gene based gwas identifies reproducible dna markers for metabolic pyrethroid resistance from standing genetic variation in east african anopheles gambiae. *Scientific reports*, 8(1):2920–12. [5](#)
- [169] Westbrook, J., Fleischer, S., Jairam, S., Meagher, R., and Nagoshi, R. (2019). Multigenerational migration of fall armyworm, a pest insect. *Ecosphere (Washington, D.C)*, 10(11):n/a. [50](#), [52](#), [76](#)
- [170] Wickham, H. (2016). ggplot2 elegant graphics for data analysis. [56](#), [81](#)
- [171] Wiegmann, B. M. and Richards, S. (2018). Genomes of diptera. *Current Opinion in Insect Science*, 25:116 – 124. Insect genomics * Development and regulation. [4](#)
- [172] Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., Karch, H., Reeves, P. R., Maiden, M. C. J., Ochman, H., and Achtman, M. (2006). Sex and virulence in escherichia coli: an evolutionary perspective. *Molecular Microbiology*, 60(5):1136–1151. [7](#), [18](#), [20](#)
- [173] Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). Pear: a fast and accurate illumina paired-end read merger. *Bioinformatics (Oxford, England)*, 30(5):614–620. [79](#)
- [174] Zieleszinski, A., Vinga, S., Almeida, J., and Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome biology*, 18(1):186–186. [9](#)

Vita

Katrina A. Schlum was born in Lockport, NY. She obtained her Bachelor's degree in Bioinformatics at St. Bonaventure University. Before she came to Knoxville, she spent two years in healthcare technology field where she worked with parsing data from insurance plan documents and electronic medical records. This experience showed her the data you work with matters. She came to University of Tennessee at Knoxville to advance her understanding of bioinformatics and work with the data she truly cares about, genomics.