Genotyping and Admixture Mapping of North American *Fragaria chiloensis* Populations

by
Jennifer Devin

A THESIS

submitted to

Oregon State University

University Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Chemical Engineering
(Honors Scholar)

Presented May 19, 2016
Commencement June 2016

AN ABSTRACT OF THE THESIS OF

Jennifer N. Devin for the degree of Honors Baccalaureate of Science in Chemical Engineering presented on May 19, 2016. Title: Genotyping and Admixture Mapping of North American *Fragaria chiloensis* Populations.

Abstract approved:

_____

Aaron Liston

Assembly and genotyping of polyploid genomes poses many problems due to the complex nature of polyploid data, including the inability to determine allelic dosage and mixed inheritance patterns. This study was the first application of POLiMAPS genotyping on whole genome sequences without using cross information. In this study, whole genome sequencing data from twenty samples of the octoploid *Fragaria chiloensis*, collected from the northwest coast of North America, were assembled using BBTools. The sequences were genotyped using a POLiMAPS script and analyzed using principal component analysis in the R programming language. STRUCTURE was used to analyze the degree of admixture of the *F. chiloensis* individuals with *F. virginiana*. Principal component analysis and admixture mapping results show clustering of individuals from the same geographical area. Differences in read depth likely account for observations inconsistent with the expected results from population and species boundaries. An increased number of SNPs were called in sequences with higher read depth, which can cause unrelated sequences to appear similar, and for individuals of the same population to show differing clustering patterns.

Key Words: *Fragaria*, Strawberry, Genotype, Admixture, POLiMAPS

Corresponding e-mail address: devinjeni@hotmail.com

Genotyping and Admixture Mapping of North American *Fragaria Chiloensis* Populations

by
Jennifer N. Devin

A THESIS

submitted to

Oregon State University

University Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Chemical Engineering
(Honors Scholar)

Presented May 19, 2016
Commencement June 2016

Honors Baccalaureate of Science in Chemical Engineering project of Jennifer N. Devin presented on May 19, 2016.

APPROVED:

Aaron Liston, Mentor, representing Botany and Plant Pathology

Jacob Tennessen, Committee Member, representing Integrative Biology

Markus Dillenberger, Committee Member, representing Botany and Plant Pathology

Toni Doolen, Dean, University Honors College

I understand that my project will become part of the permanent collection of Oregon State University, University Honors College. My signature below authorizes release of my project to any reader upon request.

Jennifer N. Devin, Author

ACKNOWLEDGEMENTS

I would like to thank Aaron Liston for his guidance and support throughout this project. You were always quick to help when I had questions, and explain the theory behind the work I was doing. I really appreciate the time you took to help me out with the projects I've worked on over the years.

I would also like to thank Jacob Tennessen and Markus Dillenberger for their assistance with this project. You were so helpful in teaching me about the programs I needed to use for this project. I extend this thanks to everyone else I have worked with in the lab over the years, especially Shannon Straub and Kevin Weitemier.

In addition, I would like to thank my parents for their continuing support throughout my years in college. With your help and encouragement throughout all my time in school, I will graduate with my degree in chemical engineering. I look forward to seeing you at graduation.

Finally, I would like to thank everyone who has listened to my ramblings about strawberry DNA over the last three years. This is an interesting subject to me, and has been a major part of my college experience. Thank you so much for encouraging me to talk about and pursue this research.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

Genotyping and Admixture Mapping of North American *Fragaria chiloensis* Populations

INTRODUCTION

Fragaria chiloensis

The strawberries (genus Fragaria) are a diverse group of plants conducive to evolutionary genetic studies. *Fragaria chiloensis*, *F. virginiana*, and *F.* × *ananassa* are octoploid species, containing eight copies of each chromosome. *Fragaria* × *ananassa* subsp. *ananassa*, the primary cultivated species of strawberry, is a hybrid of *F. chiloensis* and *F. virginiana* (Liston et al, 2014).

Native *F. chiloensis* (Chilean strawberry) populations are found on the beaches and mountains of central and southern Chile, Hawaii, and the coast of North America from central California to the Aleutian Islands. Native *F. virginiana* is found in the woodlands and meadows of the United States and Canada (Hancock et al, 1999). The ancestral octoploid species likely differentiated into *F. chiloensis* and *F. virginiana* as it moved south and adapted to coastal and mountain environments (Hancock et al, 1999).

The relationship between diploid *Fragaria* species and parental genomes of polyploid species, as well as the evolutionary origin of the octoploid strawberry species, has been explored using dense linkage maps (Tennessen et al 2014). Through this study, the origins of the four subgenomes of octoploid species *F. virginiana* and *F. chiloensis* were identified. One of the four subgenomes was found to originate with diploid *Fragaria vesca*, one with the diploid *Fragaria iinumae*, and two with an unknown ancestor similar to *F. iinumae*. This suggests that a *F. vesca* diploid hybridized with a *F. iinumae* diploid to form an allotetraploid, which then hybridized

with the unknown *F. innumae* – like autotetraploid to form the octoploid ancestor to *F. virginiana* and *F. chiloensis* (Tennessen et al, 2014).

Genotyping Polyploids

Assembly and genotyping of polyploid genomes poses many problems due to the complex nature of polyploid data, including multiple alleles and mixed inheritance patterns (Durfresne et al, 2014). As a result, tools developed for diploid population genetics often cannot be applied to polyploid populations, or exhibit problems with difficulty resolving allelic dosage and uneven amplification of alleles (Durfresne et al, 2014).

For example, polyploids can exhibit allopolyploidy, where polyploidy originates from hybridization of different species and resultant genome doubling (Dufresne et al, 2014). The chromosomes from different ancestral species (homeologues) may pair in meiosis and produce viable gametes. As a result, allopolyploid individuals may show a mixture of disomic and polysomic inheritance patterns, while most genotyping methods assume a single mode of inheritance (Dufresne et al, 2014).

Also, applications of PCR-based techniques, such as Sanger sequencing and microsatellite data, on polyploid DNA pose several problems due to bias inherent in the method. For example, allelic dosage cannot be determined due to recombination during the PCR process. Recombination can produce artifacts, including uneven amplification and missing or null alleles, which can occur because some alleles amplify less strongly than others (Dufresne et al, 2014). This bias causes issues in

distinguishing real recombination from PCR based artefacts. This bias can be resolved using segregation analyses and cloning (Mable et al, 2004).

Current methods of genotyping polyploids include multivariate analyses, such as K-means clustering (Hartigan and Wong, 1979), and discriminant analysis of principal components (Jombart et al, 2010). These methods allow clustering of polyploid populations using SNPs. Genetic distance measures include amplified fragment length polymorphism markers (AFLPs), though this method can result in loss of information because it does not take into account allele dosage for polyploid heterozygotes (Hol et al, 2008).

Additional methods of genotyping polyploids involve the use of microsatellite loci, nuclear genomes, and custom simulations, such as the isolation-with-migration and approximate Bayesian computation model used to analyze the mode of polyploidization of *C. bursa-pastoris* (St-Onge et al, 2012). For example, tetraploid population simulations with test for different evolutionary scenarios have been developed. Custom models based on these preliminary simulations have been developed to take reproduction modes, mutation rates, and demographic history into account (Arnold et al, 2012).

As next generation sequencing methods become widespread, the quantity of polyploid genetic data increases, and developing tools for polyploid population genetics becomes increasingly important.

Principal Component Analysis

The basic objective of principal component analysis is to reduce the rank of data by replacing the original variables with linear combinations of the data. This reduces the effect of noise in the data and allows for easier interpretation and visualization of the data by reducing the number of variables while retaining the informative aspects of the data (Bro and Smilde, 2014). In univariable analysis, co-variation with other variables, such linkage between SNPs due to geographic region, is ignored, which often leads to important features being ignored. Principal component analysis forms new variables from a linear combination of the original variables, a "weighted average" of sorts, preserving the size of the original variables (Bro and Smilde, 2014).

Admixture Mapping

Admixture refers to the intermixing of genetic information between two closely related taxa. *Fragaria chiloensis* and *F. virginiana* diverged 0.19 to 0.86 million years ago, so admixed individuals from *F. chiloensis* populations may reflect recent divergence of the species (Salamone et al, 2013). Admixture can also be the result of backcrossing and interbreeding among species. Analyzing admixture provides context for interpreting other parts of the evolutionary history of the species, such as sex-chromosome evolution (Salamone et al, 2013).

Structure is used to analyze differences in the distribution of genetic variants among populations. Structure, and the related admixture program InStruct, are designed to accommodate analysis of different ploidy levels (Dufresne et al, 2014).

This program identifies populations from data and assigns individuals to the populations representing the best fit for variant patterns present in the individual's genotype. Structure uses the Bayesian clustering approach to admixture mapping, randomly assigning individuals to a specified number of groups, then reassigning individuals based on estimates of variant frequencies for each group (Porras-Hurtada et al, 2013). Several thousand iterations of this approach are typically used for each user-defined number of populations (Porras-Hurtada et al, 2013).

This study presents a method of genotyping the octoploid species *F. chiloensis*, *F. virginiana*, and *F.* × *ananassa*. The genotyping results are analyzed using principal component analysis and admixture mapping to determine if the genotypic data collected matches the relationships among populations and species that are expected based on previous studies.

MATERIALS AND METHODS

Fragaria Samples

Whole genome shotgun DNA sequences of 20 *F. chiloensis* individuals were obtained from populations on the west coast of North America. These 20 individuals will be referred to as WGS *F. chiloensis* samples throughout this paper. Published whole genome shotgun sequences of *F.* × *ananassa*, *F. virginiana*, and *F. chiloensis* were obtained from the NCBI SRA. These six individuals will be referred to as SRR samples throughout this paper. The sequence of a *F. chiloensis* subsp. *lucida* individual, HM1, was used as a control. Two independent WGS *F. chiloensis*

individuals from the same accession, HM1 and chil1691, were included in the analysis. Table 1 summarizes these *Fragaria* samples.

Sample	Plant ID	Species	Location
lane4-s017-index--GTCCGCAC-GTCAGTAC-17_S17_L004	PTR14.3	*F. chiloensis* subsp. *lucida*	Point Reyes, California, USA 38.0440 N, 122.7984 W
lane4-s001-index--AGTCAACA-ACGTCCTG-1_S1_L004	EUR2.1	*F. chiloensis* subsp. *lucida*	Eureka, California, USA 40.762 N, 124.225 W
lane4-s009-index--CGTACGTA-ACGTCCTG-9_S9_L004	PIS5.1	*F. chiloensis* subsp. *lucida*	Pistol River State Park, Oregon, USA 42.2670 N, 124.4054 W
lane4-s020-index--GTTTCGGA-GTCAGTAC-20_S20_L004	EUR3.4	*F. chiloensis* subsp. *lucida*	Eureka, California, USA 40.762 N, 124.225 W
lane4-s019-index--GTGGCCTT-GTCAGTAC-19_S19_L004	EUR13.2	*F. chiloensis* subsp. *lucida*	Eureka, California, USA 40.762 N, 124.225 W
lane4-s012-index--ATTCCTTT-ACGTCCTG-12_S12_L004	SAL4.1	*F. chiloensis* subsp. *lucida*	Salishan Road, Oregon, USA 44.916 N, 124.027 W
lane4-s007-index--GTGGCCTT-ACGTCCTG-7_S7_L004	HM16.1	*F. chiloensis* subsp. *lucida*	Honeyman State Park, Oregon, USA 43.930 N, 124.107 W
lane4-s013-index--AGTCAACA-GTCAGTAC-13_S13_L004	SAL6.1	*F. chiloensis* subsp. *lucida*	Salishan Road, Oregon, USA 44.916 N, 124.027 W
lane4-s014-index--AGTTCCGT-GTCAGTAC-14_S14_L004	SAL8.1	*F. chiloensis* subsp. *lucida*	Salishan Road, Oregon, USA 44.916 N, 124.027 W
lane4-s008-index--GTTTCGGA-ACGTCCTG-8_S8_L004	PIS3.1	*F. chiloensis* subsp. *lucida*	Pistol River State Park, Oregon, USA 42.2670 N, 124.4054 W
lane4-s010-index--GAGTGGAT-ACGTCCTG-10_S10_L004	PTR5.3	*F. chiloensis* subsp. *lucida*	Point Reyes, California, USA 38.0440 N, 122.7984 W
lane4-s018-index--GTGAAACG-GTCAGTAC-18_S18_L004	PTR19.2	*F. chiloensis* subsp. *lucida*	Point Reyes, California, USA 38.0440 N, 122.7984 W
lane4-s004-index--CCGTCCCG-ACGTCCTG-4_S4_L004	EUR17.1	*F. chiloensis* subsp. *lucida*	Eureka, California, USA 40.762 N, 124.225 W

lane4-s015-index--ATGTCAGA-GTCAGTAC-15_S15_L004	HM1	*F. chiloensis* subsp. *lucida*	Honeyman State Park, Oregon, USA 43.930 N, 124.107 W
lane4-s016-index--CCGTCCCG-GTCAGTAC-16_S16_L004	SAL3.1	*F. chiloensis* subsp. *lucida*	Salishan Road, Oregon, USA 44.916 N, 124.027 W
lane4-s002-index--AGTTCCGT-ACGTCCTG-2_S2_L004	EUR9.1	*F. chiloensis* subsp. *lucida*	Eureka, California, USA 40.762 N, 124.225 W
lane4-s003-index--ATGTCAGA-ACGTCCTG-3_S3_L004	EUR12.1	*F. chiloensis* subsp. *lucida*	Eureka, California, USA 40.762 N, 124.225 W
lane4-s011-index--ACTGATAT-ACGTCCTG-11_S11_L004	PTR17.1	*F. chiloensis* subsp. *lucida*	Point Reyes, California, USA 38.0440 N, 122.7984 W
lane4-s005-index--GTCCGCAC-ACGTCCTG-5_S5_L004	HM12.1	*F. chiloensis* subsp. *lucida*	Honeyman State Park, Oregon, USA 43.930 N, 124.107 W
lane4-s006-index--GTGAAACG-ACGTCCTG-6_S6_L004	HM13.1	*F. chiloensis* subsp. *lucida*	Honeyman State Park, Oregon, USA 43.930 N, 124.107 W
s_7_sequence	virg0477.2	*F. virginiana* subsp. *virginiana*	Pennsylvania, USA 41.645 N, 113.750 W
s_1_sequence	virgY33b2	*F. virginiana* subsp. *virginiana*	Pennsylvania, USA 41.645 N, 113.750 W
SRR1513873	virg1992	*F. virginiana* subsp. glauca	Canada, British Columbia 53.27223 N, 120.06457 W
SRR1513867	chil1691	*F. chiloensis* subsp. *lucida*	Florence, Oregon 43.93167 N, 124.11083 W
SRR1513866	chil1743	*F.chiloensis* subsp. *chiloensis*	Chile, Island of Lemuy 42.61667 S, 73.71667 W
SRR1513904	anan	*Fragaria* × *ananassa*	Cultivar “Sweet Charlie”

Table 1. Information on *Fragaria* sample collection. Includes plant ID used in analysis, species name, and location collected for each individual.

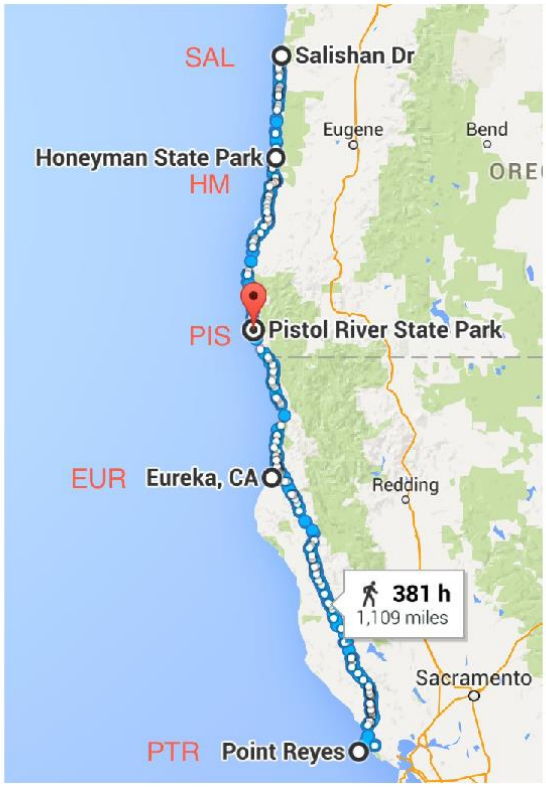

Figure 1. Map of WGS *F. chiloensis* sample collection sites on the west coast of California and Oregon. From north to south: SAL, HM, PIS, EUR, PTR.

Read Mapping

BBTools, including BBDuk, BBMerge, and BBWrap, were used to clean and assemble the reads. BBDuk refers to "Decontamination using K-mers". This program trims reads based on quality score and filters reads as specified by the K-mers present in a reference sequence (Bushnell, 2014). BBMerge processes Illumina paired reads to merge overlapping reads into a single, longer read. This process typically produces single, longer reads that are easier to assemble, and can detect errors where the reads overlap, yielding a lower error rate in the final assembly (Bushnell, 2014). BBWrap

assembles paired-end and single-end reads produced using the aforementioned tools and will write all reads to a .sam output file (Bushnell, 2014).

The WGS *Fragaria* individuals were collected at the locations shown in Figure 1 and prepared for sequencing using the Illumina paired-end library prep protocol. The WGS reads were obtained from one lane of paired-end sequencing on the Illumina HiSeq3000 PE at the Center for Genome Research and Biocomputing of Oregon State University. The SRR *Fragaria* reads were obtained from previously published data, sequenced on the Illumina HiSeq2000.

The 26 reads were aligned to the *Fragaria vesca* genome v2.0.a1 reference sequence using the BBTools aligner for DNA sequences. Reads were cleaned using BBDuk, a decontamination program using K-mers. Parameters chosen include a minimum read length of 32 bases, right and left trimming to remove 3' and 5' adapters respectively using the truseq.fa.gz adapter sequence provided in the BBTools package, K-mer size of 25 bases, allowing for K-mer lengths of 11 bases at the ends of reads, hamming distance of 1 to allow one mismatch, and quality trimming to Q10 using the Phred algorithm (Bushnell, 2014). BBMerge was used to combine overlapping paired-end reads into longer merged reads. BBWrap was used to assemble the output files from BBMerge, including both merged and unmerged reads, into a single sequence.

Genotyping

Read mapping data was converted to .bam format and single nucleotide polymorphisms (SNPs) were called using SAMtools mpileup. SAMtools mpileup takes input BAM files, then produces an output file consisting of an array of reference bases and read bases. Bases that match the reference are indicated by "." or ",", bases that differ from the reference are indicated by the letter representing the mismatched base or insertion/deletion. Resulting sequences were analyzed using mpileup to obtain statistics on read length and coverage.

The resulting data was used to genotype the sequences using POLiMAPS (Phylogenetics of Linkage-Map-Anchored Polyploid Subgenomes) for polyploid genomic data (Tennessen et al, 2014). POLiMAPS reads a pileup file and converts the file to OneMap format. In this format, "a" is assigned for ancestral-allele homozygotes and "ab" is assigned for heterozygotes or derived-allele homozygotes.

The POLiMAPS program MakeOneMapFromPileupNoParents.pl was used, with parameter –m 13 to allow for a maximum of 13 individuals, half of the individuals in the dataset, with missing data at a given loci, -o 1 to require a minimum of one heterozygote and homozygote observed in each individual, -c 1 to require a variant to be observed at least once, and –d 16 to require a sequencing depth of at least 16 (Tennessen et al, 2014). This program was then run again, changing –o and –c parameters to 2 and 2 respectively. The output was saved as a comma-delimited OneMap file.

Principal Component Analysis

Principal component analysis of the POLiMAPS genotyping output was performed in the R programming language. A sample of the code used is included in Appendix B. The POLiMAPS output file for each linkage group was formatted to create a tab delimitated file, with “a” genotypes replaced by “0”, and “ab” genotypes replaced by “1”. The formatted files were initially analyzed separately. The resulting plots were compared, and the principal components identified for each linkage group were comparable - the plots showed similar clustering patterns and principal component percentages across linkage groups. Thus, the POLiMAPS output for each linkage group were combined into one file and missing data was removed, and PCA was completed again to obtain a visual representation of the overall clustering patterns across all linkage groups. The two components that accounted for the most variation were plotted for each individual.

Admixture Mapping

Genetic admixture was modeled using Structure software (Porras-Hurtado et al, 2013). The OneMap output files from the run consisting of parameters –c 2 –o 2 was selected for this analysis, because these parameters removed unique SNPs due to variations in read depth among individuals. A subset of 10,000 loci was randomly selected from the OneMap output for each linkage group to decrease processing time. Where fewer than 10,000 SNPs were found, all loci were used in the analysis. The resulting file was formatted to replace “a” and “ab” genotypes with “0” and “1” genotypes respectively, and then the data matrix was transposed and outputted to a

.csv file. Three iterations of each K value, the number of expected population clusters, between 1 and 8 for each linkage group were run. The optimal value of K was selected using Structure Harvester for each linkage group (Earl and Vonholdt, 2012). The three iterations for the optimal K value were combined using CLUMPP (Jakobsson and Rosenberg, 2007), and the resulting data was plotted using the visual version of Structure.

RESULTS

Read Mapping

The average coverage depth for the assembled sequences is shown in Table 2. Average coverage refers to the read depth at each locus, averaged across all seven linkage groups and the unmapped scaffolds FvbUn. Covered percentage refers to the percentage of the reference sequence that was matched to the assembled contigs. Coverage data was obtained from BBTools pileup.sh on the BBWrap output files.

Plant ID	Average Coverage	Covered Percentage
PTR14.3	17.67	96.13
EUR2.1	15.05	95.78
PIS5.1	19.03	96.41
EUR3.4	18.71	96.23
EUR13.2	20.03	96.36
SAL4.1	20.78	96.51
HM16.1	22.18	96.55

SAL6.1	21.08	96.55
SAL8.1	21.60	96.55
PIS3.1	21.86	96.78
PTR5.3	23.32	96.82
PTR19.2	19.31	96.45
EUR17.1	26.79	97.11
HM1	22.52	96.71
SAL3.1	21.81	96.66
EUR9.1	24.14	96.93
EUR12.1	23.91	96.74
PTR17.1	27.57	97.12
HM12.1	29.31	97.19
HM13.1	36.75	97.46
virg0477.2	42.80	97.65
virgY33b2	63.83	97.49
virg1992	48.98	86.41
chil1691	65.31	95.54
chil1743	127.75	93.92
Anan	186.28	99.04

Table 2. Coverage data for 26 assembled sequences, obtained from BBTools pileup.sh.

The read mapping produced sequences with an average coverage of 15x to 37x for the WGS samples, and greater coverage for the six SRR sequences. The *F.* × *ananassa* sample had the greatest average coverage at 186x. Covered percentage was between 86 and 99%, with the covered percent for WGS samples in the range of 95 to 98%. Larger variations in coverage were observed for the SRR samples. These results suggest that the read mapping method using BBTools was successful for this polyploid *Fragaria* sequencing data.

Principal Component Analysis

Principal component analysis was completed on four data sets – all possible combinations of 26 or 20 individuals, and OneMap parameters of –c 1 –o 1, or –c 2 –o 2. The PCA on all 26 individuals was used to visualize the genotypic relationship between the 20 WGS *F. chiloensis* samples and the 6 SRR individuals, which consisted of *F. virginiana*, *F.* × *ananassa*, and *F. chiloensis* individuals. The PCA on only the 20 WGS samples was used to more clearly visualize the relationship between the *F. chiloensis* individuals from the five different geographic regions. –c 2 –o 2 was

The PCA plot for all 26 samples (WGS *F. chiloensis* and SRR individuals) is shown in Figure 2. The selected parameters for the OneMap runs to make this plot were c -1 and o -1, which include unique SNPs in the analysis. The first component accounted for 47.6% of the observed variation. The second component accounted for 4.8% of the observed variation. A total of 1,014,829 objects were loaded into the R programming language for PCA.

Chil1743 was observed to be genetically distant from other *F. chiloensis* samples. This sample was collected in Chile, while the other *F. chiloensis* samples were collected in northwest North America, which supports the conclusion that geographic location influences genotype. Virg1992 was genetically distant from other *F. virginiana* samples, possibly due to the considerably larger read depth on this individual, which resulted in unique SNPs. The *F.* × *ananassa* individual was located between the *F. chiloensis* and *F. virginiana* individuals for both principal components, though showed less genetic distance to *F. virginiana* samples. EUR individuals consistently showed the greatest distance from *F. virginiana* individuals in both PC 1 and PC 2.

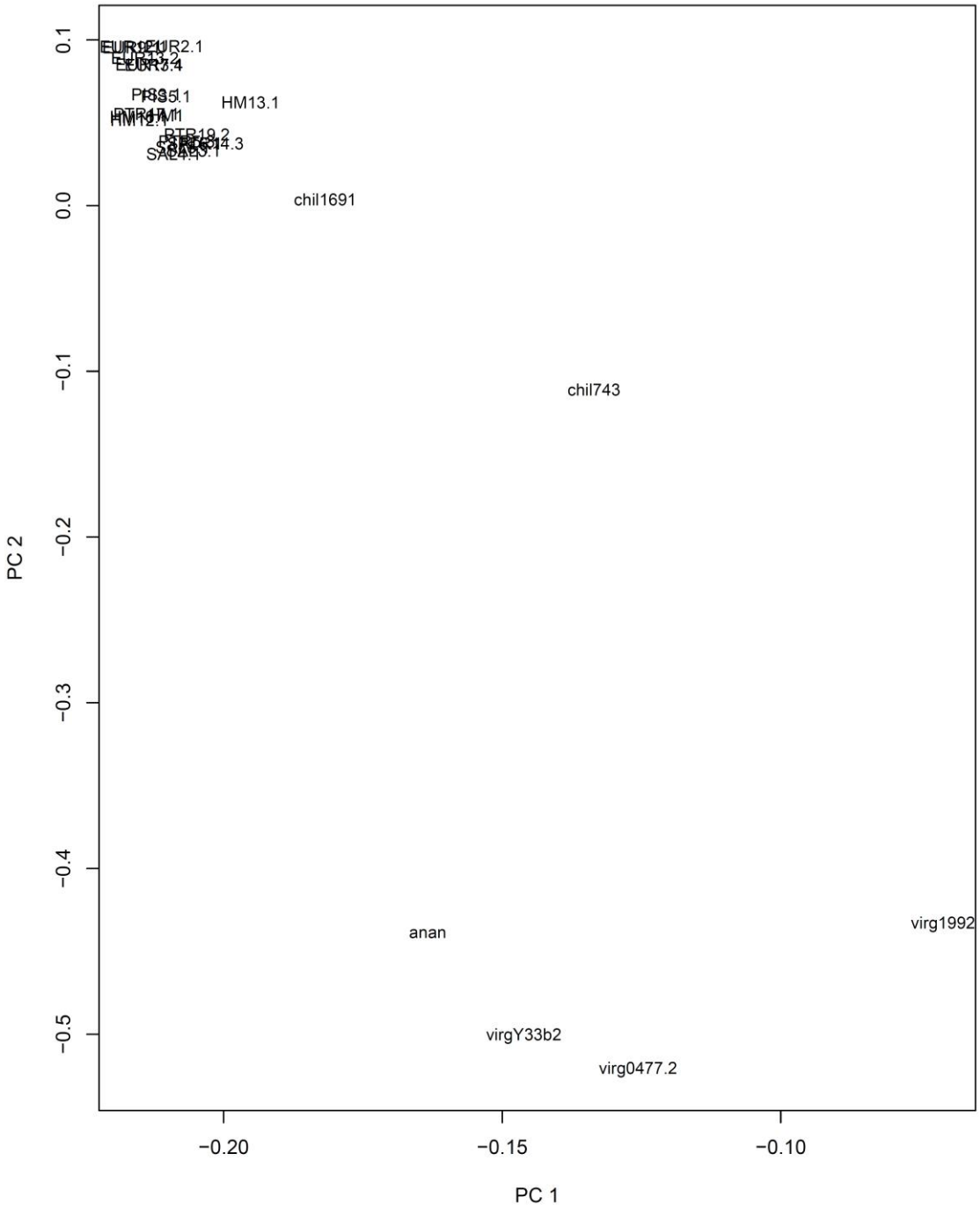

Figure 2. Principal component analysis of all 26 individuals. The two components accounting for the most variation were plotted as PC1 and PC2. Parameters chosen for OneMap analysis were –c 1, -o 1.

The PCA plot for all 26 samples (WGS *F. chiloensis* and SRR individuals) is shown in Figure 3. The selected parameters for the OneMap runs to makes this plot were c -2 and o -2, which excludes unique SNPs from the analysis. The first component accounted for 50.3% of the observed variation. The second component accounted for 6.9% of the observed variation. A total of 83,358 objects were loaded into the R programming language for PCA.

The individuals from the same location clustered together on the PCA, suggesting that genotyping was successful in identifying unique genetic characteristics within the populations. PTR individuals show the least genetic distance from *F. virginiana* individuals in PC 1, suggesting that the individuals from this area are admixed with *F. virginiana*. Virg 1992 and chil 1743 do not cluster with other individuals of the same species. Virg1992 was the most distant from the *F. chiloensis* samples in both principal components.

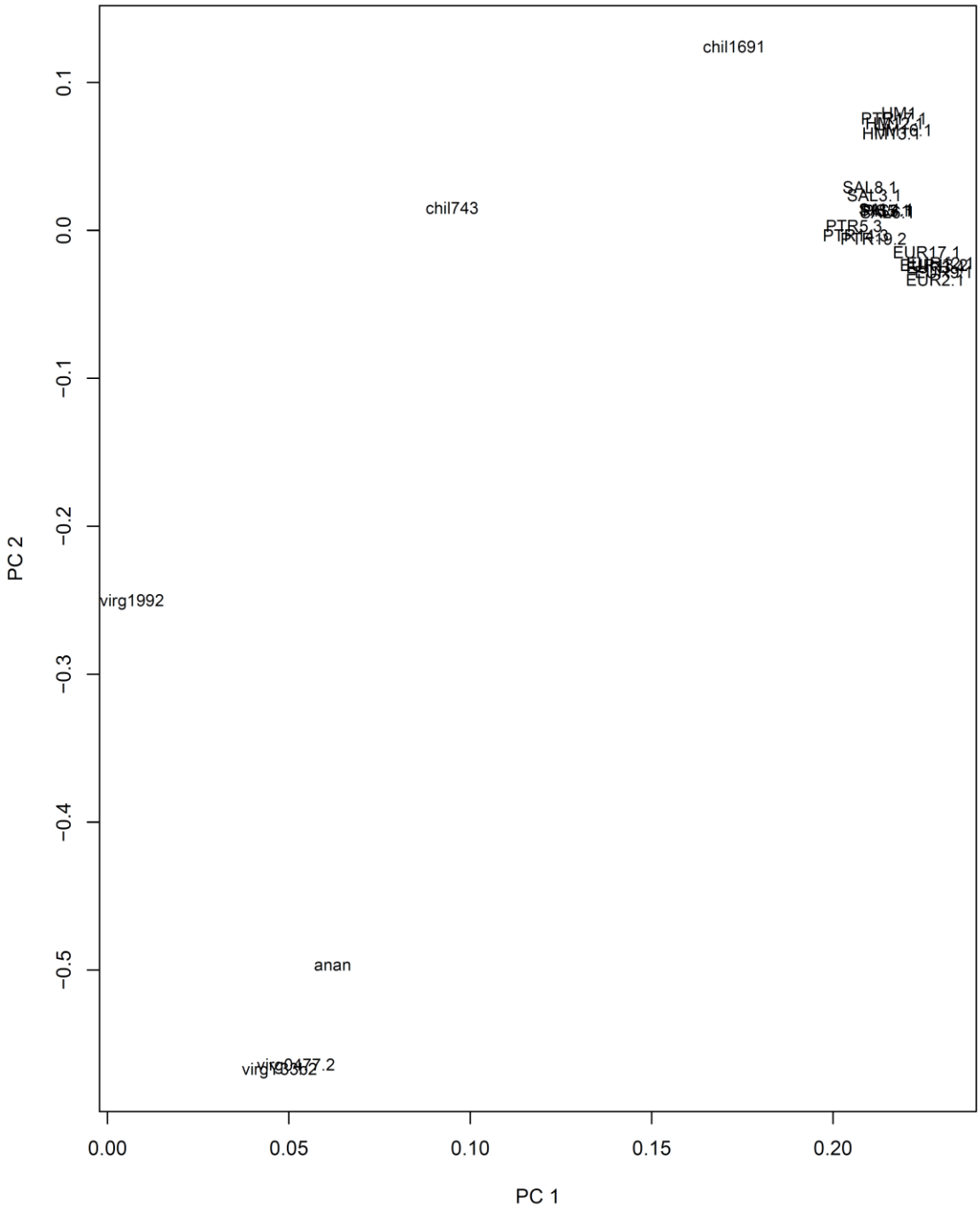

Figure 3. Principal component analysis of all 26 individuals. The two components accounting for the most variation were plotted as PC1 and PC2. Parameters chosen for OneMap analysis were –c 2, -o 2.

The PCA plot for the 20 WGS samples is shown in Figure 4. The selected parameters for the OneMap runs to makes this plot were c -1 and o -1, which include unique SNPs in the analysis. The first component accounted for 55.1% of the observed variation. The second component accounted for 4.7% of the observed variation. A total of 1,014,829 objects were loaded into the R programming language for PCA.

This figure shows that individuals clustered in clearly defined groups based on geographic location. Two individuals, HM13.1 and PTR17.1, did not cluster with their geographic group using these PCA parameters. HM13.1 has considerably higher coverage than the other HM individuals, so unique SNPs were found due to a larger number of reads and possible sequencing errors or characteristics of sequencing such as high depth in regions with large amount of SNPs. Using the parameters –c 1 –o 1 in genotyping identified these SNPs in HM13.1, causing greater genetic distance.

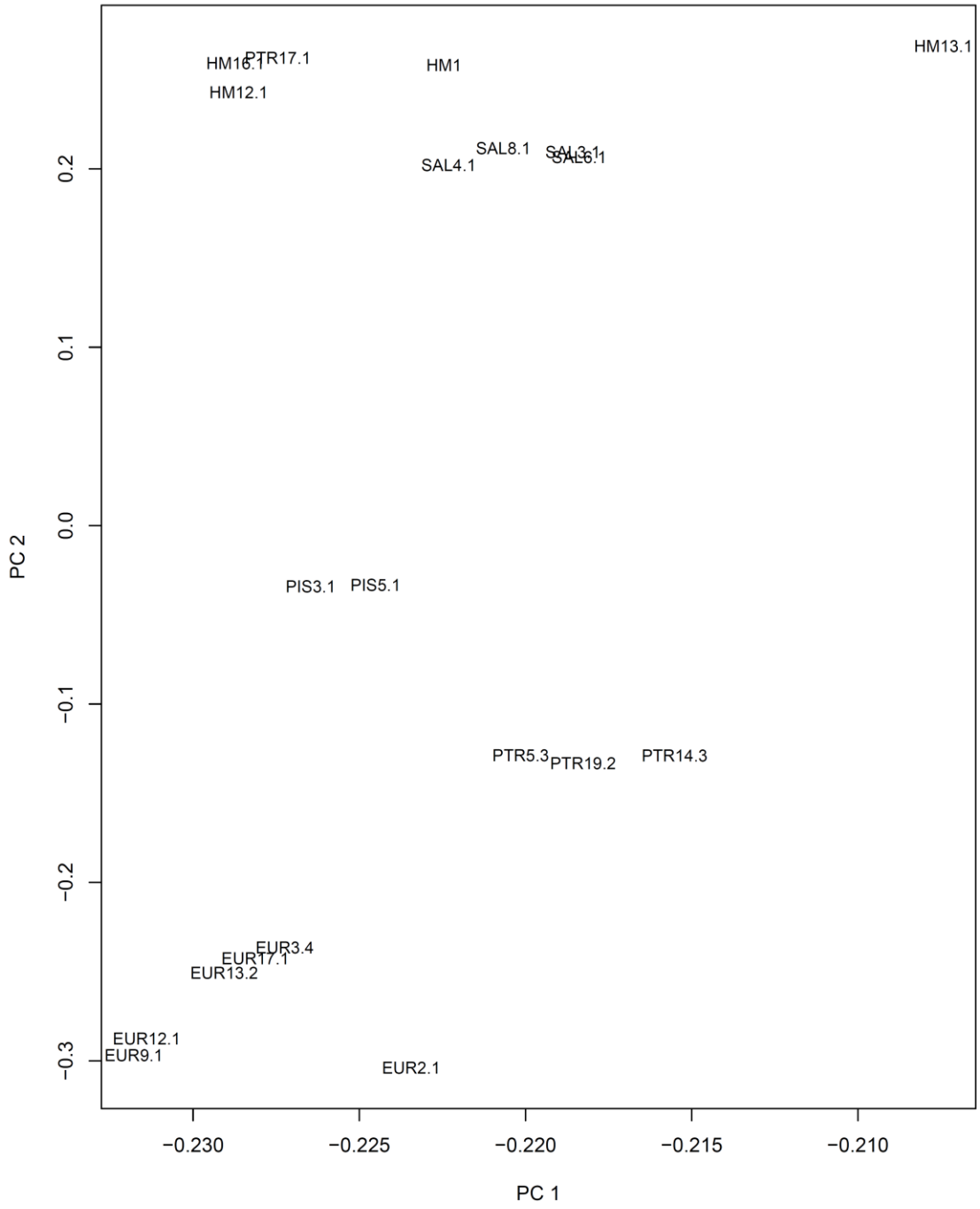

Figure 4. Principal component analysis of 20 WGS *F. chiloensis* individuals. The two components accounting for the most variation were plotted as PC1 and PC2. Parameters chosen for OneMap analysis were –c 1, -o 1.

The PCA plot for the 20 WGS samples is shown in Figure 5. The selected parameters for the OneMap runs to makes this plot were c -2 and o -2, which excludes unique SNPs from the analysis. The first component accounted for 62.4% of the observed variation. The second component accounted for 6.3% of the observed variation. A total of 83,358 objects were loaded for the PCA. There were approximately 920,000 fewer SNPs loaded using these genotyping parameters, due to the removal of unique SNPs.

Once again, the individuals within the same geographic area were shown to cluster together. Using these PCA parameters, HM13.1 clustered with the other HM individuals, suggesting that HM deviated from the population cluster for –o 1, -c 1 due to unique SNPs. PTR17.1 remained separate from the other PTR individuals, clustering more closely to the HM individuals. This clustering pattern shows some correlation with geographic location, as individuals from the northern locations, SAL and HM, tended to cluster together for both PC 1 and PC 2, while the individuals of the southern locations, PIS, EUR, and PTR, loosely clustered for PC 2.

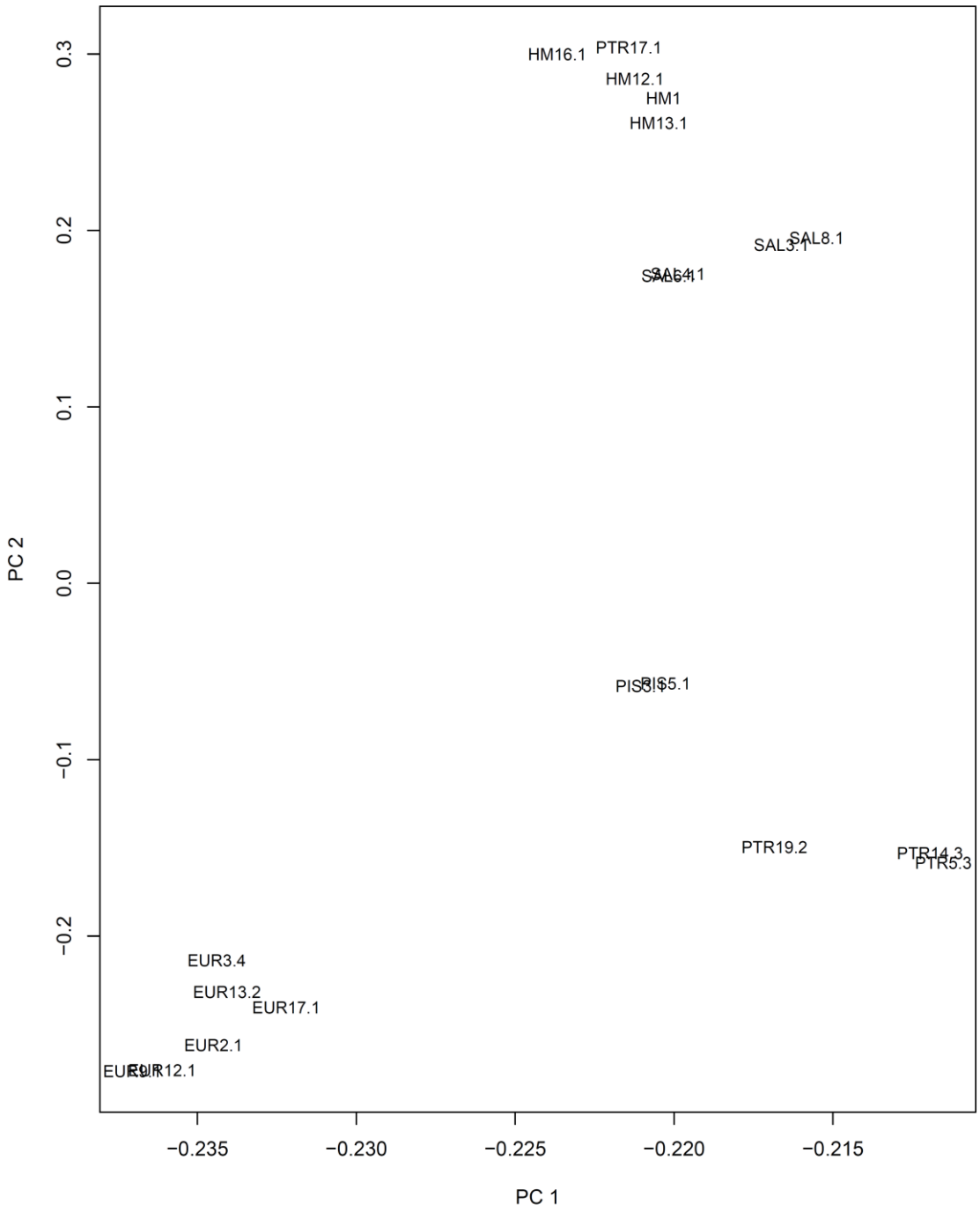

Figure 5. Principal component analysis of 20 WGS *chiloensis* individuals. The two components accounting for the most variation were plotted as PC1 and PC2. Parameters chosen for OneMap analysis were –c 2, -o 2.

Principal component analysis on genotypic data obtained from OneMap files identified two main principal components. Plotting these principal components showed overall trends that the genotypic data matches the relationships among populations and species that were expected based on previous studies. This is evidenced through the clustering of species on the PCA plots above, with *F. chiloensis* and *F. virginiana* relatively distant genetically, and *F.* × *ananassa* located intermediately between the two species. The *F. chiloensis* sequences in the same geographic region show genetic similarities. The primary deviation from this pattern is PTR17.1, which consistently does not cluster with other individuals from the same location, clustering more closely with HM individuals. *F. chiloensis* individuals from the SRR data set, particularly chil1743, show more genetic distance from the WGS individuals due to differences in subspecies and geographic location. Virg1992 consistently does not cluster with the other *virginiana* individuals, likely due to differences in geographic location (British Columbia, Canada vs. Pennsylvania, USA) and average read coverage.

Admixture Mapping

Admixture modeling of the nuclear genotypic data revealed several patterns supportive of the genotyping and PCA data. EUR individuals consistently showed the lowest percentage of admixture with clusters predominant in *F. virginiana* individuals. PTR individuals, with the exception of PTR17.1, showed the highest percentage of admixture with *F. virginiana* individuals. PTR17.1 showed patterns of clustering most similar to HM12.1 and HM13.1 individuals, while exhibiting patterns

dissimilar to other PTR individuals. Chil1691 typically showed different clustering patterns than other *F. chiloensis* samples, likely due to the geographical separation between this individual and the North American samples. Virg1992 shows unique clustering patterns compared to the other two *F. virginiana* individuals, showing clustering patterns more similar to chil1691 and chil1743. Anan shows a high percentage of membership to the clusters predominant in *F. virginiana* individuals.

The Structure plots are shown in Figures 6-13 below. The number following "Fvb" refers to the linkage group number. "FvbUn" refers to unmapped scaffolds. The number of populations (K value) selected using Structure Harvester is indicated in parentheses. A random subset of 10,000 loci was selected from the filtered data using the "shuf" command. FvbUn and Fvb7 had fewer than 10,000 loci, so all SNPs were used in the analysis.

Fvb1

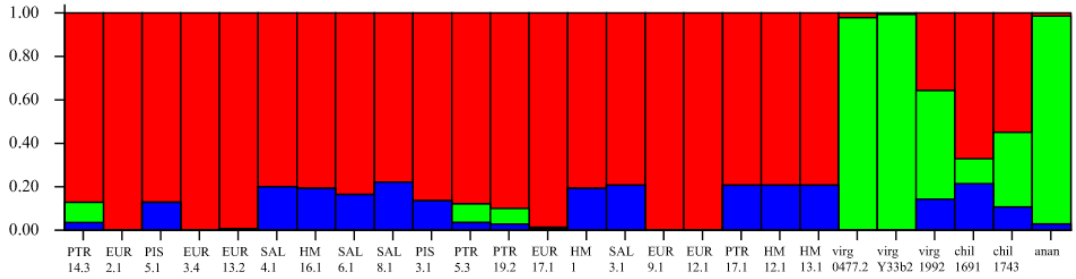

Figure 6. Structure output for 10,000 loci from linkage group 1. OneMap files with parameters –c 2, -o 2 were used for all 26 *Fragaria* individuals. Three population clusters (K=3) were selected for the analysis.

The *F. virginiana* samples contain large portions of the cluster represented by green. The WGS *F. chiloensis* samples contain varying fractions of the cluster represented by blue. PTR individuals are the only WGS *F. chiloensis* samples to contain the green cluster, supporting the PCA data that PTR individuals show admixture with *F. virginiana*. Chil1743 contains the largest fraction of the green cluster relative to other *F. chiloensis* samples. PTR 17.1 did not contain the green cluster, consistent with the PCA data that showed PTR17.1 clustering closely with HM individuals. EUR individuals show little admixture, supporting the PCA data that showed EUR as the most genetically distant from *virginiana* in PC1.

Fvb2

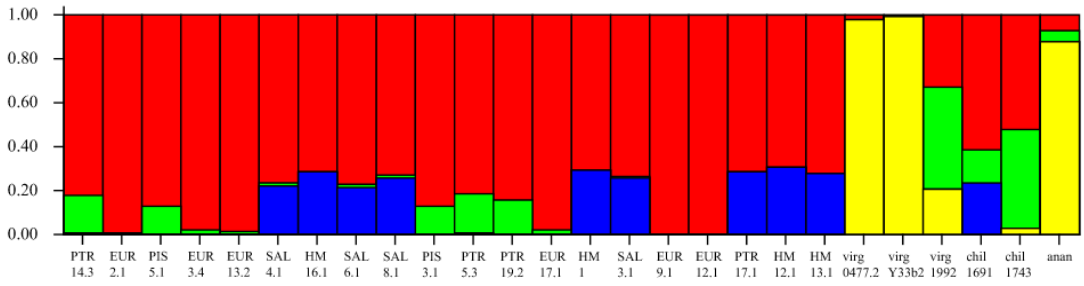

Figure 7. Structure output for 10,000 loci from linkage group 2. OneMap files with parameters –c 2, -o 2 were used for all 26 *Fragaria* individuals. Four population clusters (K=4) were selected for the analysis.

None of the 20 North American *F. chiloensis* samples contain the yellow cluster, suggesting *F. virginiana* lineage. PTR and PIS contain a larger fraction of the green cluster than other WGS *F. chiloensis* samples. PTR and PIS contain similar amounts of green cluster, as do virg 1992 and chil 1743. SAL and HM samples

contain similar amounts of blue cluster, as does chil1691. EUR samples show little admixture, with primary membership only in the red (*F. chiloensis*) cluster.

Fvb3

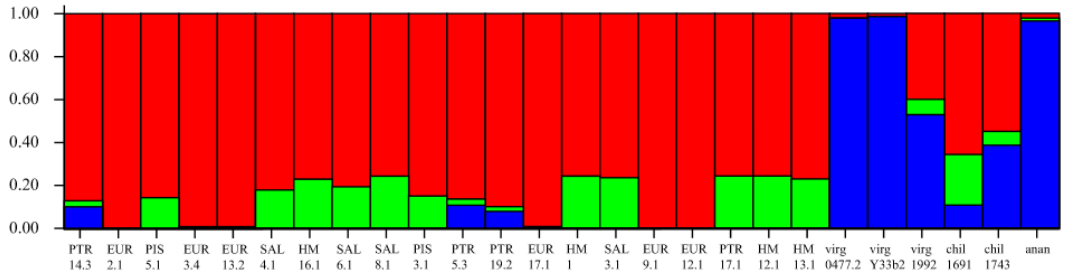

Figure 8. Structure output for 10,000 loci from linkage group 3. OneMap files with parameters –c 2, -o 2 were used for all 26 *Fragaria* individuals. Three population clusters (K=3) were selected for the analysis.

The only WGS *F. chiloensis* samples showing membership in the blue cluster are PTR individuals. This cluster is likely *F. virginiana* in origin due to the large fraction of the blue cluster in two *F. virginiana* individuals. The other four SRR *F. chiloensis* individuals contain larger fractions of blue cluster as well. Green and red clusters are present primarily in *F. chiloensis* samples, with small fractions in *F. virginiana* and *F.* × *ananassa* individuals. Again, EUR individuals show membership only in one cluster.

Fvb4

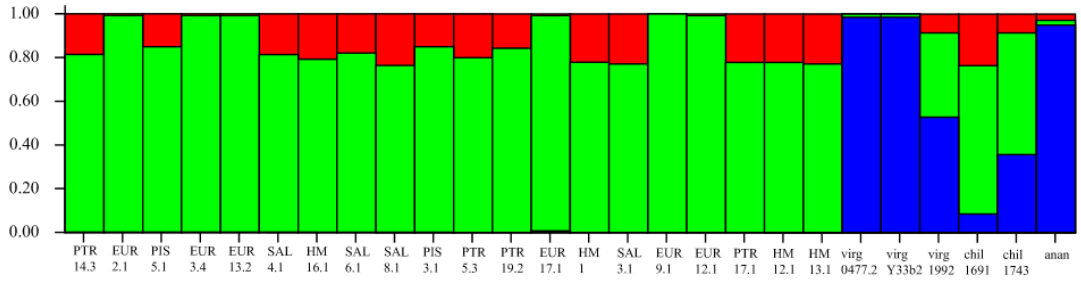

Figure 9. Structure output for 10,000 loci from linkage group 4. OneMap files with parameters –c 2, -o 2 were used for all 26 *Fragaria* individuals. Three population clusters (K=3) were selected for the analysis.

The 20 WGS *F. chiloensis* samples do not contain the cluster represented by blue. SRR *F. chiloensis* samples contained considerable fractions of blue cluster. Chil1743 shows similar clustering patterns to virg1992. The WGS samples, with the exception of EUR individuals, contain similar fractions of the red cluster, which is also found in small quantities in SRR individuals.

Fvb5

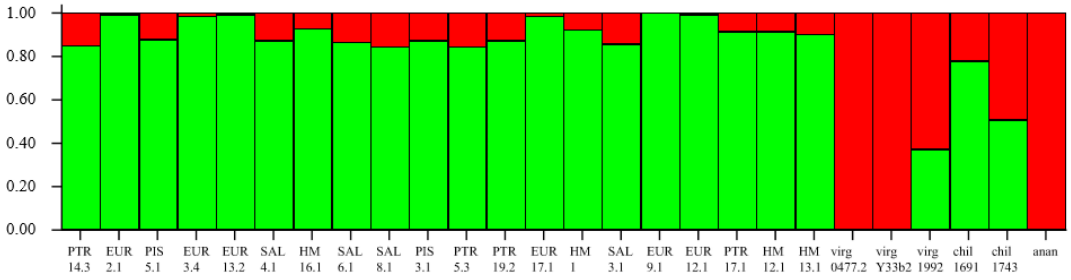

Figure 10. Structure output for 10,000 loci from linkage group 5. OneMap files with parameters –c 2, -o 2 were used for all 26 *Fragaria* individuals. Two population clusters (K=2) were selected for the analysis.

When admixture is analyzed assuming only two populations, as in this linkage group, patterns of admixture with *F. virginiana* are more apparent. *Fragaria virginiana* individuals virg0477.2 and virgY33b2 contain only the red cluster, suggesting that this cluster is *F. virginiana* in origin. Thus, the WGS *F. chiloensis* with an increased membership in the red cluster are likely admixed with *F. virginiana*. Using this result, EUR individuals show the least admixture with *F. virginiana*, while HM, PTR, PIS, SAL show more considerable admixture with *F. virginiana*. This observation is not correlated with relative geographic location, such as increased admixture in more northern populations, due to the absence of the red cluster in EUR individuals.

Fvb6

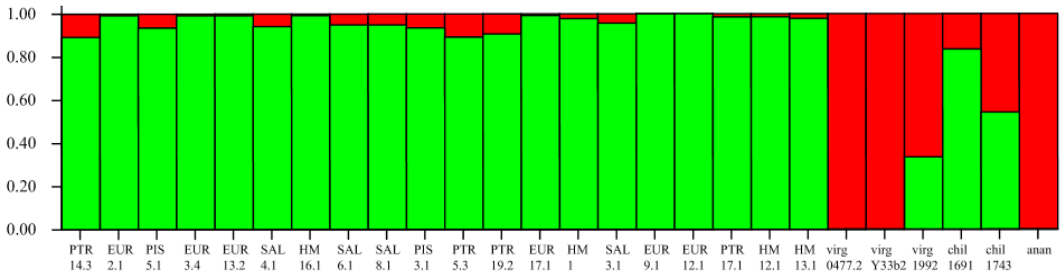

Figure 11. Structure output for 10,000 loci from linkage group 6. OneMap files with parameters –c 2, -o 2 were used for all 26 *Fragaria* individuals. Two population clusters (K=2) were selected for the analysis.

The 20 WGS *F. chiloensis* individuals showed little admixture in linkage group 6, with most individuals showing only slight fractions of the red *F. virginiana* cluster. The PTR individuals contained small fractions of the red cluster indicating *F.*

virginiana genotype. The SRR *F. chiloensis* individuals, particularly chil1743, showed larger fractions of the red cluster. Virg1992 was the only non-*F. chiloensis* individual to show membership in the green cluster.

Fvb7

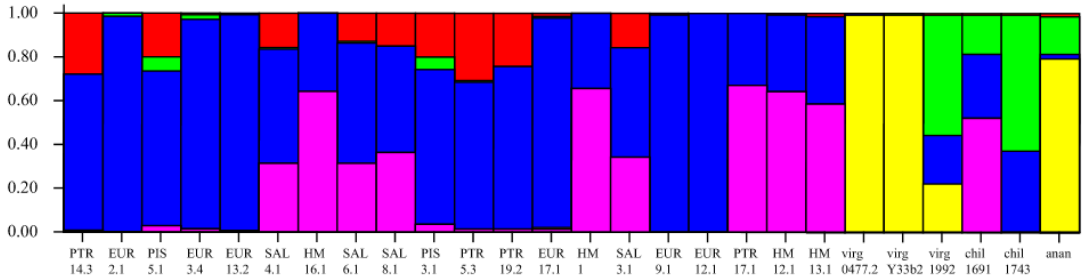

Figure 12. Structure output for 7879 loci from linkage group 7. OneMap files with parameters –c 2, -o 2 were used for all 26 *Fragaria* individuals. Five population clusters (K=5) were selected for the analysis.

HM and SAL samples, the two most northern WGS sample collection locations, show large fractions of the purple cluster, as does chil1691, which was also collected from the central Oregon coastal area. EUR individuals show little admixture, with membership primarily in the blue cluster. PTR, PIS, and SAL samples are the only individuals to contain large fractions of red cluster. The yellow cluster only found in *F. virginiana* and *F.* × *ananassa* species, suggesting *F. virginiana* origin. Large fractions of the green cluster appear only in SRR samples, particularly virg1992 and chil1743.

FvbUn

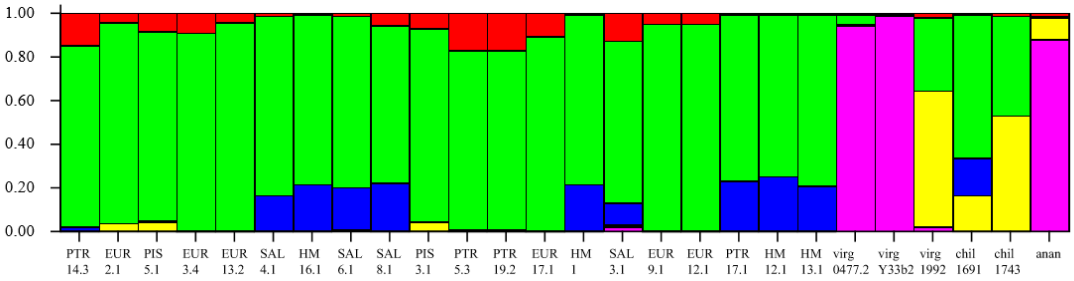

Figure 13. Structure output for 783 loci from unmapped scaffolds. OneMap files with parameters –c 2, -o 2 were used for all 26 *Fragaria* individuals. Five population clusters (K=5) were selected for the analysis.

The purple cluster is only found in *F. virginiana* and *F.* × *ananassa* individuals, suggesting *F. virginiana* origin. The yellow cluster is found primarily in SRR individuals, with small fractions in PIS and EUR WGS *F. chiloensis* individuals. This data set is the only set in which EUR individuals contain non-negligible membership in more than one cluster.

Summary

K values of 2 to 5 were selected using Structure Harvester. Patterns of admixture were consistent across linkage groups. Admixture modeling with all selected K values revealed clear "*F. virginiana*" clusters that were present only in small quantities, typically much less than 20%, in WGS *F. chiloensis* samples.

The ratios of clusters in the virgY33b2 and virg0477.2 are similar to those found in the *F.* × *ananassa* individual, while the virg1992 individual shows moderate admixture between the *F. chiloensis* individuals and the other *F. virginiana*

individuals. This suggests that the *F.* × *ananassa* individual contains large portions of *F. virginiana* ancestry. This may be the result of backcrossing North American *F.* × *ananassa* cultivars with *F. virginiana*. Additional analysis will need to be performed to ensure that this result represents determine if the genotypic data collected matches the biological relationships among the individuals, and is not the result of inadvertently switching data. The data labeling has been verified to determine that the labels were not switched for these two individuals during the analysis process described in this paper. The average read depth, ~188x for the *F.* × *ananassa* sample, compared to ~48x for virg1992 and ~60x for other two *F. virginiana* individuals, may be a possible cause of these variations. Additional analysis will be needed to verify this hypothesis.

Overall, EUR individuals show little admixture with *F. virginiana*, and tend to show membership in one primary cluster. Of the WGS *F. chiloensis* populations, PTR individuals typically show the most admixture with *F. virginiana* individuals. SRR *F. chiloensis* individuals tend to show membership in clusters not present in WGS *F. chiloensis* samples, and increased admixture with *F. virginiana* compared to the WGS *F. chiloensis* samples. Chil1691 shows more admixture than the control sample from the same accession, HM1, as evidenced by increased membership in *F. virginiana* clusters and clusters unique to SRR samples, due to differences in read depth. Admixture patterns are consistent across linkage groups, particularly linkage groups represented in the figures above by the same K value.

DISCUSSION

Next generation sequencing methods have greatly reduced the time and cost required to obtain large quantities of genomic data. This is the first application of POLiMAPS genotyping without cross information and using whole genome sequencing. This method was successful at resolving major relationships among the octoploid *Fragaria* populations included in this study. Thus, the use of POLiMAPS shows promise for population genetic studies in other polyploid species, particularly if the study design accounts for large differences in overall coverage among samples.

Overall, genotyping and admixture mapping results reflected the relationships among populations and species that were expected based on previous studies. Individuals of the same geographic area typically clustered together in PCA plots. The genetic distance between *F. chiloensis* individuals was less than the distance between *F. chiloensis* individuals and individuals of different species. *F.* × *ananassa* was located intermediately between *F. chiloensis* and *F. virginiana* clusters, representative of the hybrid heritage of the species. Admixture mapping of WGS *F. chiloensis* individuals showed varying degrees of admixture with *F. virginiana*, with individuals of the same population showing similar clustering in admixture plots.

However, several observations in this project were inconsistent with the biological expectations from population and species boundaries. In the principal component analysis, HM13.1 did not initially cluster with other individuals from the same geographic area. This was resolved by removing unique SNPs in the genotyping procedure by changing the program parameters. Also, PTR17.1 was observed to cluster with HM individuals, rather than other PTR individuals, in both PCA and

admixture mapping. These results indicate that this individual was likely misclassified at some point, and is part of the HM population. Virg1992 was consistently more distant from virg0477.2 and virgY33b2 in PC1, while anan typically clustered closer to those *F. virginiana* individuals. This result was also observed in the admixture mapping, where anan exhibited clustering patterns very similar to virg0477.2 and Y33b2, while virg1992 exhibited clustering patterns more suggestive of the known hybridization between *F. virginiana* and *F. chiloensis*. The average read depth for the SRR sequences in general, and the *F.* × *ananassa* sequence in particular, is much larger. Thus, the increased number of SNPs in the SRR sequences may cause unrelated sequences to appear similar. In addition, the POLiMAPS program was run using the parameter –d 16, which requires a sequencing depth of at least 16 to call SNPs. Thus, average depth much higher than this number will allow for an increased number of SNPs to be called. This trend is evidenced by the difference in clustering patterns between HM1 and chil1691, which are individuals from the same accession. The read depth of HM1 is 23x, compared to the read depth of 65x for chil1691. In admixture plots with K values of 2 and 3, the chil1691 individual showed increased membership in clusters identified as *virginiana* in origin compared to the HM1 individual. In admixture plots with K values of 4 and 5, chil1691 showed clustering patterns unique to other *F. chiloensis* samples. Chil1691 also showed greater genetic distance from other *F. chiloensis* individuals in both principal components in the PCA plots.

REFERENCES

ARNOLD, B., K. BOMBLIES, and J. WAKELEY. 2012. Extending coalescent theory to autotetraploids. *Genetics* 192: 195-204.

BRO, R. AND A.K. SMILDE. 2014. Principal component analysis. *Analytical Methods* 6: 2812-2831.

BUSHNELL, BRIAN. "BBMap". *SourceForge.* Feb. 2014. Web. 11 May 2016. <https://sourceforge.net/projects/bbmap/>

CLEVENGER, J., C. CHAVARRO, S.A. PEARL, P. OZIAS-AKINS, S.A. JACKSON. Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. *Molecular Plant* 8(6): 831-846

DUFRESNE, *F.*, M. STIFT, R. VERGILINO, AND B. MARBLE. 2014. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology* 23: 40-69.

EARL, D.A. AND B.M. VONHOLDT. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4(2): 359-361.

HANCOCK, J.*F.*, A. LAVIN, J.B. RETAMALES. 1999. Our Southern Strawberry Heritage: *Fragaria chiloensis* of Chile. *HortScience* 34(5):814-816.

HARTIGAN, J.A. AND M.A. WONG. 1979. Algorithm AS 136: a K Means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 28: 100-108.

HOL, W., A. VAN DER WURFF, L. SKOT, and R. COOK. 2008. Two distinct AFLP types in three populations of marram grass (Ammophila arenaria) in Wales. *Plant Genetic Resources: Characterization and Utilization* 6: 201 207.

JAKOBSSON, M. AND N. ROSENBERG. 2007. CLUMPP: a cluster matching and permutation program with label switching and multimodality in analysis of population structure. *Bioinformatics* 23(14): 1801-1806.

JOMBART, T., S. DEVILLARD, AND F. BALLOUX. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11: 94.

LIMBORG, M.T., L.W. SEEB, J.E. SEEB. 2016. Sorting duplicated loci disentangles complexities of polyploidy genomes masked by genotyping by sequencing. *Molecular Ecology*

LISTON, A., R. CRONN, AND T.-L. ASHMAN. 2014. *Fragaria*: A genus with deep historical roots and ripe for evolutionary and ecological insights. *American Journal of Botany* 101(10):1686-1699.

MABLE, B., J. BELAND, C. DIBERARDO. 2004. Inheritance and dominance of self incompatibility alleles in polyploid *Arabidopsis lyrata*. *Heredity* 93: 476-486.

PORRAS-HURTADO, L., Y. RUIZ, C. SANTOS, C. PHILLIPS, A. CARRACEDO, AND M. LAREU. 2013. An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in Genetics* 4: 1-13.

ROUSSEAU-GUEUTIN, M., E. LERCETEAU-KOHLER, L. BARROT, D.J. SARGENT, A. MONFORT, D. SIMPSON, P. ARUS, G. GUERIN, B. DENOYES-ROTHAN. 2008. Comparative Genetic Mapping Between Octoploid and Diploid *Fragaria* Species Reveals a High Level of Colinearity Between Their Genomes and the Essentially Disomic Behavior of the Cultivated Octoploid Strawberry. *Genetics* 179:2045-2060.

SAINTENAC, C., D. JIANG, S. WANG, E. AKHUNOV. 2013. Sequence-Based Mapping of the Polyploid Wheat Genome. *G3* 3(7): 1105-1114.

SALAMONE, I., R. GOVINDARAJULU, S. FALK, M. PARKS, A. LISTON, AND T.-L. ASHMAN. 2013. Bioclimatic, ecological, and phenotypic intermediacy and high genetic admixture in a natural hybrid of octoploid strawberries. *American Journal of Botany* 100(5): 939-950.

STANLEY, L., N.J. FORRESTER, R. GOVINDARAJULU, A. LISTON, T.-L. ASHMAN. 2015. Geographic patterns of genetic variation in three genomes of North American diploid strawberries with special reference to *Fragaria vesca* subsp. *bracteata*.. *Botany* 93(9): 573-588.

ST. ONGE, K.., J. FOXE, J. LI, H. LI, K. HOLM, P. CORCORAN, T. SLOTTE, M. LASCOUX, AND S. WRIGHT. 2012. Coalescent based analysis distinguishes between allo and autopolyploid origin of Shepherd's Purse (Capsella bursa pastoris). *Molecular Biology and Evolution* 29:1721-1733.

TENNESSEN, J. POLiMAPS/MakeOneMapFromPileupNoParents.pl. 2015. https://github.com/jacobtennessen/POLiMAPS/blob/master/MakeOneMapFromPileupNoParents.pl

TENNESSEN, J. A., R. GOVINDARAJULU, T.-L. ASHMAN, AND A. LISTON. 2014. Evolutionary Origins and Dynamics of Octoploid Strawberry Subgenomes Revealed by Dense Targeted Capture Linkage Maps. *Genome Biology and Evolution* 6(12): 3295-3313.

APPENDIX A – Summary of *F. chiloensis* WGS

PlantID	Sex	Index1_seq	Index2_seq	Sample	PF Clusters
EUR12.1	F	ATGTCAGA	ACGTCCTG	lane4-s003-index--ATGTCAGA-ACGTCCTG-3	19,698,411
EUR13.2	F	GTGGCCTT	GTCAGTAC	lane4-s019-index--GTGGCCTT-GTCAGTAC-19	17,565,867
EUR17.1	F	CCGTCCCG	ACGTCCTG	lane4-s004-index--CCGTCCCG-ACGTCCTG-4	21,733,292
EUR2.1	M/H	AGTCAACA	ACGTCCTG	lane4-s001-index--AGTCAACA-ACGTCCTG-1	11,596,564
EUR3.4	H	GTTTCGGA	GTCAGTAC	lane4-s020-index--GTTTCGGA-GTCAGTAC-20	16,353,365
EUR9.1	M/H	AGTTCCGT	ACGTCCTG	lane4-s002-index--AGTTCCGT-ACGTCCTG-2	20,353,431
GP33.1	F	ATGTCAGA	GTCAGTAC	lane4-s015-index--ATGTCAGA-GTCAGTAC-15	17,861,807
HM12.1	F	GTCCGCAC	ACGTCCTG	lane4-s005-index--GTCCGCAC-ACGTCCTG-5	24,914,443
HM13.1	M/H	GTGAAACG	ACGTCCTG	lane4-s006-index--GTGAAACG-ACGTCCTG-6	27,762,793
HM16.1	F	GTGGCCTT	ACGTCCTG	lane4-s007-index--GTGGCCTT-ACGTCCTG-7	19,323,418
PIS3.1	M/H	GTTTCGGA	ACGTCCTG	lane4-s008-index--GTTTCGGA-ACGTCCTG-8	19,016,136
PIS5.1	M/H	CGTACGTA	ACGTCCTG	lane4-s009-index--CGTACGTA-ACGTCCTG-9	15,925,163
PTR14.3	F	GTCCGCAC	GTCAGTAC	lane4-s017-index--GTCCGCAC-GTCAGTAC-17	14,963,326
PTR17.1	M/H	ACTGATAT	ACGTCCTG	lane4-s011-index--ACTGATAT-ACGTCCTG-11	22,855,900
PTR19.2	H	GTGAAACG	GTCAGTAC	lane4-s018-index--GTGAAACG-GTCAGTAC-18	17,671,411
PTR5.3	H	GAGTGGAT	ACGTCCTG	lane4-s010-index--GAGTGGAT-ACGTCCTG-10	19,606,460
SAL3.1	H	CCGTCCCG	GTCAGTAC	lane4-s016-index--CCGTCCCG-GTCAGTAC-16	17,856,663
SAL4.1	M/H	ATTCCTTT	ACGTCCTG	lane4-s012-index--ATTCCTTT-ACGTCCTG-12	17,859,147
SAL6.1	F	AGTCAACA	GTCAGTAC	lane4-s013-index--AGTCAACA-GTCAGTAC-13	16,521,510
SAL8.1	F	AGTTCCGT	GTCAGTAC	lane4-s014-index--AGTTCCGT-GTCAGTAC-14	17,676,591

Table 3. Summary of WGS *F. chiloensis* individuals. Includes sequencing indices and PF clusters for each sample.

APPENDIX B –Sample Code for PCA

Sample code for principal component analysis in the R programming language is shown below. The number of columns inputted into the princomp function was changed from 26 to 20 to collect data only on WGS samples.

```
OneMap_combined.new <-
read.delim("C:/Users/devinj/Desktop/bbmap/OneMap_combined.new.TXT",
na.strings="-")
View(OneMap_combined.new)
attach(OneMap_combined.new)
pcagenos<-princomp(~., data=OneMap_combined.new[,c(1:26)],cor=TRUE,
na.action=na.exclude)
summary(pcagenos,loadings=TRUE)
pcagenos$loadings
colnames(OneMap_combined.new)<-
c("PTR14.3","EUR2.1","PIS5.1","EUR3.4","EUR13.2","SAL4.1","HM16.1","SAL6.
1","SAL8.1","PIS3.1","PTR5.3","PTR19.2","EUR17.1","HM1","SAL3.1","EUR9.1",
"EUR12.1","PTR17.1","HM12.1","HM13.1",
"virg0477.2","virgY33b2","virg1992","chil1691","chil743","anan")
plot(pcagenos$loadings[,1],pcagenos$loadings[,2],cex=0,xlab="PC 1",ylab="PC 2")
text(pcagenos$loadings[,1],pcagenos$loadings[,2],colnames(OneMap_combined.new
),cex=0.8)
```