

Evaluating the adaptive genomic landscape of remnant and
backcross American chestnut populations to inform germplasm
conservation

Alexander M. Sandercock

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Genetics, Bioinformatics, and Computational Biology

Jason A. Holliday, Chair

Jared W. Westbrook

David C. Haak

Inyoung Kim

May 26, 2023

Blacksburg, Virginia

Keywords: landscape genomics, restoration, endangered species, local adaptation

Copyright 2023, Alexander M. Sandercock

Evaluating the adaptive genomic landscape of remnant and backcross American chestnut populations to inform germplasm conservation

Alexander M. Sandercock

(ABSTRACT)

The American chestnut tree (*Castanea dentata*) is a deciduous tree that largely exists in the eastern United States along the Appalachian Mountain range. Approximately 100 years ago, a fungal pathogen (*Cryphonectria parasitica*) decimated chestnut populations, resulting in the loss of billions of trees. Disease-resistant American chestnut populations have been developed, but the introgression of wild adaptive diversity into these breeding populations will be necessary to develop locally adapted and disease resistant chestnut trees for reintroduction. In this dissertation, I presented our findings which addressed previous gaps in knowledge regarding the population genomics of wild and backcross American chestnut populations. I 1) estimated the genomic diversity, population structure, and demographic history of remnant wild American chestnut populations; 2) revealed the genomic basis of local climate adaptation in American chestnut, developed a novel method to make tree sampling estimates for germplasm conservation, and defined unique seed zones for American chestnut based on climate and genotype, and 3) determined the amount of wild adaptive diversity captured by the backcross breeding program and made recommendations for their replanting region. These results will inform the development of a breeding plan for the introgression of adaptive diversity into backcross and transgenic chestnut populations.

Evaluating the adaptive genomic landscape of remnant and backcross American chestnut populations to inform germplasm conservation

Alexander M. Sandercock

(GENERAL AUDIENCE ABSTRACT)

The American chestnut tree (*Castanea dentata*) is a deciduous tree that largely exists in the eastern United States along the Appalachian Mountain range. Approximately 100 years, a fungal disease (*Cryphonectria parasitica*) decimated chestnut populations, resulting in the loss of billions of trees. The American Chestnut Foundation developed disease-resistant American chestnut backcross trees by breeding American chestnut trees with Chinese chestnut trees (*Castanea mollissima*). These trees will need additional breeding with wild American chestnut trees so that their offspring will have both the disease-resistant traits and the adaptations to the local environment where they will be replanted. This is important, because trees that are both disease-resistant and locally adapted will be most likely to survive and thrive in their replanting location. However, a comprehensive evaluation of the genomic basis for local adaptation in American chestnut populations is lacking. In this dissertation, I presented our findings which addressed previous gaps in knowledge regarding the population genomics of wild and backcross American chestnut populations. I 1) estimated the genomic diversity, number of unique populations, and population size changes over time in wild American chestnut; 2) revealed the genes related to local adaptation in American chestnut, developed a novel method to make tree sampling estimates for conserving wild American chestnut diversity, and defined unique seed zones (areas within the species range that have unique adaptations to environment) for American chestnut based on climate (ie, precipitation and temperature values) and genotype (DNA), and 3) determined the amount

of wild genomic diversity related to local adaption captured by the backcross breeding program and made recommendations for their replanting region. These results will inform the development of a breeding plan of wild American chestnut with backcross and transgenic chestnut populations to create locally adapted and disease-resistant chestnut populations for reintroduction.

Dedication

To my family, my wife, and our dog, Jafar.

Acknowledgments

I would like to take this opportunity to thank the individuals and institutions who provided me with support throughout my graduate studies and the opportunity to further my education in the bioinformatics and genomics field. Their guidance and contributions were influential in shaping my development as a researcher and the work presented in this dissertation.

First, I would like to express my heartfelt gratitude to Dr. Jason Holliday who served as my primary advisor. His constant support and guidance were invaluable throughout this process. It is with deep appreciation that I acknowledge Dr. Holliday as a professional role model, whose mentorship has directly contributed to my growth and success as a graduate student.

I am thankful for my dissertation committee members, Dr. Jared Westbrook, Dr. David Haak, and Dr. Inyoung Kim, for their valuable feedback. Their critiques and advice of my research proposal improved this dissertation. The guidance Dr. Westbrook provided was essential in improving my American chestnut knowledge and in developing the research aims of this work.

I extend my thanks to the staff and volunteers of The American Chestnut Foundation (TACF), whose tireless work to identify and conserve wild American chestnut trees provided the leaf material necessary for my research, as well as Dr. Rita Costa (Instituto Nacional de Investigação Agrária e Veterinária, Portugal) for providing *Castanea sativa* samples. Particularly, I would like to thank my collaborators, Thomas Saielli, Dr. John Scrivani, Sara Fitzsimmons, Taylor Perkins, and Dr. Hill Craddock. Their contributions toward leaf collection and insight into *Castanea* species are greatly appreciated.

I would like to recognize the members of the Genetics, Bioinformatics, and Computational Biology (GBCB) program for their support. Many thanks to Dr. Marcela Johnson and Kevin Radja for their advice and critiques as I prepared for my preliminary examination, and their continued encouragement as I finished this dissertation. I also appreciate Dennie Munson, who's guidance and attention to detail were instrumental in me completing my degree requirements.

I am grateful to the members of the Holliday Lab for their collaboration and support. Dr. Qian Zhang for her insight and assistance in plant DNA extraction, and for providing support and motivation during my research journey. I thank the undergraduate students in the lab, including Hayley Johnson, for making time to aid in sample cataloguing and DNA extraction.

I would like to acknowledge the work performed by the staff of HudsonAlpha Institute for Biotechnology and the Duke Center for Genomic and Computational Biology for the library preparation and genomic sequencing of the *Castanea* samples. Thank you to Dr. Jeremy Schmutz and Dr. Jane Grimwood for their collaboration and contributions to the study design.

I extend my deepest thanks to my family and friends for their words of encouragement and support throughout my academic journey. Their unwavering confidence in me was motivating during challenging moments.

I thank Virginia Tech Advanced Research Computing for providing and maintaining the computational resources necessary for this work to be performed. Thank you to Dr. Robert Settlege for his technical support and expert suggestions which improved the variant calling pipeline used in this dissertation.

Finally, I would like to express my appreciation for the financial support of this work provided by the United States National Institute for Food and Agriculture Projects 1018599 and 1005394, the North Carolina Biotechnology Center Flash Grant, The American Chest-

nut Foundation, and through a Graduate Fellowship awarded to me by the Virginia Tech Institute for Critical Technology and Applied Science.

Contents

List of Figures	xiv
List of Tables	xxi
1 Introduction	1
2 Frozen in time: Rangewide genomic diversity, structure, and demographic history of relict American chestnut populations	4
2.1 Introduction	4
2.2 Methods	7
2.2.1 Leaf sample collection	7
2.2.2 DNA isolation and sequencing	8
2.2.3 Bioinformatics	10
2.2.4 Species identification and estimation of hybridization in wild American chestnut populations	11
2.2.5 Population structure	12
2.2.6 Demographic history	13
2.2.7 Migration rates	14
2.2.8 Genome-wide nucleotide diversity	16

2.2.9	Detecting signatures of positive selection	17
2.3	Results	19
2.3.1	Genome resequencing	19
2.3.2	Hybridization	19
2.3.3	Population structure in <i>Castanea dentata</i>	20
2.3.4	Migration rates	22
2.3.5	Demographic history	22
2.3.6	Genomic diversity	24
2.3.7	Regions and associated genes under positive selection	26
2.4	Discussion	28
2.4.1	Hybridization is not widespread in putatively wild American chestnut genotypes	28
2.4.2	The American chestnut range can be divided into two or three popu- lations	29
2.4.3	Impact of Quaternary glaciation on extant patterns of genomic diversity	31
2.4.4	Global signatures of positive selection are enriched for disease resis- tance genes	33
2.4.5	Implications for restoration of American chestnut	35
3	Identifying and conserving the adaptive legacy of American chestnut	37
3.1	Introduction	37

3.2	Methods	39
3.2.1	Dataset preparation	39
3.2.2	Identifying climate-associated adaptive loci in American chestnut populations	41
3.2.3	Spatial distribution of adaptive genomic diversity and genomic offset	42
3.2.4	Identifying seed zones within the American chestnut natural range	43
3.2.5	Estimating the number of trees to sample to capture wild adaptive diversity for germplasm conservation	44
3.3	Results	45
3.3.1	Drivers of genomic variation across the range of American chestnut	45
3.3.2	Geographic distribution of genomic variation and maladaptation under current and future climates	48
3.3.3	Identifying seed zones	48
3.3.4	Sampling to capture wild adaptive variation	51
3.4	Discussion	52
3.4.1	Climate shapes genomic variation in American chestnut	52
3.4.2	Adaptive units for germplasm conservation and reforestation	53
3.4.3	Matching adaptive portfolios to future climates	54
3.4.4	Sampling estimates for germplasm conservation of wild adaptive genomic diversity	55
3.4.5	Conclusions	56

4	Estimating the wild adaptive genomic diversity captured within backcross populations	58
4.1	Introduction	58
4.2	Methods	59
4.2.1	Whole-genome sequencing and bioinformatics	59
4.2.2	Estimating wild adaptive diversity within breeding program	61
4.2.3	Identifying local adaptation within backcross population	61
4.3	Results	62
4.3.1	Adaptive diversity captured within backcross breeding program	62
4.3.2	Matching adaptive niche of backcross trees for restoration plantings	64
4.4	Discussion	66
4.4.1	Adaptive genomic diversity captured within the TACF backcross breeding program	67
4.4.2	Replanting locations can be predicted for backcross trees	68
4.4.3	Conclusions	69
5	Conclusions and Future Directions	70
	Bibliography	74
	Appendices	94
	Appendix A Supplementary Materials for Chapter 2 "Frozen in time: Rangewide	

genomic diversity, structure, and demographic history of relict American chestnut populations”	95
--	----

Appendix B Supplementary Materials for Chapter 3 ”Identifying and conserving the adaptive legacy of American chestnut”	110
--	-----

List of Figures

2.1	Range map of the North American <i>Castanea</i> species and the locations of the selected American chestnut and chinquapin trees.	9
2.2	The wild American chestnut population structure is best described by three or two populations. The pie charts represent the sample locations and the proportion of color within each pie chart represents population membership. The minimum CV error score and minimum BIC score for each K value (right) were used to determine most likely population number for the ADMIXTURE and DAPC analyses, respectively.	21
2.3	The Appalachian Mountain range was a barrier to gene flow for postglacial migrating American chestnut populations. A. EEMS migration rate estimates for American chestnut populations. Regions of blue are above average migration estimates, while orange regions are below average migration estimates. A loess line of Appalachian Mountain peaks was applied in <i>ggplot2</i> using peak locations obtained from https://en.wikipedia.org/wiki/List_of_mountains_of_the_Appalachians . B. ADMIXTURE analysis with $K = 4$. The pie charts represent sample location and the color represents population membership.	23

2.4	American chestnut populations underwent bottleneck and expansion events that coincide with recent North American glaciations. Greyed regions represent glaciation event times approximated from Bowen [18] (as cited in [15]). Each colored line represents the N_e for each of the three populations. Coalescent scaling was converted to years with a 30 year generation time. A. SMC++ N_e estimates from the present to 2Mya. B. SMC++ N_e estimates from the present to 100Kya. The figure inset into the upper right-hand corner is the SMC++ N_e estimates from the present to 13Kya.	24
2.5	Genome-wide sliding-window analysis of the three American chestnut populations for Tajima's D, nucleotide diversity (π), RAI SD μ -statistic, and pairwise F_{ST} . Smoothed lines were computed for all results in <i>ggplot2</i> with the default <i>gam</i> parameter.	25
2.6	Number of outlier genes identified by RAI SD within each American chestnut population.	27
3.1	Range map of <i>Castanea dentata</i> and the locations of the 356 American chestnut samples and their assigned population from Sandercock et al. [104]. . . .	40
3.2	Importance of climatic variables in patterns of allelic turnover for wild American chestnut populations. A) Ranked accuracy importance of 33 climateNA variables from a gradientForest analysis that included 353 American chestnut samples and 50k SNPs. The variables denoted with '*' were used for the analyses. B) Predictor cumulative importance plot. Cumulative change in allele composition across the 10 climate gradients and the 18,483 putatively adaptive loci. Sharp increases suggest a point in the climate gradient where allelic composition changes.	47

3.3	Adaptive genomic diversity under past and future climate conditions. (a) PC loadings from the gradientForest model of 10 uncorrelated climate variables and 18,483 putatively adaptive loci. (b) Biplot of loadings for PC1 and PC2. (c) Genomic offset under 2080 climate projections for moderate emissions (RCP4.5; left) and severe emissions (RCP8.5; right) scenarios.	49
3.4	Three seed zones for American chestnut conservation that are based on the GF-predicted genomic variation. (a) Spatial distribution of the three seed zones. (b) PCA biplot of the GF-predicted genomic variation using three clusters to represent the seed zones. (c) F_{ST} values between each seed zone at neutral loci (left) and putatively adaptive loci (right).	50
4.1	Wild adaptive diversity captured by the TACF backcross breeding program. (a) Supervised ADMIXTURE plot of ancestry. Each individual line is a single backcross sample. (b) Average proportion of ancestry within each assigned seed zone from backcross samples. (c) Allele-frequency histogram at shared adaptive loci for the wild and backcross populations. (d) Linear regression plot of allele frequencies at shared adaptive loci for the wild and backcross populations.	63
4.2	Locator origin estimates for a subset of backcross and wild American chestnut trees. 100 iterations were performed and visualized here. Blue points are locations of the 356 wild American chestnut samples that were used to train Locator, and black dots are predicted locations of a single tree for each iteration. CC-SPGCO-3-179, CC-SPGCO-3-122, CC-SPGCO-3-142, MSC-4, and JB-C5-1 are wild American chestnut trees, while the remaining samples are backcross trees.	65

4.3	Locator predicted locations for the backcross trees and their deviation from the location of the last breeding parent.	66
A.1	Leaf examples from each of the <i>Castanea</i> species included in this study. Pictured are representatives of Japanese chestnut (<i>C. crenata</i>), American chestnut (<i>C. dentata</i>), Alabama chinquapin (<i>C. pumila</i> var. <i>alabamensis</i>), Ozark chinquapin (<i>C. pumila</i> var. <i>ozarkensis</i>), European chestnut (<i>C. sativa</i>), Allegheny chinquapin (<i>C. pumila</i> var. <i>pumila</i>), and Chinese chestnut (<i>C. mollissima</i>).	96
A.2	Visualization of sequencing density and coverage in the filtered <i>Castanea dentata</i> genomic dataset. A. Genome-wide sequencing density. Greyed regions represent 95% confidence intervals. B. Mean depth per variant averaged across all samples. For visualization, 0.05% of all SNPs were excluded that had a coverage >100x. C. Mean depth per individual averaged across all variants.	97
A.3	ADMIXTURE analysis suggests that widespread hybridization does not occur within wild American chestnut populations. ADMIXTURE was performed with 384 American chestnut and 92 <i>Castanea</i> species samples. Each color bar represents the proportion of inferred species ancestry for a sample. K = 7 was determined to be the most appropriate cluster value.	98
A.4	The American chestnut population structure is characterized by an isolation-by-distance model. The vertical black line represents the observed 0.2 value that was significantly higher than 0 (p<0.001).	99

A.5	The American chestnut population boundaries follow precipitation and temperature gradients of the eastern U.S. The different shapes represent the American chestnut sample locations and population membership. A. Mean temperature of coldest quarter (WorldClim BIO11 at 2.5m resolution). B. Annual precipitation (WorldClim BIO12 at 2.5m resolution).	100
A.6	Elevation above sea level throughout the American chestnut natural range. The elevation data was retrieved from WorldClim 2.1 with 2.5m resolution. The black line represents the American chestnut estimated migration path across the Appalachian Mountain range from the EEMS analysis.	101
A.7	Estimated genetic diversity within American chestnut throughout the eastern United States. The orange regions represent areas of lower than expected diversity, and the blue regions represent areas of higher than expected diversity. A loess line of Appalachian Mountain peaks was applied in <i>ggplot2</i> using peak locations obtained from https://en.wikipedia.org/wiki/List_of_mountains_of_the_Appalachian	
A.8	Site-frequency-spectrum for each of the three wild American chestnut populations.	103
A.9	Observed heterozygosity decreases as latitude increases between the three <i>Castanea dentata</i> populations. A. The northeast population has significantly less observed heterozygosity than the central and southwest populations ($p < 0.001$, $p < 0.001$). The grey triangles represent the observed heterozygosity of individual trees within each population. B. Map of sample locations and observed heterozygosity for each sample. Each circle represents a tree sample location and the color is the observed heterozygosity ratio. . .	104

A.10	GO annotations for all RAiSD outlier genes for each of the three American chestnut populations. The gene percentage was calculated by dividing the gene count for each category by its corresponding N value.	105
B.1	Ten uncorrelated climate variables were selected for use in the genotype-environment association analyses. Pairs plot of the selected 10 environmental variables. The bottom half are regression plots and the top half are Pearson's correlations.	116
B.2	Three PC axes were used as synthetic variables for the LFMM genotype-environment association analyses. Pairs plot of the selected 10 environmental variables and the first four PCs from the PCA of the 10 climate variables. The bottom half and top half are linear regressions between the variables. . .	117
B.3	The number of genes that contained at least one adaptive loci within each <i>Castanea dentata</i> chromosome. Scaffolds that did not contain genes with adaptive loci were omitted from the figure.	118
B.4	Gradient Forest adaptive allelic turnover for each of the ten climate variables and 18,483 putatively adaptive loci. Spikes in the binned split importance reveal where changes in adaptive allele density are occurring along each climate gradient.	119
B.5	Venn diagram for the 18,483 adaptive loci identified using RDA and LFMM. For LFMM, only the SNPs from the PC1 and PC2 analyses are shown due to the PC3 analysis finding zero outlier loci.	120
B.6	AIC score evaluation to determine optimal number of seed zones.	121
B.7	BIC score evaluation to determine optimal number of seed zones.	122

B.8	WCSSE score evaluation to determine optimal number of seed zones.	123
B.9	Variance explained scores to determine optimal number of seed zones.	124

List of Tables

2.1	Average Tajima's D, nucleotide diversity, heterozygosity, and inbreeding coefficient across all samples for each American chestnut population.	26
3.1	Drivers of genomic variation in American chestnut. Four partial RDA models were performed to evaluate the influence that climate (10 environmental variables), population structure (first two PC axes from PCA on genomic dataset), and geography (geographic coordinates of each sample) had on the patterns of genomic variation throughout the American chestnut natural range. An ANOVA was performed after each partial RDA to estimate statistical significance.	45
3.2	Tree sampling estimates to capture adaptive diversity from each American chestnut seed zone.	51
A.1	GO enrichment analysis of the unique RAI _{SD} outlier gene set for the southwest American chestnut population. The GO enrichment terms were identified as being significantly overrepresented for the GO biological process (p<0.05).	106
A.2	GO enrichment analysis of the unique RAI _{SD} outlier gene set for the northeast American chestnut population. The GO enrichment terms were identified as being significantly overrepresented for the GO biological process (p<0.05). .	107

A.3	GO enrichment analysis of the unique RAiSD outlier gene set for the central American chestnut population. The GO enrichment terms were identified as being significantly overrepresented for the GO biological process ($p < 0.05$). .	108
A.4	GO enrichment analysis of the shared RAiSD outlier gene set for three American chestnut populations. The GO enrichment terms were identified as being significantly underrepresented or overrepresented for the GO biological process ($p < 0.05$).	109
B.1	The 32 climate variables from ClimateNA that were used for this study and their respective descriptions.	111
B.2	The environmental variable loadings on the three PC axes used for the LFMM analysis.	112
B.3	Number of single-nucleotide polymorphisms (SNPs) in coding and noncoding features. Three datasets were evaluated: Adaptive (18,483 putatively adaptive SNPs from the GEA analyses), Random (randomly selected SNPs with allele frequency distribution matched to the adaptive set), and Full (original filtered SNP dataset from Sandercock et al. [104] with MAF < 0.05 filter). The promoter region includes the 2kb region upstream of the mRNA intervals. Percentage of the total SNPs for each feature are shown in parentheses. . . .	113
B.4	Estimated number of samples to capture adaptive diversity from each seed zone for a two seed zone model.	114
B.5	Estimated number of samples to capture adaptive diversity under a single seed zone model. The estimate for the 99% variance explained did not complete due to excessive run time.	115

List of Abbreviations

INDEL insertion-deletion

SNP single nucleotide polymorphism

TACF The American Chestnut Foundation

WGS Whole-genome sequencing

TACF is a non-profit organization dedicated to the conservation and restoration of the American chestnut tree.

WGS is a comprehensive DNA sequencing method for examining the entire genome of an organism.

A SNP is a change in a single base pair, a nucleotide, in an organism's genome

An INDEL is the removal or insertion of one or more nucleotides in an organism's genome

Preface

The following chapters are co-authored manuscripts, which have been published or are in preparation for journal submission.

- **Chapter 2:**

Sandercock, A. M., Westbrook, J. W., Zhang, Q., Johnson, H. A., Saielli, T. M., Scrivani, J. A., Fitzsimmons, S. F., Collins, K., Perkins, M. T., Craddock, J. H., Schmutz, J., Grimwood, J., & Holliday, J. A. (2022). Frozen in time: Rangewide genomic diversity, structure, and demographic history of relict American chestnut populations. *Molecular Ecology*, 31, 4640–4655. <https://doi.org/10.1111/mec.16629>

- **Chapter 3:**

Sandercock, A. M., Westbrook, J. W., & Holliday, J. A. *in preparation*

- **Chapter 4:**

Sandercock, A. M., Westbrook, J. W., Zhang, Q., & Holliday, J. A. *in preparation*

Chapter 1

Introduction

The American chestnut (*Castanea dentata*) is a deciduous tree that is native to the eastern half of the United States, extending from Mississippi to Maine and into Ontario, Canada [100, 122]. The tree was omnipresent in American forests and served as an important economic and ecological resource [59, 72, 111]. Approximately 100 years ago, a fungal blight (*Cryphonectria parasitica*) decimated the American chestnut populations with the introduction of Chinese (*Castanea mollissima*) and Japanese (*Castanea crenata*) chestnut trees. However, the chestnut blight leaves the root systems of the trees intact. >400 million American chestnuts remain as collar sprouts that survive for several years before being killed again by the blight [86, 109, 116]. Since most of these shoots rarely reach reproductive age, the American chestnut is considered functionally extinct.

To combat the blight, researchers developed blight-resistant chestnut trees. There have been two main methods to accomplish this goal thus far. The first method was from The American Chestnut Foundation (TACF) to backcross American chestnut trees with blight-resistant Chinese chestnut trees [115]. Backcrossing is the process of initially mating an American chestnut tree with a Chinese chestnut tree, and then repeatedly “backcrossing” the subsequent offspring by mating them with American chestnut trees. This method generated trees that were 15/16 American chestnut tree and retained the Chinese chestnut blight resistance. The second method developed a transgenic American chestnut tree by inserting the wheat oxalate oxidase (OxO) gene into the American chestnut genome - allowing for

a disease-resistant tree that maintains almost all of the American chestnut genome, while obtaining blight resistance that is comparable to that of Chinese chestnut [81, 93, 132]. Trees from both methods may be integrated into a breeding program to ensure the population's genetic diversity is maintained while introducing the blight-resistant traits [126, 127].

American chestnut exhibit adaptation to their local climate, particularly to temperature variation [101, 102, 107]. Traditional methods of evaluating climate adaptations used common garden experiments, however, due to the low prevalence of natural reproduction, it is difficult to obtain enough wild seed that sufficiently captures the geographic and climatic variation throughout the native range. While breeding and planting trees that are locally adapted to a given area provides that tree with the best chance to reach reproductive maturity, a comprehensive assessment of adaptive diversity in wild American chestnut is lacking, and it is unknown how much wild adaptive diversity has been captured by the backcross breeding program. Identifying which loci are related to specific environmental variables can aid in breeding efforts for *C. dentata* so introduced trees will be adapted to their local climate. Thus, accounting for these local adaptations and maintaining high levels of genetic diversity will provide American chestnut restoration populations with an increased chance to survive the impacts of climate change [8, 65, 97].

Climate change will continue to impact the environmental conditions throughout the American chestnut natural range, altering the boundaries of the habitable landscape. The American chestnut was slow to migrate [34] and will not be able to avoid the negative effects of climate change [4, 5], so capturing wild adaptive diversity in the form of germplasm conservation orchards is essential to aid in restoration programs. Seed zones are regions within a species range that provide geographical boundaries for unique adaptations to the local climate. Seed zones can be used as a reference for replanting recommendations and regions for the conservation of locally adapted germplasm. Previous assessments of climate-adapted

regions within the American chestnut range suggested >10 different seed zones exist, but were not specific to the American chestnut and did not include genomic information in their development [19, 90]. In addition, a standardized method for making sampling recommendations to capture genomic diversity from seed zones does not exist.

To address these knowledge gaps in wild and backcross American chestnut populations, this dissertation resolved the questions regarding the neutral and adaptive genomic diversity of wild American chestnut populations. In addition, the adaptive genomic diversity captured within the backcross breeding program was evaluated and sampling recommendations for germplasm conservation of wild adaptive diversity for the development of breeding plans with backcross trees were made. Chapter 2 explored the population structure, neutral genomic diversity, and demographic history of wild American chestnut populations. Chapter 3 examined the genomic basis of local adaptation to climate in wild American chestnut, identified unique seed zones for chestnut conservation, and made sampling recommendations to capture wild adaptive diversity from each seed zone for germplasm conservation. Chapter 4 evaluated the adaptive genomic diversity captured within the backcross breeding program and provided estimates of candidate regions for replanting locally adapted backcross chestnut trees. Lastly, Chapter 5 concluded the dissertation by summarizing the key findings of Chapters 2-4 and making recommendations for future study directions.

Chapter 2

Frozen in time: Rangewide genomic diversity, structure, and demographic history of relict American chestnut populations

2.1 Introduction

Exotic pests and pathogens are among the greatest threats to forest health worldwide [45]. In the United States, >300 non-native forest pests have become established since 1860, of which approximately 100 have been classified as high impact [11]. These include such fungal pathogens as *Ophiostoma*, *Phytophthora*, and *Cryphonectria*, which have lead to catastrophic losses of *Ulmus*, *Quercus*, and *Castanea* species, respectively. As a result, an intense effort to develop disease-resistant forest tree varieties is currently being conducted. However, development of disease resistant forest tree varieties for species restoration is a fundamentally different problem than breeding for resistance to crop pathogens - requiring accounting for natural adaptive diversity matching to local abiotic and biotic environments.

Introduction of the necrotrophic fungus *Cryphonectria parasitica* to North America from

Asia in 1904, and its subsequent spread, functionally extirpated American chestnut (*Castanea dentata* (Marsh.) Borkh) by the middle of the century. The loss of American chestnut was significant from ecological, economic, and societal perspectives. Pre-blight *C. dentata* was a canopy tree reaching 18 - 36 meters in height and 1 - 1.5 meters in diameter at maturity, and was widespread in forests of the eastern United States and southeastern Canada [23, 39, 100, 122]. The chestnut's rapid growth, coupled with decay resistant wood, made it the most valuable hardwood species in the United States [59, 111], while its prodigious and reliable nut crop was an important source of food and feed [72]. Density estimates prior to the introduction of *C. parasitica* suggest that 2-4 billion trees were killed (<https://www.fao.org/3/x5348e/x5348e02.htm>), which reduced the carrying capacity of Appalachian forests for wildlife that relied on the consistent and prodigious nut crop [40]. However, as a bark pathogen that only infects above ground tree parts, >400 million American chestnuts survive as collar sprouts from uninfected roots [86, 109, 116]. Although some of these re-sprouts are able to flower, viable seed is rare due to a lack of conspecific pollen. This lack of reproduction, and concomitant negligible adaptive capacity, has led to American chestnut being identified as one of the 25 most vulnerable tree species to environmental change [92].

Two approaches to developing blight resistant chestnuts are currently in progress. First, The American Chestnut Foundation (TACF) has developed segregating populations by hybridizing Chinese chestnut (*Castanea mollissima*), which is resistant to the blight, with American chestnut, followed by repeated backcrossing to American chestnut using pollen from rare, flowering, wild trees [115]. Second, genetically modified trees have been developed that express a wheat oxalate oxidase (OxO) transgene, which confers blight resistance comparable to that of Chinese chestnut [81, 93, 132]. These methods are also being combined to develop trees that have stacked resistance loci [126, 127]. Prior to their widespread use as restora-

tion tools, however, we need a better understanding of the extent to which natural adaptive diversity has been captured in the backcross populations, and the transgenic founder lines will need to be crossed with wild germplasm to produce diverse, locally adapted reforestation populations.

The American chestnut range, like those of other eastern North American hardwoods, contracted during the Wisconsin glaciation [112, 134], and postglacial recolonization bottlenecks may have impacted contemporary genetic diversity [55, 78]. Additionally, geographic barriers to gene flow, such as the Appalachian Mountain range, could have shaped postglacial migration routes and contemporary population structure. Rangewide genomic diversity in this species has likely also been shaped by local adaptation –a common garden study using rare, wild seed found that genotypes from warmer and moderate climates grew faster but had greater winter injury than northern seed sources [102], which is consistent with many other temperate and boreal forest tree species [105]. Genetic diversity in chestnut is thus likely to be unevenly distributed across multiple genetic clusters that are influenced by past demographic and climatic events [56, 62].

Successful restoration of the American chestnut will rely on an understanding of this extant genomic diversity in the wild [54], which can be leveraged to increase effective population size and adaptability of blight-resistant populations through targeted breeding efforts [128]. The first step in this process is to define broad management units on the basis of population structure and postglacial history for the species [48], which will inform prioritization of geographic areas for *ex situ* propagation and conservation of wild trees that can be used to introgress natural genetic variation into blight resistant populations.

Previous population genetic studies used microsatellite and restriction site-associated DNA sequencing (RAD-seq) to characterize population structure and genetic diversity of American chestnut [49, 66, 80, 113]. These studies suggest the southern portions of the chestnut

range are the most genetically diverse and that two populations exist, with either a distinct northeastern population [49, 80] or a distinct southern population [113]. Furthermore, they suggest migration from a southern refugium following the last glacial maximum [49, 113]. While these studies provide our first glimpse of the extant patterns of genetic diversity in chestnut, they were limited by relatively low sampling density and incomplete coverage of the historical species range, as well as by relatively low-density genotyping methods that lack power to comprehensively characterize genome-wide diversity and signatures of selection [71, 119]. Additionally, modern demographic inference tools that approximate the ancestral recombination graph require high density genotyping that can only be achieved with whole-genome resequencing (WGS).

In this study, we leveraged a recently completed chromosome-scale reference genome for American chestnut (https://phytozome-next.jgi.doe.gov/info/Cdentata_v1_1) and $\approx 17\times$ WGS data for each of 384 American chestnut genotypes, sampled from across the entire historical species range, to (i) estimate population structure; (ii) describe demographic history and contemporary barriers to gene flow; (iii) evaluate genomic diversity across the genome and among populations; (iv) and identify signatures of selection. Results of this study will be used to identify management units for germplasm conservation, and will aid in the overall efforts to restore American chestnut to forests of eastern North America.

2.2 Methods

2.2.1 Leaf sample collection

TACF staff and volunteers sampled leaves from ≈ 1000 geo-referenced American chestnut genotypes from throughout the species range. Remnant American chestnut re-sprouts and

seedlings were identified by leaf and twig morphology (A.1;[88]). The youngest leaves were sampled from each tree in May through July over three growing seasons (2018 – 2020). During and immediately after collection, leaves were kept cool with ice and refrigeration or were desiccated and stored with silica gel. Leaves were shipped to Virginia Tech within 3 weeks of collection and stored at -80°C. The original coordinates of the sampled orchard trees were retrieved from TACF’s dentataBase (www.acf.herokuapp.com) and TreeSnap ([30]; <https://treesnap.org/>), and the coordinates of field sampled trees were obtained using GPS units.

2.2.2 DNA isolation and sequencing

We selected 384 genotypes for WGS (3.1) on the basis of geography, with the goal of including as much of the historical range as possible. Leaves were ground using a Spex 2000 Geno/-Grinder and ceramic beads with three 45 second intervals of grinding interspersed by submersion in liquid nitrogen. DNA was extracted with Qiagen’s DNAeasy Plant DNA extraction kit. For samples 1-96, a modified phenol-chloroform cleanup step, adapted from the OPS Diagnostics CTAB protocol for plants (https://opsdiagnostics.com/notes/protocols/ctab_protocol_for_plants.htm) was performed to remove organic contaminants from older leaves. Due to difficulty in extracting enough DNA necessary for sequencing, we modified the Qiagen extraction protocol for the remaining leaf samples. For samples 97 through 384, an additional 100% ethanol wash step was performed to remove organic contaminants that carried over from the previous steps and to dry the spin column membrane. DNA quality and concentration were measured by a Nanodrop One^c and Qubit 3.0 Fluorometer, respectively. When DNA concentration was low, a secondary CTAB-based extraction using the OPS Diagnostics CTAB protocol for plants was used (https://opsdiagnostics.com/notes/protocols/ctab_protocol_for_plants.htm#phenolpr). DNA was stored in a 100-200 µl AE solution in a -20°C freezer. PCR-free library preparation

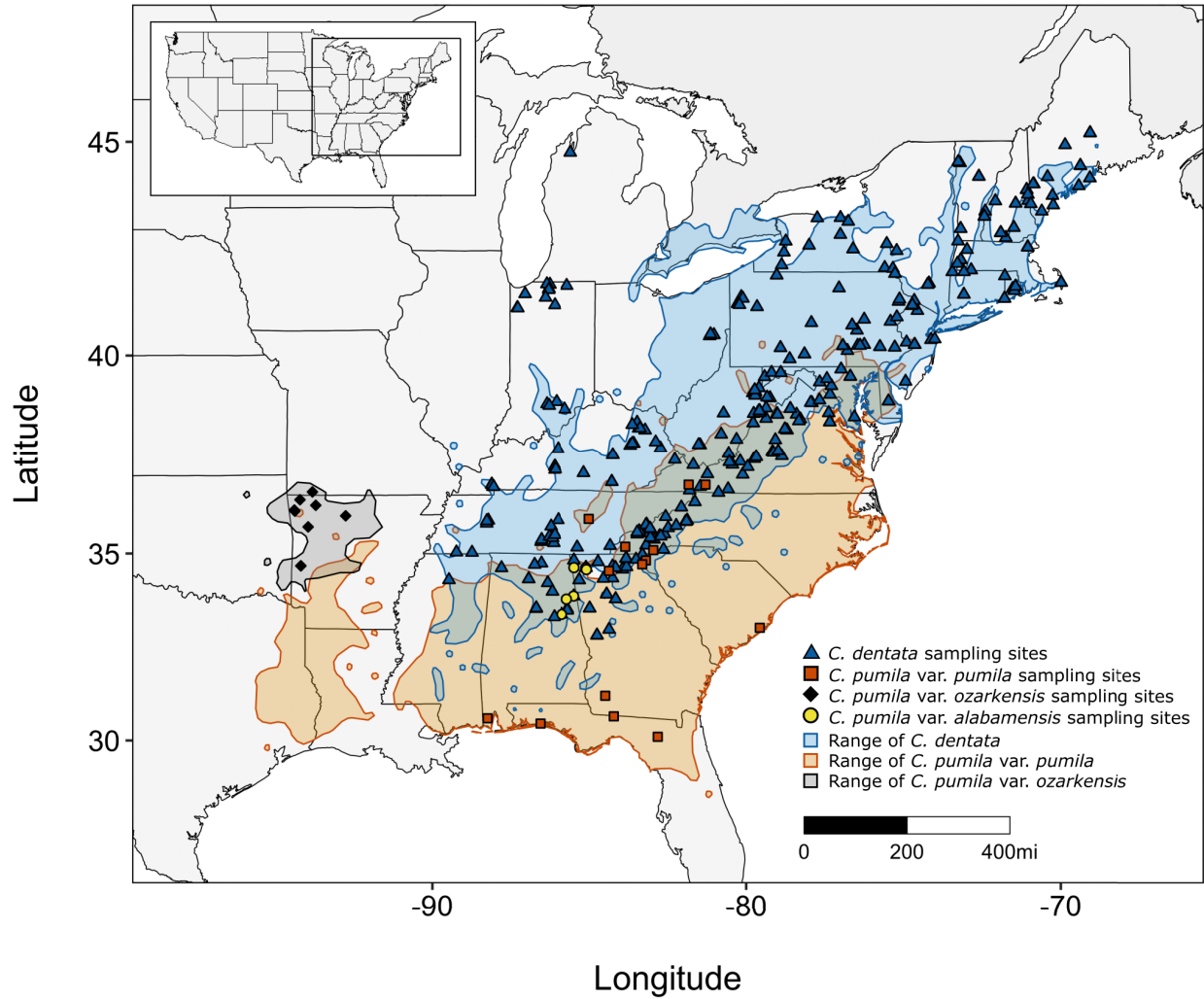


Figure 2.1: Range map of the North American *Castanea* species and the locations of the selected American chestnut and chinquapin trees.

and sequencing were conducted at the HudsonAlpha Institute for Biotechnology. Libraries were sequenced in batches of 48 on an Illumina NovaSeq 6000 instrument in 2x150bp paired-end mode with a target depth of 20x. The number of lanes to use for sequencing was determined after DNA quality and quantity checks were performed during library preparation. To achieve the desired mean coverage, three lanes were used for samples 1-96, one lane for samples 97-192, and two lanes for samples 193-384.

We also performed WGS on a set of 95 *Castanea* samples that included representatives from most *Castanea* species to validate the ancestry of our putative wild *C. dentata* samples. This was important, because *C. dentata* is sympatric with *C. pumila* var. *pumila* and *C. pumila* var. *alabamensis*, and several introduced species, including *C. mollissima*, *C. crenata*, and *C. sativa*. This 'reference' panel included 19 *C. sativa*, 15 *C. pumila* var. *pumila*, 10 *C. pumila* var. *ozarkensis*, six *C. pumila* var. *alabamensis*, four *C. dentata*, one *C. seguinii*, two *C. henryi*, 18 *C. crenata*, and 20 *C. mollissima*. Leaf samples from *C. mollissima* and *C. crenata* were collected from trees planted in the United States; genomic DNA for *C. sativa* was provided by Dr. R. Costa from trees growing in Portugal; and *C. pumila* samples were collected from wild trees in the United States. DNA was extracted as above. Libraries were prepared with an Illumina Nextera kit at the Duke University Center for Genomic and Computational Biology and sequenced using two lanes on an Illumina NovaSeq 6000 S4 flow cell (48 samples per lane) in 2x150bp paired-end mode. The resulting fastq files for all samples were deposited in the NCBI SRA (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA804196>).

2.2.3 Bioinformatics

Bioinformatic analyses were performed on Virginia Tech's Advanced Research Computing system. Individual fastq files were aligned using the BWA mem algorithm with the -M and

-R flags and the *C. dentata* genome as a reference (*Castanea dentata* v1.1; <http://phytozome-next.jgi.doe.gov/>). Resulting SAM files were converted to BAM format, sorted, and indexed using SAMtools [68, 69]. Where samples were sequenced in multiple lanes, these were combined into a single BAM file using samtools merge. For the 95 *Castanea* species samples where PCR was performed in the library preparation, PCR duplicates were removed from the BAM files using MarkDuplicates in picard.

Single nucleotide polymorphisms (SNPs) were called using a custom pipeline adapted from the Broad Institute’s Genome Analysis Toolkit (GATK v3.8) best practices [75, 91, 120]. Chromosome GVCF files for each individual were combined with GatherVcfs and merged with GenotypeGVCFs. Polymorphisms were quality filtered with GATK VariantFiltration algorithm following GATK best practices, and further filtered in VCFtools v0.1.17 [32] to include only biallelic SNPs and SNPs with <10% missing data per site. Individuals with >10% missing data were removed. Unless otherwise noted, a minor allele frequency (MAF) filter was applied ($MAF = 1/2n$, where n = number of samples in the VCF file). VCFtools was used to calculate the sequencing coverage and depth. Genome-wide scans for SNP density were determined using the `-SNPdensity` option with 50kb windows. Mean sequencing depth per variant was calculated using the `-site-mean-depth` option and mean sequencing depth per sample was determined using `-depth`. SNP density for each chromosome was visualized in *ggplot2* `geom_smooth` with default parameters [121].

2.2.4 Species identification and estimation of hybridization in wild American chestnut populations

To avoid biasing subsequent analyses, we used the *Castanea* species reference dataset with ADMIXTURE to test for evidence of introgression and identify samples for exclusion. All

Castanea species, except *C. seguinii* and *C. henryi*, were included, as there were only one and two samples of those species, respectively. The *C. dentata* and species reference datasets were combined using bcftools merge and filtered with bcftools to retain biallelic SNPs and remove singletons [33]. Both SNPs and INDELs were retained for these analyses. ADMIXTURE was run with the `-cv` flag to perform a five-fold cross-validation for $K = 1-9$ [9]. The most likely K value was selected as the number of clusters where each *Castanea* species first separated into at least one distinct group (*C. pumila* varieties were considered as a single cluster). Putative *C. dentata* samples were classified as hybrids or misidentified if they had cluster membership $>10\%$ with a different species. We used a $>10\%$ threshold since low levels of interspecific ancestry are likely to be artefacts [17] and not recent hybridization events.

2.2.5 Population structure

We assessed population structure within *C. dentata* to identify barriers to gene flow and to partition the species into genetic sub-populations for use in subsequent analyses. Population structure was estimated using ADMIXTURE and Discriminant Analysis of Principal Components (DAPC) [61]. ADMIXTURE estimates ancestry using a model-based approach, like STRUCTURE, to estimate ancestry, but is more computationally efficient [9]. DAPC combines PCA and a Discriminant Analysis (DA) to identify demographically independent clusters by minimizing within group variance and maximizing between group variances [61]. The VCF file was first converted to a BED file and LD-pruned to include SNPs with R^2 values < 0.1 within 50 SNP sliding windows (step size 10 SNPs) in PLINK v1.9 [95]. ADMIXTURE was run on the pruned BED file for K values 1-9. The `-cv` and `-j120` options were enabled to allow for five-fold cross validation and multithreading, respectively. Population membership for each sample was determined by highest ancestry proportion from the most likely ADMIXTURE model.

DAPC analysis was performed in R using the *adegenet* package [60]. The pruned BED file was converted back to a VCF file in PLINK v1.9 using the `–recode vcf` option and the `–ref-from-fa` option, and converted to genlight format using the *vcfR* package [63]. The optimal number of clusters was determined with the `find.clusters` function in *adegenet*, and BIC values were calculated for a maximum of five clusters and 355 principal components (PCs). The `optim.a.score` was assessed for each DAPC run to determine the optimal number of PCs to retain.

To test for isolation-by-distance (IBD), we performed Mantel tests using the R package *ade4* [41]. The Mantel test required the input of genetic and geographic distance matrices. Due to computational constraints, a 100k SNP subset was used for the genetic distance matrix. The pruned VCF file from DAPC was imported into R and converted to the *genind* format using the *vcfR* package [63]. The *genind* file was converted to the *genpop* format with the *genind2genpop()* function in *adegenet* [60]. A genetic distance matrix was generated with the *dist.genpop()* function using Edward’s distance [60], and a Euclidian geographic distance matrix was estimated using the individual sample locations and the *dist()* function. Finally, we tested for IBD with the *mantel.randtest()* function with 999 permutations [41].

2.2.6 Demographic history

Demographic history analyses were performed to describe the broad patterns of population size changes that may have influenced current patterns of genomic diversity. The sequential Markovian coalescent approach in SMC++ was used to estimate the historical effective population sizes (N_e) of *C. dentata* from whole genome data [118]. SMC++ required the input VCF file to not have any filtering for LD or MAF, so the input VCF file was only filtered for high missing sites (>10% missing data) and biallelic SNPs. After

removing 18 samples for high missing data and 10 samples for lacking >90% *C. dentata* ancestry, the input VCF file contained 356 samples. Sites missing in the *C. dentata* reference genome (*Castanea dentata* v1.1; <http://phytozome-next.jgi.doe.gov/>), which could be erroneously interpreted as long runs of homozygosity, were masked with a bed file generated with a conversion script (<https://www.danielecook.com/generate-a-bedfile-of-masked-ranges-a-fasta-file/>). The SMC files were generated for each of the 12 chromosomes using the `vcf2smc` function in SMC++. The `estimate` function was then used with a mutation per generation rate set to 5.2×10^{-8} from estimates of pedunculate oak (*Quercus robur*) [108]. A 30-year generation time was assumed to convert coalescent events to years. A CSV file of the results was generated with the `plot` function and `-c` option for plotting in R [96].

2.2.7 Migration rates

We estimated migration rates to further characterize contemporary geographic barriers to gene flow. Migration rates for *C. dentata* were estimated using Estimated Effective Migration Surfaces (EEMS) [89]. EEMS estimates and visualizes effective migration and diversity across a given geographic range using genetic data from known locations [89]. The EEMS program requires three input files: an average pairwise differences matrix (DIFFS), list of sample geographic locations (COORD), and a list of habitat boundary coordinates (OUTER). To reduce computational load, the LD pruned *C. dentata* dataset from the population structure analyses was used. The pruned BED file was converted to the EEMS DIFFS file using the EEMS program `bed2diffs`. The habitat boundary was a large single polygon and the coordinates of the perimeter were manually selected using a google maps like online tool (<http://www.birdtheme.org/useful/v3tool.html>). The habitat boundary coordinates were individually selected with the goal of encompassing the perimeter of all of the US states that contained at least one of the American chestnut samples used in this study. The final

habitat boundary included the entirety of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New Jersey, New York, Pennsylvania, Michigan, Ohio, Indiana, Delaware, Maryland, West Virginia, Virginia, Kentucky, Tennessee, North Carolina, South Carolina, Georgia, Alabama, and Mississippi.

An initial run was performed with 8 million iterations, 1 million burn-in, 600 nDemes, 9999 thinning, and the hyperparameters in their default setting. EEMS relies on properly adjusted proposal variances. The proposal variances are adjustable values used by the EEMS algorithms, and which influence the predicted migration and diversity rates. Another important EEMS metric is the output proposal acceptance frequency, which reports the frequency of the EEMS algorithm accepting proposals of different types. Through their own testing, the EEMS authors determined that during the analysis, each of the output proposal acceptance frequencies should be between 20%-30%, but it is sufficient for them to range from 10%-40% (<https://github.com/dipetkov/eems/blob/master/Documentation/EEMS-doc.pdf>). Following the initial EEMS run, there were two proposal frequencies that were either less than 10% or greater than 40%. To account for this, the `mEffectProposals2` parameter was increased from 0.1 to 0.2 and `mrRateMuProposals2` was decreased from 0.01 to 0.002. Four values of demes, 200, 350, 500, and 650 were used to evaluate the number of demes for best fit. Three chains, each with 20 million iterations, four-million burn-in, and 9999 thinning were performed for each deme and their outputs were combined in R using the `rEEMSplots2` program. The posterior trace plot was evaluated to determine if the MCMC chain converged. Any EEMS run that did not converge was restarted from the final parameter state in the previous run. The results were mapped in R using `rEEMSplots2`. A linear regression line was fitted to the default 'Dissimilarities between pairs of sampled demes' plot for each deme value. The R^2 values from this analysis were similar for the 650 deme model ($R^2 = 0.92$) and 500 deme model ($R^2 = 0.93$), and we thus present results for the 650 deme model

due its increased resolution. Appalachian Mountain peak locations were obtained from Wikipedia (https://en.wikipedia.org/wiki/List_of_mountains_of_the_Appalachians) and a smoothed loess line of these locations was added to the EEMS figure using *ggplot2*.

2.2.8 Genome-wide nucleotide diversity

We evaluated genome-wide patterns of nucleotide diversity and divergence within and between each population to understand how demography may have shaped these parameters, and to help prioritize geographic areas for germplasm conservation. We used ANGSD (Analysis of Next Generation Sequencing Data) to estimate these parameters, as its use of genotype likelihoods provides less biased estimates than those based only on variant calls [64, 83]. The sorted BAM files were used as the initial input for ANGSD to estimate the site allele frequency (SAF) likelihood for each population (-doSaf 1) on a per-chromosome basis and then merged using realSFS cat. Filtering parameters used for each SAF file were adapted from ANGSD recommendations and other hardwood tree studies [56, 103]. These parameters were: adjust mapping quality for excessive mismatches (-C 50), minimum base quality score 20 (-minQ 20), minimum mapping quality 30 (-minMapQ 30), discard reads that do not uniquely map (-uniqueOnly 1), only retain sites where the pair could be mapped (-only_proper_pairs 1), and remove 'bad' reads (-remove_bads 1). To polarize the site-frequency-spectrum (SFS), an ancestral reference fasta was generated in ANGSD using 11 *C. mollissima* BAM files with the following parameters: -doFasta 2 and -doCounts 1. Finally, the SFS was estimated using realSFS.

Nucleotide diversity, tests of neutrality, and population differentiation were subsequently estimated with ANGSD. Thetas were first estimated using realSFS saf2theta, and for each chromosome using thetaStat do_stat. Per site nucleotide diversity was estimated by dividing

the pairwise theta (tP) by the number of sites ($-nSites$) for each genomic region. Pairwise F_{ST} was estimated using the per chromosome SAF files for each pair of populations per chromosome to generate pairwise 2D-SFS files. F_{ST} was then estimated for each pairwise 2D-SFS file using realSFS fst index. These statistics were summarized in 50kb windows with a 10 kb slide.

VCFtools was used to estimate observed heterozygosity (H_e) for each sample by subtracting observed homozygosity (H_o) from the number of sites and dividing that total by the number of sites. H_o for each population was the average of this per-sample mean for all samples within a population. Expected heterozygosity was similarly estimated for each sample and population. Significant differences in H_o between populations were tested with the `aov()` function in R and for pairs of populations with the function `TukeyHSD()`.

2.2.9 Detecting signatures of positive selection

Genome-wide scans to detect evidence of positive selection within *C. dentata* populations were used to identify both local (within-population) and global (rangewide) targets of selection. Specifically, we used RAI_{SD} (Raised Accuracy in Sweep Detection) to identify genomic regions with evidence of recent positive selection in each population [7]. RAI_{SD} estimates a composite statistic, ω , which integrates LD, the SFS, and diversity. The filtered *C. dentata* VCF file was subset for each of the three populations as input. Regions that were missing data in the *C. dentata* reference genome were masked. Default parameters were used, except that we assigned a seed for the random number generator. After filtering, RAI_{SD} retained 8,833,026 SNPs for the northeast population, 7,545,294 SNPs for the central population, and 11,737,005 SNPs for the southwest population. We used the quantile function in R [96] to identify the 0.1% outlier regions in each of the RAI_{SD} site reports per-population.

To identify genes proximal to the 0.1% RAI_{SD} outlier regions, we annotated those regions using the American chestnut genome feature file (Cdentata_673_v1.1.gene.gff3.gz; <http://phytozome-next.jgi.doe.gov/>). Genes were assigned to outlier regions if they were within 1 Kb of the region. Only unique genes were retained. We then identified *Arabidopsis* orthologs for each outlier gene using BioMart on Phytozome [50, 110]. The list of *A. thaliana* orthologs were entered into the TAIR gene ontology tool to obtain their functional categories ([16]; <https://www.arabidopsis.org/tools/bulk/go/index.jsp>). Finally, we performed GO enrichment analyses for biological function on four gene sets. These sets were the unique genes belonging to the southwest, central, and northeast populations, and the set of genes that are shared between all three populations. The *A. thaliana* orthologs were retrieved for each gene set using the previously described methods, and were submitted to the TAIR GO Term Enrichment for Plants tool (https://www.arabidopsis.org/tools/go_term_enrichment.jsp), which sends the data to the PANTHER Classification System [77]. The summary for each gene set analysis using the PANTHER Classification System was as follows. Analysis type: PANTHER Overrepresentation Test (Released 20210224), Annotation Version and Release Date: GO Ontology database DOI: 10.5281/zenodo.5228828 Released 2021-08-18, Reference list: *Arabidopsis thaliana* (all genes in database), Annotation Data Set: GO biological process complete, Test Type: Fisher's Exact, Correction: Bonferonni correction for multiple testing for $p < 0.05$.

2.3 Results

2.3.1 Genome resequencing

As an effectively obligate outcrosser with a historically large effective population size, we expected *C. dentata* to have high heterozygosity [130]. As such, we targeted 20X coverage of each sample to enable high-quality genotype calls across most of the genome. Of the 384 samples sequenced, 86 had greater than 20x coverage, 242 had 10-20x coverage, and 56 had less than 10x coverage. The 384 sample *C. dentata* VCF file contained 23,720,251 SNPs and INDELs. Eighteen samples with greater than 10% missing data were removed, and 10 additional samples were removed that had > 10% cluster membership with one or more of the *Castanea* species reference samples. The final *C. dentata* dataset contained 356 individuals with an average coverage of $\approx 17x$ and 21,136,994 high-quality SNPs (A.2). The pruned *C. dentata* dataset contained 3,539,550 SNPs. The *Castanea* species reference dataset contained 92 samples and 62,647,079 SNPs that passed the filtering criteria. The combined and filtered *C. dentata* and *Castanea* species reference dataset contained 76,378,648 SNPs and INDELs.

2.3.2 Hybridization

Because non-native *Castanea* species have been widely planted in the United States, we sought to exclude samples with evidence of ancestry from one or more of these congeners using whole-genome resequencing (WGS) data from a panel of multiple genotypes from most *Castanea* species. ADMIXTURE analysis with the combined *C. dentata* and *Castanea* species reference panel suggested seven clusters best explained the data, with each *Castanea* species as an individual cluster in addition to three clusters within *C. dentata* (A.3). Overall,

we did not find evidence of widespread introgression in our *C. dentata* dataset. Of the 366 putative American chestnut individuals that passed quality filtering, 340 had >99% assignment to the *C. dentata* clusters., 16 had 91.6% - 98.2% *C. dentata* ancestry, and 10 had >10% ancestry from other *Castanea* species. The latter were removed from further analyses. These include three samples identified as *C. pumila*, four that were *C. sativa* x *C. dentata* hybrids, and three that were *C. mollissima* x *C. dentata* hybrids. In addition, there were 16 *C. dentata* trees with low levels of apparent ancestry from other *Castanea* species, including 14 samples from Kentucky and Tennessee that had 1-8% *C. crenata* ancestry and two samples from Maryland and New York that had 3-8% *C. mollissima* ancestry. We also observed possible hybridization or incomplete lineage sorting among the North American *Castanea* species. Among the *C. pumila* varieties, all of the *C. pumila* var. *ozarkensis* individuals had moderate levels of membership with the southwest *C. dentata* cluster (10-13%) and three of the *C. pumila* var. *alabamensis* samples shared ancestry with the southwest *C. dentata* cluster (6-7%). Conversely, two *C. dentata* that were sympatric with the *C. pumila* var. *alabamensis* range had *C. pumila* ancestry of 1-2%.

2.3.3 Population structure in *Castanea dentata*

We hypothesized that the *C. dentata* range could be partitioned into several broad populations reflecting a glacial refugium, the species' history of post-glacial recolonization, and contemporary gene flow. Population structure within *C. dentata* was best explained by a two or three population model as identified by the DAPC BIC plot and ADMIXTURE CV error plot, respectively (2.2). Additionally, dispersal was low throughout the species range, with the presence of isolation-by-distance (IBD) found to be significant (A.4). The three population ADMIXTURE model was characterized by a southwest, central, and northeast cluster (2.2); with population boundaries following precipitation and temperature gradients (A.5).

The southwest and central population separated in northern Georgia and eastern Tennessee, while the central and northeast population have an area of admixture in Pennsylvania before becoming more distinctly separated in southern New York. The central cluster showed low levels of admixture with the northeast cluster, and vice versa, whereas the southwest cluster was largely distinct. The two population DAPC model included the same southern population and boundary as ADMIXTURE, but the central and northeastern populations were merged. There was >98% agreement in population membership assignments between the DAPC and ADMIXTURE analyses at the same K values.

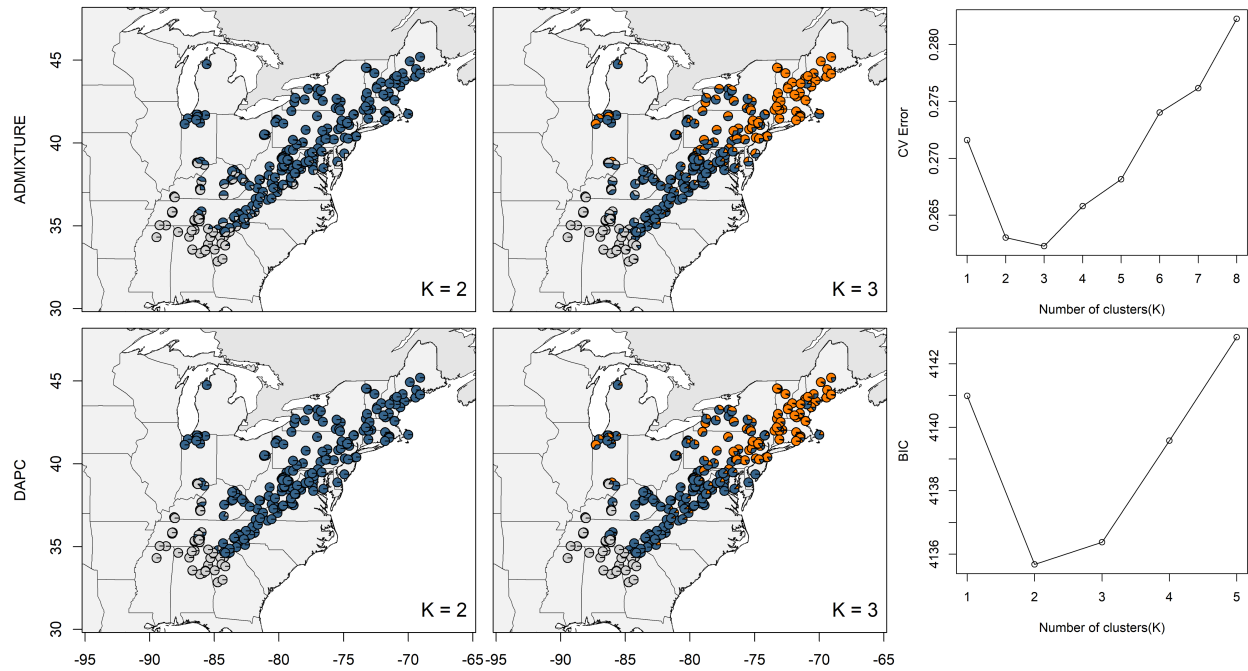


Figure 2.2: The wild American chestnut population structure is best described by three or two populations. The pie charts represent the sample locations and the proportion of color within each pie chart represents population membership. The minimum CV error score and minimum BIC score for each K value (right) were used to determine most likely population number for the ADMIXTURE and DAPC analyses, respectively.

2.3.4 Migration rates

The Appalachian Mountains are the primary geographic feature of the *C. dentata* range, and we hypothesized that these mountains may represent an east-west barrier to gene flow. EEMS analysis supported this, with migration running from southeast to northeast on either side of the Appalachians (2.3a). A single region of above average effective migration rate and relatively lower elevation was shown in southern West Virginia that crosses the Appalachian Mountain range (2.3a; A.6). The ADMIXTURE K=4 model agrees with the EEMS estimates and suggests a further subdivision of the central population on either side of the Appalachian Mountains (2.3b). EEMS also estimates a diversity parameter (q), which reflects genetic dissimilarity between individuals within the same deme, and can thus be thought of as the within-population component of genetic variance. This diversity parameter was generally high throughout the range, though more so in the central and southern portion of the range, and somewhat lower in the northeast and northwest (A.7). Pockets of lower diversity, and thus higher intra-deme genetic similarity, on the outer edges of the species native range may reflect areas of more recent expansion-associated bottlenecks (A.7).

2.3.5 Demographic history

We used SMC++ to understand the effects of cyclical Quaternary glaciation on effective population size in chestnut throughout its range. Estimates of N_e over time suggest that each population underwent contractions and expansions in N_e , beginning ≈ 2 Mya. All populations followed a similar pattern of demographic history, however, the southwest population lagged the central and northeastern populations' events by ≈ 25 -100K years. N_e rapidly increased for all three populations approximately 6.7-11.7Kya, after which the central population underwent an additional contraction within the past 7,000 years (2.4). The southwest popula-

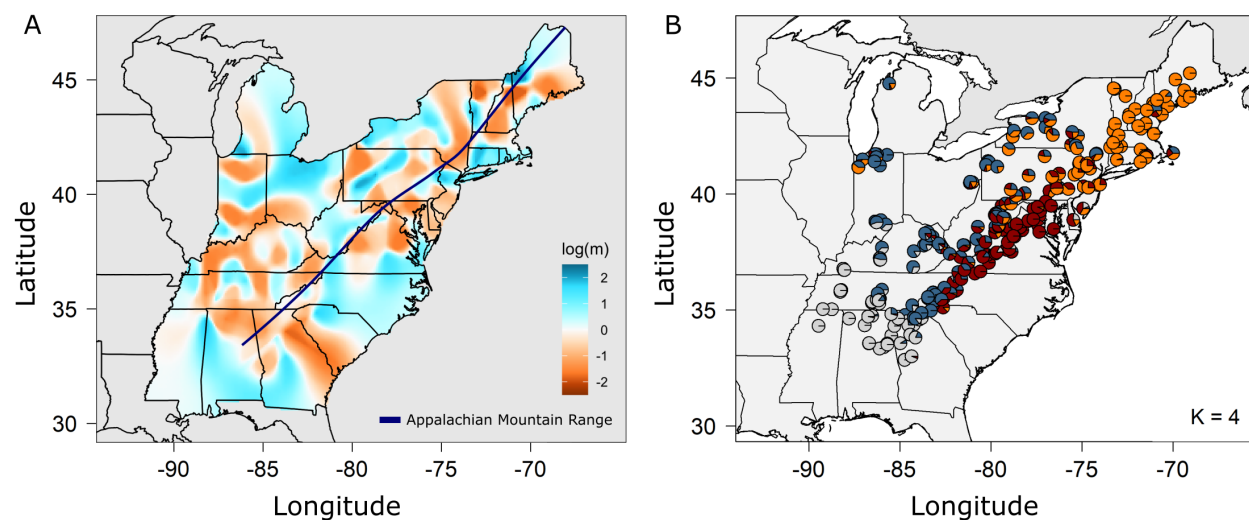


Figure 2.3: The Appalachian Mountain range was a barrier to gene flow for post-glacial migrating American chestnut populations. **A.** EEMS migration rate estimates for American chestnut populations. Regions of blue are above average migration estimates, while orange regions are below average migration estimates. A loess line of Appalachian Mountain peaks was applied in *ggplot2* using peak locations obtained from https://en.wikipedia.org/wiki/List_of_mountains_of_the_Appalachians. **B.** AD-MIXTURE analysis with $K = 4$. The pie charts represent sample location and the color represents population membership.

tion had the highest contemporary N_e ($N_{e(\text{southwest})}=20,306$, $N_{e(\text{central})}= 8,347$, $N_{e(\text{northeast})}=13,078$).

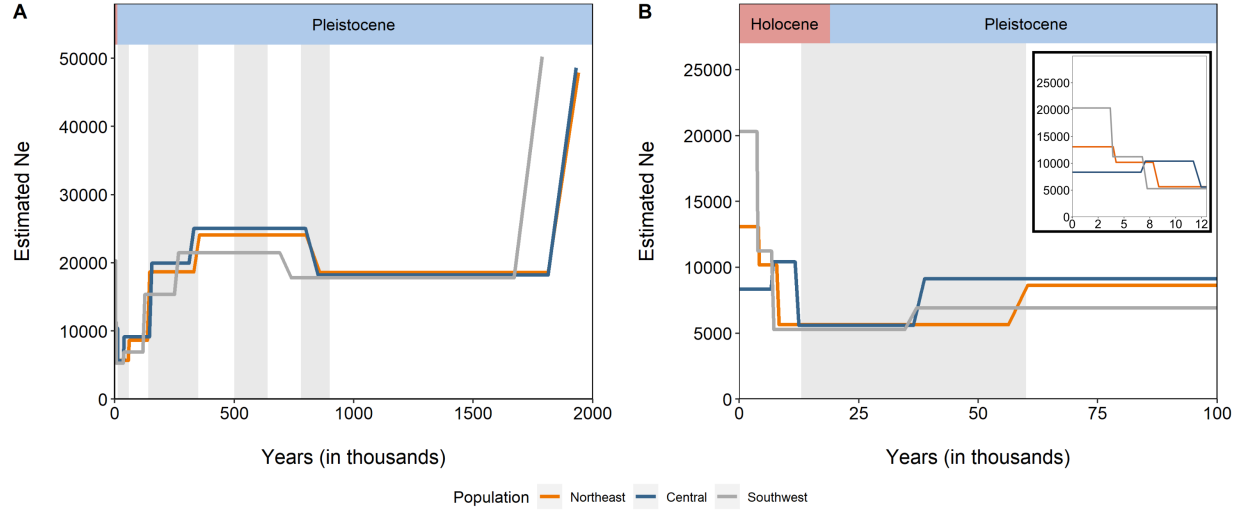


Figure 2.4: American chestnut populations underwent bottleneck and expansion events that coincide with recent North American glaciations. Greyed regions represent glaciation event times approximated from Bowen [18] (as cited in [15]). Each colored line represents the N_e for each of the three populations. Coalescent scaling was converted to years with a 30 year generation time. **A.** SMC++ N_e estimates from the present to 2Mya. **B.** SMC++ N_e estimates from the present to 100Kya. The figure inset into the upper right-hand corner is the SMC++ N_e estimates from the present to 13Kya.

2.3.6 Genomic diversity

To understand how demography and genomic context has shaped diversity, we used ANGSD to estimate the global and local SFS, and associated summary statistics. The ANGSD analyses revealed regions within each chromosome that contained negative Tajima's D and reduced nucleotide diversity (2.5). Several of these regions, such as on chromosome six, nine, and ten, were characterized by elevated F_{ST} values (2.5).

The southwest population had the greatest nucleotide diversity, followed by the central and northeast populations ($\pi_{\text{southwest}} = 0.0069$; $\pi_{\text{central}} = 0.0064$; $\pi_{\text{northeast}} = 0.0058$). All

populations had negative average Tajima's D, which were similarly clinal ($D_{\text{southwest}} = -1.083$; $D_{\text{central}} = -1.016$; $D_{\text{northeast}} = -0.335$). Consistent with these negative values for Tajima's D, the SFS plot revealed that for each population, there was an excess of rare and high frequency variants (A.8). Sliding window analyses revealed heterogeneous genome-wide Tajima's D, nucleotide diversity, and F_{ST} (2.5). Throughout the genome, the southwest population had the most negative Tajima's D values, followed by the central and northeast populations (2.5). The southwest population had the highest nucleotide diversity values throughout the genome, with decreasing values for the central and northeast populations (2.5). F_{ST} estimates between population pairs were relatively low for all comparisons, with the highest divergence between the southwest and northeast populations ($F_{ST(\text{southwest-northeast})} = 0.1076$, $F_{ST(\text{southwest-central})} = 0.0705$, $F_{ST(\text{central-northeast})} = 0.0268$).

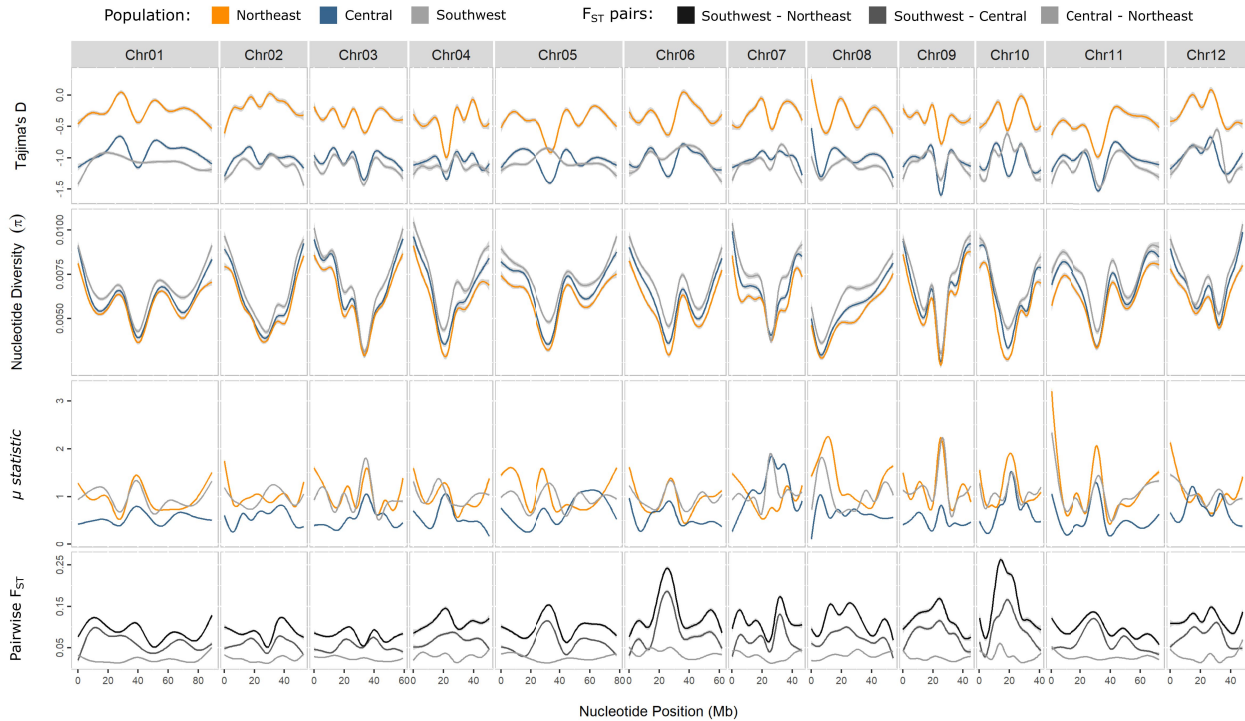


Figure 2.5: Genome-wide sliding-window analysis of the three American chestnut populations for Tajima's D, nucleotide diversity (π), RAI SD μ -statistic, and pairwise F_{ST} . Smoothed lines were computed for all results in *ggplot2* with the default gam parameter.

Consistent with the pattern for nucleotide diversity, heterozygosity was highest in the southwest (2.1). Northeast H_o was significantly lower than the central and the southwest population ($p < 0.001$, $p < 0.001$) (A.9a).

Table 2.1: Average Tajima’s D, nucleotide diversity, heterozygosity, and inbreeding coefficient across all samples for each American chestnut population.

Population	No. of trees	Tajima’s D	π	H_o	H_e	F
Southwest	68	-1.0827	0.0069	0.1515	0.1660	0.0882
Central	179	-1.0156	0.0064	0.1492	0.1660	0.1016
Northeast	109	-0.3345	0.0058	0.1353	0.1660	0.1851

2.3.7 Regions and associated genes under positive selection

As the chestnut populations migrated and diverged into new geographic regions, we hypothesized that genomic regions within each population would be subjected to recent positive selection. Every chromosome for each population contained outlier genomic regions identified by RAI_{SD} (2.5). The southwest population had the greatest number of significant regions with 11,733, followed by the northeast population with 8,832, and the central population with 7,534. Within significant outlier regions, the southwestern population contained 617 outlier genes, which was the most out of the three populations (2.6). Among these outlier genes, 402, 387, and 323 were unique to the southwestern, central, and northeastern populations, respectively, while 49 genes were shared between all three populations (2.6).

GO enrichment analysis revealed that ”response to stress” was within the top four most annotated gene families for the GO biological process for each population (A.10). Among the unique genes within each population, the southwest population had the most overrepresented GO terms, with 25 (A.1), while the northeast population had 10 (A.2), and the central population contained one (A.3). The GO term ”defense response to fungus” was significantly overrepresented in the shared gene set for all three populations (Fold enrichment = 16.5, $p =$

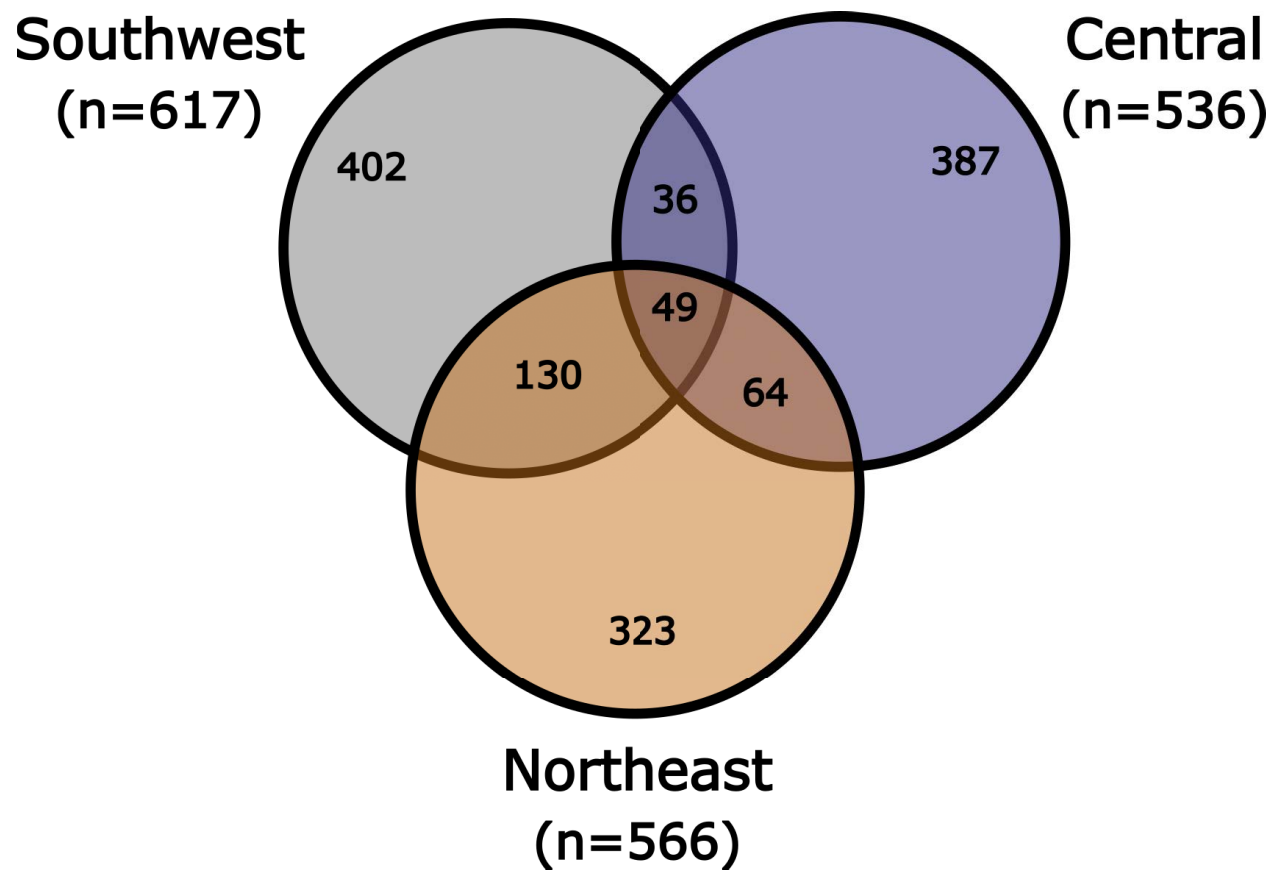


Figure 2.6: Number of outlier genes identified by RAiSD within each American chestnut population.

0.0373, [A.4](#)). The overrepresented genes for the shared gene set were Caden.02G006600 and Caden.03G057600, which are both in the apoptotic ATPase eukaryotic orthologous group (KOG).

2.4 Discussion

In this paper, our goal was to characterize population structure, postglacial history, and positive selection across the climatically diverse range of American chestnut [\[48\]](#). To do this, we re-sequenced the genomes of 384 wild chestnut genotypes. These data revealed that (i) the chestnut range can be divided into three management units, with east-west breaks roughly in Pennsylvania and Georgia; (ii) genomic diversity decreased with latitude, most likely the result of past population size reductions, and recent expansions, associated with Quaternary glaciation of North America; (iii) while the southwest cluster is the smallest both in area and census size, it holds a disproportionate level of genetic diversity; and (iv) genomic targets of natural selection are widespread throughout the genome and enriched for putative disease-resistance genes.

2.4.1 Hybridization is not widespread in putatively wild American chestnut genotypes

Our primary goal in developing a WGS dataset that included most *Castanea* species was to exclude hybrids between our wild *C. dentata* samples and congeners - a significant concern given the history of planting and naturalization of non-native *Castanea* species in the eastern U.S. [\[34\]](#). We did not detect widespread introgression with other *Castanea* species, which is consistent with previous assessments of North American *Castanea* [\[88, 113\]](#). Of the four *C.*

dentata samples in the species reference dataset that were sympatric with *Castanea pumila* var. *alabamensis*, there was little evidence of genetic ancestry with the *C. pumila* species, with only two of those *C. dentata* samples having any signatures of *C. pumila* ancestry ($<2\%$). However, as we focused on collecting from *C. dentata* genotypes with morphological characteristics representative of the species, our sample was biased against trees displaying intermediate phenotypes, and hybrids with naturalized non-native *Castanea* species, as well as the native *C. pumila*, may nevertheless be present in the wild. An additional result of the genus-wide ADMIXTURE analysis was that each of the three *C. pumila* subspecies had low levels of *C. dentata* ancestry, potentially due to residual relatedness from a common ancestor or ongoing introgression. This *C. dentata* ancestry varied among *C. pumila* subspecies, which suggests genomic differentiation among these subspecies and supports previous findings of meaningful genetic subdivisions within the *C. pumila* species complex [88, 113]. Finally, the inferred minor contribution of *C. crenata* to our Kentucky and Tennessee *C. dentata* genotypes seems unlikely to be due to recent introgression, as multiple generations of backcrossing would be required to generate the observed ancestry. More likely it is some combination of long branch attraction, incomplete lineage sorting, ancient introgression, and/or SNP calling errors [17]. A systematic exploration of the relationships among worldwide *Castanea* species is beyond the scope of this paper, and future studies will use these data to resolve the phylogenetic relationships among *Castanea* species and within the *Castanea pumila* species complex.

2.4.2 The American chestnut range can be divided into two or three populations

Previous studies based on microsatellite markers covering a smaller portion of the species distribution range (i.e., from North Carolina to Massachusetts) suggest either a single, ge-

netically diverse metapopulation [66] or a two-population model with Pennsylvania and Maryland being a transition zone [49, 80]. More recently, Spriggs and Fertakos [113] used ddRAD-seq data and found support for two populations - one in the southern portion of the range and the other comprising the central and north. Our results support a composite of this picture. While both DAPC and ADMIXTURE gave very similar cluster memberships and patterns at each K value, and both support a distinct southwestern population, they differed on whether the central and northern range was further divided. That our data support up to three sub-populations may be due to the increased genomic coverage of our study, our larger sample size and more uniform range-wide sampling, or some combination of these factors. More generally, our findings of increasing genomic divergence with increasing geographic distance are broadly similar to other temperate tree species [10, 56]. At the same time, overall among-population genomic divergence as measured by F_{ST} was relatively low, which is consistent with a multi-species assessment of northeastern North American tree species that found average F_{ST} across 20 tree species to be less than 10% [99].

While we find that isolation-by-distance is present in wild American chestnut, the three populations we identified appeared to be differentiated at latitudinal temperature and precipitation breaks across the species range. Pennsylvania marks the region of admixture between the central and northeastern populations, and is also where mean winter temperatures transition from above freezing to below freezing. In addition, the separation between the central and the southwestern populations occurs in Tennessee, which is the northern boundary for an area of the southern U.S. that receives ≈ 250 mm higher annual precipitation on average compared with eastern and northern portions of the historical chestnut range. Thus, like most temperate and boreal tree species, American chestnut exhibits local adaptation to climate. The three populations we identified may thus comprise broad chestnut ecoregions, reflecting both isolation-by-distance but also potentially isolation-by-adaptation [8, 42, 51]

as found in many temperate and boreal tree species.

2.4.3 Impact of Quaternary glaciation on extant patterns of genomic diversity

The Quaternary glaciation in North American was characterized by cooling and warming events with corresponding expansion and contraction of ice sheets beginning ≈ 2.7 Mya [18, 53]. Tree populations contracted in response to glaciation events [112, 134], and postglacial recolonization bottlenecks may have influenced genetic diversity and effective population sizes [6, 25, 55]. N_e for each of our populations declined in the distant past and subsequently each experienced repeated size changes before a rapid, recent expansion. Fluctuations in N_e for the northeastern population generally paralleled those of the central population, while the southwest population lagged by $\approx 100,000$ years, likely due to more muted impacts of climate change in this area. Curiously, following an initial increase after the last glacial maximum, N_e for the central population size again decreased. This decline may be due to a recent bottleneck event, or uncertainty in the SMC++ analysis, however, it was not due to the chestnut blight. Blight-related mortality of American chestnut populations occurred within the last century, and most of the trees we sampled are stump sprouts from root stock that predates the blight. With little generational turnover since the blight, the recent dramatic reduction of American chestnut census population size would not influence our demographic history analyses.

Among-population nucleotide diversity estimates support recolonization bottlenecks associated with the last glacial maximum. Diversity decreased as latitude increased, suggesting that the southwestern population has a higher long term effective population size, and was most likely a glacial refugium from which postglacial expansion occurred. Tajima's D was

negative in each population, which was driven by both rare and high frequency derived variants relative to neutral expectations [82]. The more negative values of Tajima's D for the southwest population, followed by the central and northeastern populations, suggest clinal variation in postglacial bottleneck timing. This same pattern of clinal variation in neutrality tests was found for Sitka spruce (*Picea sitchensis*), which recolonized its contemporary range from a southern glacial refugium [55].

Fossil pollen evidence indicates the Gulf coastal forests in Florida and Southern Alabama were a refuge for *C. dentata* during the last glacial maximum, and that the species subsequently migrated north into Tennessee ≈ 15 kya, continuing northeastward along the Appalachian Mountains at a rate of approximately 100m/year, eventually arriving in Connecticut $\approx 2,000$ ya [34, 35, 36, 37, 38, 125]. However, refugia may have existed further north, and the rate of migration may have been slower than what fossil pollen suggests [76]. As populations migrated northeastward and diverged, bottlenecks likely occurred, leading to less genomic diversity in the northern populations than those further south.

When we allowed for four clusters in the ADMIXTURE analysis, the central population separated on either side of the Appalachian Mountain range. The Appalachian Mountain range may thus serve as a barrier to gene flow throughout the American chestnut range, with migration paths northeastward on either side. The American chestnut's range is also limited by unsuitable heavy clay especially to the east of the Appalachian Mountains and extreme cold temperatures at the northern range limit [52, 100]. With four clusters, we observed a region of mixing between eastern and western clusters in southwest Virginia and southeast West Virginia, which the EEMS analysis showed was also an area of enhanced gene flow. A migration route across the Appalachian Mountain range may have existed in this region, likely due to it being an area of relatively lower elevation. This lower elevation region may have been easier to cross for the mammals and birds responsible for chestnut seed dispersal

[34]. Additionally, the presence of this migration route could explain the observation by Gailing and Nelson [49] that chestnuts in Ontario, Canada were more genetically similar to the North Carolina population than to their northeastern neighbors.

2.4.4 Global signatures of positive selection are enriched for disease resistance genes

While genomewide diversity and allele frequency spectra are largely driven by demographic factors, natural selection may have also played a role [123]. Because Tajima's D does not take the ancestral state of mutations into account, it is influenced by excesses of both high and low frequency alleles. The frequency spectrum among our populations suggest that negative Tajima's D in the southwest and central populations is caused more by low frequency alleles compared with the northeast, which had comparatively fewer low frequency alleles and more high frequency alleles (Figure S4). While this pattern is consistent with variable bottleneck age, it may also reflect recent hitchhiking associated with adaptation to the colder environment in the northeast.

The ANGSD sliding window analyses revealed several broad genomic regions of negative Tajima's D and reduced nucleotide diversity. Some of these regions likely reflect reduced recombination and associated decreased nucleotide diversity near centromeres. However, several intervals, such as on chromosome six and chromosome ten, also had elevated F_{ST} for the southwest-central and southwest-northeast population pairs, suggesting that these regions may have experienced recent selection related to environmental adaptation to more northern climates. RAI SD identified several thousand outlier regions that may be targets of selection, and were enriched for genes related to "response to stress" and "response to chemical". Further, GO enrichment analysis identified "defense response to fungus" as the

only overrepresented biological pathway for the set of shared genes between all populations. This suggests that selection related to pathogen pressure is a key feature of global adaptation in American chestnut.

A few of the American chestnuts we sampled were large surviving trees that may have higher levels of blight resistance relative to other American chestnut trees. However, most of our samples were blight killed resprouts. In addition to blight, American chestnuts from across the species range are highly susceptible to *Phytophthora cinnamomi* (the Oomycete responsible for *Phytophthora* root rot) - a contemporary agent of American chestnut decline [29, 127]. To date, the only known source of *P. cinnamomi* resistance for *C. dentata* has been introgression from Asian *Castanea* species, such as *C. mollissima* and *C. crenata* [87, 127, 133]. Thus, the enrichment for defense genes among selection targets we observed is unlikely to be related to these two contemporary pathogens that threaten the species. Rather, it likely reflects historical interactions with, and adaptations to, native fungal pathogens.

This is not to say, however, that the "defense response to fungus" selection targets could have no role in responding to chestnut blight or *Phytophthora* root rot. In response to blight inoculation, both resistant *C. mollissima* and susceptible *C. dentata* show transcriptional responses, though more genes potentially related to blight resistance were upregulated in *C. mollissima* [12]. The two genes associated with "defense response to fungus" were both members of the apoptotic ATPase KOG group. This is interesting considering Barakat et al. [12] found ATPases to be differentially expressed in chestnut blight canker tissues. In addition, blight resistance in American chestnut backcross populations is polygenic, with multiple loci contributing to resistance [129]. The lack of blight resistance in American chestnut may be due to an inadequate or inappropriate transcriptional response to infection. As such, further evaluation of genomic targets of selection related to pathogen defense will be necessary to determine the cause of their overrepresentation and their possible relationship

to blight resistance in wild populations.

2.4.5 Implications for restoration of American chestnut

Our long-term goal is to develop disease resistant, locally adapted American chestnut populations for restoration of the species. While millions of American chestnut collar sprouts still persist throughout the botanical range of *C. dentata*, the ecological niches occupied by chestnut have been eroded since the blight was introduced, and increasing deer populations are a major threat to survival of remaining sprout clumps [21, 22]. There is therefore an urgent need to conserve the evolutionary legacy that these remnant chestnut populations represent. Results of this study, and future work, will therefore be applied to apportion sampling of wild germplasm from across the historical chestnut range on the basis of management and adaptive units, by dividing the range into provenances that are homogeneous with respect to their genome-wide and adaptive allelic content, respectively. Here, we primarily addressed the first of these goals, which revealed that the chestnut range can be divided into three populations separated along a latitudinal gradient. We also found that while geographically smaller, the southwest population had the highest levels of genetic diversity. Thus, we expect to propagate a disproportionate number of genotypes from this region relative to the proportion of the species range it encompasses. The genomic targets of natural selection we identified, particularly the over-representation of disease-related loci, will be incorporated into models that define the adaptive units, in addition to information from future analysis of genotype-environment relationships.

Of the two methods being used to develop blight-resistant chestnuts, the transgenic lines are most in need of genetic diversification, having originated from only two founder trees. Previous simulations indicated that N_e for the transgenic lines can be increased to > 500

individuals, and the average inbreeding coefficient reduced to < 0.01 , by outcrossing these founders over three generations with 50, 150, and 450 (650 total) wild-type American parents, and using three progeny per cross as parents in subsequent generations [128]. Methods for propagating wild trees will include grafting wild scion and collecting rare wild seed. Pollen collection will be accelerated using high light treatment, which has been shown to result in flowering in the first growing season [128]. The American Chestnut Foundation's backcross breeding program contains much more genetic diversity than the transgenic lines, having incorporated pollen from many wild trees. These backcross trees may be crossed with transgenic trees to both combine quantitative and transgene resistance, and to further diversify the transgenic populations. Our results will be used to evaluate the proportion of adaptive variation that has been captured in these backcross families, and the extent to which further introgression of wild germplasm should be undertaken to improve their climatic resilience.

Chapter 3

Identifying and conserving the adaptive legacy of American chestnut

3.1 Introduction

American chestnut (*Castanea dentata*) is one of the tree species most vulnerable to climate change in the United States [92]. A large deciduous tree native to the Appalachian Mountains of eastern North America, the American chestnut was decimated by a fungal blight beginning in the early 20th century that killed billions of trees. While approximately 400 million chestnut stems remain throughout the range, these are almost exclusively root collar sprouts that rarely achieve more than a few years of growth before being reinfected with blight [86, 109]. This repetitive cycle of growth and die-back and the lack of widespread reproduction makes American chestnut functionally extinct in the wild [86].

Efforts to develop blight resistant American chestnut through introgression from Chinese chestnut (*Castanea mollissima*) [115] and insertion of an oxalate oxidase (OxO) transgene may soon enable widespread restoration plantings [81, 93, 132]. However, particularly in the case of the genetically-modified lines, a lack of genetic diversity in general, and adaptive diversity in particular, necessitates an outcrossing strategy to both increase effective population size and target specific families to the climate of local planting sites. And while the hybridization and backcrossing program has incorporated some standing variation via polli-

nation from wild trees, the extent of adaptive variation segregating in backcross populations is unknown.

Adaptation to local climate in temperate forest trees is determined by multiple abiotic constraints, including the correct timing of growth and dormancy transitions and associated cold acclimation, maximum winter cold hardiness, water-use efficiency, and drought tolerance or avoidance [57, 105]. Similar to many temperate tree species, American chestnut has migrated from glacial refugia to occupy extant ranges spanning thousands of kilometers from southern to northern limits [104]. Differentiation in locally-adaptive traits related to temperature variation has been documented in common gardens [101, 102, 107], which suggest adaptive clines similar to other widespread temperate tree species. However, because of the rarity of wild reproduction, it is exceptionally difficult or impossible to assemble chestnut common gardens that encompass locally dense and geographically widespread sampling of the historical range. Hence, identifying the genomic targets of climate-related selection is a key step in understanding the underlying mechanism of local adaptation in this species, as well as for developing strategies for capturing and incorporating the adaptive legacy of the species. One way to approach this objective is through the concept of seed zones, whereby a species range is divided into the smallest number of regions that effectively describe how climate and ecoregions are partitioned. Ideally these zones also reflect reaction norms from reciprocal common gardens, although such data is usually only available for species of high economic value. Existing climate and ecological data suggest >10 different seed zones within the historical American chestnut range [19, 90], but these are not specific to chestnut, and may or may not reflect how adaptive traits and genomic diversity is arrayed across the landscape for this species. Similarly, these zones do not inform the sampling intensity required per zone to capture standing adaptive diversity.

Here, we describe the genomic basis of local adaptation in American chestnuts and make

sampling recommendations to inform germplasm conservation efforts. We leverage a ≈ 21 million SNP whole-genome sequencing (WGS) dataset developed by Sandercock et al. [104] to (i) identify loci significantly associated with climate; (ii) define seed zones for germplasm conservation; and (iii) develop a novel method for making sampling recommendations to capture wild adaptive diversity. These results will inform American chestnut restoration efforts, and provide a blueprint for similar conservation efforts in the future.

3.2 Methods

3.2.1 Dataset preparation

We used the 21,136,994 filtered SNPs and 356 *C. dentata* individuals that were previously reported by Sandercock et al. [104] (3.1). Briefly, 384 leaf samples were obtained from throughout the American chestnut natural range and then their DNA was sequenced using whole-genome sequencing (WGS) at HudsonAlpha Institute for Biotechnology. After filtering the 384 individuals based on missing data and ancestry tests, 356 *C. dentata* samples were included in the final dataset. Missing sites were imputed using BEAGLE v5.4 [20] and a minor-allele frequency (MAF) filter was applied using PLINK v1.9 to remove SNPs with $MAF < 0.05$ [94]. The final genomic dataset contained 11,526,713 SNPs and 356 samples.

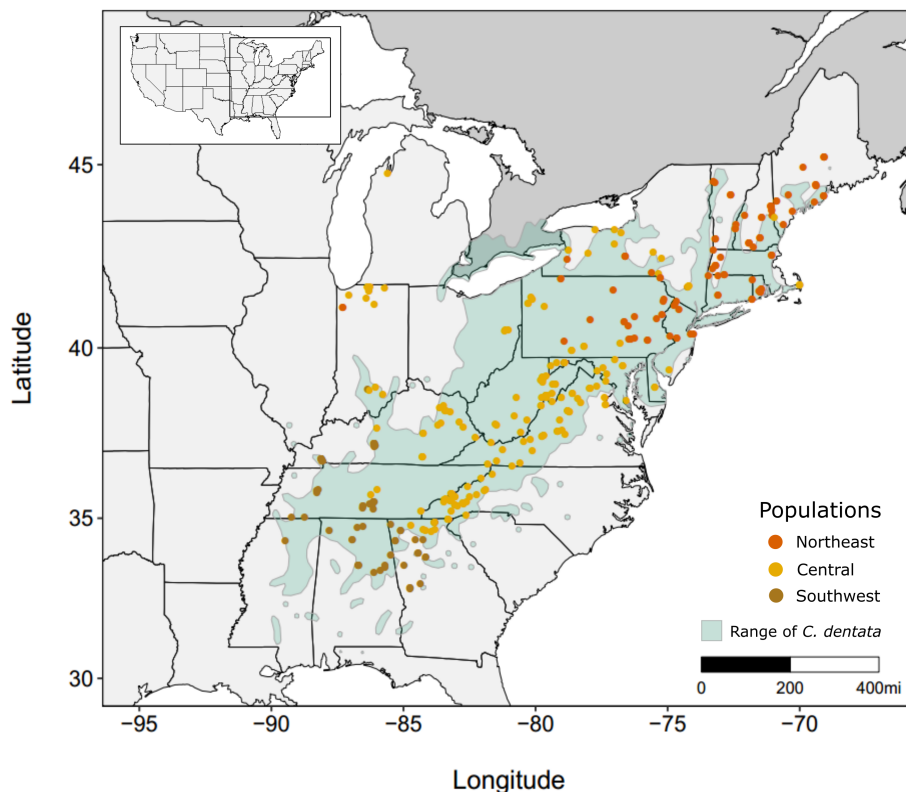


Figure 3.1: Range map of *Castanea dentata* and the locations of the 356 American chestnut samples and their assigned population from Sandercock et al. [104].

Environmental data was obtained from climateNA [124] and comprised 32 monthly and seasonal climate variables (B.1). The environmental variables were the average for 1961-1990 and the ensemble projections under the CMIP6 future climate scenarios for 2050 and 2080 [2, 74]. Climate data was not available for three of the samples, so they were removed from the analysis. Collinearity between climate variables was assessed to exclude highly correlated variables that could increase false-positive rates in genotype-environment relationships [24]. We first used gradientForest to rank the importance of each of the climate variables using a random subset of 50k SNPs [43, 44]. Pearson’s correlations between each pair of climate variables were estimated using the `pairs.panel()` function from the *psych* R package [98]. Climate variables were removed if they were highly correlated ($R^2 > 0.7$) with another variable

of higher importance from the gradientForest predictor importance results.

3.2.2 Identifying climate-associated adaptive loci in American chest-nut populations

Univariate and multivariate genotype-environment association (GEA) tests were performed to identify putatively adaptive loci that were associated with climate. Redundancy analysis (RDA) is a multivariate genotype-environment association method (GEA) that accounts for all environmental variation simultaneously. RDA has been shown to have higher true positive and lower false positive rates than univariate and several other multivariate approaches [46], and enables assessment of the contributions of multiple factors (climate, geography, and population structure) to patterns of genetic variation. For the RDA analysis, we used the vegan R package [85]. First, we used a partial RDA (pRDA) to evaluate the contributions of each of climate, geography, and population structure to rangewide diversity. Specifically, we performed four tests: a full model with all factors, a climate only model, a geography only model, and a population structure only model [24], and summarized the influence of each factor on genetic variation with an ANOVA in R. For the pRDA analyses, we used a linkage-disequilibrium (LD) pruned dataset due to computational efficiency. An LD-prune was applied to the ≈ 11.5 million SNP dataset using PLINK v1.9 to include SNPs with R^2 values < 0.5 within 50 SNP sliding windows (step size 10 SNPs) – resulting in a reduced dataset of 2,934,239 SNPs [94]. We then performed a GEA scan of the ≈ 11.5 k SNP dataset and the 10 uncorrelated climate variables following Capblancq and Forester [24]. As correcting for population structure can cause the RDA outlier tests to be overly conservative, we report only the uncorrected model [46, 70]. We retained three RDA axes, and the top 0.1% of p-values were identified as candidate locally adaptive sites.

As a complementary method to detect genotype-environment relationships, we used latent factor mixed models (LFMM) as implemented in the R package LEA [27, 47]. LFMM2 uses univariate regression to identify linear associations between loci and climatic variables, while accounting for confounding factors such as background allelic covariance arising from population structure [27]. We first summarized the 10 climatic variables with PCA, retaining the first three PC axes, which explained $\approx 89\%$ of variance within the climate variables dataset [85]. The `lfmm2()` function in the LEA package in R was used to run LFMM2 three times (once for each environmental PC axis) with $K=2$ latent factors to correct for population structure [47]. P-values, z-scores, and the genomic-inflation factors (GIF) were estimated using the `lfmm2.test()` function with default parameters [47]. To account for multiple testing, p-values were converted to q-values using the `qvalue()` function in R [117]. Loci that had a q-value < 0.1 ($FDR < 10\%$) were retained as candidate adaptive loci. Finally, outlier loci identified by the RDA and the three LFMM runs were combined for subsequent analyses.

To assess whether the putatively adaptive loci were enriched for coding or noncoding regions of the genome, the locations of the outlier SNPs were cross-referenced with the American chestnut genome feature file (Cdentata_673_v1.1.gene_exons.gff3.gz; <http://phytozome-next.jgi.doe.gov/>). Introns were not included in the GFF file, and were added using the ATAG biopython package [31]. Promoters were also added to the GFF file as the 2 kb region upstream of the mRNA intervals.

3.2.3 Spatial distribution of adaptive genomic diversity and genomic offset

To characterize the spatial distribution of adaptive genomic variation throughout the chestnut range, we used the Gradient Forest (GF) machine learning algorithm [44]. GF estimates

nonlinear turnover functions of allele frequencies across environmental gradients. Uniquely, the shape of the nonlinear turnover functions estimated by GF enable rescaling of the environmental gradients to reflect the specific ways in which the multidimensional genomic variation is constrained by the climate space. To do this, we first estimated a gradientForest model using the 18,483 adaptive SNPs from the LFMM and RDA analyses as response, and the climate information from the 10 uncorrelated climate variables for each sampling location as the predictors. This trained model was then used to predict the spatial distribution of genomic variation over the entire chestnut range. A PCA was then performed on the GF model prediction and the top three principal components (PCs) were used to visualize the multidimensional adaptive genomic variation across the American chestnut range. To understand how climate change may impact the relationship between genome and environment, we performed the same procedure outlined above using two future climate predictions for 2080 (RCP4.5 and RCP8.5). We then calculated the Euclidean distance between the current climate GF model and the future climate model, which reflects ‘genomic offset’ – an estimate of future maladaptiveness due to climate change.

3.2.4 Identifying seed zones within the American chestnut natural range

A major objective of this research was to partition the chestnut range into geographic seed zones that reflect relatively homogeneous areas with respect to multivariate adaptive genomic variation, which can subsequently be used to both conserve germplasm *ex situ*, and guide restoration plantings of blight-resistant chestnut families/lines with similar genomic compositions. To do this, the PCA loadings from the GF model above were clustered using the `Optimal_Clusters_KMeans()` function from the `ClusterR` package in R to determine the optimal number of seed zones [79][131]. Between K=1 and K=12 clusters were evaluated

for fit to the data using AIC, sum of within-cluster-sum-of-squares-of-all-clusters (WCSSE), variance explained, and BIC. We then used the `clara()` function from the `ClusterR` package with $K = 3$ [73]; the resulting clusters are the seed zones – representing a genetically similar region relative to its adaptive variation.

3.2.5 Estimating the number of trees to sample to capture wild adaptive diversity for germplasm conservation

To estimate the number of trees required to sample from each seed zone to recapitulate the multivariate allele frequencies at adaptive loci and inform germplasm conservation, we created a custom Python script. The Python script takes the input of a VCF file and optionally a list of SNP and sample information, and performs bootstrap re-sampling of the dataset to compare allele frequencies (AF) for all SNPs from the bootstrap samples with those of the full population using linear regression. Bootstrap sampling of the full population continues until the R^2 score from the linear regression meets or exceeds the user-defined R^2 score (a proxy for the percent diversity to capture), then the bootstrap sampling ends and the number of subsampled trees is recorded. The process is then repeated for a predefined number of iterations (in our case, 100), and the number of trees to sample is averaged across all iterations. The final output is the mean number of trees to sample from a population to meet the desired coefficient of determination between the adaptive multivariate AFs of the sample and that of the full dataset for each seed zone.

3.3 Results

3.3.1 Drivers of genomic variation across the range of American chestnut

We previously showed that the American chestnut range can be divided into three populations with longitudinal boundaries that roughly divide the species into southern, central, and northern groups [104]. As these boundaries also track temperature and precipitation gradients, we predicted that population structure and climate would have the largest influence on genomic variation. The partial RDA models show that climate had the largest proportion of explainable variance (0.38), followed by population structure (0.13), and geography (0.08), and that all pRDA models were significant ($p < 0.001$, 3.1). The full pRDA model accounted for 11% of the total variance.

Table 3.1: Drivers of genomic variation in American chestnut. Four partial RDA models were performed to evaluate the influence that climate (10 environmental variables), population structure (first two PC axes from PCA on genomic dataset), and geography (geographic coordinates of each sample) had on the patterns of genomic variation throughout the American chestnut natural range. An ANOVA was performed after each partial RDA to estimate statistical significance.

Partial RDA models	Inertia	Adj. R ²	p (>F)	Proportion of explainable variance	Proportion of total variance
Full model	79190	0.063	0.001***	1	0.11
Pure climate	30117	0.012	0.001***	0.38	0.04
Pure structure	10018	0.008	0.001***	0.13	0.01
Pure geography	6368	0.003	0.001***	0.08	0.01
Confounded climate/structure/geography	32687			0.41	0.05
Total unexplained	632195				0.89
Total inertia	711385				1.00

*** $p \leq 0.001$

We identified 10 uncorrelated climate variables to test for genotype-environment associations:

RH (mean annual relative humidity (%)), **Eref** (Hargreaves reference evaporation (mm)), **PPT_wt** (winter precipitation (mm)), **EXT** (extreme maximum temp. over 30 years ($^{\circ}\text{C}$)), **TD** (temp. difference between mean warmest month and mean coldest month ($^{\circ}\text{C}$)), **MAR** (mean annual solar radiation ($\text{MJ m}^{-2} \text{ d}^{-1}$)), **PPT_sm** (summer precipitation (mm)), **CMI** (Hogg's climate moisture index (mm)), **AHM** (annual heat-moisture index (mean annual temp.+10)/(mean annual precip./1000)), and **DD_18** (degree-days below 18°C) (3.2a, B.1). Since LFMM is a univariate method, the LFMM analyses utilized three PC axes as synthetic predictors in place of the 10 environmental variables due to computational demand and to reduce the number of corrections needed to account for multiple testing. For LFMM, we identified 4,988 SNPs associated with PC1(MAR, PPT_wt, TD), 2,923 SNPs associated with PC2(AHM, CMI, EXT), and 0 SNPs associated with PC3(RH) (B.2). In total, LFMM identified 7,911 outlier SNPs ($\text{FDR} < 0.1$) and the RDA analysis revealed 11,526 outlier SNPs (0.1% outliers), with 954 of these SNPs shared between the two GEA methods (B.5), for a combined total of 18,483 putatively adaptive SNPs.

Among climate variables, allelic turnover was relatively linear for temperature differentials (TD), annual heat:moisture indices (AHM), summer precipitation (PPT_sm), Hogg's climate:moisture indices (CMI), and mean annual relative humidity (RH) (3.2b). By contrast, sharp turnovers occurred at evaporative demand (Eref) between 1150 and 1200 mm, extreme temperatures (EXT) between 39 and 40 $^{\circ}\text{C}$, cooling degree-days (DD_18) between 2000 and 2500 days, and winter precipitation (PPT_wt) between 300 and 350 mm (3.2b). The remaining climate variable, mean annual solar radiation (MAR), showed a slight increase in allele turnover rate between 14 and 15 ($\text{MJ m}^{-2} \text{ d}^{-1}$) (3.2b).

Throughout the genome, 210 genes contained at least one adaptive site, with most of the 'adaptive' genes located on chromosome one, five, and 10 (B.3). Approximately 18.6% of the putatively adaptive SNPs were located in promoter and gene regions, which was lower

than in the random dataset ($\approx 28.6\%$), and the initial filtered dataset ($\approx 28.3\%$) (B.3).

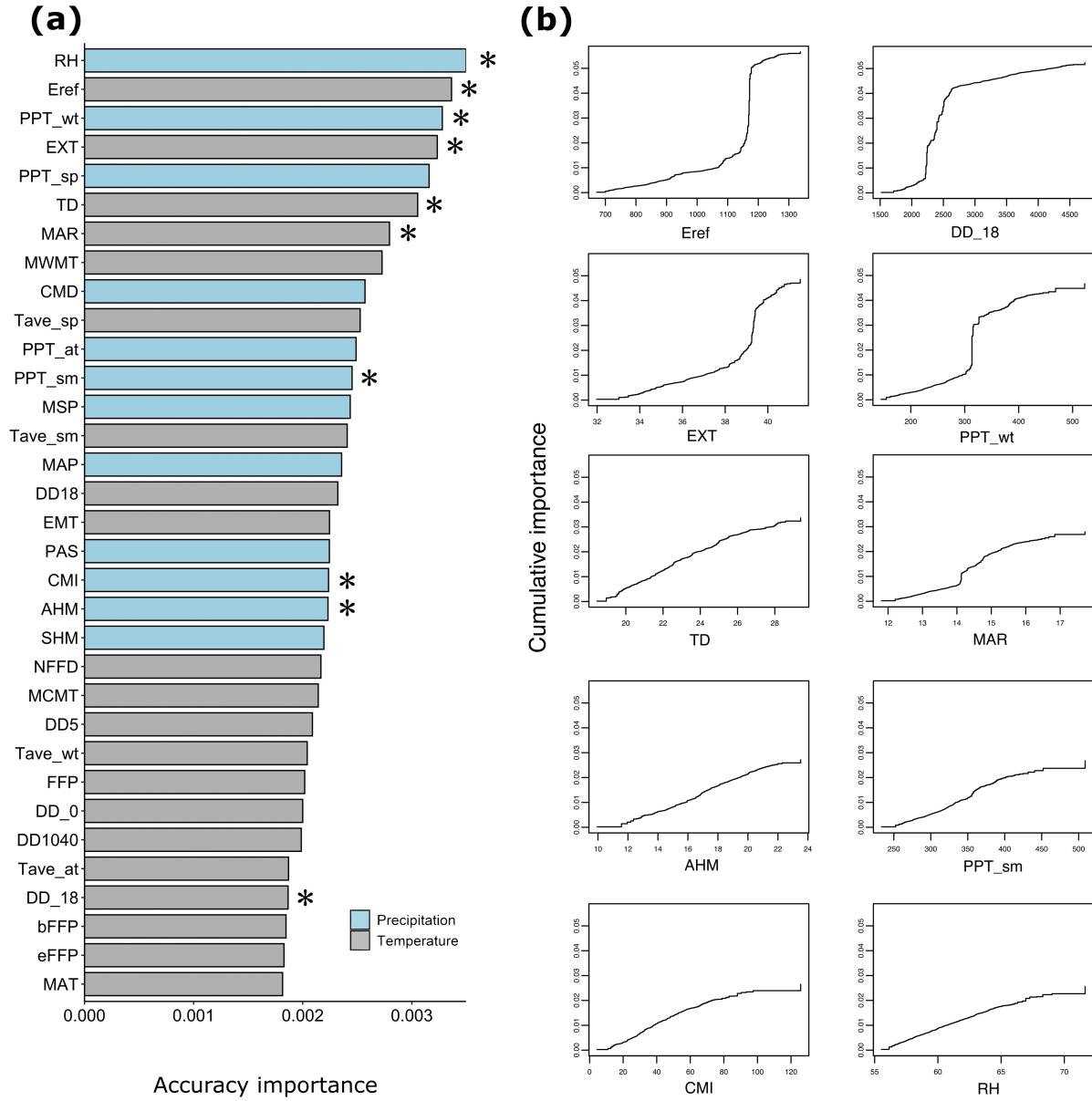


Figure 3.2: Importance of climatic variables in patterns of allelic turnover for wild American chestnut populations. A) Ranked accuracy importance of 33 climateNA variables from a gradientForest analysis that included 353 American chestnut samples and 50k SNPs. The variables denoted with '*' were used for the analyses. B) Predictor cumulative importance plot. Cumulative change in allele composition across the 10 climate gradients and the 18,483 putatively adaptive loci. Sharp increases suggest a point in the climate gradient where allelic composition changes.

3.3.2 Geographic distribution of genomic variation and maladaptation under current and future climates

Our genotype-environment analysis suggests differences in the climatic drivers of adaptive genomic variation between northern and southern portions of the American chestnut range. The northern portion of the range was associated with temperature and precipitation (TD, DD_18, AHM), while the southern portion of the range was associated with evaporative demand and extreme temperatures (Eref and EXT) (3.3a,b). An additional area in western North Carolina and central portions of the Appalachian Mountains was associated with precipitation and solar radiation (CMI, MAR, PPT_sm, and PPT_wt) (3.3a,b). To understand how climate change will impact the match between current and future adaptive matches, we estimated genomic offset for 2080 for both middle and high emissions climate scenarios (RCP4.5 and RCP8.5, respectively). Under both climate change scenarios, the central portions of the American chestnut range were predicted to experience elevated maladaptation, effects that were exaggerated for the high emissions scenario (3.3c).

3.3.3 Identifying seed zones

As climate was the greatest driver of genomic variation in American chestnut (3.1), we hypothesized that we could partition the range into seed zones with boundaries reflecting precipitation and temperature gradients in the eastern US. GradientForest model selection from K=1-12 climate-based seed zones suggested that two or three were the most likely number of zones (B.6,B.7,B.8,B.9). With three such zones, the range was split into north, central, and south regions, with latitudinal boundaries at approximately 36N and 39N (3.4a,b). Genetic differentiation between seed zones was approximately 5.5-6.5 fold higher at adaptive loci compared with neutral loci (3.4c).

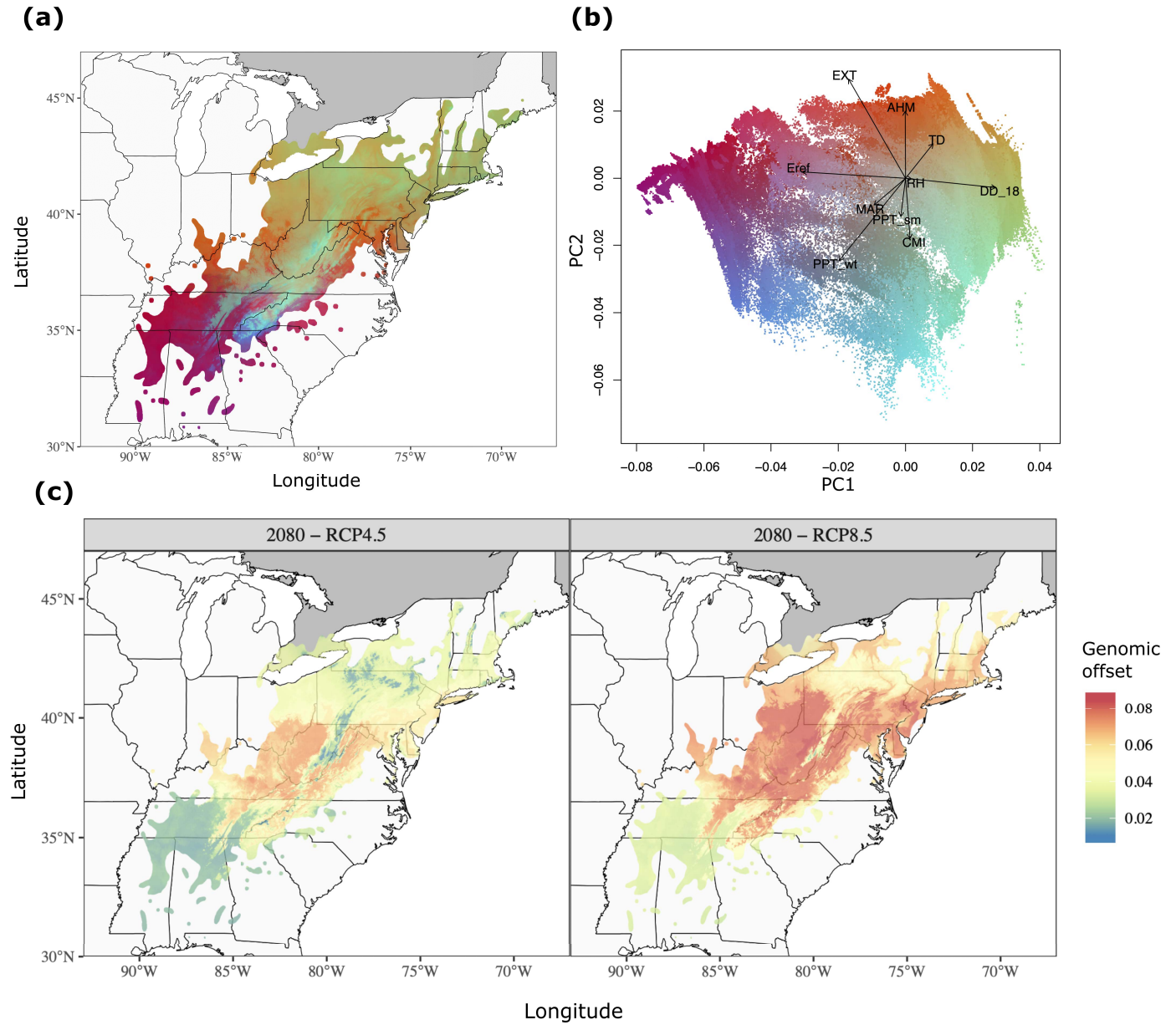


Figure 3.3: Adaptive genomic diversity under past and future climate conditions. (a) PC loadings from the gradientForest model of 10 uncorrelated climate variables and 18,483 putatively adaptive loci. (b) Biplot of loadings for PC1 and PC2. (c) Genomic offset under 2080 climate projections for moderate emissions (RCP4.5; left) and severe emissions (RCP8.5; right) scenarios.

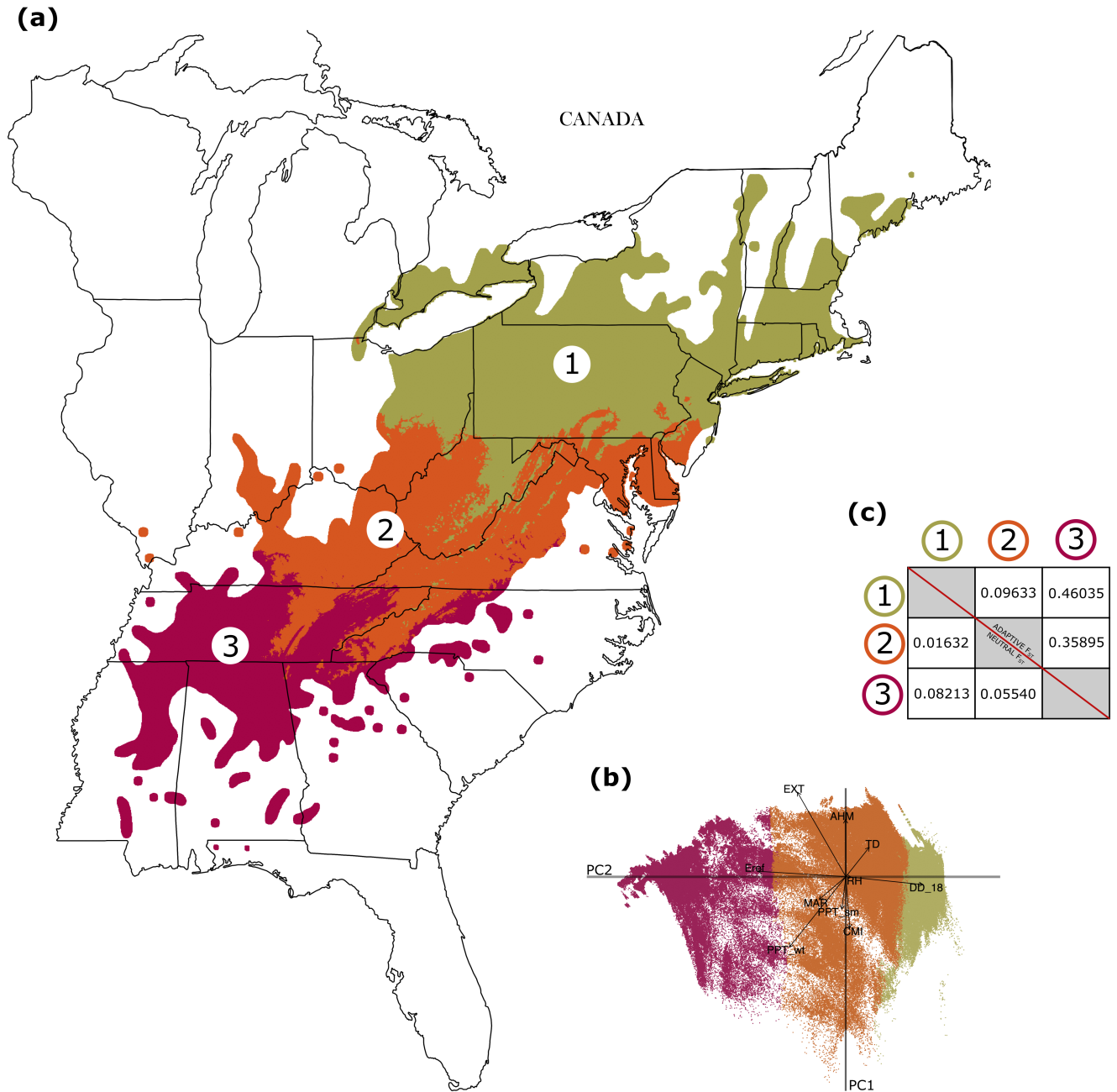


Figure 3.4: Three seed zones for American chestnut conservation that are based on the GF-predicted genomic variation. (a) Spatial distribution of the three seed zones. (b) PCA biplot of the GF-predicted genomic variation using three clusters to represent the seed zones. (c) F_{ST} values between each seed zone at neutral loci (left) and putatively adaptive loci (right).

3.3.4 Sampling to capture wild adaptive variation

Previous estimates of genomic diversity in American chestnut show that the southern portion of the range had the highest levels of heterozygosity, which decreased as latitude increased [104]. As such, we expected that sampling intensity required to capture most adaptive diversity would be lower in the more southern Zone 3, and would increase with each successive latitudinal zone. Consistent with this, we found that fewer trees will need to be sampled from Zone 3, and more trees will need to be obtained from Zone 1 for the 90% and 95% diversity matching targets (3.2). While the fewest trees will need to be sampled from Zone 3 for the 99% diversity matching target, most trees will instead need to be sampled from Zone 2. Estimates for the central Zone 2 were closer to the Zone 1 estimates across all diversity matching targets (3.2). An approximately 10-fold difference in sampling intensity will be required to match allele frequencies at 99% compared with 90% coefficients of determination (3.2). Using a three seed zone model will require more trees to be sampled overall (90%=38.15 trees, 95% = 76.75 trees, 99% = 370.35 trees) compared with the two seed zone model (90%=27.78 trees, 95% = 56.03 trees, 99% = 274.93 trees) or the one seed zone model (90%=26.86 trees, 95% = 55.89 trees, 99% = DNF) (3.2,B.4,B.5).

Table 3.2: Tree sampling estimates to capture adaptive diversity from each American chestnut seed zone.

Seed zone	% variance explained	# of trees to sample	95% CI
1	90%	14.50	(13.35, 15.65)
2	90%	13.42	(12.27, 14.57)
3	90%	10.23	(9.41, 11.05)
1	95%	28.30	(26.33, 30.27)
2	95%	27.74	(25.63, 29.85)
3	95%	20.71	(19.32, 22.10)
1	99%	132.33	(124.61, 140.05)
2	99%	136.34	(127.60, 145.08)
3	99%	101.68	(94.76, 108.60)

3.4 Discussion

The goals of this study were to describe the genomic basis of local adaptation in American chestnuts, define seed zones for germplasm conservation, and develop a quantitative framework for sampling recommendations. To do this, we used the SNP dataset generated by Sandercock et al. [104] to perform genome-wide scans for signatures of adaptation, and to model the partitioning of adaptive variation across the landscape. These analyses revealed polygenic adaptation to climate that can be subdivided into three seed zones from which a relatively small number of trees can be sampled to capture 95% of adaptive diversity.

3.4.1 Climate shapes genomic variation in American chestnut

Like many widespread temperate tree species, climate has significantly shaped genomic variation in American chestnut. Although well connected by gene flow, this species exhibits isolation-by-distance across a natural range with highly variable abiotic environments, in particular precipitation and temperature gradients. The geography of population structure across this range was recently shown to track transitions to mean below freezing winter temperatures in the north, and to a 25% higher annual precipitation in the south [104]. Among ≈ 11.5 million SNPs tested, we identified ≈ 18.5 k associated with climate between RDA and LFMM. The importance of our tested climate variables suggests precipitation, its seasonality, and its interaction with temperature in the summer are primary drivers of adaptation in chestnut, comprising the top three and 11 of the top 15 climate variables in the gradient-Forest analysis. This contrasts with, for example, lodgepole pine and red spruce, for which variables related to winter temperature are most important to adaptation [26, 131]. Many of the most studied forest tree species occupy ranges that begin and end further north than that of American chestnut. This dichotomy may reflect a different cold hardiness strategy

for chestnut that is insufficient in areas with extremely cold winter temperatures, as well as adaptations to the high summer evaporative demand in the southern portion of its range.

A limitation of both these approaches is that they assume linear relationships between climate and allele frequencies[24, 70]. Methods to identify nonlinear relationships between genotypes and environment have been developed, such as gradientForest [44], but are not computationally efficient as screening tools for large genomic datasets. Modeling such nonlinear genotype-environment interactions may be of interest in chestnut and other species, particularly given the nonlinear breaks in temperature and precipitation at population boundaries estimated with genome-wide data, and the correspondence of these breaks to seed zones estimated with our adaptive loci.

3.4.2 Adaptive units for germplasm conservation and reforestation

Chestnut restoration efforts will need to account for local adaptation by matching blight resistant families to climatic niches in which they exhibit appropriate timing of growth and dormancy, water use efficiency, drought tolerance, and more. Characterization of the pattern and extent of such local adaptations in forest trees is typically achieved through reciprocal common gardens. As these are not possible for American chestnut, we sought to use correlations between multivariate genotypes to the climate space in which they are found as a proxy for adaptive traits. In both scenarios, the concept of seed zones has been an essential tool for land managers to identify unique regions of climate adaption and to inform candidate areas for reintroduction of seed, often from orchards of known provenance. For chestnut, this matching is further complicated by the hybrid or transgenic nature of the planting stock, which makes provenance a more nebulous concept even if phenotypic local adaptation was well described. Using both genomic and climate information, we found that

the American chestnut range can be most parsimoniously divided into three seed zones, which best represents the American chestnut adaptive boundaries. These zones roughly match background population structure in the species [104], which suggests our putative adaptive cohort of SNPs may be substantially impacted by false positives that simply recapitulate populations structure. We view this as unlikely for three reasons. First, we applied very strict outlier cutoffs, which if anything, biased the dataset toward the exclusion of true positives rather than inclusion of false positives. While the number of adaptive loci in our dataset was very large, it is consistent with the polygenic nature of the traits that usually comprise local adaptation in temperate trees species [105, 106]. Second, F_{ST} for the adaptive SNP cohort was 5.5-6.5-fold higher than the genome-wide background across the seed zones, suggesting divergent selection. Finally, while the three seed zones roughly track background population structure, there were important differences. For example, seed zone one extended southward through high elevation areas of the state of West Virginia, which is not reflected in the neutral population structure map, and there was an extension of Seed Zone 2 southward and Seed Zone 3 northward along the border of Tennessee and North Carolina, which reflects a demarcation between the Great Smokey Mountains in the east (Seed Zone 2) and the much warmer Tennessee Valley to the west (Seed Zone 3).

3.4.3 Matching adaptive portfolios to future climates

American chestnut has clearly adapted to regional precipitation and temperature regimes, and while the above seed zones provide a spatial delineation of how genotypes match geography, these estimates reflect historical climate. Under climate change, the coupling of climate and geography in the context of our modeling will break down, and matching genotypes to planting sites will require reconciling current climates that interact with seedling fitness, and future climates that impinge on competitive ability, reproduction, recruitment,

and adaptation of subsequent generations. Our results of areas that are adapted to drier conditions will be expected to have greater maladaptation due to future climate is consistent with previous assessments ([84]). Though, these previous assessments did not include genomic data when estimating the impact of climate change on American chestnut distributions. Nevertheless, genomic offset is predicted to be higher for all areas of the range under the more severe RCP8.5 climate predictions – which is consistent with recent evaluations of American chestnut habitat suitability under climate change conditions [3].

Climate change will shift the habitable range northward [84], thus areas farther north should also be considered for reintroduction. American chestnuts are a slow migrating species – traveling approximately 100m per year after the Last Glacial Maximum [34]. As such, assisted migration will be necessary for maladapted regions [5]. Chestnuts planted northward outside of the parent range will suffer winter injury, but perform better than other assisted migration species from the eastern United States [28]. Germplasm from the central and northern seed zones (Seed Zones 2 and 1), exhibits less winter injury and should be considered for use in the most northern replanting locations [107]. We recommend additional evaluations of the three seed zones identified here in common garden and controlled environment experiments to determine how germplasm from each of these seed zones will respond to assisted migration and future climate conditions. The additional information will ensure that locally adapted American chestnut trees will be replanted in regions where they are adapted to thrive in.

3.4.4 Sampling estimates for germplasm conservation of wild adaptive genomic diversity

Ultimately, we sought to estimate the number of trees to sample from each seed zone to capture most of the adaptive diversity present in the wild. These sampled trees will be

used for germplasm conservation, and to introgress adaptive variation into disease-resistant populations. To make these sampling recommendations, we developed a novel computational method since an existing tool/pipeline was not available. Consistent across the three diversity percentages, 90%, 95%, and 99%, we found that the fewest amount of trees will need to be sampled from the southern Seed Zone 3, and the most from the northern Seed Zone 1. These findings were expected considering that genomic diversity is highest in the southern chestnut range and decreases as latitude increases [49, 80, 104, 113]. Thus, fewer trees would need to be sampled from regions of higher heterozygosity and nucleotide diversity to capture the desired percentage of adaptive diversity.

3.4.5 Conclusions

The American chestnut is a functionally extinct species facing both biotic and abiotic threats. Developing disease-resistant chestnut populations is an important first step, but these populations require sufficient adaptive variation to survive and thrive in their replanting locations. However, information regarding the adaptive genomic landscape of the American chestnut was lacking and methods were not available to estimate the number of trees necessary to sample from wild populations to inform germplasm and breeding plans. Therefore, the seed zones identified here, which represent areas of homogeneity relative to the chestnut adaptive genotypes, will be used as sources of germplasm and regions for reintroduction. We found that 18,483 SNPs were significantly associated with climate, that three seed zones best separates the chestnut range into unique regions of adaptive diversity, and that fewer trees will need to be sampled from the southern seed zones to capture the same proportion of adaptive diversity as the more northern seed zones – likely due to the more heterozygosity present in the southern portion of the chestnut range. Future work will need to estimate genomic diversity already present in the backcross breeding program, and to develop a breeding strategy

to introgress additional adaptive diversity from the subsampled individuals into the back-cross program. Additionally, evaluations should be done to estimate candidate regions for reintroduction of these locally adapted, disease-resistant populations to survive current and future climate conditions. Future breeding simulations will be performed which utilize the sampling estimates and adaptive genomic dataset developed here – leading to the development of locally adapted and disease-resistant populations for the restoration of the American chestnut.

Chapter 4

Estimating the wild adaptive genomic diversity captured within backcross populations

4.1 Introduction

The American chestnut (*Castanea dentata*) is a deciduous tree native to the eastern United States and southeastern Canada. Introduction of an invasive fungal pathogen, chestnut blight (*Cryphonectria parasitica*), decimated wild populations, which are now considered functionally extinct. Since the late 1980's, The American Chestnut Foundation (TACF) has worked to develop disease-resistant American chestnut populations to rescue the species. These backcross populations were created by crossing disease-resistant Chinese chestnut (*Castanea mollissima*) with wild American chestnut. The hybrid offspring were then backcrossed to wild American chestnut trees over several generations, with the eventual goal of trees that are 15/16 American chestnut and 1/16 Chinese chestnut, retaining most American chestnut phenotypes but having captured blight-resistance from the Chinese chestnut great-great grandparent [115]. While some families approach this goal, albeit with a greater *C. mollissima* ancestry than originally envisioned, [129], the extent of wild *C. dentata* adaptive diversity they comprise is unknown.

To support development of a breeding plan that leverages the information gained about the adaptive diversity in wild populations (Chapter 3), we sought to characterize adaptive genomic diversity already captured within the TACF backcross breeding program. Previously, genotyping-by-sequencing (GBS) of >4,500 backcross trees was completed to evaluate their ancestry and disease susceptibility [129]. While GBS is sufficient for estimating ancestry, it does not provide enough genome coverage to obtain comprehensive genomic diversity estimates [71], and for capturing the putatively adaptive loci that were previously identified in wild American chestnut (Chapter 3). Thus, whole-genome sequencing (WGS) must be performed for TACF backcross samples to properly estimate genomic diversity and the proportion of wild adaptive genomic diversity captured.

In this study, we performed WGS on 371 TACF backcross and wild chestnut trees to: 1) estimate wild adaptive diversity captured within the backcross populations and 2) estimate the best regions for reintroduction of the backcross samples based on their adaptive genotype. We used the 18,483 putatively adaptive loci as a reference for wild American chestnut diversity from Chapter 3. These results will provide the foundation for developing a breeding plan for diversifying the current disease-resistant chestnut populations with wild American chestnut genotypes – leading to locally adapted and disease-resistant chestnut populations for reintroduction.

4.2 Methods

4.2.1 Whole-genome sequencing and bioinformatics

Leaf samples were collected from 371 TACF orchard trees and the DNA was extracted using the Qiagen DNeasy Plant DNA extraction kit. Library preparation and whole-genome se-

quencing were performed at Duke University Center for Genomic and Computational Biology using an Illumina NovaSeq 6000 and four sequencing lanes. This included 269 backcross trees, 25 large-surviving (LSA) American chestnut trees, and 76 wild disease-impacted American chestnut trees.

Bioinformatic analyses were performed on Virginia Tech’s Advanced Research Computing system. SNP calling was performed using a custom pipeline based on the Genome Analysis Toolkit (GATK) best practices [120]. The fastq files were first aligned to the American chestnut reference genome using Burrows-Wheeler Aligner (BWA) mem with the -M and -R option ([67]; *Castanea dentata* v1.1; <http://phytozome-next.jgi.doe.gov/>). Then, the SAM files were converted to sorted binary files (BAM) and indexed using SAMtools [33]. To remove PCR duplicates, the Mark Duplicates tool from picard was used [1]. The GATK HaplotypeCaller algorithm was used to identify SNPs and INDELs per chromosome [75, 91], and the individual chromosome GVCF files were combined using GatherVcfs for each sample. GVCF files for all samples were then simultaneously input into the GATK GenotypeCaller algorithm to generate a raw variant call file (VCF) of SNPs and INDELs. This VCF file was filtered for quality based on GATK best practices using GATK VariantFiltration [120]. Finally, BCFtools was used to remove INDELs, multi-allelic SNPs, and sites with high missingness (>10% missing data) [33]. The remaining missing sites were imputed using BEAGLE v5.4 [20].

We then generated two datasets for subsequent analyses. 1) a dataset of the shared sites between the 18,483 adaptive loci VCF, the backcross WGS VCF, and 15 *C. mollissima* samples from the species reference VCF (Chapter 2), and 2) A dataset of the shared sites between the 18,483 adaptive loci VCF (Chapter 3) and the backcross WGS VCF. The first dataset was used to infer the ancestry of the backcross samples. For each dataset, the required individual VCF files were combined with BCFtools merge –merge all, and then any

sites with missing data were removed [33]. Thus, the final datasets only contained the shared SNPs between each VCF file.

4.2.2 Estimating wild adaptive diversity within breeding program

We first estimated the ancestry of the backcross samples at their putative adaptive loci using the 8,354 shared SNP dataset. We included *C. mollissima* samples to account for the Chinese chestnut ancestry within the backcross breeding program. A supervised ADMIXTURE analysis was performed to infer the percent association of each backcross sample with the three seed zones (Chapter 3) and *C. mollissima* [9]. The 356 wild American chestnut samples and the 15 *C. mollissima* were input with their assigned seed zone/species information and the backcross trees were left unassigned for ADMIXTURE to infer. Following the ADMIXTURE results, backcross individuals with the majority of their associations to a particular seed zone were then assigned to that seed zone.

To estimate the amount of wild adaptive diversity within the breeding program, we performed a linear regression of allele-frequency (AF) estimates at the shared 14,767 adaptive loci between the 371 backcross samples and the 356 wild American chestnut samples using the matplotlib Python package [58]. The R^2 value of allele frequencies provided an estimate for the percent of the wild adaptive diversity captured. Finally, AF values were also compared between the backcross and wild populations with a histogram of AF values.

4.2.3 Identifying local adaptation within backcross population

Since the backcross breeding population contains individuals with complex ancestry drawing on pollen from various locations in some families, we used the deep-learning method, Locator, to infer the unknown locations of backcross trees on the basis of the known geographic

origins of the wild trees [13]. Geo-locations of the 356 wild American chestnut trees and the $\approx 14k$ shared adaptive SNPs were used as input with default setting. Accuracy of the Locator results were estimated by comparing the known locations of the wild trees within the backcross dataset with their estimated origin.

4.3 Results

4.3.1 Adaptive diversity captured within backcross breeding program

The American Chestnut Foundation (TACF) and affiliated state chapters have breeding orchards located throughout the species' historical range, but pollen collections for the core breeding program were biased toward flowering wild trees from the central portion (Seed Zone 2; see Chapter 3), and germplasm resulting from these crosses was shared to varying extents with state chapter breeding programs. We therefore hypothesized that there would be a bias toward Seed Zone 2 ancestry within the backcross samples. Indeed, ancestry for the core TACF and central state chapter breeding program material was predominantly from Seed Zone 2, while the Seed Zone 3 was the least represented ancestry in the sample (mean Zone 1 ancestry = 37.4%, mean Zone 2 ancestry = 40.2%, mean Zone 3 ancestry = 8.8%, mean *C. mollissima* ancestry = 13.6%). Approximately 76% of the backcross trees contained ancestry from at least two seed zones (4.1a,b).

The allele-frequency distribution for adaptive loci in wild and backcross samples, was skewed toward medium frequency alleles in the wild population, and low frequency alleles in the backcross population (4.1c). Nevertheless, the overall allele frequency distribution correspondence was high between wild and backcross populations: $\approx 79.8\%$ of allele frequency

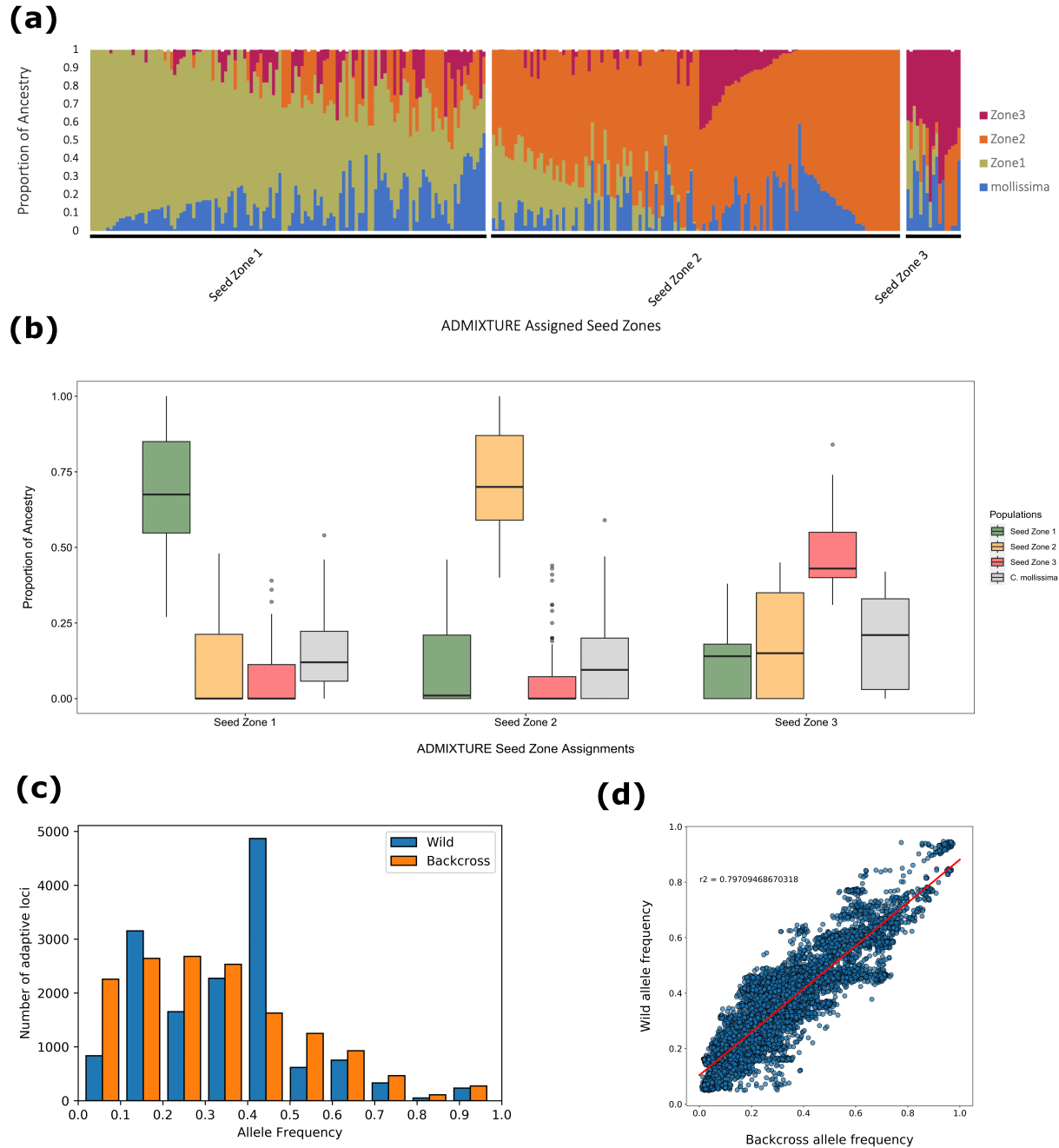


Figure 4.1: Wild adaptive diversity captured by the TACF backcross breeding program. (a) Supervised ADMIXTURE plot of ancestry. Each individual line is a single backcross sample. (b) Average proportion of ancestry within each assigned seed zone from backcross samples. (c) Allele-frequency histogram at shared adaptive loci for the wild and backcross populations. (d) Linear regression plot of allele frequencies at shared adaptive loci for the wild and backcross populations.

variation in the wild population can be explained by that of the backcross population (4.1d).

4.3.2 Matching adaptive niche of backcross trees for restoration plantings

Due to the complex ancestry of the backcross trees, it is difficult to predict from pedigree information alone the geography and climate to which a given tree would be best adapted. We therefore used Locator software and the adaptive loci to estimate the seed zone to which each tree would be best adapted. Using wild-type trees of known provenance, Locator estimates were on average within 169.82 km of their origin (4.2). The predicted locations for the backcross trees were also compared to the known location of the most recent wild *C. dentata* parent, and had an average error of 313.16 km (4.3). Assignment of backcross trees to seed zones mostly agreed with the most recent *C. dentata* parent seed zone origin (85.71 %)

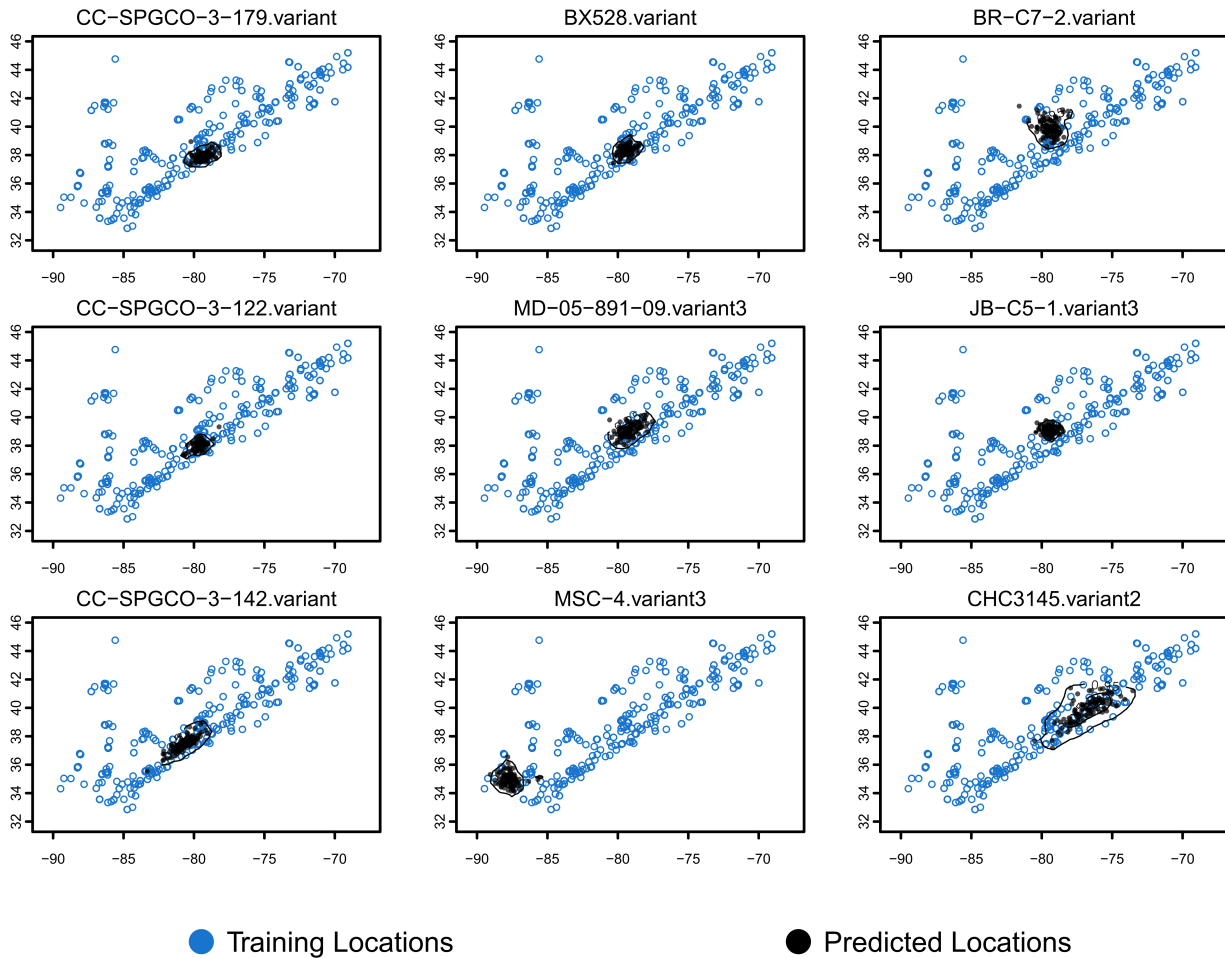


Figure 4.2: Locator origin estimates for a subset of backcross and wild American chestnut trees. 100 iterations were performed and visualized here. Blue points are locations of the 356 wild American chestnut samples that were used to train Locator, and black dots are predicted locations of a single tree for each iteration. CC-SPGCO-3-179, CC-SPGCO-3-122, CC-SPGCO-3-142, MSC-4, and JB-C5-1 are wild American chestnut trees, while the remaining samples are backcross trees.

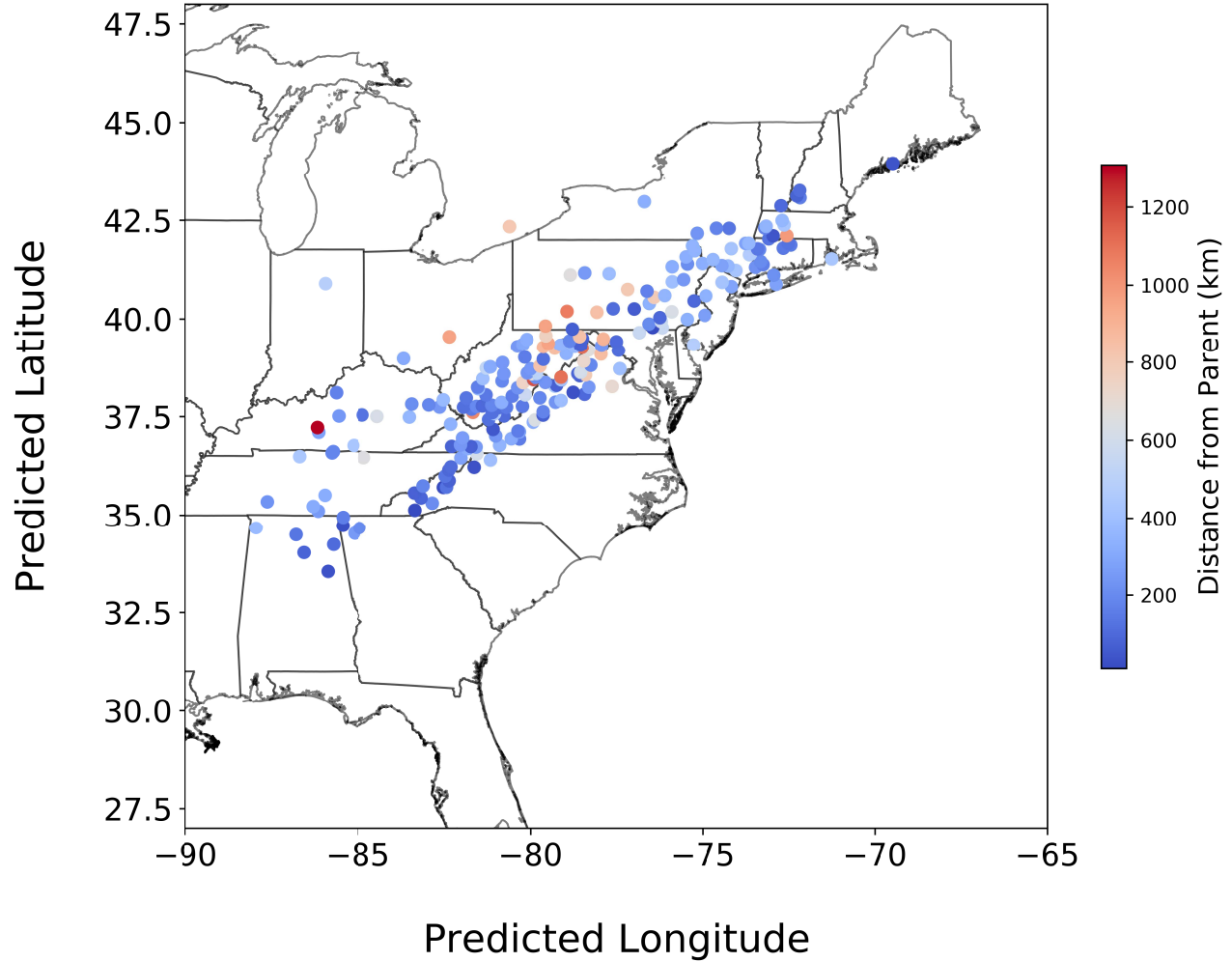


Figure 4.3: Locator predicted locations for the backcross trees and their deviation from the location of the last breeding parent.

4.4 Discussion

In this study, our goal was to evaluate the wild adaptive genomic diversity within The American Chestnut Foundation’s backcross breeding program, and to predict which region the backcross trees are most adapted to inform breeding plans. To accomplish these aims, we re-sequenced the genomes of 371 TACF orchard trees, of which 269 were backcross trees. The

results showed that 1) adaptive genomic diversity from all three seed zones has been captured within the backcross breeding program, but the least amount has been captured from the most genetically diverse Seed Zone 3; 2) adaptive alleles tend to be at lower frequency in the backcross compared with wild populations, although overall the allele frequency distributions of wild population and backcross population are well matched; and 3) replanting regions can be inferred using these adaptive loci for the backcross populations.

4.4.1 Adaptive genomic diversity captured within the TACF backcross breeding program

Beginning in the late 1980's, The American Chestnut Foundation's backcross breeding program has aimed to develop disease-resistant American x Chinese chestnut hybrid trees that retain the disease-resistant traits of the Chinese chestnut, but are not easily distinguishable from American chestnut in other respects. The process of developing a disease-resistant tree involves several generations of backcross breeding these initial F1 hybrids with pollen from different American chestnut father trees [115]. While pollen used in backcrossing was sourced from rare, wild, flowering trees, this haphazard sampling of wild diversity may not have sufficiently represented the adaptive genomic content of the wild American chestnut range. Our results suggest that the backcross breeding strategy has in fact captured much of this wild adaptive genomic diversity. However, two aspects of our results warrant further consideration. First, we observed some residual variance ($\approx 20\%$) in the adaptive allele frequency distribution between wild and backcross groups. Part of this was due to *C. mollissima* ancestry, an inevitable consequence of its hybrid origins and of the hitchhiking effect that is a consequence of selection for blight resistance. But this was also driven by a bias toward medium-frequency alleles in the wild and rare alleles in the backcross group, and the potential for fixation of these low-frequency alleles may need to be considered in future selections

and crosses. Second, only $\approx 8.8\%$ of the adaptive ancestry was associated with Seed Zone 3. This seed zone is located in the southern region of the historical chestnut range – which was also the region with the highest overall genomic diversity [49, 104, 113]. Future efforts to sample adaptive diversity through grafting and pollen collection should pay special focus to this underrepresented region, as it represents a reservoir of historical diversity that has been less affected by cyclical bottlenecks associated with Quaternary glaciation (Chapter 2), but more specifically because it is the warmest part of the contemporary range, and likely comprises unique alleles that will be important under climate change [114]. While Seed Zone 3 has lower numbers of trees recommended to be sampled (Chapter 3), it would be prudent to err on the side of caution and oversample this region when developing a breeding plan for diversifying the backcross (and transgenic) populations.

4.4.2 Replanting locations can be predicted for backcross trees

The backcross population contains trees with diverse ancestry, with most arising from at least two seed zones. These individuals contain ancestors from throughout the American chestnut range, as such, determining where these trees are likely to be best adapted is a challenge. The Locator tool was designed to infer the original location of plant and animal samples using SNP data and a training set of individuals with known locations [13]. Our use of Locator with a putative adaptive loci dataset should provide estimates of a tree’s origin based on their local adaptation – leading to an estimate of the geographic origin of each tree’s multilocus adaptive genotype, which can serve as a proxy for the best fit replanting environment. While we found high agreement between the last American chestnut parent seed zone and the Locator estimated seed zone for the backcross trees, we suggest that the Locator results take precedence. This is due to the polygenic nature of climate adaptation in American chestnut (Chapter 3), which may not be accounted for if we only consider the

origin of the most recent parent tree. However, if genetic information is not available for a given backcross tree, using the most recent American chestnut parent seed zone would be the next best resource.

4.4.3 Conclusions

Our focus in this study was in evaluating the wild adaptive genomic diversity captured within the TACF backcross breeding program to provide information for the development of a breeding plan to diversity the backcross population. Despite the backcross population containing adaptive genomic diversity from throughout the natural range, the least amount was from the southern Seed Zone 3. Ensuring that restoration populations contain the Seed Zone 3 adaptations to warmer weather and higher genomic diversity (Chapter 2; Chapter 3), when combined with the chestnut species' preadaptation to drought conditions [14], will lead to disease-resistant populations that are better suited to the impacts of climate change. Future common garden studies should be performed to fully evaluate the phenotypic response and fitness implications for the backcross trees based on its replanting location.

Chapter 5

Conclusions and Future Directions

In this dissertation, I presented our findings, which addressed previous gaps in knowledge regarding the population genomics of wild and backcross American chestnut populations. I 1) estimated the genomic diversity, population structure, and demographic history of remnant wild American chestnut populations; 2) revealed the genomic basis of local climate adaptation in American chestnut, developed a novel method to make tree sampling estimates for germplasm conservation, and defined unique seed zones for American chestnut based on climate and genotype, and 3) determined the amount of wild adaptive diversity captured by the backcross breeding program and estimated the most likely regions for backcross tree reintroduction. These results will inform the development of a breeding plan for the introgression of adaptive diversity into backcross and transgenic chestnut populations.

In Chapter 2, I evaluated the population structure, neutral genomic diversity, and demographic history of wild American chestnut populations. The two genomic datasets that were developed here will serve as important resources for American chestnut and *Castanea* species studies, which revealed that the American chestnut range is best partitioned into two or three populations. The contemporary spatial arrangement of genomic diversity within American chestnut was likely influenced by post-glacial migration bottleneck events - leading to the highest diversity in the southeastern United States, and decreasing diversity as latitude increases. Population structure and migration analyses suggest that the Appalachian Mountains served as a barrier to gene flow, with a potential migration route across the mountains

located in northwestern North Carolina and southeast West Virginia. We identified genes under positive selection, which suggests that defense against fungal pathogens is a common target of selection across the American chestnut range. The three populations defined here should serve as management units for conservation plans, with a focus on preserving germplasm from the more genetically diverse southern population.

In Chapter 3, I examined the genomic basis of local adaptation to climate in wild American chestnut, identified unique seed zones for chestnut conservation, and made sampling recommendations to capture wild adaptive diversity from each seed zone for germplasm conservation. Whole genome re-sequencing of 384 trees coupled with genotype-environment association methods suggest the species range can be most parsimoniously subdivided into three seed zones characterized by relatively homogeneous allele frequencies relative to rangewide adaptive diversity. Using these data, we developed a method to estimate the number of samples required from each seed zone to recapitulate standing adaptive diversity in each seed zone, and found that on average, 21-29 trees will need to be conserved to capture 95% of the wild adaptive diversity. Taken together, these results will inform the development of an *ex situ* germplasm conservation and breeding plan to develop locally adapted blight-resistant American chestnut populations, and provide a blueprint for developing restoration plans for other imperiled tree species.

In Chapter 4, I evaluated the adaptive genomic diversity captured within the backcross breeding program and predicted which regions within the American chestnut range to which individual backcross trees are most adapted. To accomplish these objectives, we re-sequenced the genomes of 371 trees within the backcross orchards, which represented individuals from unique backcross lines, large-surviving American chestnut trees, and disease-impacted American chestnut trees. We found that this approach incorporated adaptive genomic diversity from throughout the natural range, and we can use this information to predict replanting

zones that maximize the match between backcross families and adaptive genomic content. We also found that adaptive diversity from the southern region of the chestnut natural range is underrepresented in the backcross population, and future breeding plans should prioritize additional sampling from this area.

Taken together, our results support the development of a breeding plan to create locally-adapted and disease-resistant American chestnut trees. Following this work will be studies that utilize this genomic information to define a breeding plan for the introgression of adaptive traits into the backcross and transgenic chestnut populations. In addition, the genomic datasets for both the American chestnut and the other *Castanea* species will be an important resource for future studies regarding the genomic relationships within *Castanea* and exploring potential genomic adaptations that other chestnut species have for defense to chestnut blight.

Future range-wide common garden studies should be performed to validate these results. Representative trees can be obtained from each of the three seed zones and grown in reciprocal common gardens in each zone. Existing common garden experiments for American chestnut found phenotypic evidence of local adaptation and explored assisted migration for the species, however, the common gardens were located in the northern limits of the natural range [102, 107], thus were unable to evaluate the response to hotter temperatures from different regions of the range. Climate change will lead to increasing temperatures and precipitation changes within the chestnut range. So, common garden experiments will evaluate how chestnut trees throughout the range will cope with higher temperatures. The additional common garden experiments would serve two purposes: 1) They would provide phenotypic evidence for the response of American chestnut populations to varying climate conditions (importantly, warmer conditions), and 2) determine the importance of replanting disease-resistant chestnut trees into the seed zone which they are predicted to be most adapted.

While chestnut trees can experience damage when grown outside of their local range [28], long-term observations can determine how detrimental this damage can be toward a tree's fitness, and the impact of more rare extreme climate conditions can have on replanting location. Ultimately, an understanding of how each seed zone responds to climate conditions outside of their range will improve breeding plan development and the long-term restoration of American chestnut populations.

Bibliography

- [1] Picard toolkit. <https://broadinstitute.github.io/picard/>, 2019.
- [2] AdaptWest Project. Gridded current and projected climate data for North America at 1km resolution, generated using the ClimateNA v7.30 software (T. Wang et al., 2022). <http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm>, 2022. Available at adaptwest.databasin.org.
- [3] Segun M. Adeyemo and Joshua J. Granger. Habitat suitability model and range shift analysis for American chestnut (*Castanea dentata*) in the United States. *Trees, Forests and People*, 11:100360, 2023. ISSN 2666-7193. doi: <https://doi.org/10.1016/j.tfp.2022.100360>. URL <https://www.sciencedirect.com/science/article/pii/S2666719322001662>.
- [4] Sally N Aitken and Jordan B Bemmels. Time to get moving: assisted gene flow of forest trees. *Evolutionary applications*, 9(1):271–290, 2016.
- [5] Sally N Aitken, Sam Yeaman, Jason A Holliday, Tongli Wang, and Sierra Curtis-McLane. Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary applications*, 1(1):95–111, 2008.
- [6] Mohammad A Al-Rabab’ah and Claire G Williams. An ancient bottleneck in the Lost Pines of central Texas. *Molecular Ecology*, 13(5):1075–1084, 2004.
- [7] Nikolaos Alachiotis and Pavlos Pavlidis. RAI_{SD} detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, 1(1):1–11, 2018.

- [8] Florian J Alberto, Sally N Aitken, Ricardo Alía, Santiago C González-Martínez, Heikki Hänninen, Antoine Kremer, François Lefèvre, Thomas Lenormand, Sam Yeaman, Ross Whetten, and Outi Savolainen. Potential for evolutionary responses to climate change—evidence from tree populations. *Global Change Biology*, 19(6):1645–1661, 2013.
- [9] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.
- [10] Mary V Ashley, Saji T Abraham, Janet R Backs, and Walter D Koenig. Landscape genetics and population structure in Valley Oak (*Quercus lobata* née). *American Journal of Botany*, 102(12):2124–2131, 2015.
- [11] Juliann E Aukema, Deborah G McCullough, Betsy Von Holle, Andrew M Liebhold, Kerry Britton, and Susan J Frankel. Historical accumulation of nonindigenous forest pests in the continental United States. *BioScience*, 60(11):886–897, 12 2010. ISSN 0006-3568. doi: 10.1525/bio.2010.60.11.5. URL <https://doi.org/10.1525/bio.2010.60.11.5>.
- [12] Abdelali Barakat, Denis S DiLoreto, Yi Zhang, Chris Smith, Kathleen Baier, William A Powell, Nicholas Wheeler, Ron Sederoff, and John E Carlson. Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC Plant Biology*, 9(1):51, 2009.
- [13] Christopher J Battey, Peter L Ralph, and Andrew D Kern. Predicting geographic location from genetic variation with deep neural networks. *Elife*, 9:e54507, 2020.
- [14] William L Bauerle, G Geoff Wang, Joseph D Bowden, and Christina M Hong. An analysis of ecophysiological responses to drought in American chestnut. *Annals of Forest Science*, 63(8):833–842, 2006.

- [15] DF Belknap. Quaternary, April 2020. URL <https://www.britannica.com/science/Quaternary/The-Quaternary-environment>.
- [16] Tanya Z Berardini, Suparna Mundodi, Leonore Reiser, Eva Huala, Margarita Garcia-Hernandez, Peifen Zhang, Lukas A Mueller, Jungwoon Yoon, Aisling Doyle, Gabriel Lander, Nick Moseyko, Danny Yoo, Iris Xu, Brandon Zoeckler, Mary Montoya, Neil Miller, Dan Weems, and Seung Y Rhee. Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiology*, 135(2):745–755, 2004.
- [17] Johannes Bergsten. A review of long-branch attraction. *Cladistics*, 21(2):163–193, 2005.
- [18] David Q Bowen. *Quaternary geology: a stratigraphic framework for multidisciplinary work*. Pergamon Press Inc., 1978.
- [19] Andrew D Bower, J Bradley St Clair, and Vicky Erickson. Generalized provisional seed zones for native plants. *Ecological Applications*, 24(5):913–919, 2014.
- [20] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.
- [21] Katie L Burke. The effects of logging and disease on American chestnut. *Forest Ecology and Management*, 261(6):1027–1033, 2011.
- [22] Katie L Burke. Niche contraction of American chestnut in response to chestnut blight. *Canadian Journal of Forest Research*, 42(3):614–620, 2012.
- [23] Philip Laurence Buttrick. *Chestnut and the chestnut blight in North Carolina*. Number 56. 1925.

- [24] Thibaut Capblancq and Brenna R Forester. Redundancy analysis: A swiss army knife for landscape genomics. *Methods in Ecology and Evolution*, 12(12):2298–2309, 2021.
- [25] Thibaut Capblancq, John R Butnor, Sonia Deyoung, Ethan Thibault, Helena Munson, David M Nelson, Matthew C Fitzpatrick, and Stephen R Keller. Whole-exome sequencing reveals a long-term decline in effective population size of red spruce (*Picea rubens*). *Evolutionary Applications*, 13(9):2190–2205, 2020.
- [26] Thibaut Capblancq, Susanne Lachmuth, Matthew C. Fitzpatrick, and Stephen R. Keller. From common gardens to candidate genes: exploring local adaptation to climate in red spruce. *New Phytologist*, 237(5):1590–1605, 2023. doi: <https://doi.org/10.1111/nph.18465>. URL <https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.18465>.
- [27] Kevin Caye, Basile Jumentier, Johanna Lepeule, and Olivier François. Lfmm 2: fast and accurate inference of gene-environment associations in genome-wide studies. *Molecular Biology and Evolution*, 36(4):852–860, 2019.
- [28] Peter W Clark, Alissa J Freeman, Anthony W D’Amato, Paul G Schaberg, Gary J Hawley, Kevin S Evans, and Christopher W Woodall. Restoring a keystone tree species for the future: American chestnut assisted migration plantings in an adaptive silviculture experiment. *Forest Ecology and Management*, 523:120505, 2022. ISSN 0378-1127. doi: <https://doi.org/10.1016/j.foreco.2022.120505>. URL <https://www.sciencedirect.com/science/article/pii/S0378112722004996>.
- [29] Bowen S Crandall, GF Gravatt, and Margaret Milburn Ryan. Root disease of *Castanea* species and some coniferous and broadleaf nursery stocks, caused by *Phytophthora cinnamomi*. *Phytopathology*, 35:162–80, 1945.

- [30] Ellen Crocker, Bradford Condon, Abdullah Almsaeed, Benjamin Jarret, C Dana Nelson, Albert G Abbott, Doreen Main, and Margaret Staton. TreeSnap: A citizen science app connecting tree enthusiasts and forest scientists. *Plants, People, Planet*, 2(1):47–52, 2020.
- [31] J Dainat. *AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format*, 2022.
- [32] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, aug 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr330. URL <https://doi.org/10.1093/bioinformatics/btr330>.
- [33] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), 02 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008. URL <https://doi.org/10.1093/gigascience/giab008>. giab008.
- [34] Donald Edward Davis. *The American chestnut: An environmental history*. University of Georgia Press, 2021.
- [35] Margaret B Davis. Quaternary history of deciduous forests of eastern North America and Europe. *Annals of the Missouri Botanical Garden*, 70(3):550–563, 1983.
- [36] Hazel R Delcourt. Late Quaternary vegetation history of the eastern Highland Rim and adjacent Cumberland Plateau of Tennessee. *Ecological Monographs*, 49(3):255–280, 1979.

- [37] Paul A Delcourt. Goshen Springs: late Quaternary vegetation record for southern Alabama. *Ecology*, 61(2):371–386, 1980.
- [38] Paul A Delcourt, Hazel R Delcourt, Ronald C Brister, and Laurence E Lackey. Quaternary vegetation history of the Mississippi embayment 1, 2. *Quaternary Research*, 13(1):111–132, 1980.
- [39] Samuel Detwiler. The American chestnut tree: Identification and characteristics. *American Forestry*, 21(262):957–959, 1915.
- [40] Seth J Diamond, Robert H Giles, Jr, Roy L Kirkpatrick, and Gary J Griffin. Hard mast production before and after the chestnut blight. *Southern Journal of Applied Forestry*, 24(4):196–201, 11 2000. ISSN 0148-4419. doi: 10.1093/sjaf/24.4.196. URL <https://doi.org/10.1093/sjaf/24.4.196>.
- [41] Stéphane Dray and Anne-Béatrice Dufour. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22:1–20, 2007.
- [42] Andrew J Eckert, Andrew D Bower, Santiago C González-Martínez, Jill L Wegrzyn, Graham Coop, and David B Neale. Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology*, 19(17):3789–3805, 2010.
- [43] Nick Ellis, Stephen J Smith, and C Roland Pitcher. Gradient forests: calculating importance gradients on physical predictors. *Ecology*, 93(1):156–168, 2012.
- [44] Matthew C Fitzpatrick and Stephen R Keller. Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology letters*, 18(1):1–16, 2015.
- [45] Food and Agriculture Organization of the United Nations. *Global review of forests*

- pests and diseases*. Food & Agriculture Organization of the United Nations (FAO), Rome, Italy, 2009.
- [46] Brenna R Forester, Jesse R Lasky, Helene H Wagner, and Dean L Urban. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology*, 27(9):2215–2233, 2018.
- [47] Eric Frichot and Olivier Francois. LEA: an R package for Landscape and Ecological Association studies. *Methods in Ecology and Evolution*, 2015. URL <http://membres-timc.imag.fr/Olivier.Francois/lea.html>.
- [48] W Chris Funk, John K McKay, Paul A Hohenlohe, and Fred W Allendorf. Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, 27(9):489–496, 2012.
- [49] Oliver Gailing and C Dana Nelson. Genetic variation patterns of American chestnut populations at EST-SSRs. *Botany*, 95(8):799–807, 2017.
- [50] David M. Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, and Daniel S. Rokhsar. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1):D1178–D1186, 2012.
- [51] Paul F Gugger, Sorel T Fitz-Gibbon, Ana Albarrán-Lara, Jessica W Wright, and Victoria L Sork. Landscape genomics of *Quercus lobata* reveals genes involved in local climate adaptation at multiple spatial scales. *Molecular Ecology*, 30(2):406–423, 2021.
- [52] Kendra M Gurney, Paul G Schaberg, Gary J Hawley, and John B Shane. Inadequate cold tolerance as a possible limitation to American chestnut restoration in the northeastern United States. *Restoration Ecology*, 19(1):55–63, 2011.

- [53] Gerald H Haug, Andrey Ganopolski, Daniel M Sigman, Antoni Rosell-Mele, George EA Swann, Ralf Tiedemann, Samuel L Jaccard, Jörg Bollmann, Mark A Maslin, Melanie J Leng, and Geoffrey Eglinton. North Pacific seasonality and the glaciation of North America 2.7 million years ago. *Nature*, 433(7028):821–825, 2005.
- [54] Sean Hoban, Michael W Bruford, W Chris Funk, Peter Galbusera, M Patrick Griffith, Catherine E Grueber, Myriam Heuertz, Margaret E Hunter, Christina Hvilsom, Belma Kalamujic Stroil, Francine Kershaw, Colin K Khoury, Linda Laikre, Margarida Lopes-Fernandes, Anna J MacDonald, Joachim Mergeay, Mariah Meek, Cinnamon Mittan, Tarek A Mukassabi, David O’Brien, Rob Ogden, Clarisse Palma-Silva, Uma Ramakrishnan, Gernot Segelbacher, Robyn E Shaw, Per Sjögren-Gulve, Nevena Veličković, and Cristiano Vernesi. Global commitments to conserving and monitoring genetic diversity are now necessary and feasible. *BioScience*, 71(9):964–976, 2021.
- [55] JA Holliday, M Yuen, K Ritland, and SN Aitken. Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Molecular Ecology*, 19(18):3857–3864, 2010.
- [56] Zhe Hou, Ang Li, and Jianguo Zhang. Genetic architecture, demographic history, and genomic differentiation of *Populus davidiana* revealed by whole-genome resequencing. *Evolutionary Applications*, 13(10):2582–2596, 2020.
- [57] Glenn T Howe, Sally N Aitken, David B Neale, Kathleen D Jermstad, Nicholas C Wheeler, and Tony HH Chen. From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. *Canadian Journal of Botany*, 81(12):1247–1266, 2003.
- [58] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

- [59] Richard A Jaynes and Arthur H Graves. Connecticut hybrid chestnuts and their culture. *Connecticut Agricultural Experiment Station*, 1963.
- [60] Thibaut Jombart and Ismaïl Ahmed. *adeigenet 1.3-1*: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27(21):3070–3071, 2011.
- [61] Thibaut Jombart, Sébastien Devillard, and François Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11(1):1–15, 2010.
- [62] Stephen R Keller, Matthew S Olson, Salim Silim, William Schroeder, and Peter Tiffin. Genomic diversity, population structure, and migration following rapid range expansion in the Balsam Poplar, *Populus balsamifera*. *Molecular Ecology*, 19(6):1212–1226, 2010.
- [63] Brian J Knaus and Niklaus J Grünwald. vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, 17(1):44–53, 2017.
- [64] Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. Angsd: analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1):356, 2014.
- [65] Koen Kramer, Bernd Degen, Jutta Buschbom, Thomas Hickler, Wilfried Thuiller, Martin T Sykes, and Wim de Winter. Modelling exploration of the future of European beech (*Fagus sylvatica* L.) under climate change—range, abundance, genetic diversity and adaptive response. *Forest Ecology and Management*, 259(11):2213–2222, 2010.
- [66] Thomas L Kubisiak and James H Roberds. Genetic structure of American chestnut populations based on neutral DNA markers. In *Restoration of American chestnut to forest lands: proceedings of a conference and workshop. Asheville, NC, USA: National Park Service*, pages 109–122, 2006.

- [67] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.
- [68] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.
- [69] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.
- [70] Katie E Lotterhos. Does the paradigm of genotype-environment associations need to be re-assessed? the paradox of adaptive phenotypic clines with non-clinal patterns in causal alleles. *bioRxiv*, 2022.
- [71] David B Lowry, Sean Hoban, Joanna L Kelley, Katie E Lotterhos, Laura K Reed, Michael F Antolin, and Andrew Storfer. Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, 17:142–152, 2017.
- [72] Ralph H Lutts. Like manna from God: The American chestnut trade in southwestern Virginia. *Environmental History*, 9(3):497–525, 2004.
- [73] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2021. URL <https://CRAN.R-project.org/package=cluster>. R package version 2.1.2 — For new features, see the ‘Changelog’ file (in the package source).
- [74] Colin R Mahony, Tongli Wang, Andreas Hamann, and Alex J Cannon. A global climate model ensemble for downscaled monthly climate normals over North America. *International Journal of Climatology*, 2022.

- [75] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.
- [76] Jason S McLachlan, James S Clark, and Paul S Manos. Molecular indicators of tree migration capacity under rapid climate change. *Ecology*, 86(8):2088–2098, 2005.
- [77] Huaiyu Mi, Anushya Muruganujan, Xiaosong Huang, Dustin Ebert, Caitlin Mills, Xinyu Guo, and Paul D Thomas. Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v. 14.0). *Nature Protocols*, 14(3):703–721, 2019.
- [78] M Mimura and SN Aitken. Local adaptation at the range peripheries of Sitka spruce. *Journal of Evolutionary Biology*, 23(2):249–258, 2010.
- [79] Lampros Mouselimis. *ClusterR: Gaussian Mixture Models, K-Means, Mini-Batch-Kmeans, K-Medoids and Affinity Propagation Clustering*, 2022. URL <https://CRAN.R-project.org/package=ClusterR>. R package version 1.2.9.
- [80] Markus Müller, C Dana Nelson, and Oliver Gailing. Analysis of environment-marker associations in American chestnut. *Forests*, 9(11):695, 2018.
- [81] Andrew E Newhouse, Linda D Polin-McGuigan, Kathleen A Baier, Kristia ER Valletta, William H Rottmann, Timothy J Tschaplinski, Charles A Maynard, and William A Powell. Transgenic American chestnuts show enhanced blight resistance and transmit the trait to T1 progeny. *Plant Science*, 228:88–97, 2014.

- [82] Rasmus Nielsen. Molecular signatures of natural selection. *Annual Review of Genetics*, 39(1):197–218, 2005.
- [83] Rasmus Nielsen, Thorfinn Korneliussen, Anders Albrechtsen, Yingrui Li, and Jun Wang. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PloS One*, 7(7):e37558, 2012.
- [84] Paul H Noah, Nicolette L Cagle, Jared W Westbrook, and Sara F Fitzsimmons. Identifying resilient restoration targets: Mapping and forecasting habitat suitability for *Castanea dentata* in Eastern USA under different climate-change scenarios. *Climate Change Ecology*, 2:100037, 2021.
- [85] Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O’Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, and Helene Wagner. *vegan: Community Ecology Package*, 2020. URL <https://CRAN.R-project.org/package=vegan>. R package version 2.5-7.
- [86] Frederick L Paillet. Chestnut: history and ecology of a transformed species. *Journal of Biogeography*, 29(10-11):1517–1530, 2002.
- [87] M Taylor Perkins, Anna Claire Robinson, Martin L Cipollini, and J Hill Craddock. Identifying host resistance to *Phytophthora cinnamomi* in hybrid progeny of *Castanea dentata* and *Castanea mollissima*. *HortScience*, 54(2):221–225, 2019.
- [88] M Taylor Perkins, Tetyana Zhebentyayeva, Paul H Sisco, and J Hill Craddock. Genome-wide sequence-based genotyping supports a nonhybrid origin of *Castanea alabamensis*. *Systematic Botany*, 46(4):973–984, 2021.

- [89] Desislava Petkova, John Novembre, and Matthew Stephens. Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, 48(1): 94–100, 2016.
- [90] Carolyn Pike, Kevin M Potter, Paul Berrang, Barbara Crane, Joanne Baggs, Laura Leites, and Tom Luther. New seed-collection zones for the eastern United States: the eastern seed zone forum. *Journal of Forestry*, 118(4):444–451, 2020.
- [91] Ryan Poplin, Valentin Ruano-Rubio, Mark A DePristo, Tim J Fennell, Mauricio O Carneiro, Geraldine A Van der Auwera, David E Kling, Laura D Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J Daly, Ben Neale, Daniel G MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, page 201178, 2018.
- [92] Kevin M Potter, Barbara S Crane, and William W Hargrove. A United States national prioritization framework for tree species vulnerability to climate change. *New Forests*, 48(2):275–300, 2017.
- [93] William A Powell, Andrew E Newhouse, and Vernon Coffey. Developing blight-tolerant American chestnut trees. *Cold Spring Harbor Perspectives in Biology*, 11(7):a034587, 2019.
- [94] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007. ISSN 0002-9297. doi: <https://doi.org/10.1086/519795>. URL <https://www.sciencedirect.com/science/article/pii/S0002929707613524>.

- [95] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.
- [96] R Core Team. R: A language and environment for statistical computing. 2021. URL <https://www.R-project.org/>.
- [97] Orly Razgour, Brenna Forester, John B Taggart, Michaël Bekaert, Javier Juste, Carlos Ibáñez, Sébastien J Puechmaille, Roberto Novella-Fernandez, Antton Alberdi, and Stéphanie Manel. Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. *Proceedings of the National Academy of Sciences*, 116(21):10418–10423, 2019.
- [98] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2022. URL <https://CRAN.R-project.org/package=psych>. R package version 2.2.3.
- [99] Samuel Royer-Tardif, Laura Boisvert-Marsh, Julie Godbout, Nathalie Isabel, and Isabelle Aubin. Appendices: Finding common ground: Toward comparable indicators of adaptive capacity of tree species to a changing climate. *Ecology and Evolution*, 11(19):13081–13100, 2021.
- [100] Emily WB Russell. Pre-blight distribution of *Castanea dentata* (marsh.) borkh. *Bulletin of the Torrey Botanical Club*, 114(2):183–190, 1987.
- [101] Thomas M Saielli, Paul G Schaberg, Gary J Hawley, Joshua M Halman, and Kendra M Gurney. Nut cold hardiness as a factor influencing the restoration of American chestnut

- in northern latitudes and high elevations. *Canadian Journal of Forest Research*, 42(5): 849–857, 2012.
- [102] Thomas M Saielli, Paul G Schaberg, Gary J Hawley, Joshua M Halman, and Kendra M Gurney. Genetics and silvicultural treatments influence the growth and shoot winter injury of American chestnut in Vermont. *Forest Science*, 60(6):1068–1076, 2014.
- [103] Jarkko Salojärvi, Olli-Pekka Smolander, Kaisa Nieminen, Sitaram Rajaraman, Omid Safronov, Pezhman Safdari, Airi Lamminmäki, Juha Immanen, Tianying Lan, Jaakko Tanskanen, et al. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics*, 49(6):904–912, 2017.
- [104] Alexander M. Sandercock, Jared W. Westbrook, Qian Zhang, Hayley A. Johnson, Thomas M. Saielli, John A. Scrivani, Sara F. Fitzsimmons, Kendra Collins, Matthew Taylor Perkins, James Hill Craddock, Jeremy Schmutz, Jane Grimwood, and Jason A. Holliday. Frozen in time: Rangewide genomic diversity, structure, and demographic history of relict American chestnut populations. *Molecular Ecology*, 31(18):4640–4655, 2022. doi: <https://doi.org/10.1111/mec.16629>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16629>.
- [105] Outi Savolainen, Tanja Pyhäjärvi, and Timo Knürr. Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, 38(1):595–619, 2007.
- [106] Outi Savolainen, Martin Lascoux, and Juha Merilä. Ecological genomics of local adaptation. *Nature Reviews Genetics*, 14(11):807–820, 2013.
- [107] Paul G Schaberg, Paula F Murakami, Kendra M Collins, Christopher F Hansen, and Gary J Hawley. Phenology, cold injury and growth of American chestnut in a Range-Wide provenance test. *Forest Ecology and Management*, 513:120178, 2022.

- [108] Emanuel Schmid-Siebert, Namrata Sarkar, Christian Iseli, Sandra Calderon, Caroline Gouhier-Darimont, Jacqueline Chrast, Pietro Cattaneo, Frédéric Schütz, Laurent Farinelli, Marco Pagni, Michel Schneider, Jérémie Voumard, Michel Jaboyedoff, Christian Fankhauser, Christian S Hardtke, Laurent Keller, John R Pannell, Alexandre Reymond, Marc Robinson-Rechavi, Ioannis Xenarios, and Philippe Reymond. Low number of fixed somatic mutations in a long-lived oak tree. *Nature Plants*, 3(12): 926–929, 2017.
- [109] John Scrivani. Forest inventory and analysis. *Journal of the American Chestnut Foundation*, 25:17–18, 2011.
- [110] Damian Smedley, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz, Mohammad Hamza Awedh, Richard Baldock, Giulia Barbera, Philippe Bardou, Tim Beck, Andrew Blake, Merideth Bonierbale, Anthony J Brookes, Gabriele Bucci, Iwan Buetti, Sarah Burge, Cédric Cabau, Joseph W. Carlson, Claude Chelala, Charalambos Chrysostomou, Davide Cittaro, Olivier Collin, Raul Cordova, Rosalind J. Cutts, Erik Dassi, Alex Di Genova, Anis Djari, Anthony Esposito, Heather Estrella, Eduardo Eyra, Julio Fernandez-Banet, Simon Forbes, Robert C Free, Takatomo Fujisawa, Emanuela Gadaleta, Jose M Garcia-Manteiga, David Goodstein, Kristian Gray, José Afonso Guerra-Assunção, Bernard Haggarty, Dong-Jin Han, Byung Woo Han, Todd Harris, Jayson Harshbarger, Robert K Hastings, Richard D Hayes, Claire Hoede, Shen Hu, Zhi-Liang Hu, Lucie Hutchins, Zhengyan Kan, Hideya Kawaji, Aminah Keliet, Arnaud Kerhornou, Sunghoon Kim, Rhoda Kinsella, Christophe Klopp, Lei Kong, Daniel Lawson, Dejan Lazarevic, Ji-Hyun Lee, Thomas Letellier, Chuan-Yun Li, Pietro Lio, Chu-Jun Liu, Jie Luo, Alejandro Maass, Jerome Mariette, Thomas Maurel, Stefania Merella, Azza Mostafa Mohamed, Francois Moreews, Ibounyamine Nabihoudine, Nelson Ndegwa, Céline Noirot, Cristian

- Perez-Llamas, Michael Primig, Alessandro Quattrone, Hadi Quesneville, Davide Rambaldi, James Reecy, Michela Riba, Steven Rosanoff, Amna Ali Saddiq, Elisa Salas, Olivier Sallou, Rebecca Shepherd, Reinhard Simon, Linda Sperling, William Spooner, Daniel M Staines, Delphine Steinbach, Kevin Stone, Elia Stupka, Jon W Teague, Abu Z Dayem Ullah, Jun Wang, Doreen Ware, Marie Wong-Erasmus, Ken Youens-Clark, Amonida Zadissa, Shi-Jian Zhang, and Arek Kasprzyk. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1):W589–W598, 2015.
- [111] David M Smith. American chestnut: Ill-fated monarch of the eastern hardwood forest. *Journal of Forestry*, 98(2):12–15, 2000.
- [112] Douglas E Soltis, Ashley B Morris, Jason S McLachlan, Paul S Manos, and Pamela S Soltis. Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, 15(14):4261–4293, 2006.
- [113] Elizabeth L Spriggs and Matthew E Fertakos. Evolution of *Castanea* in North America: restriction-site-associated DNA sequencing and ecological modeling reveal a history of radiation, range shifts, and disease. *American Journal of Botany*, 108(9):1692–1704, 2021.
- [114] John Bradley St Clair and Glenn Thomas Howe. Strategies for conserving forest genetic resources in the face of climate change. *Turkish Journal of Botany*, 35(4):403–409, 2011.
- [115] Kim C Steiner, Jared W Westbrook, Frederick V Hebard, Laura L Georgi, William A Powell, and Sara F Fitzsimmons. Rescue of American chestnut with extraspecific genes following its destruction by a naturalized pathogen. *New Forests*, 48(2):317–336, 2017.

- [116] Steven L Stephenson, Harold S Adams, and Michael L Lipford. The present distribution of chestnut in the upland forest communities of Virginia. *Bulletin of the Torrey Botanical Club*, 118(1):24–32, 1991.
- [117] John D. Storey, Andrew J. Bass, Alan Dabney, and David Robinson. *qvalue: Q-value estimation for false discovery rate control*, 2021. URL <http://github.com/jdstorey/qvalue>. R package version 2.26.0.
- [118] Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, 2017.
- [119] Peter Tiffin and Jeffrey Ross-Ibarra. Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, 29(12):673–680, 2014.
- [120] Geraldine A Van der Auwera and Brian D O’Connor. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. O’Reilly Media, 2020.
- [121] Randle Aaron M Villanueva and Zhuo Job Chen. *ggplot2: elegant graphics for data analysis*, 2019.
- [122] G Geoff Wang, Benjamin O Knapp, Stacy L Clark, and Bryan T Mudder. The silvics of *Castanea dentata* (Marsh.) Borkh., American chestnut, Fagaceae (beech family). *In: Publication: Gen. Tech. Rep. SRS-GTR-173. Asheville, NC: US Department of Agriculture Forest Service, Southern Research Station*, page 18, 2013.
- [123] Jing Wang, Nathaniel R Street, Douglas G Scofield, and Pär K Ingvarsson. Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics*, 202(3):1185–1200, 2016.

- [124] Tongli Wang, Andreas Hamann, Dave Spittlehouse, and Carlos Carroll. Locally down-scaled and spatially customizable climate data for historical and future periods for North America. *PloS one*, 11(6):e0156720, 2016.
- [125] William A Watts, Barbara CS Hansen, and Eric C Grimm. Camel Lake: A 40 000-yr record of vegetational and forest history from northwest Florida. *Ecology*, 73(3):1056–1066, 1992.
- [126] J Westbrook, V Lakoba, and I Golden. The marriage of biotechnology and breeding: Enhancing disease resistance and genetic diversity in the American chestnut restoration population. *The Journal of the American Chestnut Foundation*, 35(3):23–26, 2021.
- [127] Jared W Westbrook, Joseph B James, Paul H Sisco, John Frampton, Sunny Lucas, and Steven N Jeffers. Resistance to *Phytophthora cinnamomi* in American chestnut (*Castanea dentata*) backcross populations that descended from two Chinese chestnut (*Castanea mollissima*) sources of resistance. *Plant Disease*, 103(7):1631–1641, 2019.
- [128] Jared W Westbrook, Jason A Holliday, Andrew E Newhouse, and William A Powell. A plan to diversify a transgenic blight-tolerant American chestnut population using citizen science. *Plants, People, Planet*, 2(1):84–95, 2020.
- [129] Jared W Westbrook, Qian Zhang, Mihir K Mandal, Eric V Jenkins, Laura E Barth, Jerry W Jenkins, Jane Grimwood, Jeremy Schmutz, and Jason A Holliday. Optimizing genomic selection for blight resistance in american chestnut backcross populations: A trade-off with american chestnut ancestry implies resistance is polygenic. *Evolutionary applications*, 13(1):31–47, 2020.
- [130] Huan Xiong, Feng Zou, Sujuan Guo, Deyi Yuan, and Genhua Niu. Self-sterility may be due to prezygotic late-acting self-incompatibility and early-acting inbreeding depres-

- sion in Chinese chestnut. *Journal of the American Society for Horticultural Science*, 144(3):172–181, 2019.
- [131] Yue Yu, Sally N. Aitken, Loren H. Rieseberg, and Tongli Wang. Using landscape genomics to delineate seed and breeding zones for lodgepole pine. *New Phytologist*, 235(4):1653–1664, 2022. doi: <https://doi.org/10.1111/nph.18223>. URL <https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.18223>.
- [132] Bo Zhang, Allison D Oakes, Andrew E Newhouse, Kathleen M Baier, Charles A Maynard, and William A Powell. A threshold level of oxalate oxidase transgene expression reduces *Cryphonectria parasitica*-induced necrosis in a transgenic American chestnut (*Castanea dentata*) leaf bioassay. *Transgenic Research*, 22(5):973–982, 2013.
- [133] Tetyana N Zhebentyayeva, Paul H Sisco, Laura L Georgi, Steven N Jeffers, M Taylor Perkins, Joseph B James, Frederick V Hebard, Christopher Saski, C Dana Nelson, and Albert G Abbott. Dissecting resistance to *Phytophthora cinnamomi* in interspecific hybrid chestnut crosses using sequence-based genotyping and QTL mapping. *Phytopathology*, 109(9):1594–1604, 2019.
- [134] John WR Zinck and Om P Rajora. Post-glacial phylogeography and evolution of a wide-ranging highly-exploited keystone forest tree, eastern white pine (*Pinus strobus*) in North America: single refugium, multiple routes. *BMC Evolutionary Biology*, 16: 56, 2016.

Appendices

Appendix A

Supplementary Materials for Chapter 2 "Frozen in time: Rangewide genomic diversity, structure, and demographic history of relict American chestnut populations"

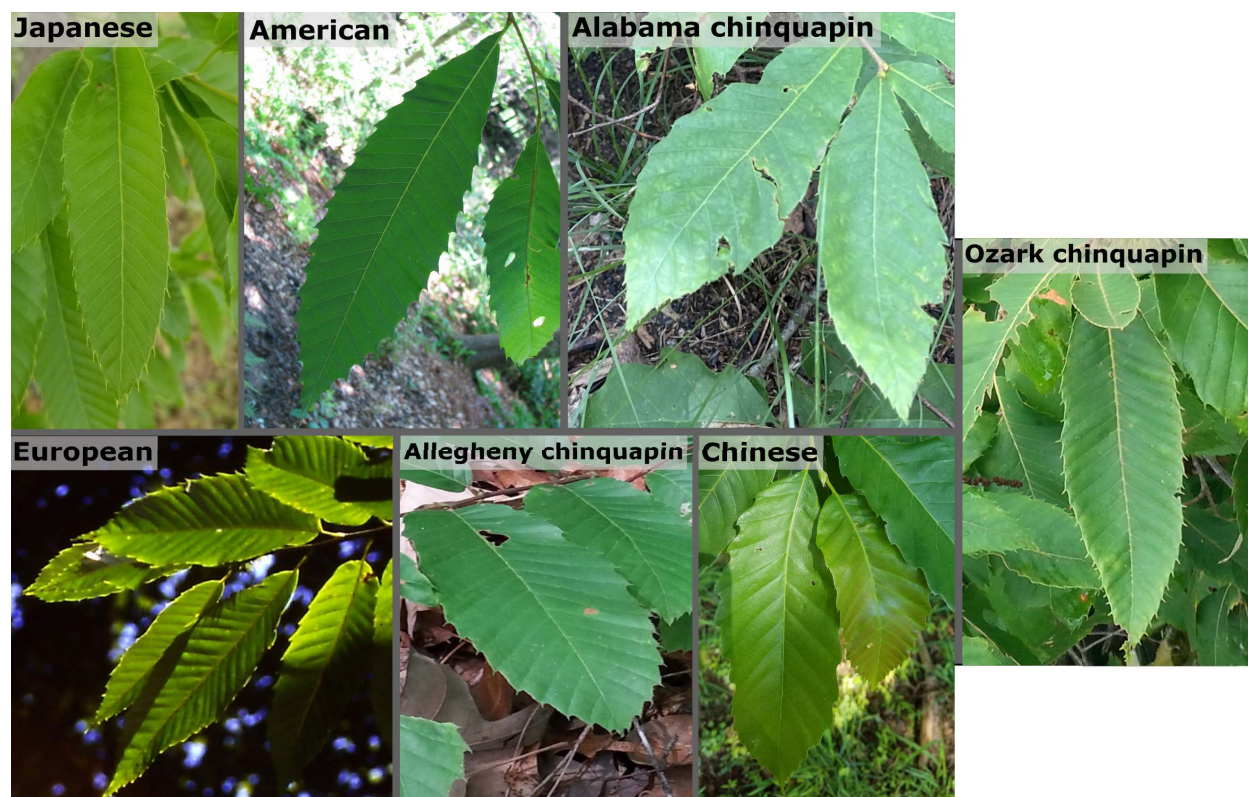


Figure A.1: Leaf examples from each of the *Castanea* species included in this study. Pictured are representatives of Japanese chestnut (*C. crenata*), American chestnut (*C. dentata*), Alabama chinquapin (*C. pumila* var. *alabamensis*), Ozark chinquapin (*C. pumila* var. *ozarkensis*), European chestnut (*C. sativa*), Allegheny chinquapin (*C. pumila* var. *pumila*), and Chinese chestnut (*C. mollissima*).

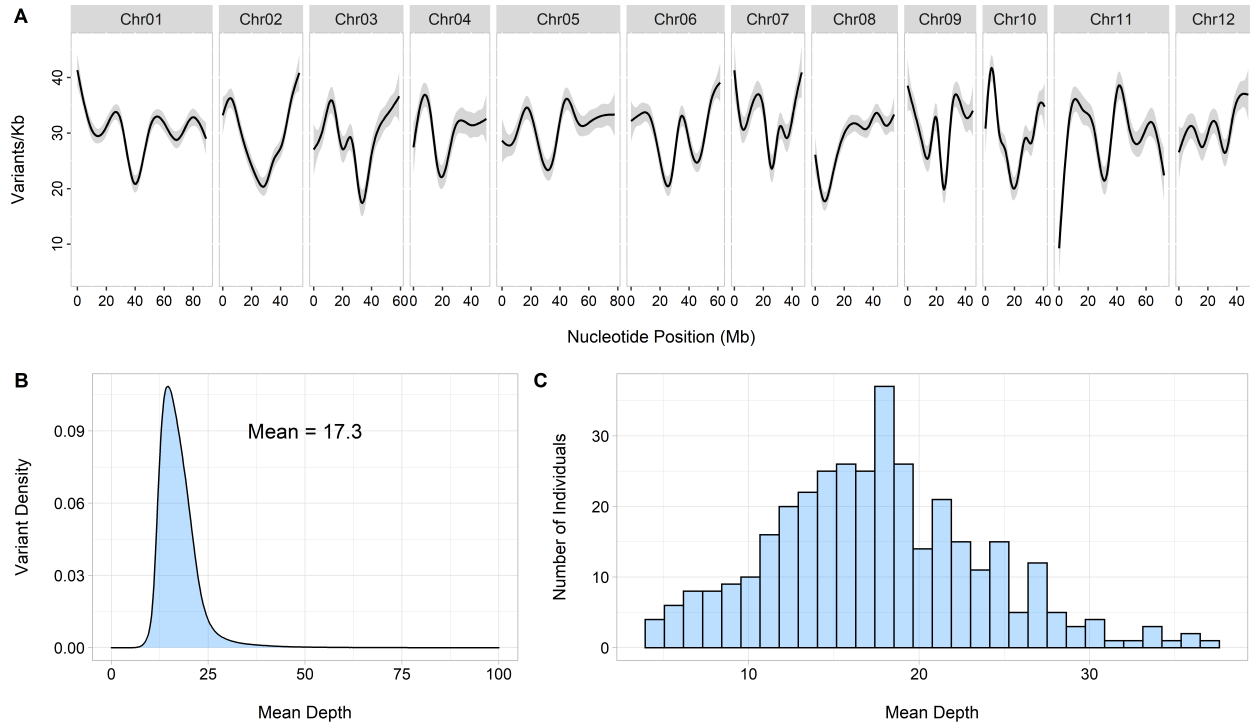


Figure A.2: Visualization of sequencing density and coverage in the filtered *Castanea dentata* genomic dataset. **A**. Genome-wide sequencing density. Greyed regions represent 95% confidence intervals. **B**. Mean depth per variant averaged across all samples. For visualization, 0.05% of all SNPs were excluded that had a coverage >100x. **C**. Mean depth per individual averaged across all variants.

APPENDIX A. SUPPLEMENTARY MATERIALS FOR CHAPTER 2 "FROZEN IN TIME: RANGEWIDE
GENOMIC DIVERSITY, STRUCTURE, AND DEMOGRAPHIC HISTORY OF RELICT AMERICAN CHESTNUT

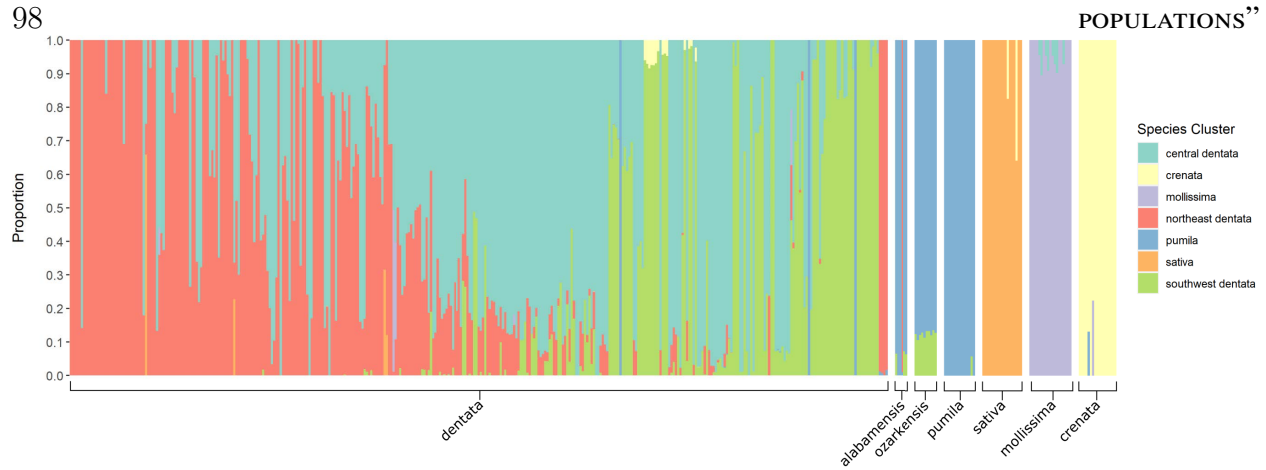


Figure A.3: ADMIXTURE analysis suggests that widespread hybridization does not occur within wild American chestnut populations. ADMIXTURE was performed with 384 American chestnut and 92 *Castanea* species samples. Each color bar represents the proportion of inferred species ancestry for a sample. $K = 7$ was determined to be the most appropriate cluster value.

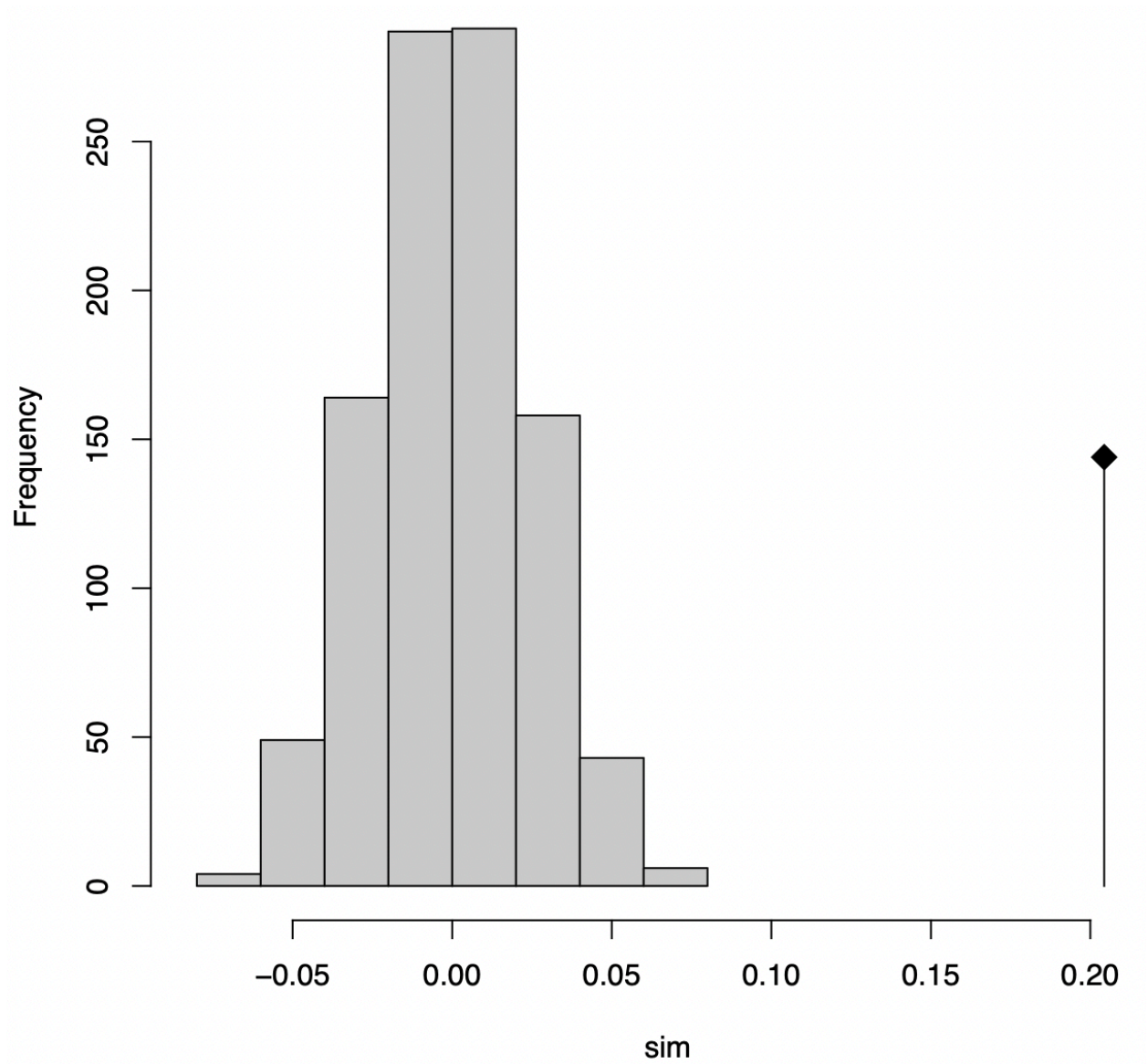


Figure A.4: The American chestnut population structure is characterized by an isolation-by-distance model. The vertical black line represents the observed 0.2 value that was significantly higher than 0 ($p < 0.001$).

APPENDIX A. SUPPLEMENTARY MATERIALS FOR CHAPTER 2 "FROZEN IN TIME: RANGEWIDE GENOMIC DIVERSITY, STRUCTURE, AND DEMOGRAPHIC HISTORY OF RELICT AMERICAN CHESTNUT POPULATIONS"

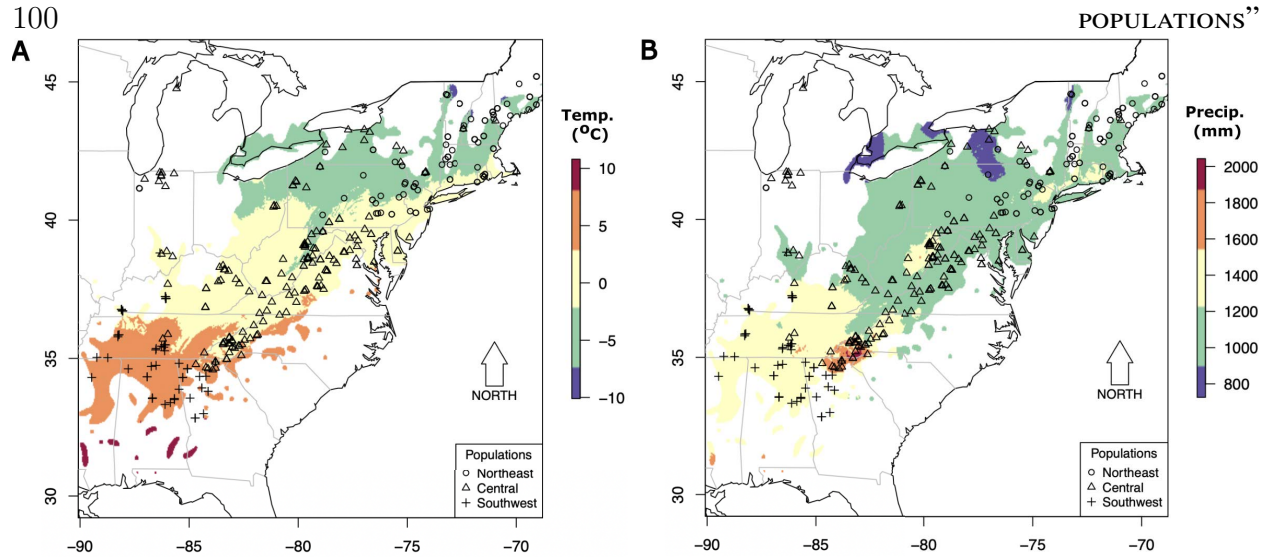


Figure A.5: The American chestnut population boundaries follow precipitation and temperature gradients of the eastern U.S. The different shapes represent the American chestnut sample locations and population membership. **A.** Mean temperature of coldest quarter (WorldClim BIO11 at 2.5m resolution). **B.** Annual precipitation (WorldClim BIO12 at 2.5m resolution).

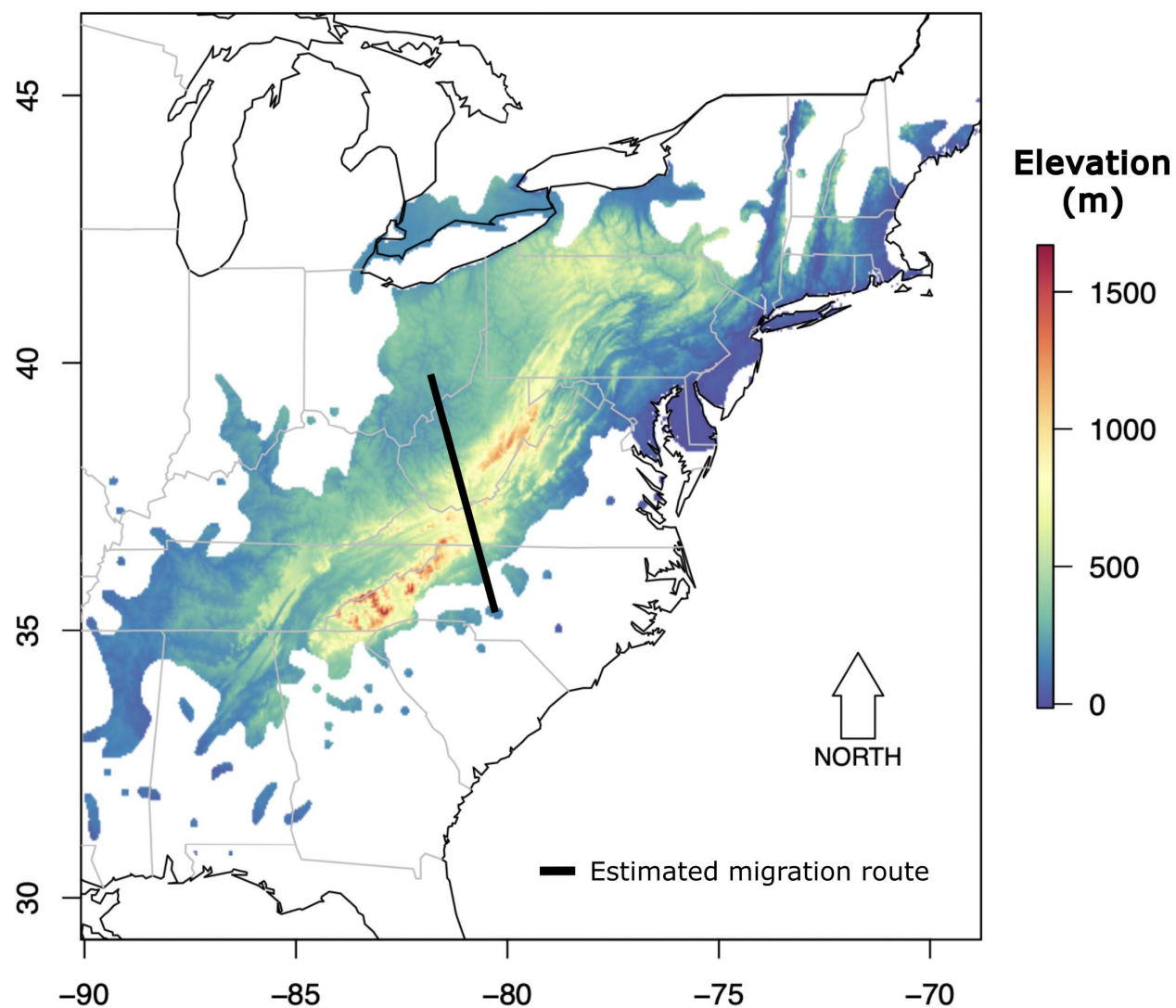


Figure A.6: Elevation above sea level throughout the American chestnut natural range. The elevation data was retrieved from WorldClim 2.1 with 2.5m resolution. The black line represents the American chestnut estimated migration path across the Appalachian Mountain range from the EEMS analysis.

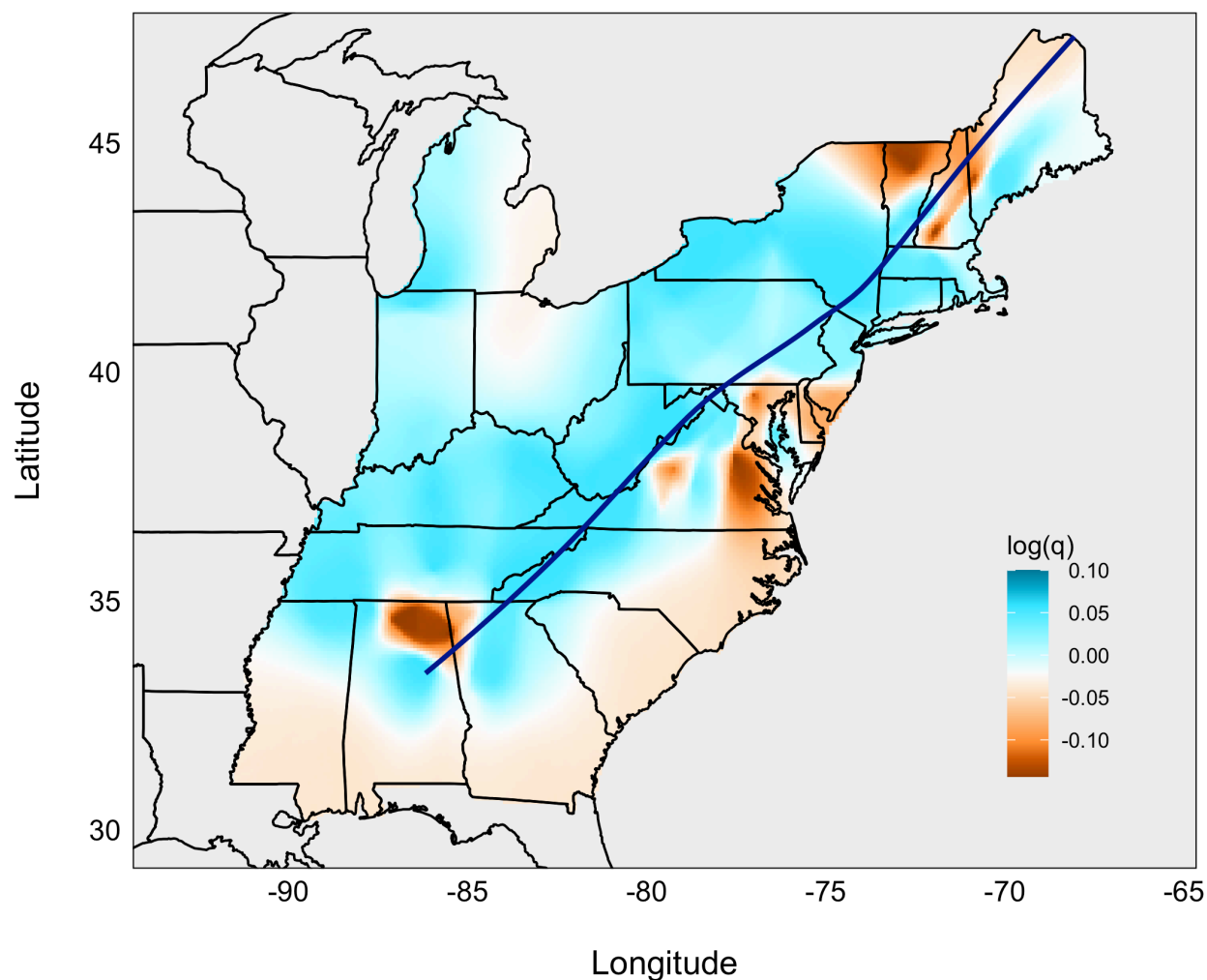


Figure A.7: Estimated genetic diversity within American chestnut throughout the eastern United States. The orange regions represent areas of lower than expected diversity, and the blue regions represent areas of higher than expected diversity. A loess line of Appalachian Mountain peaks was applied in *ggplot2* using peak locations obtained from https://en.wikipedia.org/wiki/List_of_mountains_of_the_Appalachians.

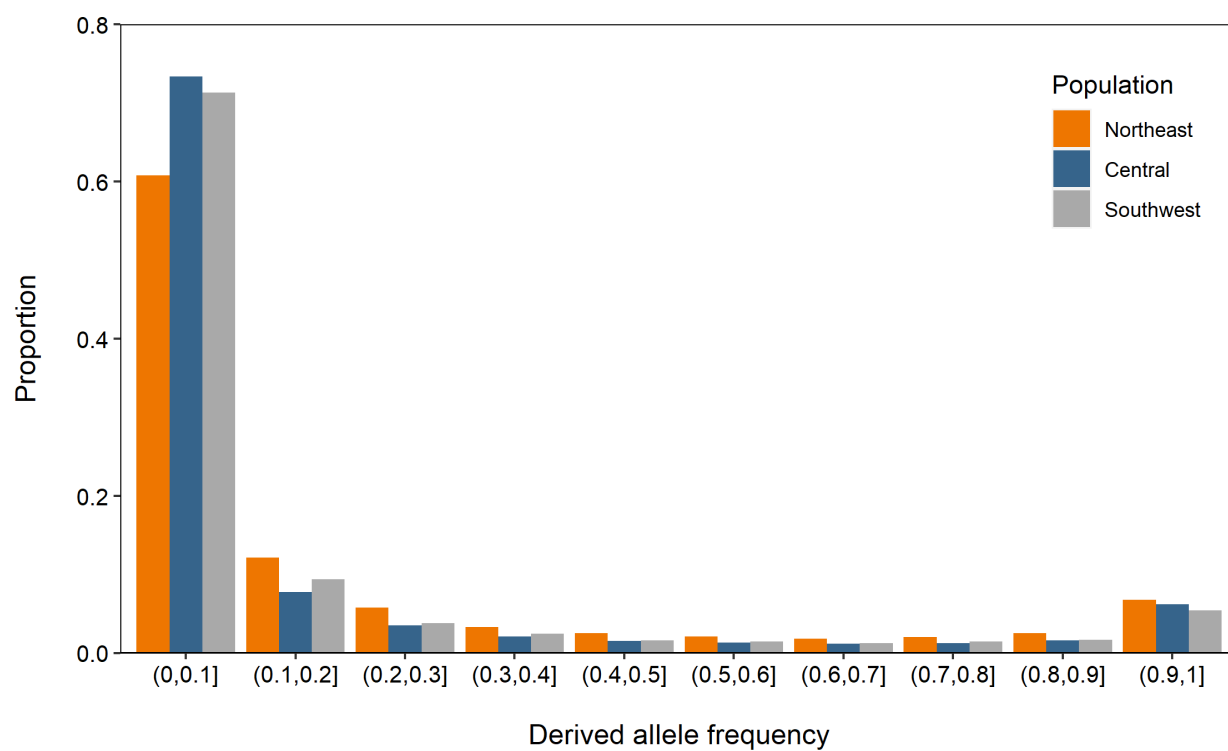


Figure A.8: Site-frequency-spectrum for each of the three wild American chestnut populations.

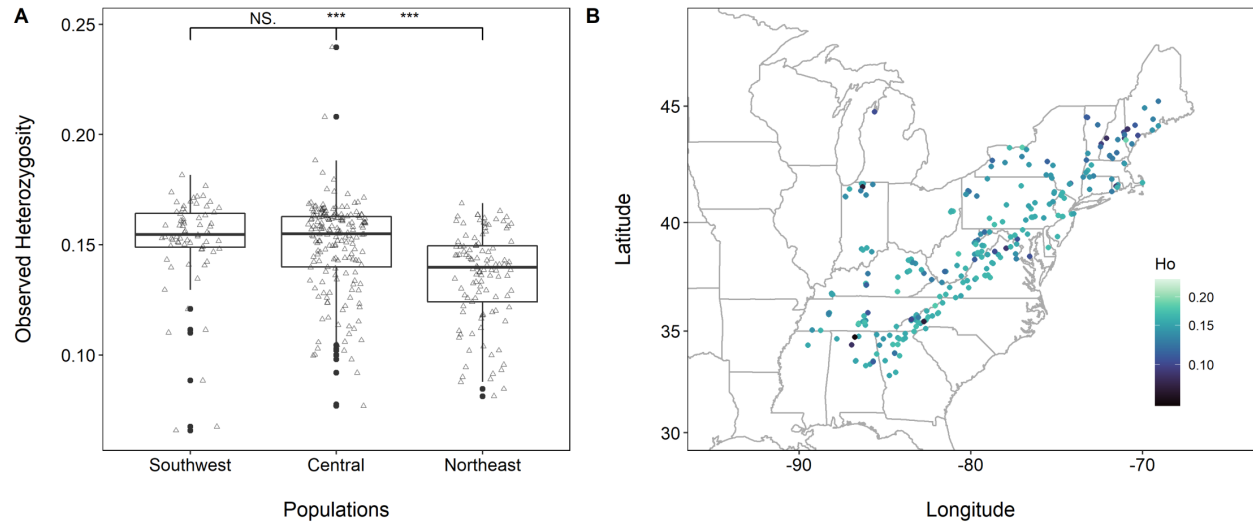


Figure A.9: Observed heterozygosity decreases as latitude increases between the three *Castanea dentata* populations. **A.** The northeast population has significantly less observed heterozygosity than the central and southwest populations ($p < 0.001$, $p < 0.001$). The grey triangles represent the observed heterozygosity of individual trees within each population. **B.** Map of sample locations and observed heterozygosity for each sample. Each circle represents a tree sample location and the color is the observed heterozygosity ratio.



Figure A.10: GO annotations for all RAiSD outlier genes for each of the three American chestnut populations. The gene percentage was calculated by dividing the gene count for each category by its corresponding N value.

Table A.1: GO enrichment analysis of the unique RAiSD outlier gene set for the southwest American chestnut population. The GO enrichment terms were identified as being significantly overrepresented for the GO biological process ($p < 0.05$).

GO ID	GO Description	Gene Count	Expected Count	Fold Enrichment	P-value
GO:0007166	cell surface receptor signaling pathway	20	0.38	53.28	1.64E-27
GO:0050832	defense response to fungus	10	1.6	6.25	1.90E-02
GO:0006468	protein phosphorylation	37	6.32	5.85	1.52E-15
GO:0009620	response to fungus	12	2.19	5.48	8.68E-06
GO:0016310	phosphorylation	38	6.94	5.47	4.31E-14
GO:0007165	signal transduction	46	8.58	5.36	1.70E-16
GO:0023052	signaling	47	8.82	5.33	7.32E-16
GO:0007154	cell communication	51	10.39	4.91	3.12E-15
GO:0006955	immune response	18	4.93	3.65	8.69E-06
GO:0002376	immune system process	18	5	3.60	1.00E-02
GO:0006796	phosphate-containing compound metabolic process	38	10.83	3.51	3.49E-03
GO:0098542	defense response to other organism	17	4.89	3.47	3.12E-02
GO:0006793	phosphorus metabolic process	38	11.11	3.42	7.22E-05
GO:0051716	cellular response to stimulus	53	15.87	3.34	5.62E-15
GO:0043207	response to external biotic stimulus	21	6.78	3.10	1.57E-03
GO:0051707	response to other organism	21	6.78	3.10	1.57E-03
GO:0009607	response to biotic stimulus	21	6.79	3.09	1.61E-03
GO:0044419	biological process involved in interspecies interaction between organisms	21	6.86	3.06	1.85E-03
GO:0036211	protein modification process	42	15.32	2.74	4.43E-06
GO:0006464	cellular protein modification process	42	15.32	2.74	4.43E-06
GO:0043412	macromolecule modification	42	17.32	2.42	1.76E-04
GO:0044267	cellular protein metabolic process	43	20.83	2.06	9.14E-05
GO:0050896	response to stimulus	69	33.69	2.05	2.24E-05
GO:0019538	protein metabolic process	44	21.77	2.02	1.33E-03
GO:0050794	regulation of cellular process	55	29.16	1.89	5.08E-05

Table A.2: GO enrichment analysis of the unique RAiSD outlier gene set for the northeast American chestnut population. The GO enrichment terms were identified as being significantly overrepresented for the GO biological process ($p < 0.05$).

GO ID	GO Description	Gene Count	Expected Count	Fold Enrichment	P-value
GO:1900459	positive regulation of brassinosteroid mediated signaling pathway	4	0.06	69	3.51E-03
GO:0007166	cell surface receptor signaling pathway	8	0.44	18.1	1.02E-04
GO:0002376	immune system process	24	5.89	4.07	2.72E-05
GO:0006955	immune response	23	5.8	3.96	9.27E-05
GO:0098542	defense response to other organism	22	5.76	3.82	3.51E-04
GO:0007165	signal transduction	35	10.1	3.46	5.36E-07
GO:0023052	signaling	35	10.38	3.37	1.10E-06
GO:0007154	cell communication	41	12.23	3.35	3.02E-08
GO:0006952	defense response	22	7.28	3.02	1.53E-02
GO:0051716	cellular response to stimulus	46	18.69	2.46	3.45E-05

Table A.3: GO enrichment analysis of the unique RAIiSD outlier gene set for the central American chestnut population. The GO enrichment terms were identified as being significantly overrepresented for the GO biological process ($p < 0.05$).

GO ID	GO Description	Gene Count	Expected Count	Fold Enrichment	P-value
GO:0050896	response to stimulus	84	54.02	1.56	4.89E-02

Table A.4: GO enrichment analysis of the shared RAiSD outlier gene set for three American chestnut populations. The GO enrichment terms were identified as being significantly underrepresented or overrepresented for the GO biological process ($p < 0.05$).

GO ID	GO Description	Gene Count	Expected Count	Fold Enrichment	P-value
GO:0050832	defense response to fungus	5	0.3	16.5	3.73E-02
GO:0008150	biological_process	13	24.72	0.53	2.64E-02

Appendix B

Supplementary Materials for Chapter 3 "Identifying and conserving the adaptive legacy of American chestnut"

Table B.1: The 32 climate variables from ClimateNA that were used for this study and their respective descriptions.

Abbreviation	Description
Annual variables directly calculated from monthly variables	
MAT	mean annual temperature (°C)
MWMT	mean warmest month temperature (°C)
MCMT	mean coldest month temperature (°C)
TD	temperature difference between MWMT and MCMT, or continentality (°C)
MAP	mean annual precipitation (mm)
AHM	annual heat-moisture index $((MAT+10)/(MAP/1000))$
SHM	summer heat-moisture index $((MWMT)/(MSP/1000))$
Annual variables derived from monthly variables	
DD_0	degree-days below 0°C, chilling degree-days
DD5	degree-days above 5°C, growing degree-days
DD_18	degree-days below 18°C, cooling degree-days
DD18	degree-days above 18°C, heating degree-days
NFFD	the number of frost-free days
FFP	frost-free period
bFFP	the day of the year on which FFP begins
eFFP	the day of the year on which FFP ends
PAS	precipitation as snow (mm). For individual years, it covers the period between August in the previous year and July in the current year
EMT	extreme minimum temperature over 30 years (°C)
EXT	extreme maximum temperature over 30 years (°C)
Eref	Hargreaves reference evaporation (mm)
CMD	Hargreaves climatic moisture deficit (mm)
MAR	mean annual solar radiation (MJ m ⁻² d ⁻¹)
RH	mean annual relative humidity (%)
CMI	Hogg's climate moisture index (mm)
DD1040	degree-days above 10°C and below 40°C
Seasonal variables directly calculated from monthly variables	
Tave_wt	winter mean temperature (°C)
Tave_sp	spring mean temperature (°C)
Tave_sm	summer mean temperature (°C)
Tave_at	autumn mean temperature (°C)
PPT_wt	winter precipitation (mm)
PPT_sp	spring precipitation (mm)
PPT_sm	summer precipitation (mm)
PPT_at	autumn precipitation (mm)

Table B.2: The environmental variable loadings on the three PC axes used for the LFMM analysis.

	PC1	PC2	PC3
RH	-0.048	-0.034	0.937
Eref	0.378	0.324	-0.111
PPT_wt	0.416	-0.116	0.053
EXT	0.039	0.472	-0.073
TD	-0.437	0.01	-0.134
MAR	0.443	0.006	0.027
PPT_sm	0.351	-0.256	-0.156
CMI	0.132	-0.498	0.111
AHM	-0.119	0.487	0.156
DD_18	-0.373	-0.326	-0.148

Table B.3: Number of single-nucleotide polymorphisms (SNPs) in coding and noncoding features. Three datasets were evaluated: **Adaptive** (18,483 putatively adaptive SNPs from the GEA analyses), **Random** (randomly selected SNPs with allele frequency distribution matched to the adaptive set), and **Full** (original filtered SNP dataset from Sandercock et al. [104] with MAF<0.05 filter). The promoter region includes the 2kb region upstream of the mRNA intervals. Percentage of the total SNPs for each feature are shown in parentheses.

	Adaptive	Random	Full
Number of SNPs in CDS	264 (1.43%)	619 (3.35%)	377161 (3.27%)
Number of SNPs in exon	435 (2.35%)	991 (5.36%)	611402 (5.30%)
Number of SNPs in five_prime_UTR	69 (0.37%)	141 (0.76%)	89577 (0.78%)
Number of SNPs in three_prime_UTR	104 (0.56%)	236 (1.28%)	150477 (1.31%)
Number of SNPs in mRNA	2120 (11.47%)	3462 (18.73%)	2102039 (18.24%)
Number of SNPs in promoter	1309 (7.08%)	1830 (9.90%)	1154261 (10.01%)
Number of SNPs in gene	2120 (11.47%)	3462 (18.73%)	2102039 (18.24%)
Number of SNPs in intron	1712 (9.26%)	2537 (13.73%)	1528332 (13.26%)
Total SNPs in dataset	18483	18483	11526713

Table B.4: Estimated number of samples to capture adaptive diversity from each seed zone for a two seed zone model.

Seed zone	% variance explained	# of trees to sample	95% CI
1	90%	16.73	(15.69, 17.77)
2	90%	11.05	(10.13, 11.97)
1	95%	34.12	(31.90, 36.34)
2	95%	21.91	(20.11, 23.71)
1	99%	167.56	(156.37, 178.75)
2	99%	107.37	(100.20, 114.54)

Table B.5: Estimated number of samples to capture adaptive diversity under a single seed zone model. The estimate for the 99% variance explained did not complete due to excessive run time.

Seed zone	% variance explained	# of trees to sample	95% CI
1	90%	26.86	(24.55, 29.17)
1	95%	55.89	(51.01, 60.77)
1	99%	DNF	DNF

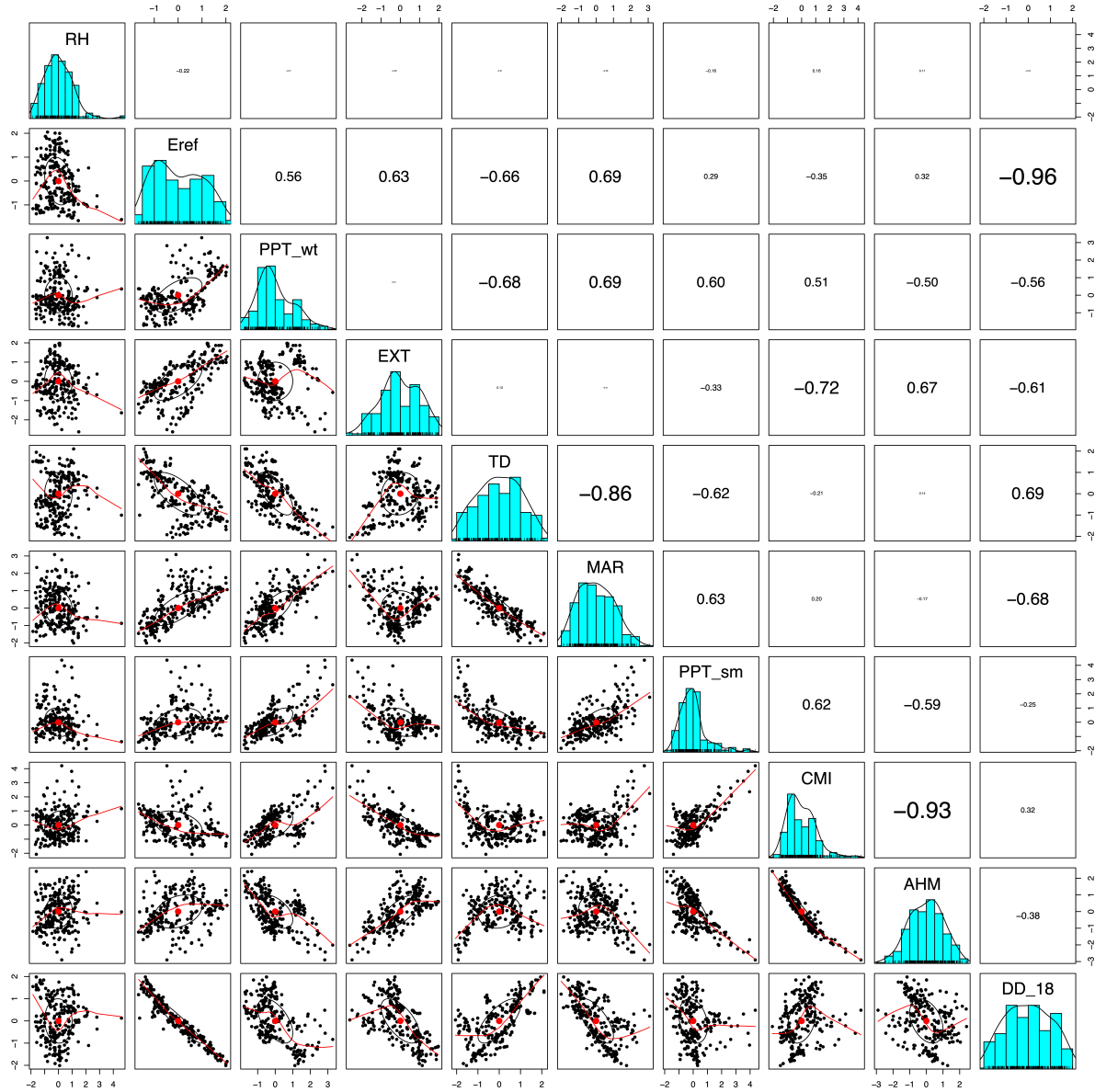


Figure B.1: Ten uncorrelated climate variables were selected for use in the genotype-environment association analyses. Pairs plot of the selected 10 environmental variables. The bottom half are regression plots and the top half are Pearson's correlations.

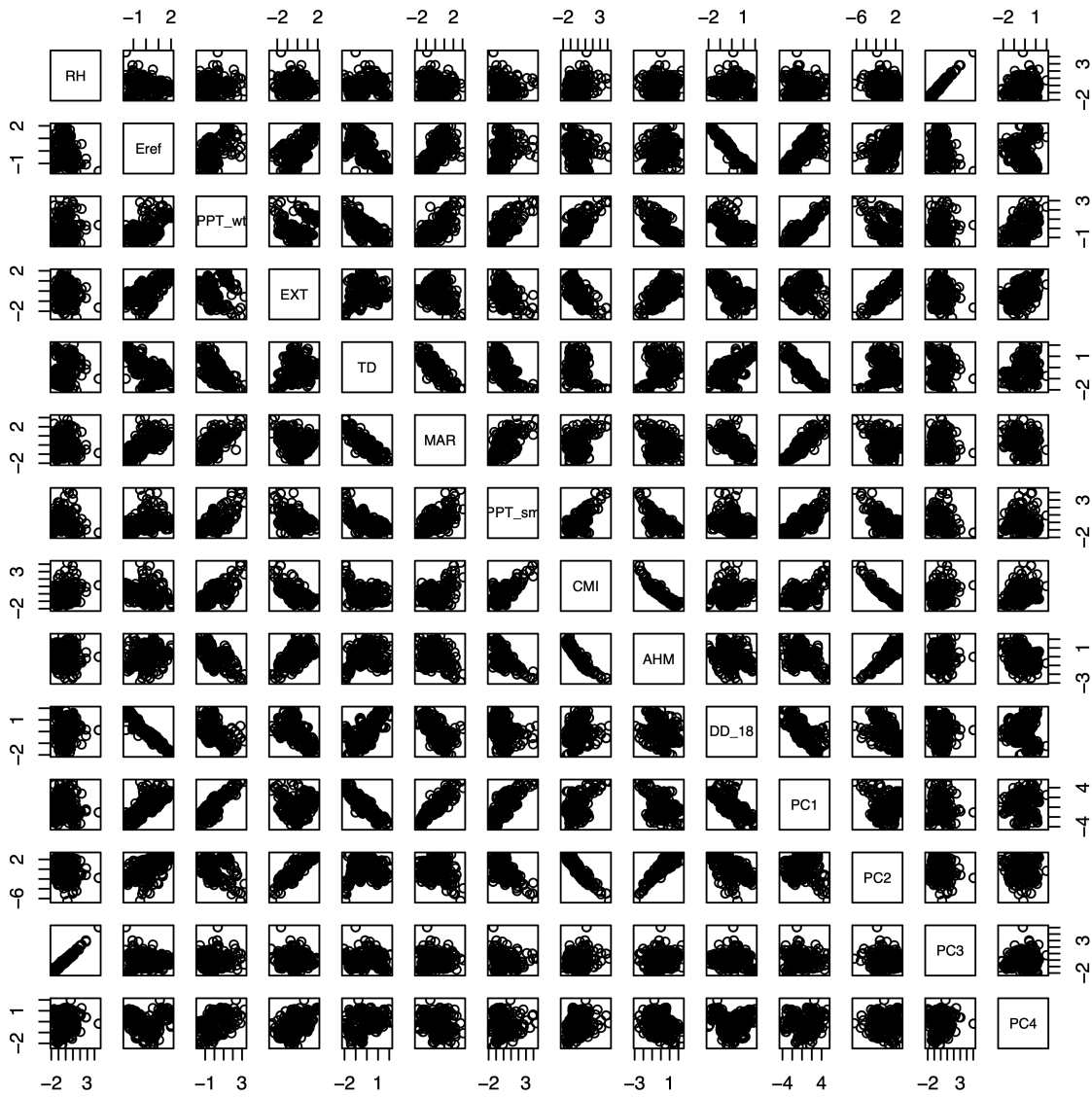


Figure B.2: Three PC axes were used as synthetic variables for the LFMM genotype-environment association analyses. Pairs plot of the selected 10 environmental variables and the first four PCs from the PCA of the 10 climate variables. The bottom half and top half are linear regressions between the variables.

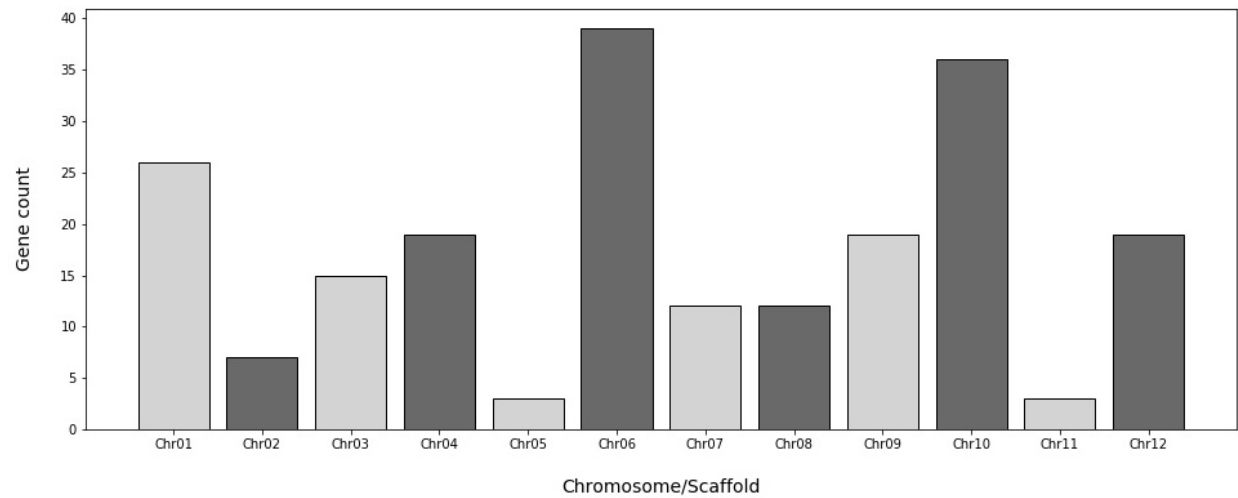


Figure B.3: The number of genes that contained at least one adaptive loci within each *Castanea dentata* chromosome. Scaffolds that did not contain genes with adaptive loci were omitted from the figure.

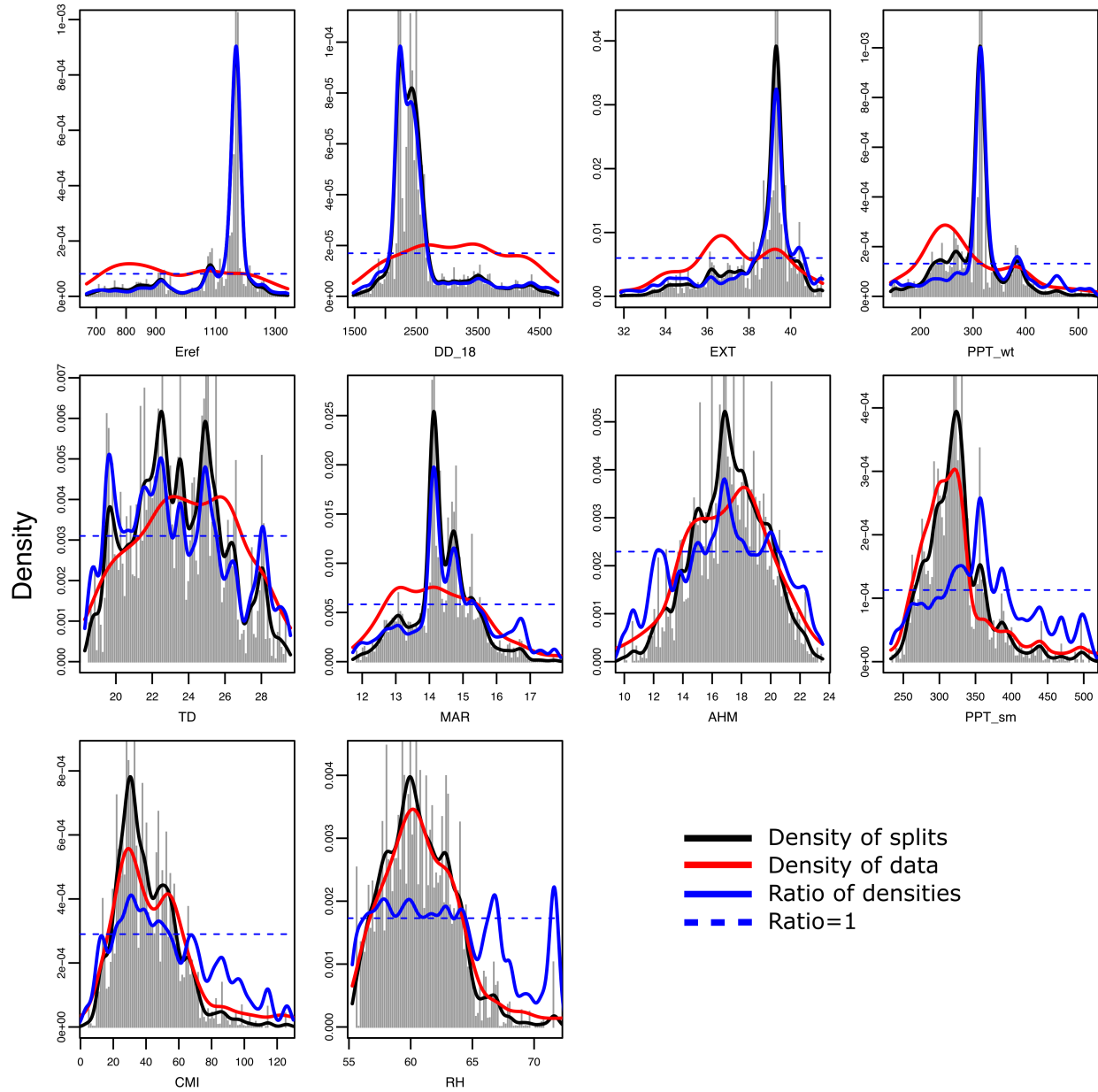


Figure B.4: Gradient Forest adaptive allelic turnover for each of the ten climate variables and 18,483 putatively adaptive loci. Spikes in the binned split importance reveal where changes in adaptive allele density are occurring along each climate gradient.

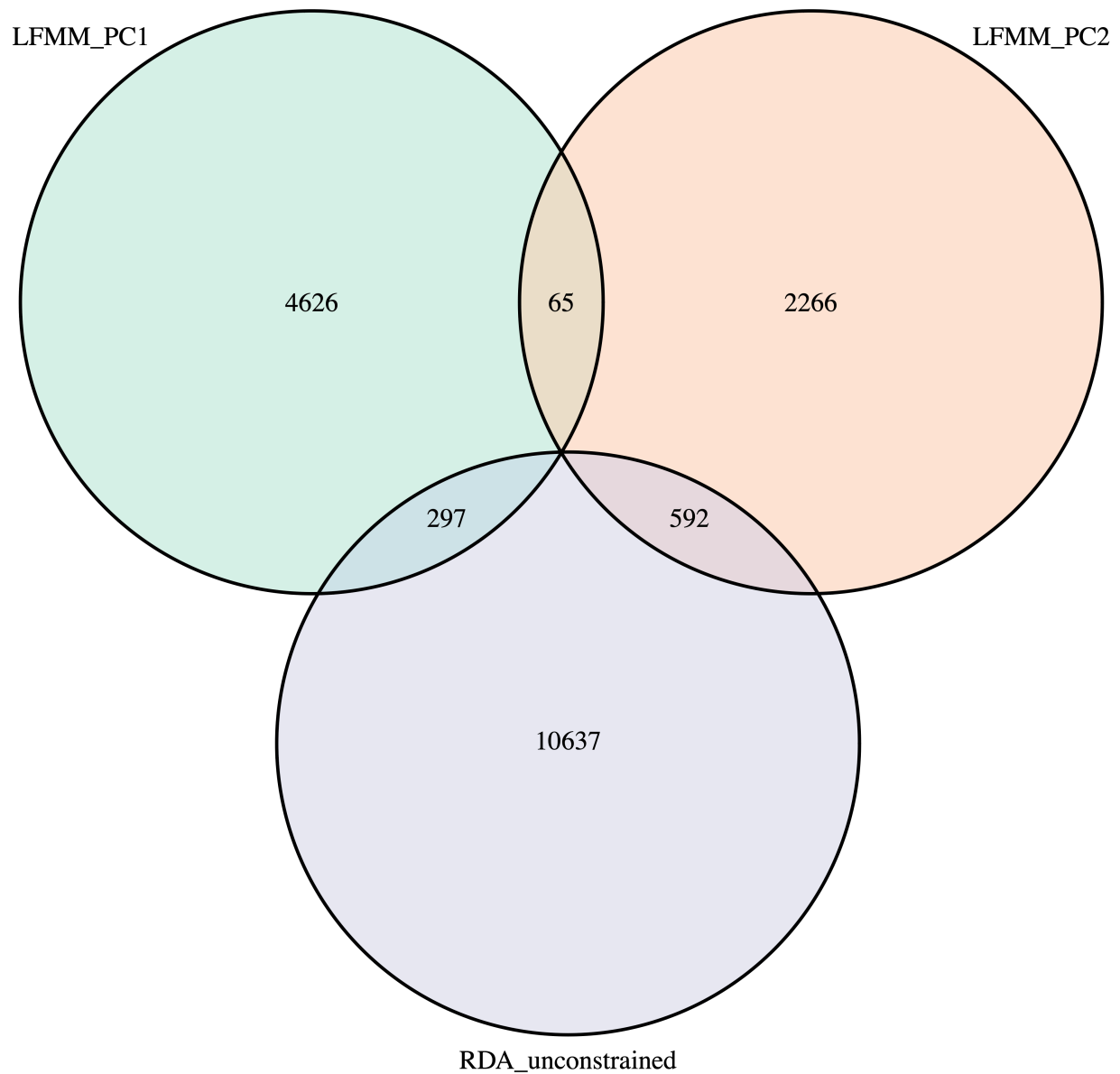


Figure B.5: Venn diagram for the 18,483 adaptive loci identified using RDA and LFMM. For LFMM, only the SNPs from the PC1 and PC2 analyses are shown due to the PC3 analysis finding zero outlier loci.

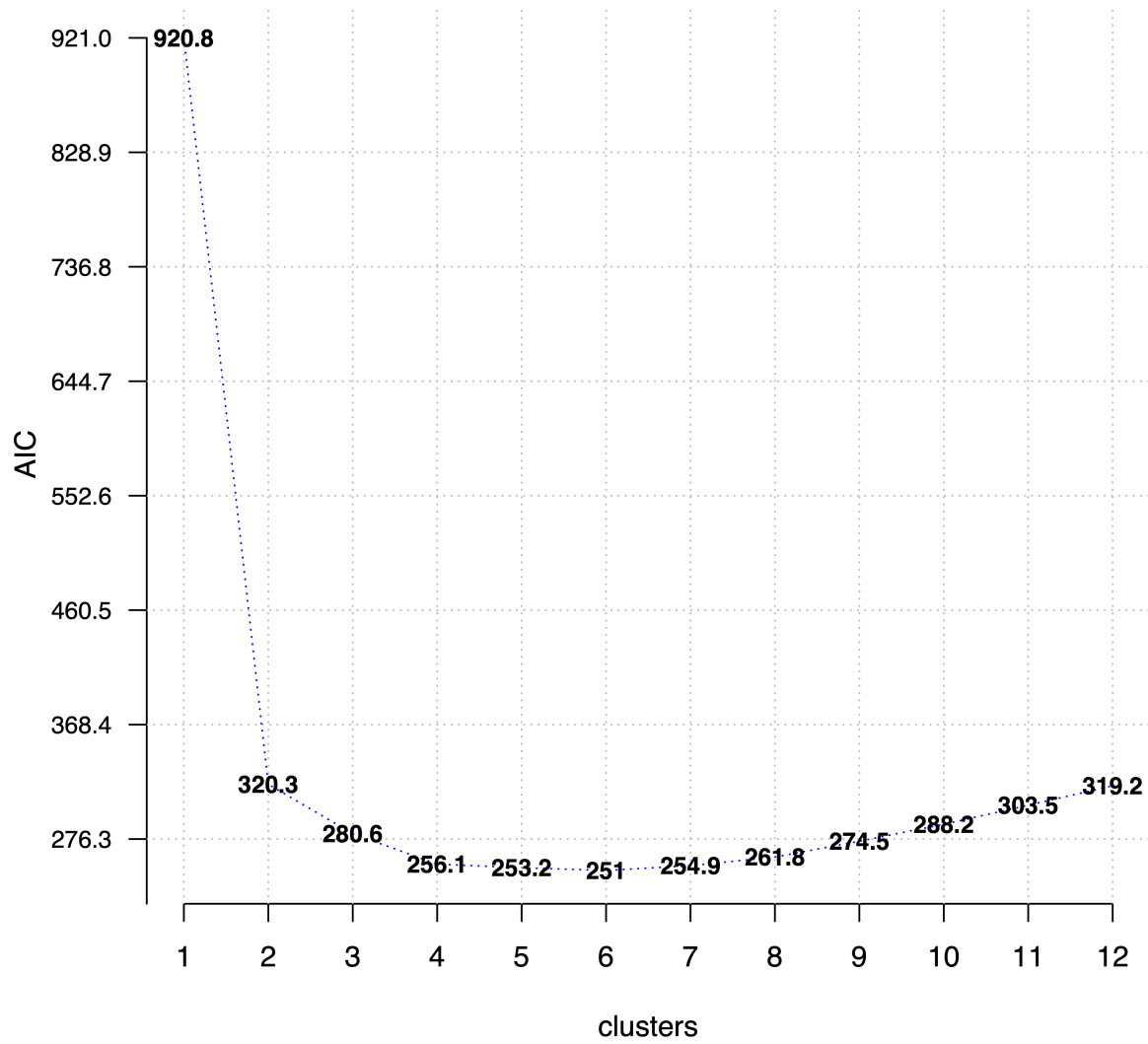


Figure B.6: AIC score evaluation to determine optimal number of seed zones.

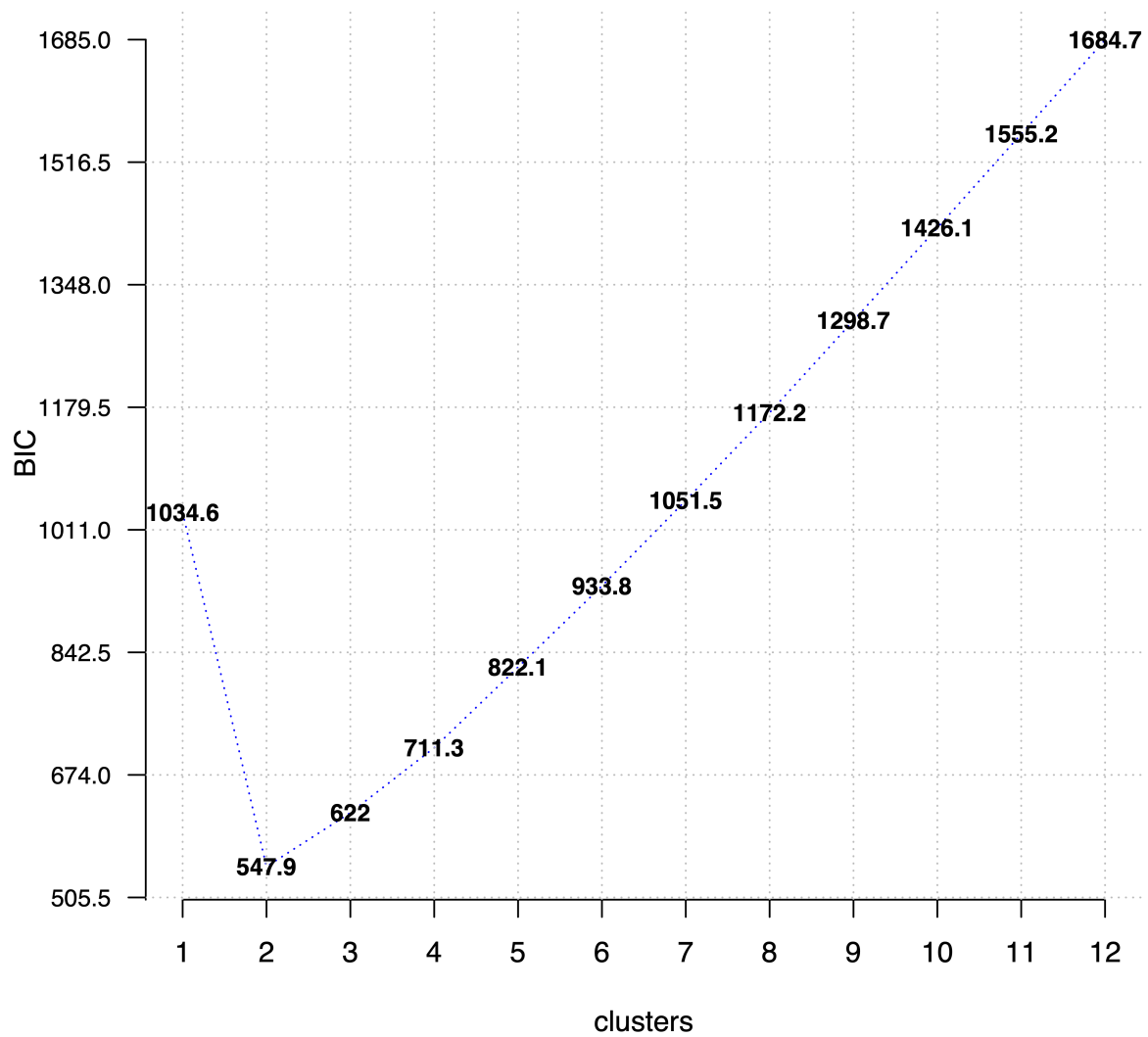


Figure B.7: BIC score evaluation to determine optimal number of seed zones.

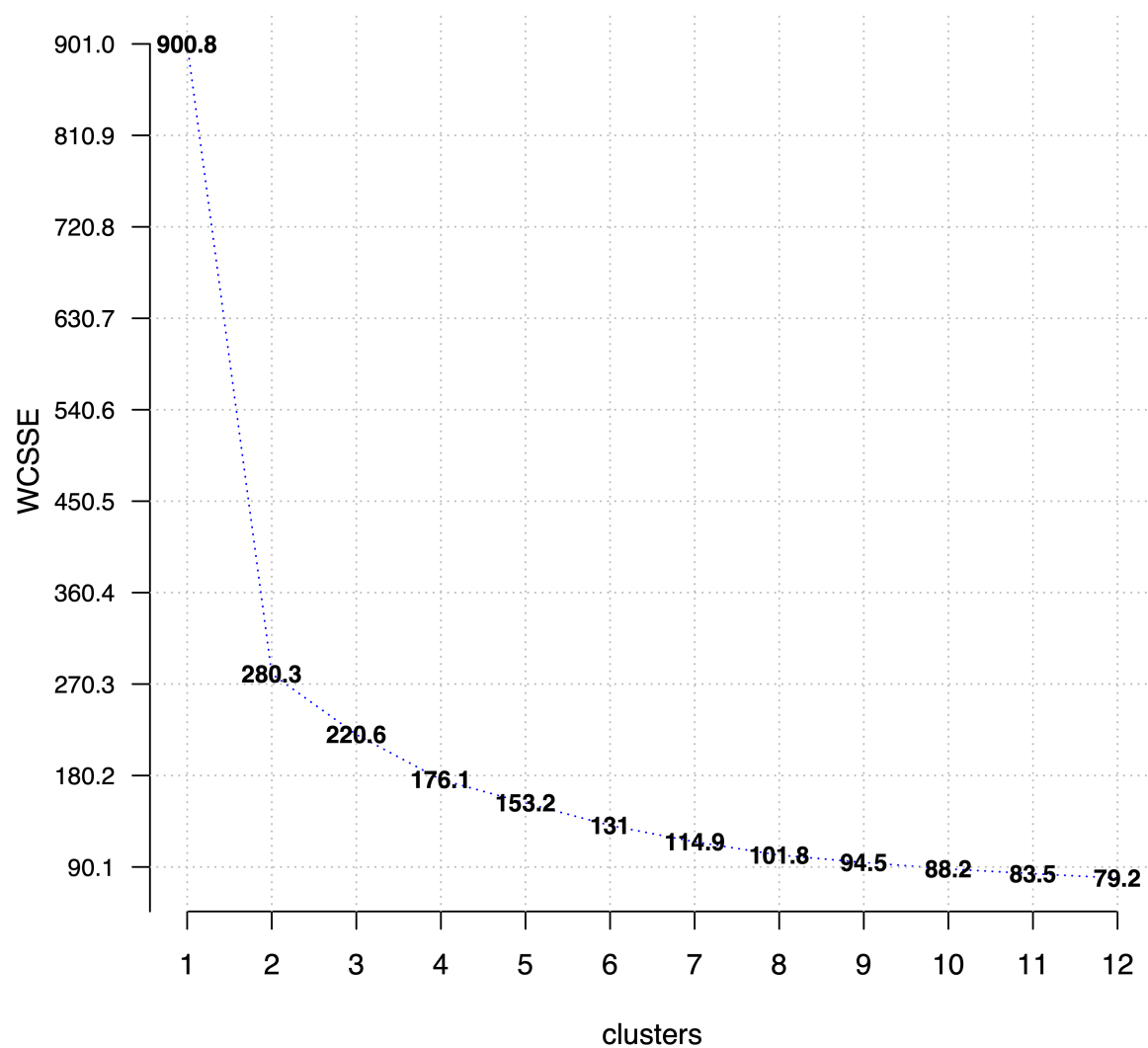


Figure B.8: WCSSE score evaluation to determine optimal number of seed zones.

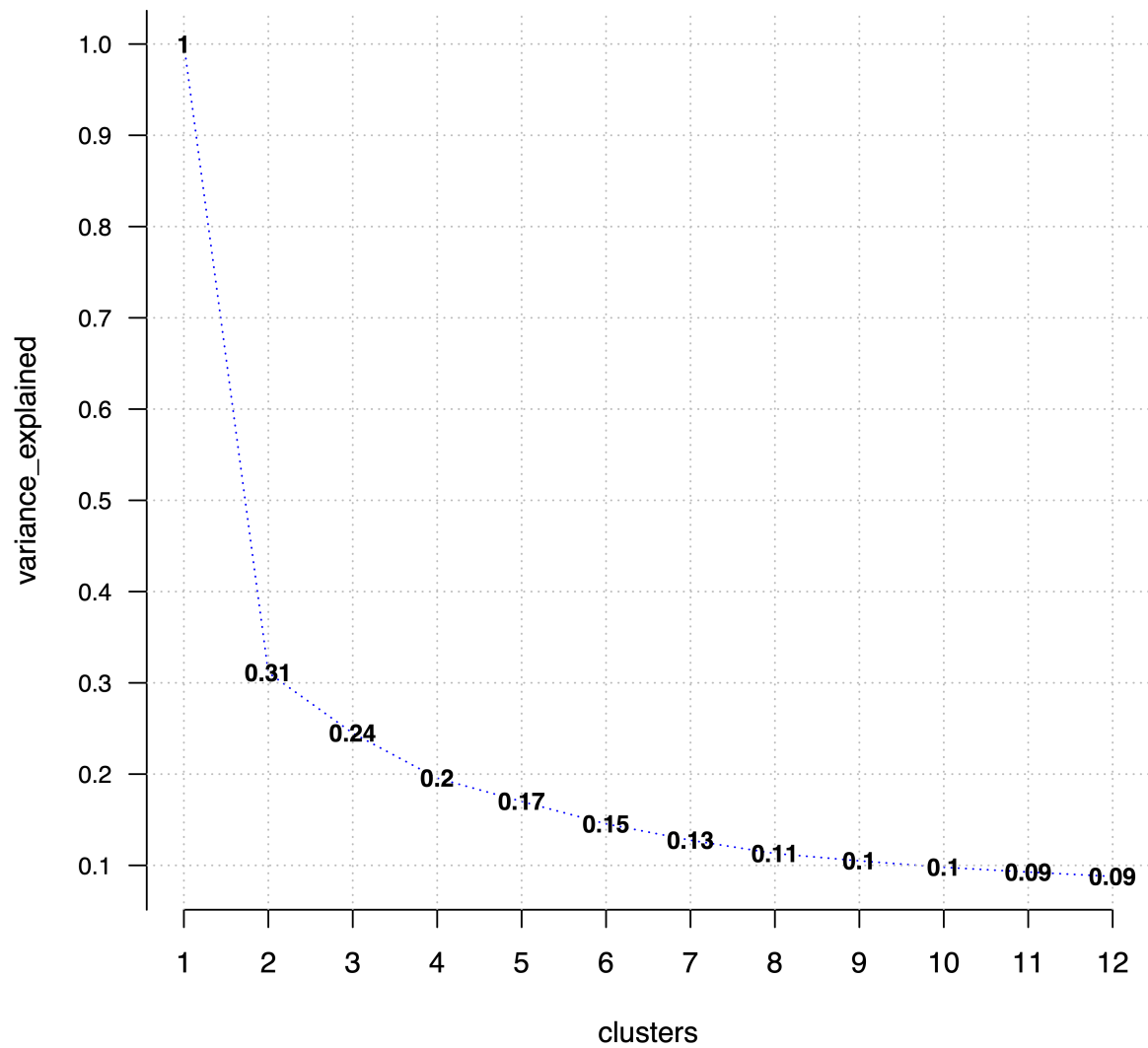


Figure B.9: Variance explained scores to determine optimal number of seed zones.