



VCU

Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations


Graduate School

2022

Quantifying contributions of climate, geography, and gene flow to divergence: a case study for three North American pines

Constance E. Bolte
Virginia Commonwealth University

Follow this and additional works at: <https://scholarscompass.vcu.edu/etd>

 Part of the [Botany Commons](#), [Ecology and Evolutionary Biology Commons](#), and the [Genomics Commons](#)

© Constance E. Bolte

Downloaded from

<https://scholarscompass.vcu.edu/etd/7110>

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

© Constance Ellen Bolte 2022

All Rights Reserved

**Quantifying contributions of climate, geography, and gene flow to
divergence: a case study for three North American pines**

By Constance Ellen Bolte (PhD, MS, MT)

A Dissertation

Submitted in Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy in

Integrative Life Sciences

Virginia Commonwealth University

Richmond

June, 2022

Andrew J. Eckert, PhD. Chair

Rodney J. Dyer, PhD

Christopher M. Gough, PhD

Catherine M. Hulshof, PhD

Daniel J. McGarvey, PhD

Peter E. Smouse, PhD

Acknowledgements

This dissertation is dedicated to my son, Miles. Through discoveries of how to be the best mom for you, there existed a profound recognition of my capacity to overcome obstacles, embrace challenges, and accept change. Indeed, the only constant in this world is change, and you continue to change me for the better. To my husband, Wouter, thank you for your unwavering support. This dissertation, along with our son's growth and happiness, is our joint achievement. To my parents, thank you for instilling in me a 'can do' attitude. While my invincibility drove you mad at times, you helped me channel that toward scholastics. Thank you for being hands-on parents. Your early attention to my well-being helped foster self-confidence. To my advisor, Dr. Andrew Eckert, thank you for your mentorship. You held me to high academic standards but allowed flexibility so I could maintain work-life balance. In many cases, I felt like one of your science peers. We were working toward a common goal, good science. Your flexibility was interpreted as trust and very important to my success. The same sentiment can be extended to my committee members. Our discussions were highlights during my tenure as a PhD student and candidate. I look forward to any opportunity to collaborate with you again. To my friends and family who helped keep things light and full of laughter, thank you. Laughter is the best medicine. It has healed my heart and kept me sane countless times during graduate school.

Table of Contents

List of Tables.....	v
List of Figures.....	vii
Abstract.....	xii
Introduction.....	1
Chapter 1: Divergence amid recurring gene flow: the complex demographic histories inferred for <i>Pinus pungens</i> and <i>P. rigida</i> align with a growing expectation for forest trees.....	11
Appendix 1.....	62
Chapter 2: Potential drivers in the differential development of reproductive isolation for three cryptically related North American pine species (<i>Pinus pungens</i> , <i>P. rigida</i> , and <i>P. taeda</i>).....	72
Appendix 2.....	123
Chapter 3: The extent of genetic diversity and hybridization within sympatric stands of two closely related pine species (<i>Pinus pungens</i> and <i>P. rigida</i>) in the southern Appalachian Mountains.....	135
Appendix 3.....	166

List of Tables

Table 1.1 Location of sampled populations, number of trees (n) that were sampled, and the observed heterozygosity (H_o) versus the expected heterozygosity ($H_e = 2pq$) for *Pinus pungens* and *P. rigida* populations.

Table 1.2 Summary statistics of genetic differentiation for the sampled populations of *P. rigida* and *P. pungens*. Expected (H_e) and observed heterozygosity (H_o) values are the averages across 2168 SNPs averaged across populations.

Table 1.3 Results of model fitting for thirteen representative demographic models of divergence. Models are ranked by the number of parameters (k). Log-likelihood ($\log L$) and Akaike information criterion (AIC) are provided for each model. Model details are given in the footnote.

Table 1.S1 Parameters and the estimates associated with the best run of each model type. The model with three time intervals (PSCMIGCsT3) is not included in this table. Those parameters are summarized in Table 1.S2.

Table 1.S2 Parameter estimates from the best run (lowest AIC) for the model allowing 3 time intervals (PSCMIGCsT3).

Table 2.1. Summaries redundancy analyses with climate and geographic as predictors of genetic variation. Adjusted r^2 represents the individual contribution of the predictor with all others removed and the proportion of variance explained (PVE) represents the overall contribution without controlling for interactive effects among the predictors. An asterisk denotes model significance ($p < 0.01$).

Table 2.2 Results from Fisher's Exact Tests for seven EggNOG descriptions associated with attributes of the *P. taeda* genome. Descriptions with p -values < 0.5 have an asterisk.

Table 2.S1 Location of sampled populations, number of trees (n) that were sampled for *Pinus pungens* (PU), *P. rigida* (RI), and *P. taeda* (TA) populations. Averaged ancestry assignments (with $K = 3$) for each population are in the last three columns.

Table 2.S2 Parameter estimates and 95% Confidence Intervals (CI) for the two-species models with the lowest AIC scores for each pairwise species inference. Values are unscaled. The eps value in the FIM uncertainty test is the relative step size used when taking numerical derivatives.

Table 2.S3 Summary results from pairwise analysis of F_{ST} across SNPs that hit within genic regions and information extracted from the *P. taeda* annotated genome files (Query Sequence and EggNOG Description) for each match. The lower the blastn e-value the better the match.

Table 3.1 Genetic diversity estimates expected heterozygosity (H_E) and observed heterozygosity (H_O), for each species at each sympatric stand. Estimates of genetic differentiation across species (F_{ST}) at each sympatric stand are also provided.

Table 3.2 Counts of RADtag sequences (i.e., contigs) and how they mapped to the *P. taeda* genome for each F_{ST} and distance category.

Table 3.S1 DIC scores from analysis of structure across four replicate runs (# of chains) of each cluster assignment (K).

Table 3.S2 Summary of BLAST results for RADtag_ID matches to the *P. taeda* (PITA) genome that were within a gene. Annotations based on EggNOG descriptions and GO terms were sourced directly from the annotation file that accompanies the genome on treegenomesdb.org.

Table 3.S3 Summary of BLAST results for RADtag_ID matches to the *P. taeda* (PITA) genome and within 20kbp of a gene. Annotations based on EggNOG descriptions and GO terms were sourced directly from the annotation file that accompanies the genome on treegenomesdb.org.

Table 3.S4 Summary of BLAST results for RADtag_ID matches to the *P. taeda* (PITA) genome and over 20kbp from a gene. Annotations based on EggNOG descriptions and GO terms were sourced directly from the annotation file that accompanies the genome on treegenomesdb.org.

List of Figures

Figure i Conceptualizing factors involved in speciation and the interconnectivity among the factors often considered during investigations related to niche evolution, adaptation, and speciation. This is a modified figure from Bolte & Eckert (2020).

Figure ii Hypotheses related to speciation rate in relation to ecological divergence scenarios. a) Simple, 2-dimensional schematic showing the relationship between the realized niche (i.e., where the species is known to occur), the fundamental niche (i.e., where the species has the capacity to occur) and the hypothesized importance of divergent selection in the time needed for reproductive isolation to develop when all other factors from Figure i are held constant. For the top two diagrams (i.e., stabilizing selection versus directional selection) imagine the niche spaces for two species are stacked on top of each other after completion of reproductive isolation. b) Hypothesized relationship between environmental complexity and speciation rate. Open circles meet expectations. Closed circles may have life history traits or genetic architectures that allow deviation from expectations. c) Hypothesized relationship between combined factors of standing genetic variation and environmental complexity on the probability for niche divergence. In environments with low complexity the probability of niche divergence is low regardless of standing genetic variation. In homogeneous environments it is hypothesized that niche stasis or niche directional shifts are more likely to occur than niche divergence. This is a modified figure from Bolte & Eckert (2020).

Figure 1.1 Known geographical distribution of focal species, a) *Pinus pungens* and b) *P. rigida*, (Little 1971) in relation to populations sampled (black dots) for genetic analysis; Phenotypic characterization of each species was illustrated by Pierre-Joseph Redouté (Michaux 1819).

Figure 1.2 Measures of genetic differentiation and diversity among sampled trees of *P. pungens* and *P. rigida*: a) Principal components analysis of 2168 genome-wide single nucleotide polymorphism (SNPs) for *Pinus pungens* (blue, left side of PC1) and *P. rigida* (orange, right side of PC1); b) log-likelihood values across ten replicate runs in fastSTRUCTURE for $K = 2$ through $K = 7$; c) results of averaged $K = 2$ ancestry (Q) assignments for each sample arranged latitudinally in each species.

Figure 1.3 Redundancy analysis (RDA) of the multilocus genotypes for each tree with climate and geographic predictor variables (full model). Direction and length of arrows on each RDA plot correspond to the loadings of each variable.

Figure 1.4 Hypotheses associated with each SDM - GCM model prediction versus the ensemble SDM prediction based on relative grid cell counts of high habitat suitability (> 0.5) for *P. rigida*, *P. pungens*, and overlap across four time periods (LIG, LGM, HOL, and PD). Bolded text were statements supported by the best-fit model of demographic inference.

Figure 1.5 The best-fit model (PSCMIGCs) and unscaled parameter estimates from $\partial\alpha\partial i$ analysis. Time intervals (T_i) are represented in millions of years and associated with lineage population sizes (N_i) and a specific rate of symmetrical gene flow (M_i).

Figure 1.S1 The thirteen divergence scenarios tested within the program $\partial\alpha\partial i$.

Figure 1.S2 Principal component analysis (PCA) of 300 *P. rigida* and *P. pungens* trees labeled by population assignment.

Figure 1.S3 Individual based assignments of admixture from analysis of *fastSTRUCTURE* for $K = 2$ through $K = 7$. The plot associated with each value of K represents the average assignments for each individual across 10 replicate runs.

Figure 1.S4 Distribution of missing data across the sampled trees in relation to ancestral coefficients (from $K = 2$). Blue circles to the right are samples of *P. pungens*. Orange circles to the left are samples of *P. rigida*.

Figure 1.S5 Species distribution model (SDM) predictions across four time points for *P. pungens* and *P. rigida*. Measures of raster overlap in terms of Schoener's D and Warren's I index between the models of each species, and at each time point, are presented in the bottom right corner of the prediction plots for *P. rigida*. Venn diagrams illustrate the number of grid cells with moderate to high habitat suitability scores (> 0.5) for each SDM at a given time point, as well as the number of shared, or overlapping, grid cells. Blue Venn diagram ovals show grid cell counts from the *P. pungens* SDM, and orange Venn diagram ovals show grid cell counts from the *P. rigida* SDM for the aligning time point (denoted on the left side). Habitat suitability distributions for LGM and HOL depict ensembled predictions. Glacial extent data (labeled ice in LGM plots) for 18 kya was provided by Dyke (2003).

Figure 1.S6 Last Glacial Maximum (LGM, ~21 kya) model predictions from each GCM (CCSM4, MIROC-ESM, and MPI-ESM).

Figure 1.S7 Mid-Holocene (~6 kya) model predictions from each GCM (CCSM4, MIROC-ESM, and MPI-ESM).

Figure 1.S8 Presentation of data-model fit to the PSCMIGCs model run with highest log likelihood.

Figure 2.1 Known geographical distribution of focal species, a) *P. pungens*, b) *P. rigida*, c) *P. taeda* (Little 1971) in relation to populations sampled (black dots) for genetic analysis.

Figure 2.2 Measures of genetic differentiation and diversity among sampled trees of *P. pungens*, *P. rigida*, and *P. taeda*: a) Principal components analysis of 5051 genome-wide single nucleotide polymorphism (SNPs) for *Pinus pungens* (blue, right side of PC1), *P. rigida* (orange, right side of PC1), and *P. taeda* (green, left side of PC1); b) log-likelihood

values across ten replicate runs in fastSTRUCTURE for $K = 3$ through $K = 7$; c) results of averaged $K = 3$ ancestry (Q) assignments for each sample arranged by population name in Table 2.S1.

Figure 2.3 Redundancy analysis (RDA) of the multilocus genotypes for each tree with a) climate and geographic predictor variables (full model), b) climate predictor variables (geography removed), and c) geographic predictor variables (climate removed). Panels d-e present redundancy analysis of the ancestral coefficients from structure analysis ($K = 3$) for each tree with d) climate and geographic predictor variables (full model), e) climate predictor variables (geography removed), and f) geographic predictor variables (climate removed). Direction and length of arrows on each RDA plot correspond to the loadings of each variable.

Figure 2.4 SDM predictions a) across four time points for *P. pungens*, *P. rigida*, and *P. taeda*. Occurrence records for each species (black dots) overlay habitat suitability predictions. Venn diagrams illustrate the number of grid cells with moderate to high habitat suitability scores (> 0.5) for each SDM at a given time point, as well as the number of overlapping grid cells. Blue ovals show counts for *P. pungens*, orange ovals show counts for *P. rigida*, and green ovals show counts for the *P. taeda* SDM predictions at each aligning time point. SDM Glacial extent data (labeled ice in LGM plots) for 18 kya was provided by Dyke (2003). Panel b illustrates pairwise comparisons of raster overlap across each time period.

Figure 2.5 Relative distributions of asymmetrical background similarity tests (gray bars) to niche overlap (red arrow). Panels from left to right illustrate the niche relationships between *P. pungens* and *P. rigida*, *P. pungens* and *P. taeda*, and *P. rigida* and *P. taeda*, respectively. An arrow to the left of a background similarity distribution indicates niche divergence, while an arrow to the right indicates niche conservatism.

Figure 2.6 Demographic inference workflow where a) two models with the lowest AIC from the first round of inferences were used in b) to force deeper divergence time inferences through manipulation of lower and upper bounds of parameter space. Two population models to test species relationships and topology are presented in panels c - e. GF stands for gene flow. The acronyms symRT and asymRT stands for allowing symmetrical or asymmetrical gene flow between *P. rigida* and *P. taeda* during T2 (time interval 2). Respectively, N_A , N_P , N_R , and N_T are the effective population sizes of *P. taeda* at the end of T1, then *P. pungens*, *P. rigida*, and *P. taeda* at the end of T2.

Figure 2.7 Description of blastn hits to the *P. taeda* draft genome. Panel a) shows the number of hits after data was filtered down to one RADseq contig per scaffold with max three unique scaffold IDs allowed per hit and how those relate to matched attributes (i.e., annotations) and locations to genes. Values associated with some bars are nested within bars to the left. Panel b) shows the distribution of F_{CT} values associated with our 5051 SNPs. The number of unique RADseq-scaffold hits and corresponding F_{CT} value ranges are shown in c) for those outside 20k bp of a gene, d) for those within 20k bp of a gene, and e) for those that hit within the gene. The third and fourth bars in panels c-e are nested

components of the second bar. In parentheses are the number of unique RADtags (i.e., RADseq IDs) defining the number of hits. The distribution of F_{ST} values from pairwise species comparisons (PR, comparing variation between *P. pungens* and *P. rigida*; PT, between *P. pungens* and *P. taeda*; RT, between *P. rigida* and *P. taeda*) for SNPs that are f) relatively far from a gene, g) relatively close to a gene, and h) within a gene.

Figure 2.S1 Seven demographic models that were tested in the first round of model selection. SGF is speciation with gene flow. SC is secondary contact. GF allowed gene flow at T1 (first time interval) and T2 (second time interval). The acronym sym means the model inferred symmetrical gene flow. The acronym asym means the model inferred asymmetric gene flow.

Figure 2.S2 Bioclimatic variable associations with a) occurrence data used in SDMs and b) SDM permutation importance and percent contribution to each model. Blue bars correspond to *P. pungens*. Orange bars correspond to *P. rigida*. Green bars correspond to *P. taeda*.

Figure 2.S3 Geographical distributions of *P. pungens* (blue), *P. rigida* (orange), *P. taeda* (green), as described in Little (1971). Five populations with the most admixture present between *P. taeda* and *P. rigida* are plotted (black dots) and labeled. The dashed line illustrates distance between the closest region of geographical overlap between natural stands of *P. taeda* in Louisiana and Mississippi in relation to suitable habitat of *P. rigida*.

Figure 2.S4 Folded site frequency spectrum for the data (top row) and symmetrical gene flow model (second row). Residuals are plotted in the last two rows and correspond to the three- species model run with the lowest AIC.

Figure 2.S5 Folded site frequency spectrum for the data and asymmetrical gene flow model for the *P. pungens* and *P. rigida* two-species model. Residuals are plotted in the bottom row and correspond to the two-species model run with the lowest AIC.

Figure 2.S6 Folded site frequency spectrum for the data and strict isolation model for the *P. pungens* and *P. taeda* two-species model. Residuals are plotted in the bottom row and correspond to the two-species model run with the lowest AIC.

Figure 2.S7 Folded site frequency spectrum for the data and strict isolation model for the *P. rigida* and *P. taeda* two-species model. Residuals are plotted in the bottom row and correspond to the two-species model run with the lowest AIC.

Figure 3.1 Distribution of sampled sympatric populations a) in relation to each other geographically and across the described geographic range of each species in Little (1975). The trees sampled within each population are shown in for b) Brown Mountain of Shenandoah National Park, c) Laurel Falls of Great Smoky Mountains National Park, and d) Dragon Tooth of Jefferson National Forest. Blue circles indicate samples morphologically identified as *P. pungens*. Orange triangles are samples indicative of *P. rigida*.

Figure 3.2 Species level genetic differentiation for 194 sampled trees across three sympatric stands (map, panel a). Principal component analysis results based on multilocus genotypes across 6343 SNPs are provided in panel b. Inference of structure from ($K = 2$) is provided in panel c.

Figure 3.3 Population-level genetic differentiation across 6343 SNPs illustrated in a) principal component analysis (PCA) for *P. pungens* and *P. rigida* sampled trees, and b) Pairwise population level comparisons for *P. pungens* (top row, blue) and *P. rigida* (bottom row, orange) where BM is Brown Mountain, DT is Dragon Tooth, and LF is Laurel Falls. Dashed line is the realized pairwise F_{ST} in each plot. Distributions are permutations of F_{ST} based on random selection of individuals. If the dashed line is to the right of the distribution, then populations are more different than expected by random chance.

Figure 3.4 Counts of SNPs based on two categories of F_{ST} , low ($0.3 > F_{ST} > 0.1$; panel a) and high ($F_{ST} \geq 0.8$; panel b) for Brown Mountain (BM), Dragon Tooth (DT), and Laurel Falls (LF). The last bar in each plot represents the number of SNP IDs (dDocent contigs) that were shared between BM, DT, and LF.

Figure 3.S2 PCA of *P. pungens* populations with the hybrid sample removed.

Quantifying contributions of climate, geography, and gene flow to divergence: a case study for three North American pines

By Constance Ellen Bolte (PhD, MS, MT)

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University, 2022.

Major Director: Dr. Andrew J. Eckert, Department of Biology

Abstract

Long-lived species of trees, especially conifers, often display weak patterns of reproductive isolation, but clear patterns of local adaptation and phenotypic divergence. Discovering the evolutionary history of these patterns is paramount to a generalized understanding of speciation and the processes that confer population persistence versus those that compromise adaptive potential under rapidly changing environments. Forest trees have long generation times and low migratory potential making them especially vulnerable to population fragmentation and reductions of genetic diversity due to insufficient tracking of niche optima and adaptational lags. Within clades of the genus *Pinus*, evolutionary histories appear to be riddled with hybridization (i.e., interspecific gene flow), periods of isolation, and effective population size changes that co-occur with major shifts in climate. Quantifying the relative contributions of each of these factors to divergence and changes to genetic diversity requires a multidisciplinary approach involving historical species distributional modeling, demographic inference, and associations of genetic structure to climate and geography.

This dissertation focuses on identifying drivers of divergence and explaining differing levels of reproductive isolation across three ecologically and economically valuable North American pine species (*Pinus pungens*, *P. rigida*, and *P. taeda*). First, we

inferred demographic histories and found the recurrence of interspecific gene flow between *P. pungens* and *P. rigida*, as well as population size reductions during the last glacial period, to be important contributors to the mode and tempo of previously documented reproductive isolation between these species. Seasonality and elevation associated with both genetic and distributional differences indicating ecological divergence was also important to the divergences among the three focal species, but the relationship of *P. taeda* to the other two species remains enigmatic. Next, we illustrate how genomic patterns of differentiation across genic and intergenic regions can explain differing levels of reproductive isolation through pairwise assessments and mapping RADseq contigs to the annotated genome of *P. taeda*. Finally, in estimating the extent of hybridization and genetic diversity in shared forest stands of *P. pungens* and *P. rigida*, we discovered a general lack of hybridization at present and low genetic diversity in southern, trailing edge populations.

Striking congruences across results, various methods employed, and work previously performed for the genus *Pinus* all provide support for emerging hypotheses related to forest tree speciation and biodiversity. This dissertation also presents useful information for forest conservation and management planning. At present, the adaptive potential of *P. pungens*, a montane pine with highly fragmented populations, is low based on genetic diversity estimates, its current distribution, and restricted levels of interspecific gene flow.

Introduction

It is increasingly evident that the process of speciation does not strictly adhere to a simple model of vicariance among geographically isolated populations. Divergence often proceeds with varying levels of gene flow, natural selection, and geographic isolation. Over the last few decades, an array of tools has been developed allowing us to more thoroughly investigate the multitude of ways in which species arise and the varying ways in which reproductive isolation evolves. For many lineages, there is a strong role of ecologically driven adaptation contributing to the evolution of reproductive isolation and hence the origin of new species (Hendry et al. 2007). Yet for others, geographically and ecologically separated populations comprise single species taxonomically housed within monotypic genera (e.g., Kou et al. 2019). Different degrees of gene flow, isolation, population size change, and local adaptation among populations may explain variations in observed diversification rate (Liu et al. 2014; Kou et al. 2019; Kremer and Hipp 2019; Wu et al. 2022). Here, we consider general mechanisms of speciation for conifers; the timing of which is particularly apt given the explosion of genomic data for these charismatic plants.

Mechanisms driving speciation for conifers are not as well characterized as in other groups of plants despite a long history of crossing and common garden experiments. This is likely driven by their long generation times, large genome sizes, historical lack of genomic resources, and propensities to hybridize (Petit and Hampe 2006). Of the few detailed examples available (e.g., Mao and Wang 2011), there is a complex interplay among gene flow across populations (including hybridization), demographic processes

within populations, and local adaptation to the formation of new conifer species. For conifers, we think this complexity is best thought of within models of ecological speciation (Rundle and Nosil 2005).

Ecological divergence plays a major role in the establishment and maintenance of reproductive isolation in plants (Hendry et al. 2007), which suggests ecological speciation as a major generator of plant biodiversity. This model of speciation requires the buildup of reproductive isolation through ecological divergence among populations driving the development of prezygotic and postzygotic isolating mechanisms (Harvey et al. 2019). For conifers, prezygotic isolating mechanisms are often related to differential timing of phenological events (e.g., Zobel 1969), while postzygotic isolating mechanisms are centered on hybrid inferiority due to genomic conflict among the mixing of genetic material from ecologically diverged lineages (e.g., Manley and Ledig 1979). In all cases, ecological divergence can be thought of in the context of the relationship between the fundamental and realized niche and how these evolve across populations, species, and lineages. We argue, as does Pearman et al. (2007), that the relative time scales required for evolutionary processes to occur may be better understood if we looked through the kaleidoscopic lens of niche dynamics within and across lineages, as well as current and historical landscapes (Figure i; Figure ii).

The rate of adaptation, niche evolution, and speciation are often affected by the same suite of interconnected factors (Figure i). For example, a reduction in realized niche breadth during founder events (Pearman et al. 2007), has constraints on niche evolution

due to limited genetic variation (Schiffers et al. 2014). Likewise, niche evolution within a more homogeneous environment (e.g., low landscape complexity with gradual, unidirectional changes in climate) may be restricted to directional instead of divergent shifts when tracking fitness optima (Figure ii.a,b). Additional influencers of niche evolution could include the presence of biotic interactions (e.g., competition; Pearman et al. 2007) and the underlying genetic architecture of traits under selection (Schiffers et al. 2014), which affects movement of the realized niche within the space defined by the fundamental niche. Due to these interconnections and the scope of variation housed within each factor, it is unlikely that generalized predictions towards the rate of speciation and the development of reproductive isolation will emerge without further empirical and theoretical work (Figure ii.b and Figure ii.c are hypotheses respectively posed in Kou et al. 2020 and Bolte and Eckert 2020). We do anticipate though that with a focused comparison of taxa sharing similar demographic histories, life history traits, and geographical distributions, trends will emerge.

Fortunately, a multitude of methods and data types have been developed and collected over the last decade allowing us to now begin rigorously linking concepts of niche evolution, ecological speciation, and evolutionary genetics to further our understanding of macroevolutionary trends within clades of plants, like conifers, where this knowledge is limited. As argued above, we think one of the major keys to understanding mechanisms of conifer speciation is to think about niche evolution and its multifarious influences within a model of ecological speciation (Figure ii). This is not to say that all speciation within conifers requires adaptive evolution, but that a modeling framework that explicitly

acknowledges this often noted attribute of conifer lineages may be more illuminating than one without it, especially if the goal is to estimate the relative importance of factors contributing to species formation.

The genus *Pinus* is the most diverse group of conifers with over 110 species that inhabit an array of geographic regions and climatic regimes, providing an extensive resource for comparative investigation into conifer speciation and the development of reproductive isolation (Zukowska and Wachowiak 2016; Jin et al. 2021). Much of the genomic, evolutionary-based research performed in the genus *Pinus* has used economically valuable species as focal taxa. As a result, many species that do not hold reasonably high economic value have been largely ignored regardless of their high ecological importance. One such species is Table Mountain pine (*Pinus pungens* Lamb.). While conservation efforts are being made to restore populations of this montane conifer (Jetton et al. 2015), no genetic data, especially genome-wide data, have been collected. The phylogenetic relationships between *P. pungens* Lamb. and two other related species, *P. rigida* Mill. and *P. taeda* L., have been notoriously difficult to resolve (Hernández-León et al. 2013; Saladin et al. 2017; Gernandt et al. 2018). Hybridization challenges phylogenetic inference and may explain the lack of consensus in defining the relationships across these three species. Employing a demographic inference framework that uses genome-wide nuclear data and range-wide samples of each species is an appropriate next step to estimate the extent of intraspecific gene flow, the timing of gene flow, and the role of gene flow in the maintenance of species boundaries. All of which is considerably important information to predicting outcomes of forest management plans.

In this dissertation, we focused on inferring the divergence histories for three related pine species of eastern North America and analyzing niche and genetic differentiation through geographic and climate variable associations to elucidate potential drivers in differential developments of reproductive isolation. Chapter 1 describes a complex divergence history involving gene flow and population size changes for *P. pungens* and *P. rigida* and identifies potential drivers, such as seasonality and fire regime, involved in the development of reproductive isolation. The gene flow dynamics between these two focal species inspired Chapter 2, which expanded demographic inference to include a third related species, *P. taeda*, which actively hybridizes with *P. rigida* at present (Smouse and Saylor 1973). While the relationship of *P. taeda* relative to *P. pungens* and *P. rigida*, remains enigmatic post-demographic inference, we were able to describe the genomic distribution of our RADseq data by mapping contigs to the annotated genome of *P. taeda*. We observed contrasting levels of differentiation in pairwise species comparisons across contigs associated with genic and intergenic regions. We found that the higher levels of differentiation (F_{ST}) in comparisons with *P. pungens* correspond to greater strength of reproductive isolation (as described in ecological assays and artificial crossing experiments; Zobel 1969; Critchfield 1963). In Chapter 3, we focused more closely on the development of reproductive isolation between *P. pungens* and *P. rigida* by examining the extent of current hybridization across three sympatric stands and mapping RADseq contigs to the *P. taeda* genome (as performed in Chapter 2). We provide convincing evidence that species boundaries have been maintained through reduced hybrid fitness in sympatric stands (reinforcement) and ecological character displacement. From

population genetic summaries, we also observed lower genetic diversity in southern, trailing edge populations. We took our evidence of reproductive isolation across species and genetic differences across populations and contextualized them for relevance to forest conservation and management planning.

Throughout this work, we examine metrics associated with niche and distributional overlap across time and landscape to explain patterns in genetic data and the development of reproductive isolation in terms of both tempo and mode for three species of North American pines. Our findings illustrate how high rates of interspecific gene flow, likely in tandem with disruptive selection acting on ecological traits, can promote the rapid development of reproductive isolation. Whether the speciation histories and drivers of divergence are unique to the focal species of this dissertation or part of a larger pattern will remain unknown until more clade-specific investigations are performed for coniferous species. Given conifers are foundational species to many forest ecosystems, we foresee a heightened interest in genetically-based inferences for these taxa, as well as are hopeful for how this knowledge can contribute to the general understanding of when, why, and how reproductive isolation evolves in long-lived tree species.

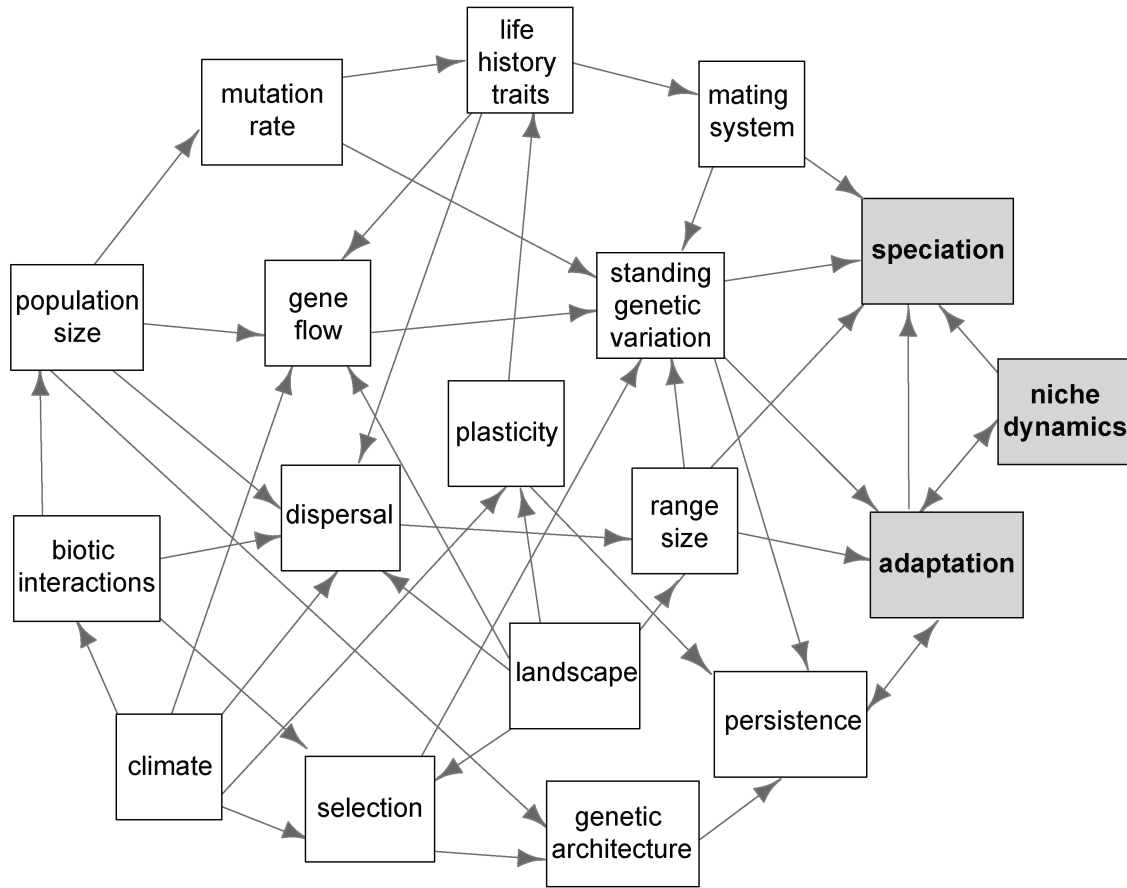


Figure i Conceptualizing factors involved in speciation and the interconnectivity among the factors often considered during investigations related to niche evolution, adaptation, and speciation. This is a modified figure from Bolte and Eckert (2020).

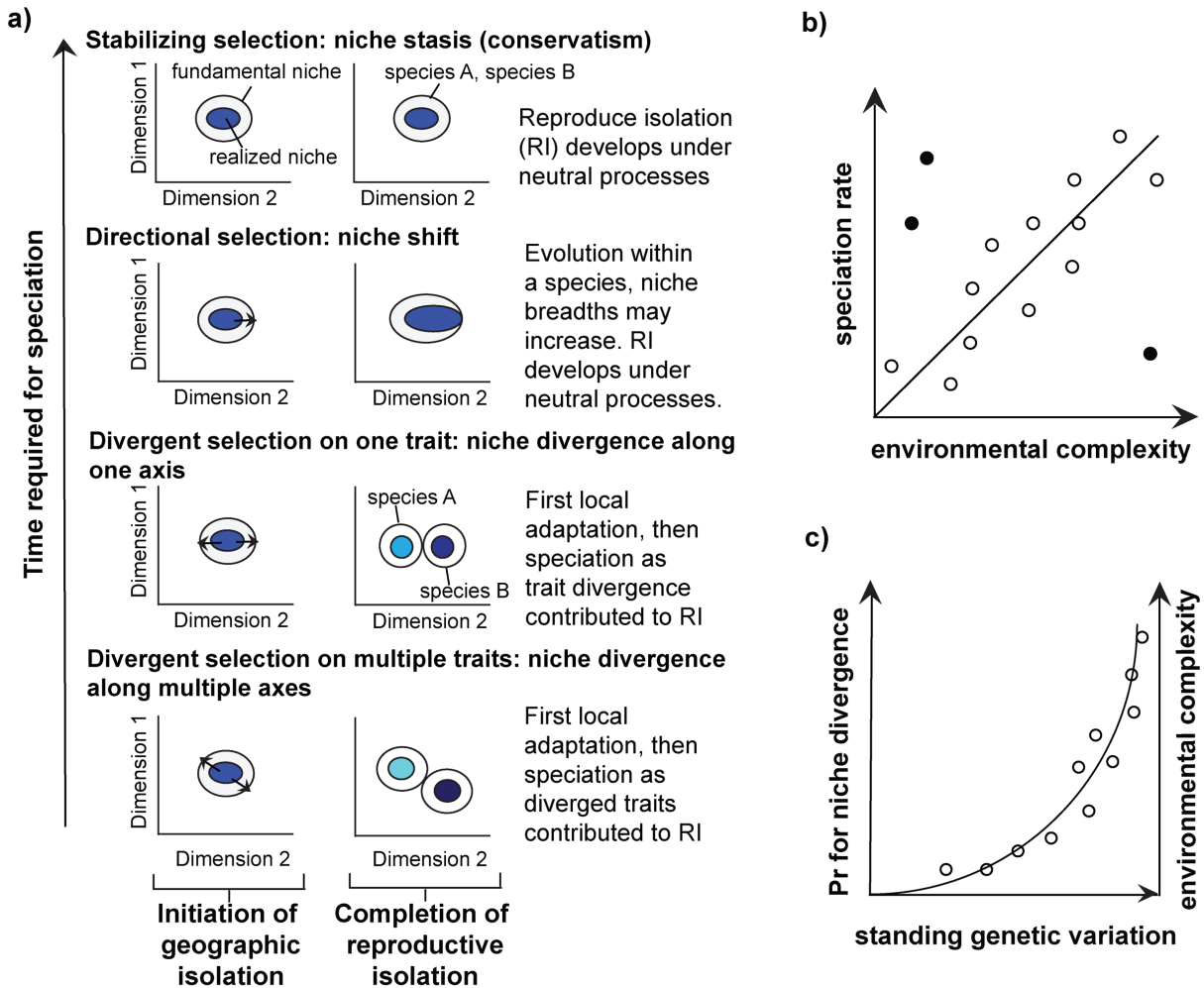


Figure ii Hypotheses related to speciation rate in relation to ecological divergence scenarios. a) Simple, 2-dimensional schematic showing the relationship between the realized niche (i.e., where the species is known to occur), the fundamental niche (i.e. where the species has the capacity to occur) and the hypothesized importance of divergent selection in the time needed for reproductive isolation to develop when all other factors from Figure i are held constant. For the top two diagrams (i.e., stabilizing selection versus directional selection) imagine the niche spaces for two species are stacked on top of each other after completion of reproductive isolation. b) Hypothesized relationship between environmental complexity and speciation rate. Open circles meet expectations. Closed circles may have life history traits or genetic architectures that allow deviation from expectations. c) Hypothesized relationship between combined factors of standing genetic variation and environmental complexity on the probability for niche divergence. In environments with low complexity the probability of niche divergence is low regardless of standing genetic variation. In homogeneous environments it is hypothesized that niche stasis or niche directional shifts are more likely to occur than niche divergence. This is a modified figure from Bolte and Eckert (2020).

Literature Cited

- Bolte CE & Eckert AJ. 2020. Determining the when, where and how of conifer speciation: a challenge arising from the study 'Evolutionary history of a relict conifer *Pseudotsuga chienii*.' *Annals of Botany*, 125(1), v–vii.
- Critchfield WB. 1963. The Austrian x red pine hybrid. *Silvae Genetica*12:187-191.
- Gernandt DS, Aguirre-Dugua X, Vázquez-Lobo A, et al. 2018. Multi-locus phylogenetics, lineage sorting, and reticulation in *Pinus* subsection *Australes*. *American Journal of Botany* 105:711–725.
- Harvey MG, Singhal S, Robasky DL. 2019. Beyond reproductive isolation: demographic controls on the speciation process. *The Annual Review of Ecology, Evolution, and Systematics* 50: 75-95.
- Hendry AP, Nosil P, Rieseberg LH. 2007. The speed of ecological speciation. *Functional Ecology* 21: 455-464.
- Hernández-León S, Gernandt DS, Pérez de la Rosa J, Jardón-Barbolla L. 2013. Phylogenetic relationships and species delimitation in *Pinus* section *Trifoliae* inferred from plastid DNA. *PLoS One* 8:1–14.
- Jetton RM, Crane BS, Whittier WA, Dvorak WS. 2015. Genetic resource conservation of Table Mountain pine (*Pinus pungens*) in the central and southern Appalachian Mountains. *Tree Plant Notes* 58:42–52.
- Jin WT, Gernandt DS, Wehenkel C, et al. 2021. Phylogenomic and ecological analyses reveal the spatiotemporal evolution of global pines. *Proceedings of the National Academy of Science USA* 118.
- Kou Y, Zhang L, Fan D, et al. 2019. Evolutionary history of a relict conifer *Pseudotsuga chienii* (Taxaceae) in southeast China during the late Neogene: Old lineage, young populations. *Annals of Botany*
- Kremer A & Hipp AL. 2020. Oaks: an evolutionary success story. *New Phytologist*, 226: 987–1011.
- Liu L, Hao ZZ, Liu YY, Wei XX, Cun YZ, & Wang XQ. 2014. Phylogeography of *Pinus armandii* and its relatives: Heterogeneous contributions of geography and climate changes to the genetic differentiation and diversification of Chinese white pines. *PLoS ONE*, 9: 1–12.
- Manley SAM, Ledig FT. 1979. Photosynthesis in black and red spruce and their hybrid derivatives: ecological isolation and hybrid adaptive inferiority. *Canadian Journal of Botany* 57: 305-314.

Mao JF, Wang XR. 2011. Distinct niche divergence characterizes the homoploid hybrid speciation of *Pinus densata* on the Tibetan Plateau. *The American Naturalist* 177: 424-439.

Pearman PB, Guisan A, Broennimann O, Randin CF. 2008. Niche dynamics in space and time. *Trends in Ecology & Evolution* 23: 149-158.

Petit RJ, Hampe A. 2006. Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution, and Systematics* 37: 187-214.

Rundle HD, Nosil P. 2005. Ecological speciation. *Ecology letters* 8: 336-352.

Saladin B, Leslie AB, Wüest RO, et al. 2017. Fossils matter: improved estimates of divergence times in *Pinus* reveal older diversification. *BMC Evolutionary Biology* 17:95.

Schiffers K, Schurr F, Travis J, et al. 2014. Landscape structure and genetic architecture jointly impact rates of niche evolution. *Ecography* 37: 1218-1229.

Smouse PE & Saylor LC. 1973. Studies of the *Pinus rigida*-*Serotina* Complex II. Natural hybridization among the *Pinus rigida-serotina* complex, *P. taeda* and *P. echinata*. *Annals of the Missouri Botanical Garden*, 60(2), 192-203.

Wu S, Wang Y, Wang Z, Shrestha N & Liu J. 2022. Species divergence with gene flow and hybrid speciation on the Qinghai–Tibet Plateau. *New Phytologist*: 392–404.

Zobel DB. 1969. Factors affecting the distribution of *Pinus pungens*, an Appalachian endemic. *Ecological Monographs* 39: 303-333.

Zukowska WB, Wachowiak W. 2016. Utility of closely related taxa for genetic studies of adaptive variation and speciation: Current state and perspectives in plants with focus on forest tree species. *Journal of Systematics and Evolution*, 54(1), 17–28.

Chapter 1

Divergence amid recurring gene flow: the complex demographic histories inferred for *Pinus pungens* and *P. rigida* align with a growing expectation for forest trees

Abstract

Long-lived species of trees, especially conifers, often display weak patterns of reproductive isolation, but clear patterns of local adaptation and phenotypic divergence. Discovering the evolutionary history of these patterns is paramount to a generalized understanding of speciation for long-lived plants. We focus on two closely related yet phenotypically divergent pine species, *Pinus pungens* and *P. rigida*, that co-exist along high elevation ridgelines of the southern Appalachian Mountains. In this study, we performed historical species distribution modeling (SDM) to form hypotheses related to population size change and gene flow to be tested in a demographic inference framework. We further sought to identify drivers of divergence by associating climate and geographic variables with genetic structure within and across species boundaries. Population structure within each species was absent based on genome-wide RADseq data, however signals of admixture were present range-wide, and species-level genetic differences associated with precipitation seasonality and elevation. When combined with information from contemporary and historical species distribution models, these patterns are consistent with a complex evolutionary history of speciation influenced by Quaternary climate. This was confirmed using inferences based on the multidimensional site-frequency spectrum, where demographic modeling inferred recurring gene flow since

divergence (2.74 million years ago) and population size reductions that occurred during the last glacial period (~35.2 thousand years ago). This suggests that phenotypic and genomic divergence, including the evolution of divergent phenological schedules leading to partial reproductive isolation, as previously documented for these two species, can happen rapidly, even between long-lived species of pines.

Introduction

The process of speciation has been characterized as a continuum of divergence underpinned with the expectation that reproductive isolation strengthens over time leading to increased genomic conflict between species (Seehausen et al. 2014). While the term continuum suggests linear directionality, it is better thought of as a multivariate trajectory that is nonlinear, allowing stalls and even breakdown of reproductive barriers in the overall progression toward complete reproductive isolation (Cannon and Petit 2020; Kulmuni et al. 2020). Indeed, speciation can occur with or without ongoing gene flow and demographic processes such as expansions, contractions, isolation, and introgression leave detectable genetic patterns within and among populations of species that affect the evolution of reproductive isolation (Nosil 2012; e.g., Gao et al. 2012). Divergence histories with gene flow are an emerging pattern for species of forest trees with reproductive isolation often developing through prezygotic isolating mechanisms and reinforced by environmental adaptation (Abbott 2017; Cavender-Bares 2019). Together, these two

processes can facilitate the development of genomic incompatibilities over time (Baack et al. 2015).

Climate and geography are well-established drivers of demographic processes and patterns (Hewitt 2001). For the past 2.6 million years, Quaternary climate has oscillated between glacial and interglacial periods causing changes in species distributions, but the significance of these changes and their influence on population differentiation has varied by region and taxon (Hewitt 2004; Lascoux et al. 2004). In North America, the effects of Quaternary climate on tree species distributions and patterns of genetic diversity have been profound but more drastic for species native to northern (i.e., previously glaciated) and eastern regions. For instance, the geographical distribution of white oak (*Quercus alba* L.), a native tree species to eastern North America, experienced greater shifts since the last interglacial period (LIG), approximately 120 thousand years ago (kya), compared to the distributional shifts of valley oak (*Quercus lobata* Née) in California (Gugger et al. 2013). For the latter, distributional, and hence niche, stability was correlated with higher levels of genetic diversity.

Given the climate instability of eastern North America since the LIG, a host of phylogeographic studies have reported genetic diversity estimates for taxa of this region and the genetic structuring of populations due to geographic barriers such as the Appalachian Mountains and Mississippi River (Soltis et al. 2006) as well as postglacial expansion (e.g., Gougherty et al. 2020). The vast majority of tree taxa in these studies, however, were angiosperms, with the divergence history of only one closely related pair

of conifer species native to this region, *Picea mariana* (Mill.) Britton, Sterns, & Poggenb. and *P. rubens* Sarg., being fully characterized (Perron et al. 2000; Lafontaine et al. 2015). The relative differences in geographical distributions and genetic diversities across *P. mariana* and *P. rubens*, as well as models of demographic inference, suggest a progenitor-derivative species relationship that initiated approximately 110 kya through population contractions and geographical isolation. Despite this history, these two species actively hybridize today. In general, speciation among conifer lineages remains an enigmatic process (Bolte and Eckert 2020), largely because there is a mismatch between species-level taxonomy and the existence of reproductive isolation, so that hybridization among species is common both naturally as well as artificially (Critchfield 1986). The ability to hybridize, moreover, is idiosyncratic, with examples ranging from well-developed incompatibilities among populations within species (e.g., *P. muricata* D. Don; Critchfield 1967) to the almost complete lack of incompatibilities among diverged and geographically distant species (*P. wallichiana* A. B. Jacks. from central Asia and *P. monticola* Douglas ex D. Don from western North America; Wright 1959). Thus, the tempo and mode for the evolution of reproductive isolation for conifers remains largely unexplained despite decades of research into patterns of natural hybridization, crossing rates, and the mechanisms behind documented incompatibilities (McWilliam 1959; Kriebel 1972; Hagman 1975; Critchfield 1986; Vasilyeva and Goroshkevich 2018).

The key to understanding the evolution of reproductive isolation, and hence a more developed explanation of the process of speciation for conifers, is the role of demography and gene flow during the divergence among lineages. Analytical approaches have been

developed to infer past demographic processes from population genomic data, which can now easily be generated even for conifers (Parchman et al. 2018). While many studies have used demographic inference methods to describe the phylogeographic history of a single species (e.g., Gugger et al. 2013; Li et al. 2013; Bagley et al. 2020; Ju et al. 2019; Park and Donoghue 2019; Capblancq et al. 2020; Yang et al. 2020; Labiszak et al. 2021), some of these established methods have also been used to infer divergence histories between two or three species (e.g., Zou et al. 2013; Christe et al. 2017; Kim et al. 2018; Menon et al. 2018). Single species inferences have found that the last glacial maximum (LGM; ~21 kya) affected distributional shifts and intraspecific gene flow dynamics, while multispecies studies have focused almost solely on how these climatic oscillations drove periods of increased and decreased interspecific gene flow which contributed to the formation of environmentally dependent hybrid zones, ancient periodical introgression, or adaptive divergence in the development of reproductive isolation.

The number of potential divergence histories underlying even a modest number of species is vast. The preemptive formation of a hypothesis from historical species distribution modeling (SDM), however, can aid in defining a more realistic set of models from which to make inference, as well as to examine the impact of climate change on genetic diversity and demographic processes (Carstens and Richards 2007). For example, Lima et al. (2017) modeled distributional changes for *Eugenia dysenterica* DC. between the LGM and today which led to a hypothesis that range stability was more likely than range expansion or contraction in this South American region. Their SDM informed hypothesis was supported by range-wide, *E. dysenterica* genetic data. Likewise, SDMs

across several time points allows for estimation of habitat suitability change (i.e., a proxy for contraction or expansion) and distributional overlap of multiple species (i.e., potential gene flow). With these quantified changes, testable hypotheses emerge, leading to more deliberate investigations of speciation through justified parameter selection (Richards et al. 2007). Of course, there are inherent limitations associated with SDMs and interpreting historical distributions should be done cautiously but using SDMs to complement demographic inference is now common in the field of phylogeography (Hickerson et al. 2010; Gavin et al. 2014; Peterson and Anamza 2015). For example, where a species occurs is determined to some degree by its traits and thus at least partially its genetics, so that non-optimal inference can occur by ignoring putative adaptation within lineages during SDM formation and testing. Indeed, Ikeda et al. (2017) found that SDM predictions under future climate scenarios improved with acknowledgement of local adaptation in *Populus fremontii* S. Watson (i.e., three identified genetic clusters across the full species distributional range were modeled independently).

Here, we focus on two closely related, yet phenotypically diverged, pine species, Table Mountain pine (*Pinus pungens* Lamb.) and pitch pine (*Pinus rigida* Mill.). Recent estimates from multiple time-calibrated phylogenies across nuclear and plastid DNA have placed the time of divergence in the range of 1.5 to 17.4 million years ago (mya; Hernandez-Leon et al. 2013; Saladin et al. 2017; Gernandt et al. 2018; Jin et al. 2021), with these studies either placing them as sister species (e.g., Hernandez-Leon et al. 2013; Saladin et al. 2017) or as part of a clade with *P. serotina* Michx. as the sister to *P. rigida* (e.g., Gernandt et al. 2018; Jin et al. 2021). Changes in climate, fire regime, and

geographic distributions have likely influenced species divergence (Keeley 2012). This is plausible given that *P. pungens* populations are restricted to high elevations of the Appalachian Mountains, while the much larger distribution of *P. rigida* ranges from Georgia into portions of eastern Canada. It is particularly interesting that these recently diverged species are found in sympatry, yet hybridization has rarely been observed in the field (Zobel 1969), although they can be reciprocally crossed to yield viable offspring (Critchfield 1963). An ecological study of three sympatric *P. pungens* and *P. rigida* populations indicated that the timing of pollen release was separated by approximately four weeks, enough to sustain partial reproductive isolation at these sites (Zobel 1969), which is a common contributor to prezygotic isolation among conifer species (Dorman and Barber 1956; Critchfield 1963). It was also noted that while *P. pungens* was most densely populated on arid, rocky, steep southwestern slopes, *P. rigida* was less confined to these areas (Zobel 1969), thus suggesting environmental adaptation through ecological character displacement may also be important in the divergence of these two closely related species.

Considering the dynamic interplay of climate, topography, and ecology potentially involved in the divergence of these two pine species, we asked three questions: 1) Which demographic processes were involved in the divergence of *P. pungens* and *P. rigida*? 2) Does the timing of demographic events align with shifts in climate? 3) To what extent are climate and geographic variables associated with genetic differentiation? To answer these three questions, we hypothesized that *P. pungens* and *P. rigida* experienced divergence with gene flow followed by population contraction and isolation (i.e., different

refugia) initiated during the LGM as an explanation for strongly diverged traits and phenological schedules. From historical SDM predictions across four time points since the LIG, we formed additional hypotheses to be tested within a demographic inference framework. Three hypotheses corresponded to SDM predictions from specific general circulation models (GCMs) and were compared to a fourth hypothesis formed from ensembled SDM predictions. We then used the multidimensional, folded site frequency spectrum from 2168 genome-wide, unlinked single nucleotide polymorphisms (SNPs) across 300 trees to infer demographic processes and timing of divergence. Our best-fit demographic model inferred initial divergence at 2.74 mya, aligning with the start of the Quaternary Period, and described divergence as occurring with ongoing gene flow and drastic population size reductions during the last glacial period (~35.2 kya). SDM hypotheses were partially supported, especially for ongoing gene flow and population size reductions during the LGM. We conclude that climatic oscillations, differential adaptation to seasonality, and gene flow influenced the divergence of *P. pungens* and *P. rigida* and present evidence from SDM, genetic association analyses, and demographic inference as support.

Methods

Sampling

Range-wide samples of needle tissue were obtained from 14 populations of *Pinus pungens* and 19 populations of *Pinus rigida* (Figure 1.1). Each population consisted of 4-

12 trees with each sampled tree distanced by approximately 50 m from the next to avoid potential kinship (Table 1.1). Needle tissue was dried using silica beads, then approximately 10 mg of tissue was cut and lysed for DNA extraction.

DNA sequence data

Genomic DNA was extracted from all 300 sampled trees using DNeasy Plant Kits (Qiagen) following the manufacturer's protocol. Four ddRADseq libraries (Peterson et al. 2012), each containing up to 96 multiplexed samples, were prepared using the procedure from Parchman et al. (2012). EcoRI and MseI restriction enzymes were used to digest all four libraries before performing ligation of adaptors and barcodes. After PCR, agarose gel electrophoresis was used to separate then select DNA fragments between 300-500 bp in length. The pooled DNA was isolated using a QIAquick Gel Extraction Kit (Qiagen). Single-end sequencing was conducted on Illumina HiSeq 4000 platform by Novogene Corporation (Sacramento, CA). Raw fastq files were demultiplexed using GBSX (Herten et al. 2015) version 1.2, allowing two mismatches (-mb 2). The dDocent bioinformatics pipeline (Puritz et al. 2014) was subsequently used to generate a reference assembly and call variants. The reference assembly was optimized using shell scripts and documentation within dDocent (cutoffs: individual = 6, coverage = 6; clustering similarity: -c 0.92), utilizing cd-hit-est (Fu et al. 2012) for assembly. The initial variant calling produced 87,548 single nucleotide polymorphisms (SNPs) that were further filtered using *vcftools* (Danecek et al. 2011) version 0.1.15. We retained only biallelic SNPs with sequencing data for at least 50% of the samples, minor allele frequency (MAF) > 0.01, summed depth across samples > 100 and < 10000, and alternate allele call quality ≥ 50 .

Additionally, stringent filtering steps were taken to minimize the potential misassembly of paralogous genomic regions. Removing loci with excessive coverage and retaining only loci with two alleles present, as above, should ameliorate the influence of misassembled paralogous loci in our data (Hapke and Thiele 2016; McKinney et al. 2018). Lastly, we retained loci with $F_{IS} > -0.5$, as misassembly to paralogous genomic regions can lead to abnormal levels of heterozygosity (Hohenlohe et al. 2013; McKinney et al. 2017). To account for linkage disequilibrium among the 20,932 SNPs that passed quality controls, which if not properly acknowledged can lead to erroneous inferences of demographic history (Gutenkunst et al. 2009), we thinned the dataset to one SNP per contig (--thin 100). The reduced 2168 SNP dataset was used in all analyses.

Population structure and genetic diversity

Patterns of genetic diversity and structure within and between *P. pungens* and *P. rigida* were assessed using a suite of standard methods. Overall patterns of genetic structure were investigated using principal component analysis (PCA), as employed in the `prcomp` function of the *stats* version 4.0.4 package, on centered and scaled genotypes following Patterson et al. (2006), in R version 3.6.2 (R Development Core Team, 2021). Genetic diversity within each species was examined using multilocus estimates of observed and expected heterozygosity (H_o and H_e) for each population using a custom R script (www.github.com/boltece/Speciation_2pines). An individual-based assignment test was conducted using *fastSTRUCTURE* (Raj et al. 2014) with cluster assignments ranging from $K = 2$ to $K = 7$. Ten replicate runs of each cluster assignment were conducted. The cluster assignment with the highest log-likelihood value was determined to be the best fit.

Individual admixture assignments were then aligned and averaged across the 10 runs using the *pophelper* version 1.2.0 (Francis 2017) package in R. Third, multilocus, hierarchical fixation indices (F -statistics) were defined by nesting trees into populations and populations into species, with F_{CT} describing differentiation between species and F_{SC} describing population differentiation within species (Yang 1998). F -statistics and associated confidence intervals (95% CIs) from bootstrap resampling ($n = 100$ replicates) were calculated in the *hierfstat* version 0.5-7 package (Goudet and Jombart 2020) in R.

To assess influences on within-species genetic structure, Mantel tests (Mantel 1967) were used to examine Isolation-by-Distance (IBD; Wright 1943) and Isolation-by-Environment (IBE; Wang and Bradburd 2014). In these analyses, the Mantel correlation coefficient (r) was calculated between linearized pairwise F_{ST} , estimated with the method of Weir and Cockerham (1984) using the *hierfstat* package in R, and either geographical (IBD) or environmental (IBE) distances. For geographical distances, latitude, and longitude records for each tree in a population were averaged to obtain one representative coordinate per population. Geographic distances among populations were then calculated using the Vincenty (ellipsoid) method within the *geosphere* version 1.5-10 package (Hijmans 2019) in R. Environmental distances were calculated as Euclidean distances using extracted raster values associated with the mean population coordinates from 19 bioclimatic variables, downloaded from WorldClim at 30 arc second resolution (version 2.1; Fick and Hijmans 2017). Values associated with the mean population coordinates for were extracted using the *raster* version 2.5-7 R package. Environmental data were centered and scaled prior to estimation of distances. Additionally, we used a

Mantel test to assess correlation between population-based environmental distances and population-based geographic distances.

Associations between genetic structure and environment

To test the multivariate relationships among genotype, climate, and geography within and across species, redundancy analysis (RDA) was conducted using the *vegan* version 2.5-7 package (Oksanen et al. 2020) in R version 4.0.4 (R Core Development Team, 2021). Genotype data were coded as counts of the minor allele for each sample (i.e., 0,1, or 2 copies) and then standardized following Patterson et al. (2006). Climate raster data (i.e., 19 bioclimatic variables at 30 arc second resolutions), as well as elevational raster data from WorldClim, were extracted, as mentioned above, from geographic coordinates for each sampled tree and then tested for correlation using Pearson's correlation coefficient (r). Five bioclimatic variables that were not highly correlated ($r < |0.75|$) but known to influence diversification in the genus *Pinus* (Jin et al. 2021; Menon et al. 2018) were retained for analysis: Bio 2 (mean diurnal range), Bio 10 (maximum temperature of the warmest quarter), and Bio 11 (minimum temperature of the coldest quarter), Bio 15 (precipitation seasonality), and Bio 17 (precipitation of the driest quarter). The full explanatory data set included these five bioclimatic variables, latitude, longitude, and elevation. The multivariate relationship between genetic variation, climate, and geography was then evaluated through RDA. Statistical significance of the RDA model ($\alpha = 0.05$), as well as each axis within the model, was assessed using a permutation-based analysis of variance (ANOVA) procedure with 999 permutations (Legendre and Legendre 2012). The influence of predictor variables, as well as their confounded effects,

in RDA were quantified using variance partitioning as employed in the *varpart* function of the *vegan* package in R.

Species distribution modeling

To help formulate a testable hypothesis in the inference of demography from genomic data (see Richards et al. 2007), species distribution modeling (SDM) was performed for each species to identify areas of suitable habitat under current climate conditions and across three historical time periods (HOL, ~6 kya, interglacial; LGM, ~21 kya, glacial; and LIG, ~120 kya, interglacial). These temporal inferences were then used to help identify plausible demographic responses. For example, if overlap in modeled habitat suitability changed over time, the hypothesis for demographic inference would include changes in gene flow parameters over time. If the amount of suitable habitat changed over time, the hypothesis would also include changes in effective population size to allow for potential expansions or contractions. This in effect helps to constrain the possible parameter space for exploration.

Occurrence records for *P. pungens* were downloaded from GBIF.org (18th December 2018; GBIF occurrence download, <https://doi.org/10.15468/dl.urehu0>) and combined with known occurrences published by Jetton et al. (2015). For *P. rigida*, all occurrence records were downloaded from GBIF.org (29th December 2015; GBIF occurrence download, <http://doi.org/10.15468/dl.ak0weh>). Records were examined for presence within or close to the known geographical range of each species (Little 1971). Records far outside the known geographic range were pruned. The remaining locations were then thinned to one

occurrence per 10 km to reduce the effects of sampling bias using the *spThin* version 0.1.0.1 package (Aiello-Lammens et al. 2015) in R. The resulting occurrence dataset included 84 records for *P. pungens* and 252 records for *P. rigida* (available at www.github.com/boltece/Speciation_2pines). All subsequent analyses were performed in R version 3.6.2 (R Development Core Team, 2021).

The same bioclimatic variables (Bio2, Bio10, Bio11, Bio15, Bio17) selected for RDA were used in species distribution modeling but were downloaded from WorldClim version 1.4 (Hijmans et al. 2005) at 2.5 arc minute resolution. The change in resolution from above was necessary because paleo-climate data in 30 arc second resolution were not available for the LGM. Paleoclimate raster data for the LGM (~21 kya) and Holocene (HOL, ~6 kya) were downloaded for three General Circulation Models (GCMs; CCSM4, MIROC-ESM, and MPI-ESM). Ensembles were built by averaging the habitat suitability predictions from the three GCMs for each time period (e.g., Menon et al. 2018). SDM predictions associated with each individual GCM, for both the HOL and LGM, were analyzed for incongruences as recommended in Varela et al. (2015). Paleoclimate data for the LIG (~120 kya) were only available at 30 arc second resolution and required downscaling to 2.5 arc minute resolution using the aggregate function (fact = 5) of the *raster* package. Only one GCM is available for the LIG from WorldClim (NCAR-CCSM; Otto-Bliesner et al. 2006); therefore, no ensemble was built.

Raster layers were cropped to the same extent using the *raster* package to include the most northern and eastern extent of *P. rigida*, and the most western and southern extent

of *P. pungens*. Species distribution models (SDMs) were built using MAXENT version 3.4.1 (Phillips et al. 2017) and all possible features and parameter combinations were evaluated using the *ENMeval* version 2.0.0 R package (Kass et al. 2021). Metadata about model fitting and evaluation are available in (Bolte et al. 2022).

The selected features used in predictive modeling were those associated with the best-fit model as determined using AIC. Raw raster predictions were standardized to have the sum of all grid cells equal the value of one using the *raster.standardize* function in the *ENMTools* version 1.0.5 (Warren et al. 2021) R package. Standardized predictions were then transformed to a cumulative raster prediction with habitat suitability scaled from 0 to 1, allowing for quantitative SDM comparisons across species and time. Next, SDM cumulative raster predictions were converted into coordinate points using the *sf* version 0.9-7 R package to calculate the number of points with habitat suitability values greater than 0.5 (i.e., moderate to high suitability areas). Population size expansion or contraction was hypothesized if the number of points increased or decreased over time, respectively. Overlap (i.e., shared points across species) in SDM predictions for each time period was measured using the *inner_join* function in the *dplyr* version 1.0.5 R package. The extent of modeled species distributional overlap was also quantified using the *raster.overlap* function in *ENMTools*, thus providing measures for Schoener's *D* (1968) and Warren's *I* (Warren et al. 2008). Four testable hypotheses were formed from these quantifications. Three of which were formed from predictions associated with each GCM used in HOL and LGM SDMs. The fourth hypothesis was formed from ensembled SDM predictions for the HOL and LGM.

Demographic modeling

Demographic modeling was conducted using Diffusion Approximation for Demographic Inference (*∂α∂i* v.2.0.5; Gutenkunst et al. 2009). A model of pure divergence (SI; strict isolation) was compared against twelve other demographic models representing different potential divergence scenarios with or without gene flow and effective population size changes (Appendix 1, Figure 1.S1). Based on SDM predictions across four time points, we hypothesized that a model that allowed changes in effective population size and rate of gene flow before the LIG would best fit the genetic data. Ten replicate runs of each model were performed in *∂α∂i* with a 200 x 220 x 240 grid space and the nonlinear Broyden-Fletcher-Goldfarb-Shannon (BFGS) optimization routine. Model selection was conducted using Akaike information criterion (AIC; Akaike 1974). The best replicate run (highest log composite likelihood) for each model was then used to calculate ΔAIC ($AIC_{\text{model } i} - AIC_{\text{best model}}$) scores (Burnham and Anderson 2002). From the best supported model, upper and lower 95% confidence intervals (CIs) for all parameters were obtained using the Fisher Information Matrix (FIM)-based uncertainty analysis. Unscaled parameter estimates and their 95% CIs were obtained using a per lineage substitution rate of 7.28×10^{10} substitutions/site/year rate for *Pinaceae* (De La Torre et al. 2017) and a generation time of 25 years (Ma et al. 2006). Genome length (L) a requirement for determining N_{ref} ($= \theta/4\mu L$) from *∂α∂i* parameters, was calculated as the sum across contigs (i.e., RADtags) of the number of bp per SNP. This quantity was calculated for each contig by dividing 92 bp (i.e., the trimmed length of each contig) by the number of SNPs in the contig from the unthinned SNP dataset ($n = 20,932$ SNPs in total). This was

necessary because only a single SNP was retained per contig and counting all bp in a contig would upwardly bias the genome length (i.e., the SNPs were dropped but the bp they occupy would be counted).

Results

Population structure and genetic diversity

A clear separation at the species level was apparent along PC1, which explained 4.232% of the variation across the 2168 SNP x 300 tree data set (Figure 1.2a). Of the 2168 SNPs analyzed, 380 of them were fixed for the same allele across all samples of *P. pungens*, and 196 SNPs were fixed (i.e., not polymorphic) across samples of *P. rigida*. The other 1592 SNPs had variant calls within both species. Lack of population clustering within each species was observed when the PCA was labeled by population (Appendix 1, Figure 1.S2). Using hierarchical *F*-statistics, the estimate of differentiation between species (F_{CT}) was 0.117 (95% CI: 0.099 – 0.136) and similarly to that among all sampled populations ($F_{ST} = 0.123$, 95% CI: 0.106 – 0.143), thus highlighting structure is largely due to differences between species. Differentiation among populations within species was consequently much lower ($F_{SC} = 0.007$ (95% CI: 0.0055-0.0088) whether analyzed jointly (F_{SC}) or separately (see Table 1.2). In the analysis of structure, $K = 2$ had the highest log-likelihood values (Figure 1.2b). Admixture in small proportions (assigning to the other species by 2-10%) was observed in 41 out of the 300 samples (13.67% of samples) across both species. There were 16 trees with ancestry coefficients higher than 10%

assignment to the other species: four *P. rigida* samples (2.29% of sampled *P. rigida*) and twelve *P. pungens* samples (9.60% of sampled *P. pungens*). Admixture proportions were moderately correlated to latitude (Pearson's $r = -0.414$), longitude (Pearson's $r = -0.291$), and elevation (Pearson's $r = 0.445$). All three correlative relationships were significant ($p < 0.001$). Ancestry assignments for each tree at $K = 3$ through $K = 7$ are available in Appendix 1 (Figure 1.S3). All cluster assignments analyzed did not reveal intraspecific population structure. To be certain the signals of admixture were not artifacts of missing data, we plotted the relationship of missing data to the ancestral coefficient for each tree. For the samples with admixture present, the assigned ancestral coefficients at $K = 2$ do not appear to be artifacts of missing data (Appendix 1, Figure 1.S4). Admixture was present in trees with both low and moderate levels of missing data.

Pairwise F_{ST} estimates for *P. pungens* ranged from 0 to 0.0457, while a similar but narrower range of values (0 – 0.0257) was noted for *P. rigida*. The highest pairwise F_{ST} value across both species was between two *P. pungens* populations located in Virginia, PU_DT and PU_BB (Table 1.1). Interestingly, PU_DT in general had higher pairwise F_{ST} values (0.0146 – 0.0457) compared to all the other sampled *P. pungens* populations. For *P. rigida*, the RI_SH population located in Ohio had higher pairwise F_{ST} values for 16 out of the 18 comparisons (0.0123 – 0.0257). The two populations that had low pairwise F_{ST} values with RI_SH were geographically nearby: RI_OH located in Ohio (pairwise $F_{ST} = 0$, distance: 90.1 km) and RI_KY located in Kentucky (pairwise $F_{ST} = 0.0089$, distance: 107.7 km). The highest pairwise F_{ST} value among *P. rigida* populations was between RI_SH and RI_HH, which are geographically distant from one another. From the Mantel tests for IBD

and IBE, Pearson correlations were low (Table 1.2). The correlation with geographical distances was highest for *P. rigida* (Mantel $r = 0.176$, $p = 0.055$). From the Mantel test, the correlation between geographic distance and environmental distance was high for both *P. rigida* ($r = 0.611$, $p = 0.001$) and *P. pungens* ($r = 0.893$, $p = 0.001$).

Heterozygosity estimates for each population are listed in Table 1.1 and were only moderately correlated with geography and elevation. Observed heterozygosity of *P. pungens* ($H_o = 0.127 \pm 0.015$ SD), averaged across SNPs and populations, was higher than the average expected heterozygosity ($H_e = 0.118 \pm 0.008$ SD), both of which were higher than the almost equal values for *P. rigida* ($H_o = 0.102 \pm 0.009$ SD; $H_e = 0.104 \pm 0.005$ SD; Table 1.2). Across both species, observed heterozygosity was mildly associated with geography and elevation. For *P. rigida*, the highest correlation was with elevation ($r = 0.300$, p -value = 0.212), followed by correlation with longitude ($r = 0.113$, p -value = 0.646). Observed heterozygosity in *P. pungens* had a negative correlative relationship with elevation ($r = -0.105$, p -value = 0.721) and positive correlative relationship with longitude ($r = 0.175$, p -values = 0.549). Correlations between latitude and heterozygosity were low in both species ($r = -0.008$ for *P. rigida*; $r = 0.08$ for *P. pungens*; p -values > 0.785).

Associations between genetic structure and environment

The combined effects of climate and geography explained 1.52% (adj. r^2) to 4.16% (r^2) of the genetic variance across 2168 SNPs and 300 sampled trees. The first RDA axis accounted for the bulk of the explanatory variance (42.3%, Figure 1.3) and was the only

RDA axis with a p -value ($p < 0.001$) less than commonly accepted thresholds of significance (e.g., $\alpha = 0.05$). The first RDA was dominated by effects of elevation and Bio15 (precipitation seasonality). Average elevation associated with *P. pungens* samples was 724.68 m (± 224.17 SD), while average elevation across *P. rigida* samples was lower (399.69 m, ± 292.26 SD). The average for Bio15 (precipitation seasonality) was 11.33 (± 1.83 SD) for *P. pungens*, and higher for *P. rigida* (14.23 ± 3.97 SD). Considering the standard deviations around the mean, overlap in values for elevation and precipitation seasonality provide some context to present day overlap in species distributions along the southern Appalachian Mountains. Comparisons of predictor loadings across both RDA axes show latitude, longitude, and Bio11 (mean temperature of the coldest quarter) as also important to explaining the variance both within (RDA 2, 9.77%) and across species (RDA1).

Partitioning the effects of each predictor set revealed that climate independently (i.e., conditioned on geography) accounted for 31.93% of the explanatory variance. Geography independently (i.e., conditioned on climate) accounted for 34.10% of the explained variance. The confounded effect, due to the correlations inherent to the chosen geographic and climatic predictor variables, was 33.97%.

Species distribution modeling

Because population structure within each of the focal species was not observed from our genetic data (i.e., no clear genetic clusters were identified), we produced SDMs using occurrence records across the full distributional range of each species. The best-fit SDM

for *P. pungens* used a linear and quadratic feature class with a 1.0 regularization multiplier, while the SDM for *P. rigida* used a linear, quadratic, and hinge feature class with a regularization multiplier of 3.0. The AUC associated with the training data of the *P. pungens* and *P. rigida* SDMs was 0.929 and 0.912, respectively. Metadata, data inputs, outputs, and statistical results for model evaluation are available in Bolte et al. (2022). The climatic variables with the highest permutation importance were Bio11 (mean temperature of the coldest quarter) and Bio15 (precipitation seasonality) which contributed 41.1% and 39.7% to the *P. pungens* SDM and 19.5% and 62.4% to the *P. rigida* SDM. Of the five climate variables included in the RDA, Bio15 and Bio11 had the highest loadings along RDA axis 1, helping to explain differences across species. The tandem reporting of Bio15 and Bio11 importance to both genetic differentiation and species distributions could be indicative that these climatic variables were drivers in the divergence of these two species.

Distributional overlap was observed in all analyzed SDMs at each of the four time points, therefore all four hypotheses stated that gene flow occurred between the LIG and present day (Figure 1.4). The areas of high habitat suitability shifted substantially over time for both species though, with overlapping areas of suitable habitat exhibiting some of these fluctuations, as well. Current SDMs indicated a larger area of suitable habitat for *P. rigida* (11,128 grid cells had > 0.5 habitat suitability) compared to *P. pungens* (6,632 grid cells) with 14.1% overlap in distributional predictions (Figure 1.4). SDM ensembled predictions for HOL indicated the highest overlap (21.2% of grid cells with > 0.5 habitat suitability), while LGM ensembled predictions indicated the lowest overlap (9.1%). Likewise,

calculations of overlap from full distributional predictions were the lowest (Schoener's $D = 0.217$) for LGM followed by the LIG (Schoener's $D = 0.288$). The highest distributional overlap was associated with the current SDM (Schoener's $D = 0.612$; Figure 1.S5). Raster plots associated with the SDM predictions across the four time points (LGM and HOL ensemble predictions) and species are in Appendix 1, Figure 1.S5.

LGM predictions across the three GCMs varied substantially in terms of where and to what extent there was suitable habitat. We observed drastic reduction in suitable habitat for both species from predictions associated with the CCSM4 GCM. MPI-ESM associated predictions indicated reductions for *P. rigida*, while MIROC associated predictions indicated habitat expansion for *P. rigida* since the LIG. As found in Varela et al. (2015), the use of Bio2 and Bio15 in historical SDM modeling for the LGM led to very different predictions across GCM types making averaged predictions (i.e., the ensemble approach) potentially misleading. We have provided model predictions associated with each LGM-GCM in Appendix 1 (Figure 1.S6). Calculations of overlap from all LGM-GCM predictions (range = 2.0 - 18.3%) were lower than overlap estimates from other time periods providing some indication of consistency and usefulness to the widely implemented ensemble technique. For the HOL, predictions were more similar across GCMs with overlap varying between 13.1 and 20.5% (Appendix 1, Figure 1.S7). Hypotheses associated with each GCM and the ensemble are presented in Figure 1.4.

The ensembled prediction for *P. pungens* and *P. rigida* during the LGM shows multiple potential refugial areas that overlap (Figure 1.S5). From the MIROC-ESM GCM-based

model predictions, interspecific gene flow during the LGM may have been possible just south of the glacial extent, but CCSM4 and MPI-ESM GCM-based predictions (Figure 1.S6) indicate two, small overlapping refugial regions farther south than where either species currently occurs. Ensembled distributions for *P. pungens* and *P. rigida* during the HOL were proximal to each other, with high habitat suitability west of and along the Appalachian Mountains (Figure 1.S5). These distributions may have promoted both intraspecific and interspecific gene flow to occur ~6 kya.

Demographic modeling

The best replicate run (highest composite log-likelihood) for each of the thirteen modeled divergence scenarios, their associated parameter outputs, and ΔAIC ($AIC_{\text{model } i} - AIC_{\text{best model}}$) are summarized in Appendix 1 (Table 1.S1 and Table 1.S2). A model that allowed changes in both effective population size and rate of symmetrical gene flow across two time periods (PSCMIGCs) best fit the 2168 SNP data set (Table 1.2) and had small, normally distributed residuals (Figure 1.S8). This model was 20.84 AIC units better than the second best-fit model (PSCMIGs; Table 1.3), which inferred change in population size estimates across two time intervals but inferred only one, constant symmetrical gene flow parameter across time intervals.

Initial divergence was estimated to be 2.74 mya (95% CI: 2.25 – 3.24). The first time interval during divergence (T_1) lasted 98.7% of the total divergence time with symmetrical gene flow (M_i) occurring at a rate of 48.6 (95% CI: 33.1 – 64.1) migrants per generation (Figure 1.5). The effective size of the ancestral population (N_{ref}) was 36,137 (95% CI: 31,367 – 40,908; Figure 1.5) prior to divergence. For most of the divergence history, *P.*

pungens had an effective population size of $N_{P1} = 1,024,573$ (95% CI: 140,601 - 1,908,546) while *P. rigida* had a relatively smaller, but still large, effective size of $N_{R1} = 758,920$ (95% CI: 214,423 - 1,303,417). The second time interval (T_2) during divergence was estimated to have begun 35.2 kya (95% CI: 32.9 - 37.4) when effective population sizes decreased instantaneously to 3,448 (95% CI: 3,226 - 3,669) for *P. pungens* (N_{P2}) and 3,935 (95% CI: 3,679 - 4,191) for *P. rigida* (N_{R2}). During this time interval, the relative rate of symmetrical gene flow dropped from 48.6 to 38.4 (95% CI: 35.7 – 41.1) migrants per generation.

Discussion

Using a multidisciplinary approach, we demonstrated that the divergence history of *P. pungens* and *P. rigida* involved a complex mixture of population size changes linked to changing climates, as well as changing rates of gene flow. We also demonstrated that consideration of each GCM-based SDM prediction is important to hypothesis formation for phylogeographic and demographic inference studies as the more widely employed method of ensembling historical SDM predictions can be misleading, especially when inferences include population size change. All four of our SDM hypotheses were supported in terms of gene flow occurrence since the LIG, but only Hypothesis 1 (CCSM4) for population size change since the LIG was supported by genetic data. The best-fit demographic model using 2168 SNPs as summarized using the multidimensional site

frequency spectrum indicated initial divergence to have occurred 2.74 mya, an estimate similar to the one inferred in Saladin et al. (2017; 2.66 mya). Our best-fit model also indicated a large reduction in effective population size which coincided with a reduction in gene flow during the last glacial period (~10,000 years before the last glacial maxima). A three-epoch model to test SDM observations of expansion since the LGM was included, but model fit did not improve. This could be due to the more pronounced impact of a recent bottleneck to site frequency spectrum patterns or that our data simply did not capture expansion.

Climate drives divergence

The total divergence time inferred for *P. pungens* and *P. rigida* (2.74 mya) aligns with the onset of the Quaternary Period (~2.6 mya), a time period widely recognized as driving adaptations to seasonality for many temperate species (Dobzhansky 1950; Savolainen et al. 2004; Jump and Penuelas 2005; Williams and Jackson 2007; Bonebrake and Mastrandea 2010). For *P. pungens* and *P. rigida*, Bio15 (precipitation seasonality) was important to genetic differentiation (RDA) and species distributions (SDMs) which strongly implies adaptations to seasonality were drivers of divergence. Phenological traits have been linked to seasonal variation within various plant species of North America (Jump and Penuelas 2005), and differences in seasonality requirements for *P. pungens* and *P. rigida* likely explain the observed trait differences in seed size, reproductive age, timing of pollen release, and rates of seedling establishment across these two species (Zobel 1969; Della-Bianca 1990; Ledig et al. 2015).

Using niche and trait data, the phylogenetic inference of Jin et al. (2021) also identified precipitation seasonality (Bio15) as a driver of diversification in eastern North American pines along with Bio1 (annual mean temperature), Bio8 (mean temperature of the wettest quarter), elevation, and soil silt content. Although three of these variables were not included in our RDA, the two that were (i.e., Bio15 and elevation) were most important to explaining species level genetic differences. In terms of distributional differences between these two species, narrow niche requirements for Bio15 and elevation help explain the patchy distribution of *P. pungens* along the southern Appalachian Mountains, while contrastingly, populations of *P. rigida* may have evolved a response to increased precipitation seasonality during the Quaternary period. In a study of pinyon pine diversification, Ortiz-Medrano et al. (2016) suggested the response to seasonality as potentially linked to the evolution of plasticity. This could explain *P. rigida*'s less stringent niche requirements for Bio15 and elevation, larger geographic distribution, greater trait variation, and proposed latitudinal expansion into northeastern North America (Ledig et al. 2015).

The evolution of fire-related traits in pines has been linked to the mid-Miocene period, but fire intensity and frequency in certain geographic regions have been cyclical in nature allowing the evolution of adaptive traits related to fire endurance, tolerance, or avoidance possible across multiple geologic time scales (e.g., He et al. 2012; Lafon et al. 2017; Jin et al. 2021). Fine-scale geographical distributions of our focal species are locally divergent across slope aspects in the Appalachian Mountains, with *P. pungens* primarily distributed on southwestern slopes and *P. rigida* primarily distributed on southeastern

slopes (Zobel 1969). Currently, there is higher fire frequency and intensity on western slopes. The high levels of cone serotiny and fast seedling development associated with *P. pungens* are evolved strategies that confer population persistence in more active fire regimes (Zobel 1969). Although some northern *P. rigida* populations exhibit serotiny, the populations found along the southern Appalachian Mountains, and proximal to *P. pungens*, have nonserotinous cones and other traits consistent with enduring fire (e.g., thick bark and epicormics; Zobel 1969) as opposed to relying on it (Jin et al. 2021). With these factors in mind and the correlative evidence between fire intensity and level of serotiny presented across populations of other pine species (*P. halepensis* and *P. pinaster*; Hernandez-Serrano et al. 2013), we suspect genomic regions involved in the complex, polygenic trait of serotiny (Parchman et al. 2012; Budde et al. 2014) may have also contributed to the rapid development reproductive isolation between our focal species.

Reproductive isolation can evolve rapidly during speciation

While *P. pungens* and *P. rigida* can be found on the same mountain and even established within a few meters of each other, mountains are heterogeneous, complex landscapes offering opportunity for niche evolution along multiple axes of biotic and abiotic influence for parental species and hybrids alike. The distances to disperse into novel environments are relatively short in these heterogeneous landscapes thus suggesting diversification could be more rapid as environmental complexity increases (Bolte and Eckert 2020). Mountains have rain shadow regions characterized by drought and thus more active fire

regimes (Parisien and Moritz 2009). A host of adaptive traits in trees are associated with fire frequency and intensity (Pausas and Schwilk 2012). Among those, the genetic basis of serotiny is characterized as being polygenic with large effect loci in *P. contorta* Dougl. (Parchman et al. 2012) and in *P. pinaster* Aiton (Budde et al. 2014). Such genetic architectures, even in complex demographic histories such as the one described here, can evolve relatively rapidly to produce adaptive responses to shifting optima (e.g., Stetter et al. 2018; reviewed for forest trees by Lind et al. 2018), so that it is not unreasonable to expect divergence in fitness-related traits such as serotiny to also contribute to niche divergence and reproductive isolation. Considering large effect loci associated with serotiny were also associated with either water stress response, winter temperature, cell differentiation, or root, shoot, and flower development (Budde et al. 2014), serotiny may be a trait that contributes to widely distributed genomic islands of divergence thus explaining the development of ecologically based reproductive isolation between *P. pungens* and *P. rigida* amid recurring gene flow (Nosil and Feder 2012). Given that our focal species are reciprocally crossable to yield viable offspring (Critchfield 1963), it is likely that postzygotic ecological processes, such as selection for divergent fire-related and climatic niches, limits hybrid viability in natural stands as a form of reinforcement layered on the aforementioned prezygotic divergence of phenological schedules. Indeed, hybrids are rarely identified in sympatric stands (Zobel 1969; Brown 2021). Thus, it appears that niche divergence is associated with divergence in reproductive phenologies during speciation for our focal taxa. Whether niche divergence reinforces reproductive

isolation based on pollen release timing or divergent pollen release timing is an outcome of niche divergence itself, however, remains an open question.

The rate of gene flow in our best-fit demographic model was reduced by approximately 10 migrants per generation providing evidence that prezygotic reproductive isolation may have strengthened during the glacial period. This reduction reflects a scenario of reduced effective population sizes, reduced rates of gene flow (m), or both. The rate of gene flow associated with a given time interval should not be interpreted as constant, though. Sousa et al. (2011) found that posterior distributions for the timing of gene flow parameters in demographic inference were highly variable across the simulations they performed making pulses of gene flow (i.e., a gene flow event occurring within a time frame of no active gene flow), as probable as constant, ongoing gene flow. This likely explains the high levels of gene flow inferred using $\partial\alpha\partial i$ with the empirical lack of frequent and identifiable hybrids in extant samples of each species (Figure 1.2; Brown 2021). While acknowledging this blurs interpretation of parameter estimates for gene flow, a history with recurring gene flow events fits the narrative of prezygotic isolation being labile especially when geographical distributions or reproductive phenology are the factors involved. Indeed, observations of hybridization occurring between once prezygotically isolated species have been made and suggests phenological barriers such as timing of pollen release and flowering may not be permanently established and can shift towards synchrony in warming climates (Vallejo-Marín and Hiscock 2016).

Climate instability reduces genetic diversity

Conifers often have high levels of genetic diversity and low levels of population differentiation because of outcrossing, wind-dispersion, and introgression (Petit and Hampe 2006). *Pinus pungens* and *P. rigida* both have modest levels of genetic diversity within and across the populations we sampled, and no detectable within-species population structure given our genome-wide data. Our best-fit model inferred a drastic effective population size reduction (*P. pungens*, ~99.7%; *P. rigida*, ~99.5%) 35 kya. Since then, climate has continued to oscillate between extreme warming and cooling events (Jackson and Overpeck, 2000) and for geologic time intervals too short for species with long generation times and low migratory potential to sufficiently track causing a mismatch between the breadth of a species' climatic niche and where populations are established (Svenning et al. 2015). This dynamic affects population persistence, reduces genetic variation within populations due to excessive mortality, and thus to some degree limits the potential for local adaptation in climatically unstable regions. The lack of IBD and IBE across the populations of our focal species can be explained in one of two ways, the mismatch described in Svenning et al. (2015) or the primarily nongenic regions investigated in our RADseq data reflect little to no structure. Our SDM predictions showed substantial shifts in habitat suitability since the LIG, providing evidence of high climate instability in temperate eastern North America during the Quaternary period. We acknowledge though that niche conservatism is an underlying assumption in historical SDMs, so interpretations were done cautiously. Gene flow and local adaptation affect

niche dynamics in various ways (Pearman et al. 2008), but neither of these processes were able to be accounted for in our SDMs.

From a theoretical standpoint, we anticipated the patchy, mountain top distribution of *P. pungens* to be characterized by strong patterns of population differentiation. Lack of structure in *P. pungens* could be attributed to long distance dispersal or a recent move up in elevation with genomes still housing elements of historical panmixia. Indeed, suitable habitat predictions during the HOL, just 6000 years ago, were rather contiguously distributed (Figure 1.S5 and Figure 1.S7) and may have allowed an increase in intraspecific gene flow. For *P. rigida* some structure differentiating the northern populations from those along the southern Appalachian Mountains was expected from an empirical standpoint because previously reported trait values in a common garden study led to identification of three latitudinally arranged genetic groupings (Ledig et al. 2015). Although structure analysis did not support groupings within *P. rigida*, our estimates for isolation-by-distance (IBD) yielded a correlation of 0.177 ($p = 0.055$) which is suggestive of structure. While this shows some differentiation across its distribution, pairwise F_{ST} values were small and on average smaller than those between populations of *P. pungens* suggesting higher population connectivity in *P. rigida*. The three GCM-based SDM predictions for both *P. pungens* and *P. rigida* differed substantially but did consistently show two or three disjunct refugia where gene flow dynamics intraspecifically and interspecifically may have been affected. Even though genetic differences may have accumulated in these separate refugia, the SDM predictions for the HOL were more

compact and contiguous for our focal taxa, providing greater potential for intraspecific gene flow across diverged populations and the reestablishment of interspecific gene flow under a warming climate.

Future work and conclusions

The divergence history of *P. pungens* and *P. rigida* involved a complex interplay of recurring interspecific gene flow and dramatic population size reductions associated with changes in climate. Future detailed examinations of hybridization between *P. pungens* and *P. rigida* are needed to elucidate the role hybridization plays in the maintenance of species boundaries. Ideally, future research involving these two species would use a method that sufficiently captures genic regions so population structure in both species may be revealed and investigations into genomic islands of divergence that are often associated with ecological speciation can be performed (Nosil and Feder 2012). It may also be of interest to conduct population genetic analyses from chloroplast and mitochondrial DNA to obtain resolved inferences of gene flow directionality (i.e., asymmetry) and population connectivity.

While more time, effort, and genomic resources are needed for us to accurately predict gains and losses in biodiversity or describe the development of reproductive isolation in conifer speciation, we must recognize that some montane conifer species will be disproportionately affected by future climate projections (Aitken et al. 2008) and time is of the essence in terms of capturing and understanding current levels of biodiversity. High

elevational species such as *P. pungens* may already be experiencing a tipping point, but because *P. pungens* is a charismatic Appalachian tree with populations already threatened by fire suppression practices over the last century, conservation efforts have begun through seed banking (Jetton et al. 2015) and prescribed burning experiments of natural stands (Welch and Waldrop 2001). Our contributions to these conservation efforts include genome-wide population diversity estimates for *P. pungens* and *P. rigida* and a demographic inference scenario that involves a long history of interspecific gene flow. In conifer species of the family *Pinaceae*, there are multiple accounts of introgression occurring through hybrid zones (De La Torre et al. 2014; Hamilton et al. 2015; Menon et al. 2018). The implications of introgression are far-reaching as it leads to greater genetic diversity and thus a greater capacity for adaptive evolution. Trees are often foundation species in many plant communities, so understanding a population's potential to withstand environmental changes provides some insight into the future stability of the ecological communities dominated by these charismatic plant taxa.

Table 1.1 Location of sampled populations, number of trees (n) that were sampled, and the observed heterozygosity (H_o) versus the expected heterozygosity ($H_e = 2pq$) for *Pinus pungens* and *P. rigida* populations.

Species	Code	Location	Lat	Long	n	H_o	H_e
<i>P. pungens</i>	PU_BB	Briery Branch, VA	38.48	-79.22	8	0.110	0.108
<i>P. pungens</i>	PU_BN	Buchanan State Forest, PA	39.77	-78.43	6	0.141	0.121
<i>P. pungens</i>	PU_BV	Buena Vista, VA	37.76	-79.29	11	0.124	0.120
<i>P. pungens</i>	PU_DT	Dragon's Tooth, VA	37.37	-80.16	7	0.101	0.098
<i>P. pungens</i>	PU_EG	Edinburg Gap, VA	38.79	-78.53	8	0.139	0.124
<i>P. pungens</i>	PU_EK	Elliott Knob, VA	38.17	-79.30	10	0.131	0.123
<i>P. pungens</i>	PU_GA	Walnut Fork, GA	34.92	-83.28	10	0.129	0.123
<i>P. pungens</i>	PU_LG	Looking Glass Rock, NC	35.30	-82.79	8	0.130	0.119
<i>P. pungens</i>	PU_NM	North Mountain, VA	37.82	-79.63	12	0.130	0.121
<i>P. pungens</i>	PU_PM	Poor Mountain, VA	37.23	-80.09	11	0.130	0.125
<i>P. pungens</i>	PU_SC	Pine Mountain, VA	34.70	-83.30	8	0.128	0.122
<i>P. pungens</i>	PU_SH	Shenandoah NP, VA	38.55	-78.31	5	0.160	0.128
<i>P. pungens</i>	PU_SV	Stone Valley Forest, PA	40.66	-77.95	9	0.110	0.110
<i>P. pungens</i>	PU_TR	Table Rock Mountain, NC	35.89	-81.88	12	0.113	0.114
<i>P. rigida</i>	RI_BR	Bass River State Forest, NJ	39.80	-74.41	9	0.101	0.105
<i>P. rigida</i>	RI_CT	Pachaug State Forest, CT	41.54	-71.81	10	0.096	0.107
<i>P. rigida</i>	RI_DT	Dragon's Tooth, VA	37.37	-80.16	10	0.109	0.106
<i>P. rigida</i>	RI_GA	Chattahoochee NF, GA	34.75	-83.78	9	0.096	0.103
<i>P. rigida</i>	RI_GW	George Washington NF, VA	38.36	-79.20	10	0.102	0.103
<i>P. rigida</i>	RI_HH	Hudson Highlands State Park, NY	41.44	-73.97	7	0.102	0.101
<i>P. rigida</i>	RI_JF	Jefferson NF, VA	37.15	-82.64	10	0.095	0.100
<i>P. rigida</i>	RI_KY	Daniel Boone NF, KY	37.84	-83.62	9	0.113	0.110
<i>P. rigida</i>	RI_ME	Acadia NP, ME	44.36	-68.19	10	0.107	0.106
<i>P. rigida</i>	RI_MI	Michaux State Forest, PA	39.98	-77.44	10	0.123	0.114
<i>P. rigida</i>	RI_NJ	Wharton State Forest, NJ	39.68	-74.53	9	0.098	0.101
<i>P. rigida</i>	RI_NY	Macomb State Park, NY	44.63	-73.58	9	0.101	0.104
<i>P. rigida</i>	RI_OH	South Bloomingville, OH	39.45	-82.59	8	0.093	0.096
<i>P. rigida</i>	RI_RS	Rome Sand Plains, NY	43.23	-75.56	9	0.097	0.103
<i>P. rigida</i>	RI_SH	Shawnee State Park, OH	38.75	-83.13	9	0.082	0.094
<i>P. rigida</i>	RI_SP	Sproul State Forest, PA	41.24	-77.78	9	0.106	0.105
<i>P. rigida</i>	RI_TN	Great Smoky Mountains NP, TN	35.68	-83.58	8	0.099	0.104
<i>P. rigida</i>	RI_TR	Table Rock Mountain, NC	35.89	-81.89	10	0.113	0.112
<i>P. rigida</i>	RI_VT	Bellows Falls, VT	43.11	-72.44	10	0.098	0.104

Table 1.2 Summary statistics of genetic differentiation for the sampled populations of *P. rigida* and *P. pungens*. Expected (H_e) and observed heterozygosity (H_o) values are the averages across 2168 SNPs averaged across populations.

Species	F_{ST} (95% CI)	IBD r (p -value)	IBE r (p -value)	H_e (range)	H_o (range)
<i>P. pungens</i>	0.0057 (0.0032 - 0.0084)	-0.0789 (0.638)	0.0131 (0.411)	0.118 (0.098-0.129)	0.127 (0.101-0.160)
<i>P. rigida</i>	0.0056 (0.0032 - 0.0082)	0.1758 (0.055)	-0.0669 (0.633)	0.104 (0.094-0.114)	0.102 (0.082 -0.123)

Table 1.3 Results of model fitting for thirteen representative demographic models of divergence. Models are ranked by the number of parameters (k). Log-likelihood ($\log L$) and Akaike information criterion (AIC) are provided for each model. Model details are given in the footnote.

Model	k	$\log L$	AIC
SI	3	-2254.18	4,514.37
MIGs	4	-2201.51	4,411.02
MIGa	5	-2210.81	4,431.62
SCs	5	-2213.93	4,437.86
SGFs	5	-2229.65	4,469.30
SCa	6	-2238.03	4,488.06
SGFa	6	-2241.07	4,494.14
PSC	6	-2277.78	4,567.56
PSCSCs	7	-2178.16	4,370.32
PSCMIGs	7	-1866.42	3,746.84
PSCMIGCs	8	-1853.99	3,726.00
PSCMIGa	10	-2117.91	4,251.82
PSCMIGCs_T3	12	-1925.86	3,875.71

SI, strict isolation; MIGs, symmetrical gene flow; MIGa, asymmetrical gene flow; SCs, secondary contact with symmetrical gene flow; SCa, secondary contact with asymmetrical gene flow; SGFa, speciation with asymmetrical gene flow; SGFs, speciation with symmetrical gene flow; PSC, population size change; MIGCs, change in rate of symmetrical gene flow; T3, for three time intervals. The best-fit model is in bold.

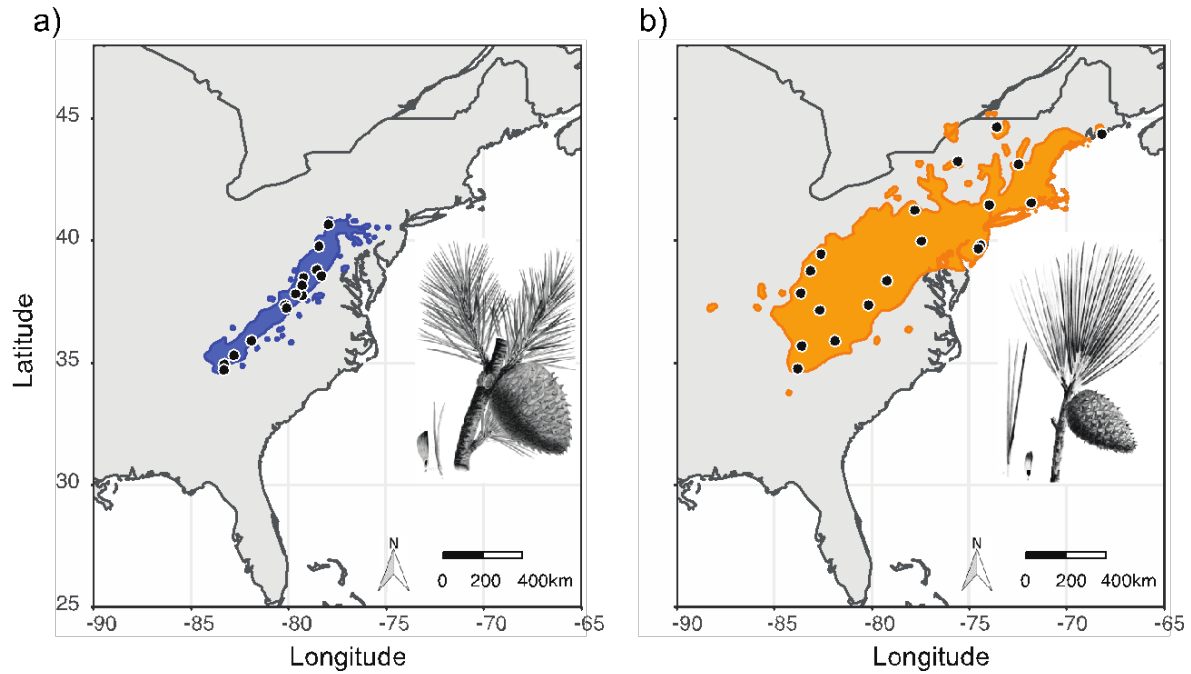


Figure 1.1 Known geographical distribution of focal species, a) *Pinus pungens* and b) *P. rigida*, (Little 1971) in relation to populations sampled (black dots) for genetic analysis; Phenotypic characterization of each species was illustrated by Pierre-Joseph Redouté (Michaux 1819).

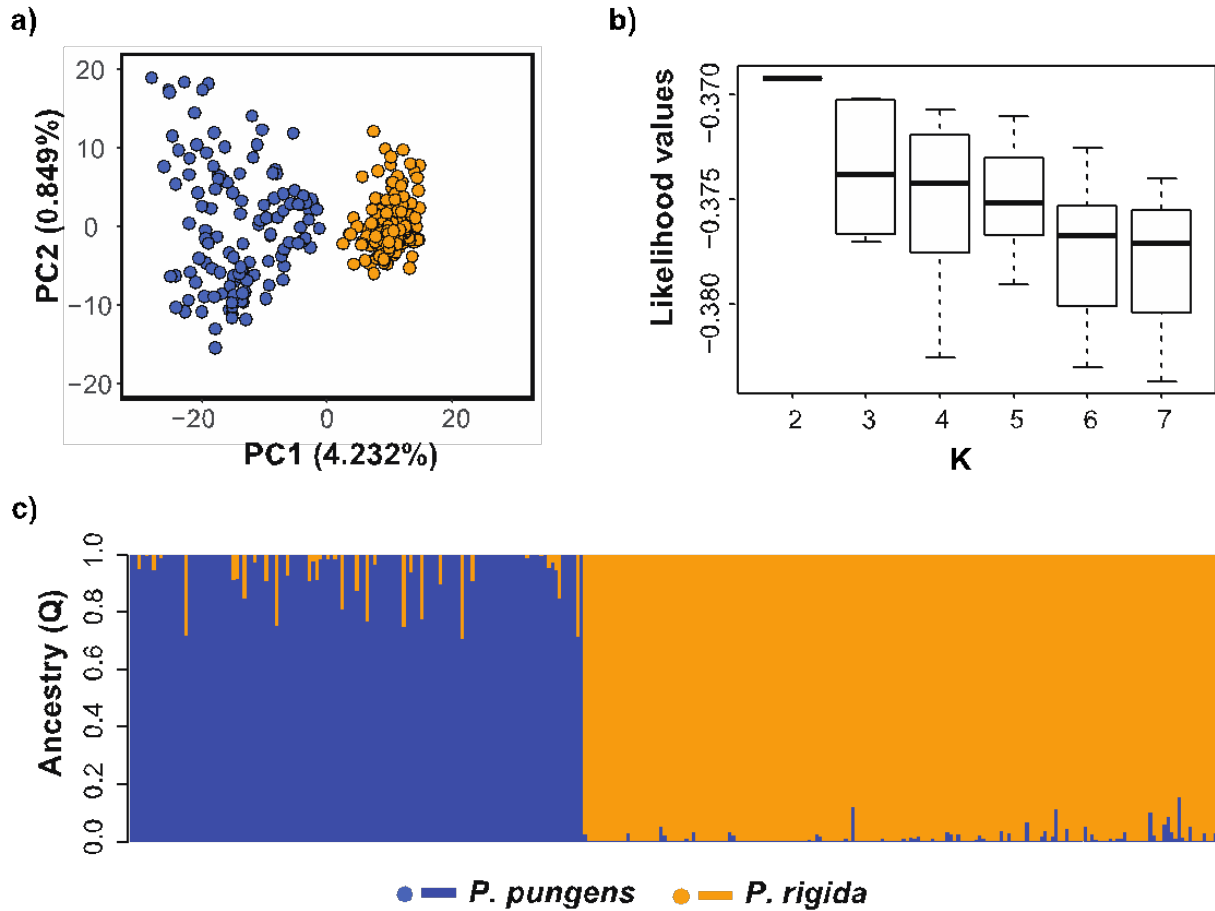


Figure 1.2 Measures of genetic differentiation and diversity among sampled trees of *P. pungens* and *P. rigida*: a) Principal components analysis of 2168 genome-wide single nucleotide polymorphism (SNPs) for *Pinus pungens* (blue, left side of PC1) and *P. rigida* (orange, right side of PC1); b) log-likelihood values across ten replicate runs in fastSTRUCTURE for $K = 2$ through $K = 7$; c) results of averaged $K = 2$ ancestry (Q) assignments for each sample arranged latitudinally in each species.

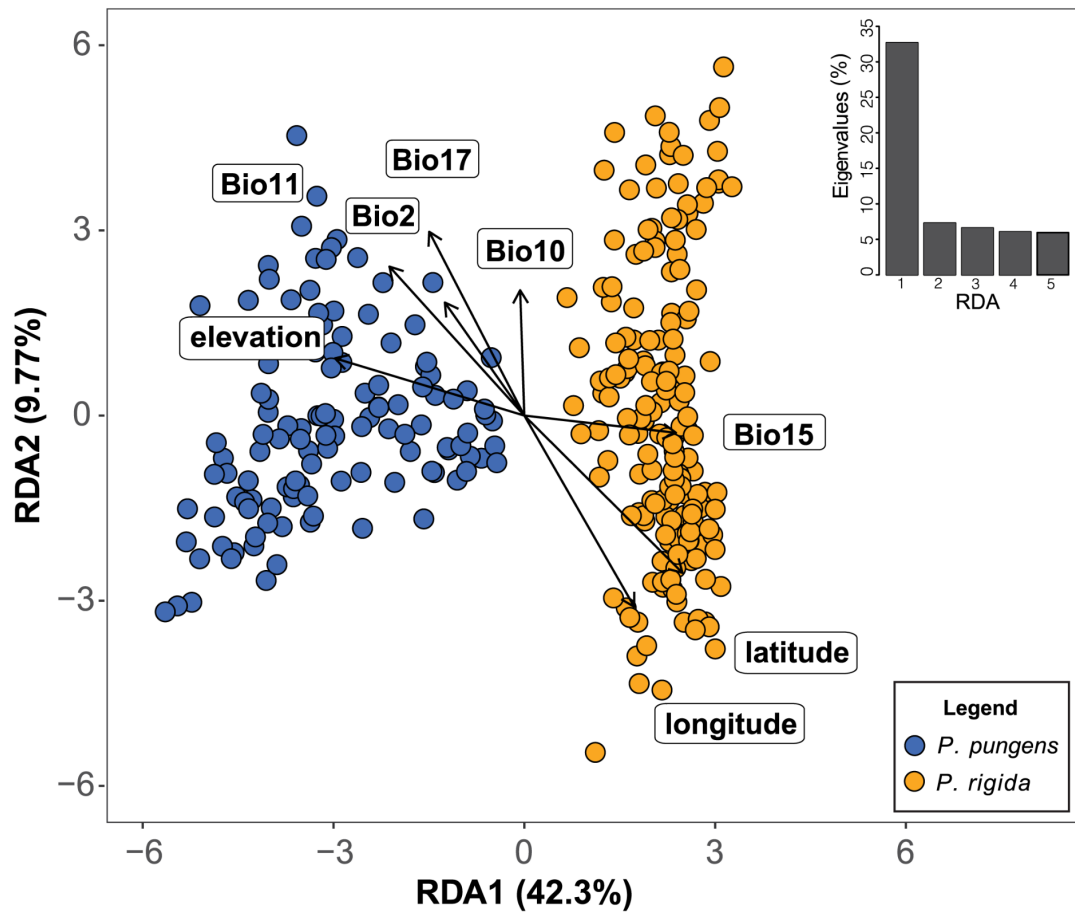


Figure 1.3 Redundancy analysis (RDA) of the multilocus genotypes for each tree with climate and geographic predictor variables (full model). Direction and length of arrows on each RDA plot correspond to the loadings of each variable.

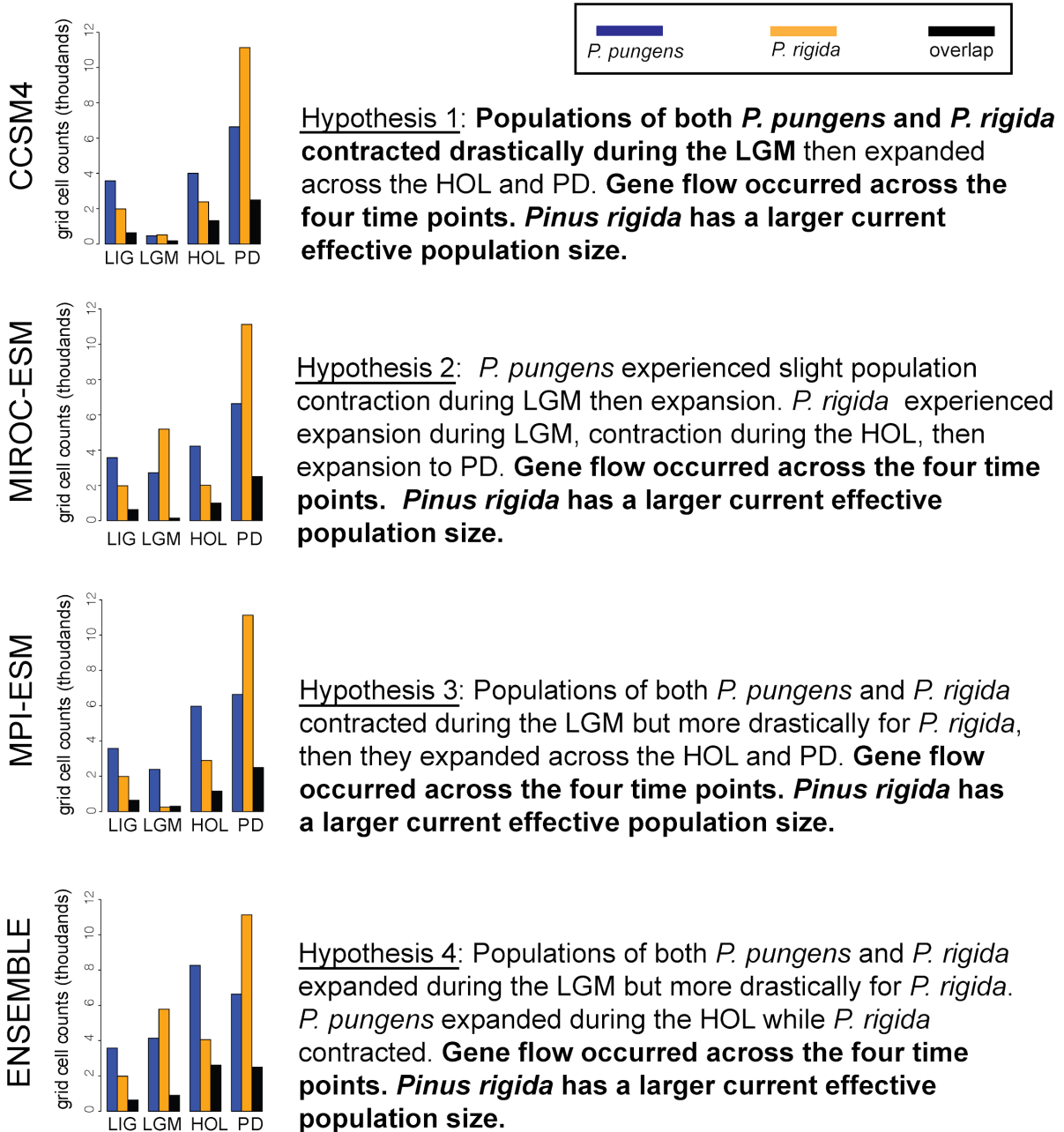


Figure 1.4 Hypotheses associated with each SDM - GCM model prediction versus the ensemble SDM prediction based on relative grid cell counts of high habitat suitability (> 0.5) for *P. rigida*, *P. pungens*, and overlap across four time periods (LIG, LGM, HOL, and PD). Bolded text were statements supported by the best-fit model of demographic inference.

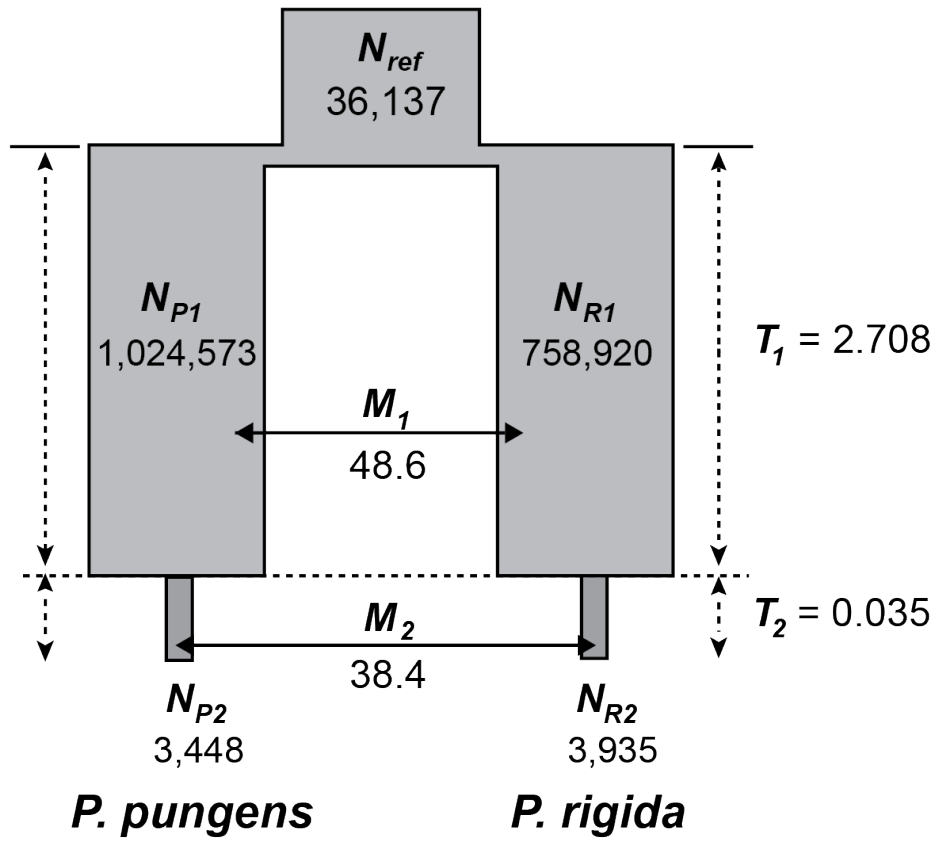


Figure 1.5 The best-fit model (PSCMIGCs) and unscaled parameter estimates from $\partial\alpha\partial i$ analysis. Time intervals (T_i) are represented in millions of years and associated with lineage population sizes (N_i) and a specific rate of symmetrical gene flow (M_i).

Data Archiving Statement

Raw reads generated during this study are available at NCBI SRA database under BioProject: PRJNA803632 (Sample IDs: SAMN25684544 – SAMN25684843). Python scripts for demographic modeling and R scripts for genetic analyses and producing SDMs are available at www.github.com/boltece/Speciation_2pines.

Literature Cited

- Abbott RJ (2017) Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. *J Syst Evol* 55:238–258. <https://doi.org/10.1111/jse.12267>
- Aiello-Lammens ME, Boria RA, Radosavljevic A, et al (2015) spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* 38:541–545. <https://doi.org/10.1111/ECOG.01132>
- Aitken SN, Yeaman S, Holliday JA, et al (2008) Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evol Appl* 1:95–111. <https://doi.org/10.1111/J.1752-4571.2007.00013.X>
- Baack E, Melo MC, Rieseberg LH, Ortiz-Barrientos D (2015) The origins of reproductive isolation in plants. *New Phytol* 207:968–984. <https://doi.org/10.1111/NPH.13424>
- Bagley JC, Heming NM, Gutiérrez EE, et al (2020) Genotyping-by-sequencing and ecological niche modeling illuminate phylogeography, admixture, and Pleistocene range dynamics in quaking aspen (*Populus tremuloides*). *Ecol Evol* 10:4609–4629. <https://doi.org/10.1002/ece3.6214>
- Bolte CE, Eckert AJ (2020) Determining the when, where and how of conifer speciation: a challenge arising from the study ‘Evolutionary history of a relict conifer *Pseudotsaxus chienii*.’ *Ann Bot* 125:v–vii. <https://doi.org/10.1093/AOB/MCZ201>
- Bonebrake TC, Mastrandrea MD (2010) Tolerance adaptation and precipitation changes complicate latitudinal patterns of climate change impacts. *Proc Natl Acad Sci U S A* 107:12581–12586. <https://doi.org/10.1073/pnas.0911841107>
- Brown AL (2021) Phenotypic characterization of Table Mountain (*Pinus pungens*) and pitch pine (*Pinus rigida*) hybrids along an elevational gradient in the Blue Ridge Mountains, Virginia. Thesis, Virginia Commonwealth University
- Cannon CH, Petit RJ (2020) The oak syngameon: more than the sum of its parts. *New Phytol* 226:978–983. <https://doi.org/10.1111/nph.16091>

- Capblancq T, Butnor JR, Deyoung S, et al (2020) Whole-exome sequencing reveals a long-term decline in effective population size of red spruce (*Picea rubens*). *Evol Appl* 13:2190. <https://doi.org/10.1111/EVA.12985>
- Carstens BC, Richards CL (2007) Integrating coalescent and ecological niche modeling in comparative phylogeography. *Evolution* 61:1439–1454. <https://doi.org/10.1111/J.1558-5646.2007.00117.X>
- Cavender-Bares J (2019) Diversification, adaptation, and community assembly of the American oaks (*Quercus*), a model clade for integrating ecology and evolution. *New Phytol* 221:669–692. <https://doi.org/10.1111/nph.15450>
- Christe C, Stölting KN, Paris M, et al (2017) Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Mol Ecol* 26:59–76. <https://doi.org/10.1111/mec.13765>
- Critchfield WB (1963) The Austrian x red pine hybrid. *Silvae Genet* 12:187-191
- Critchfield WB (1967) Crossability and relationships of the closed-cone pines. *Silvae Genet* 16:89–97
- Critchfield WB (1986) Hybridization and Classification of the White Pines (*Pinus* Section *Strobus*). *Taxon* 35:647–656
- Danecek P, Auton A, Abecasis G, et al (2011) The variant call format and VCFtools. *Bioinformatics* 27:2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- De La Torre AR, Birol I, Bousquet J, et al (2014) Insights into conifer giga-genomes. *Plant Physiol* 166:1724–1732. <https://doi.org/10.1104/pp.114.248708>
- De La Torre AR, Li Z, Van De Peer Y, Ingvarsson PK (2017) Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Mol Biol Evol* 34:1363–1377. <https://doi.org/10.1093/molbev/msx069>
- Della-Bianca L (1990) *Pinus pungens* Lamb., Table Mountain pine. In: Burns, R.M. and B.H. Honkala (eds.). *Silvics of North America. Volume 1. Conifers*. USDA Forest Service Agriculture Handbook 654, Washington, D.C.
- Dobzhansky T (1950) *Heredity, Environment, and Evolution*. *Assoc Adv Sci* 111:161-166. <https://doi.org/10.1126/science.111.2877.161>
- Dorman KW, Barber JC (1956) Time of flowering and seed ripening in southern pines. USDA Forest Service, Southeastern Forest Experiment Station, Old Station Paper SE-072, 72.

- Dyke AS, Moore A, Robertson L (2003) Deglaciation of North America, Open File 1574, Natural Resources Canada, Ottawa.
- Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol* 37:4302–4315. <https://doi.org/10.1002/JOC.5086>
- Francis RM (2017) pophelper: an R package and web app to analyse and visualize population structure. *Mol Ecol Resour* 17:27–32. <https://doi.org/10.1111/1755-0998.12509>
- Fu L, Niu B, Zhu Z, et al (2012) CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Gao J, Wang B, Mao JF, et al (2012) Demography and speciation history of the homoploid hybrid pine *Pinus densata* on the Tibetan Plateau. *Mol Ecol* 21:4811–4827. <https://doi.org/10.1111/j.1365-294X.2012.05712.x>
- Gernandt DS, Aguirre Dugua X, Vázquez-Lobo A, et al (2018) Multi-locus phylogenetics, lineage sorting, and reticulation in *Pinus* subsection *Australes*. *Am J Bot* 105:711–725. <https://doi.org/10.1002/AJB2.1052>
- Goudet J, Jombart T (2020) hierfstat, estimations and tests of hierarchical *F*-statistics. R package version 0.5-7. <https://CRAN.R-project.org/package=hierfstat>
- Gougherty A V., Chhatre VE, Keller SR, Fitzpatrick MC (2020) Contemporary range position predicts the range-wide pattern of genetic diversity in balsam poplar (*Populus balsamifera* L.). *J Biogeogr* 47:1246–1257. <https://doi.org/10.1111/jbi.13811>
- Gugger PF, Ikegami M, Sork VL (2013) Influence of late Quaternary climate change on present patterns of genetic variation in valley oak, *Quercus lobata* Née. *Mol Ecol* 22:3598–3612. <https://doi.org/10.1111/MEC.12317>
- Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5:1-11. <https://doi.org/10.1371/journal.pgen.1000695>
- Hagman M (1975) Incompatibility in forest trees. *Proc R Soc London Ser B Biol Sci* 188:313–326.
- Hamilton JA, De la Torre AR, Aitken SN (2015) Fine-scale environmental variation contributes to introgression in a three-species spruce hybrid complex. *Tree Genet Genomes* 11. <https://doi.org/10.1007/s11295-014-0817-y>

- Hapke A, Thiele D (2016) GIBPSs: a toolkit for fast and accurate analyses of genotyping-by-sequencing data without a reference genome. *Mol Ecol Resour* 16:979–990. <https://doi.org/10.1111/1755-0998.12510>
- Hernández-León S, Gernandt DS, Pérez de la Rosa J a., Jardón-Barbolla L (2013) Phylogenetic relationships and species delimitation in *Pinus* section *Trifoliae* inferred from plastid DNA. *PLoS One* 8:1–14. <https://doi.org/10.1371/journal.pone.0070501>
- Hernández-Serrano A, Verdú M, González-Martínez SC, Pausas JG (2013) Fire structures pine serotiny at different scales. *Am J Bot* 100:2349–2356. <https://doi.org/10.3732/ajb.1300182>
- Herten K, Hestand MS, Vermeesch JR, Van Houdt JKJ (2015) GBSX: A toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinformatics* 16:1-6. <https://doi.org/10.1186/s12859-015-0514-3>
- Hewitt GM (2001) Speciation, hybrid zones and phylogeography - Or seeing genes in space and time. *Mol Ecol* 10:537–549. <https://doi.org/10.1046/J.1365-294X.2001.01202.X>
- Hewitt GM (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philos Trans R Soc Lond B* 359:183–195. <https://doi.org/10.1098/rstb.2003.1388>
- Hickerson MJ, Carstens BC, Cavender-Bares J, et al (2010) Phylogeography's past, present, and future: 10 years after. *Mol Phylogenet Evol* 54:291–301. <https://doi.org/10.1016/J.YMPEV.2009.09.016>
- Hijmans RJ, Cameron SE, Parra JL, et al (2005) Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* 25:1965–1978. <https://doi.org/10.1002/JOC.1276>
- Hijmans RJ (2019). geosphere: Spherical trigonometry. R package version 1.5-10. <https://CRAN.R-project.org/package=geosphere>
- Hohenlohe PA, Day MD, Amish SJ, et al (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Mol Ecol* 22:3002–3013. <https://doi.org/10.1111/mec.12239>
- Ikeda DH, Max TL, Allan GJ, et al (2017) Genetically informed ecological niche models improve climate change predictions. *Glob Chang Biol* 23:164–176. <https://doi.org/10.1111/gcb.13470>
- Jackson, S. T., & Overpeck, J. T. (2000). Responses of Plant Populations and Communities to Environmental Changes of the Late Quaternary. 26(4), 194–220.

- Jetton RM, Crane BS, Whittier WA, Dvorak WS (2015) Genetic resource conservation of Table Mountain pine (*Pinus pungens*) in the central and southern Appalachian Mountains. *Tree Plant Notes* 58:42–52
- Jin WT, Gernandt DS, Wehenkel C, et al (2021) Phylogenomic and ecological analyses reveal the spatiotemporal evolution of global pines. *Proc Natl Acad Sci USA* 118. <https://doi.org/10.1073/PNAS.2022302118/-/DCSUPPLEMENTAL>
- Ju M-M, Feng L, Yang J, et al (2019) Evaluating population genetic structure and demographic history of *Quercus spinosa* (Fagaceae) based on specific length amplified fragment sequencing. *Front Genet* 10:965. <https://doi.org/10.3389/FGENE.2019.00965>
- Jump AS, Peñuelas J (2005) Running to stand still: Adaptation and the response of plants to rapid climate change. *Ecol Lett* 8:1010–1020. <https://doi.org/10.1111/j.1461-0248.2005.00796.x>
- Kass JM, Muscarella R, Galante PJ, et al (2021) ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions. *Methods Ecol Evol* 12:1602–1608. <https://doi.org/10.1111/2041-210X.13628>
- Keller SR, Olson MS, Salim S, et al (2010) Genomic diversity, population structure, and migration following rapid range expansion in the Balsam poplar, *Populus balsamifera*. *Mol Ecol* 19:1212–1226. <https://doi.org/10.1111/j.1365-294X.2010.04546.x>
- Keeley JE (2012) Ecology and evolution of pine life histories. *Ann For Sci* 69:445–453. <https://doi.org/10.1007/s13595-012-0201-8>
- Kim BY, Wei X, Fitz-Gibbon S, et al (2018) RADseq data reveal ancient, but not pervasive, introgression between Californian tree and scrub oak species (*Quercus* sect. *Quercus*: *Fagaceae*). *Mol Ecol* 27:4556–4571. <https://doi.org/10.1111/MEC.14869>
- Kriebel HB (1972). Embryo development and hybridity barriers in the white pines (Section *Strobus*). *Silvae Genet* 21:39-44.
- Kulmuni J, Butlin RK, Lucek K, et al (2020) Towards the completion of speciation: The evolution of reproductive isolation beyond the first barriers: Progress towards complete speciation. *Philos Trans R Soc B Biol Sci* 375:20190528. <https://doi.org/10.1098/rstb.2019.0528>
- Łabiszak B, Zaborowska J, Wójkiewicz B, Wachowiak W (2021) Molecular and paleo-climatic data uncover the impact of an ancient bottleneck on the demographic history and contemporary genetic structure of endangered *Pinus uliginosa*. *J Syst Evol* 59:596–610. <https://doi.org/10.1111/jse.12573>

- Lafontaine G de, Prunier J, Gérardi S, Bousquet J (2015) Tracking the progression of speciation: variable patterns of introgression across the genome provide insights on the species delimitation between progenitor–derivative spruces (*Picea mariana* × *P. rubens*). *Mol Ecol* 24:5229–5247. <https://doi.org/10.1111/MEC.13377>
- Lascoux M, Palmé AE, Cheddadi R, Latta RG (2004) Impact of Ice Ages on the genetic structure of trees and shrubs. *Philos Trans R Soc Lond B Biol Sci* 359:197–207. <https://doi.org/10.1098/rstb.2003.1390>
- Ledig FT, Smouse PE, Hom JL (2015) Postglacial migration and adaptation for dispersal in pitch pine (*Pinaceae*). *Am J Bot* 102:2074–2091. <https://doi.org/10.3732/AJB.1500009>
- Legendre P, Legendre L (2012) *Numerical Ecology*. Third Edition. Elsevier.
- Li L, Abbott RJ, Liu B, et al (2013) Pliocene intraspecific divergence and Plio-Pleistocene range expansions within *Picea likiangensis* (Lijiang spruce), a dominant forest tree of the Qinghai-Tibet Plateau. *Mol Ecol* 22:5237–5255. <https://doi.org/10.1111/MEC.12466>
- Lima JS, Telles MPC, Chaves LJ, et al (2017) Demographic stability and high historical connectivity explain the diversity of a savanna tree species in the Quaternary. *Ann Bot* 119:645–657. <https://doi.org/10.1093/AOB/MCW257>
- Lind BM, Menon M, Bolte CE, et al (2018) The genomics of local adaptation in trees: are we out of the woods yet? *Tree Genet Genomes* 14:29. <https://doi.org/10.1007/s11295-017-1224-y>
- Little EL Jr. (1971) *Atlas of United States trees, Vol. 1, conifers and important hardwoods*: U.S. Department of Agriculture, 1146, 9, p200
- Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research* 27:209–220.
- Ma XF, Szmidt AE, Wang XR (2006) Genetic structure and evolutionary history of a diploid hybrid pine *Pinus densata* inferred from the nucleotide variation at seven gene loci. *Mol Biol Evol*, 23:807–816. <https://doi.org/10.1093/molbev/msj100>
- McKinney GJ, Waples RK, Seeb LW, Seeb JE (2017) Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Mol Ecol Resour* 17:656–669. <https://doi.org/10.1111/1755-0998.12613>

- McKinney GJ, Waples RK, Pascal CE, et al (2018) Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. *Mol Ecol Resour* 18:570–579. <https://doi.org/10.1111/1755-0998.12763>
- McWilliam JR (1959) Interspecific Incompatibility in *Pinus*. *Am J Bot* 46:425–433
- Menon M, Bagley JC, Friedline CJ, et al (2018) The role of hybridization during ecological divergence of southwestern white pine (*Pinus strobiformis*) and limber pine (*P. flexilis*). *Mol Ecol* 27:1245–1260. <https://doi.org/10.1111/MEC.14505>
- Michaux, FA (1819) *The North American Sylva, or A description of the forest trees of the United States, Canada and Nova Scotia considered particularly with respect to their use in the arts, and their introduction into commerce; to which is added a description of the most useful of the European forest trees: illustrated by 156 coloured engravings.* C. d'Hautel, Paris <https://doi.org/10.5962/bhl.title.48807>
- Nosil, P (2012) *Ecological Speciation.* Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:osobl/9780199587100.001.0001>
- Nosil P, Feder JL (2012) Widespread yet heterogeneous genomic divergence. *Mol Ecol* 21:2829–2832. <https://doi.org/10.1111/j.1365-294X.2012.05580.x>
- Oksanen J, Blanchet FG, Friendly M, et al (2020) *vegan: Community ecology package.* R package version 2.5-7. <https://CRAN.R-project.org/package=vegan>
- Otto-Bliesner BL, Marshall SJ, Overpeck JT, et al (2006) Simulating Arctic climate warmth and icefield retreat in the last interglaciation. *Sci* 311:1751–1753. <https://doi.org/10.1126/science.1120808>
- Parchman TL, Gompert Z, Mudge J, et al (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Mol Ecol* 21:2991–3005. <https://doi.org/10.1111/j.1365-294X.2012.05513.x>
- Parchman TL, Jahner JP, Uckele KA, et al (2018) RADseq approaches and applications for forest tree genetics. *Tree Genet Genomes* 14. <https://doi.org/10.1007/s11295-018-1251-3>
- Parisien MA, Moritz MA (2009) Environmental controls on the distribution of wildfire at multiple spatial scales. *Ecol Monogr* 79:127–154. <https://doi.org/10.1890/07-1289.1>
- Park B, Donoghue MJ (2019) Phylogeography of a widespread eastern North American shrub, *Viburnum lantanoides*. *Am J Bot* 106:389–401. <https://doi.org/10.1002/AJB2.1248>

- Pausas JG, Schwilk D (2012) Fire and plant evolution. *New Phytol* 193:301–303. <https://doi.org/10.1111/j.1469-8137.2011.04010.x>
- Pearman PB, Guisan A, Broennimann O, & Randin CF (2008) Niche dynamics in space and time. *Trends Ecol Evol.* <https://doi.org/10.1016/j.tree.2007.11.005>
- Perron M, Perry DJ, Andalo C, Bousquet J (2000) Evidence from sequence-tagged-site markers of a recent progenitor-derivative species pair in conifers. *Proc Natl Acad Sci USA* 97:11331–11336. <https://doi.org/10.1073/PNAS.200417097>
- Peterson BK, Weber JN, Kay EH, et al (2012) Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7:e37135 . <https://doi.org/10.1371/journal.pone.0037135>
- Peterson AT, Anamza T (2015) Ecological niches and present and historical geographic distributions of species: A 15-year review of frameworks, results, pitfalls, and promises. *Folia Zool* 64:207–217. <https://doi.org/10.25225/FOZO.V64.I3.A3.2015>
- Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annu Rev Ecol Evol Syst* 37:187-214. <https://doi.org/10.1146/annurev.ecolsys.37.091305.110215>
- Phillips SJ, Anderson RP, Dudík M, et al (2017) Opening the black box: an open-source release of Maxent. *Ecography* 40:887–893. <https://doi.org/10.1111/ECOG.03049>
- Puritz JB, Hollenbeck CM, Gold JR (2014) dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* 2:e431. <https://doi.org/10.7717/peerj.431>
- R Core Team (2021) R: A Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- Raj A, Stephens M, Pritchard JK (2014) FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* 197:573–589. <https://doi.org/10.1534/genetics.114.164350>
- Richards CL, Carstens BC, Lacey Knowles L (2007) Distribution modelling and statistical phylogeography: An integrative framework for generating and testing alternative biogeographical hypotheses. *J. Biogeogr.* 34:1833–1845. <https://doi.org/10.1111/j.1365-2699.2007.01814.x>
- Saladin B, Leslie AB, Wüest RO, et al (2017) Fossils matter: improved estimates of divergence times in *Pinus* reveal older diversification. *BMC Evol Biol* 17:95. <https://doi.org/10.1186/S12862-017-0941-Z>

- Savolainen O, Bokma F, García-Gil R, et al (2004) Genetic variation in cessation of growth and frost hardiness and consequences for adaptation of *Pinus sylvestris* to climatic changes. *For Ecol Manage* 197:79–89.
<https://doi.org/10.1016/j.foreco.2004.05.006>
- Seehausen O, Butlin RK, Keller I, et al (2014) Genomics and the origin of species. *Nat Rev Genet* 15:176–192. <https://doi.org/10.1038/nrg3644>
- Soltis DE, Morris AB, McLachlan JS, et al (2006) Comparative phylogeography of unglaciated eastern North America. *Mol Ecol* 15:4261–4293.
<https://doi.org/10.1111/j.1365-294X.2006.03061.x>
- Sousa VC, Grelaud A, Hey J (2011) On the nonidentifiability of migration time estimates in isolation with migration models. *Mol Ecol* 20:3956–3962.
<https://doi.org/10.1111/j.1365-294X.2011.05247.x>
- Stetter MG, Thornton K, Ross-Ibarra J (2018) Genetic architecture and selective sweeps after polygenic adaptation to distant trait optima. *PLoS Genet* 14:e1007794. <https://doi.org/10.1101/313247>
- Svenning JC, Eiserhardt WL, Normand S, et al (2015) The Influence of Paleoclimate on Present-Day Patterns in Biodiversity and Ecosystems. *Annu Rev Ecol Evol Syst* 46:551–572. <https://doi.org/10.1146/annurev-ecolsys-112414-054314>
- Vallejo-Marín M, Hiscock SJ (2016) Hybridization and hybrid speciation under global change. *New Phytol* 211:1170–1187. <https://doi.org/10.1111/nph.14004>
- Vasilyeva G, Goroshkevich S (2018) Artificial crosses and hybridization frequency in five-needle pines. *Dendrobiology* 80:123–130.
<https://doi.org/10.12657/denbio.080.012>
- Wang IJ, Bradburd GS (2014) Isolation by environment. *Mol Ecol* 23:5649–5662.
<https://doi.org/10.1111/mec.12938>
- Warren DL, Glor RE, Turelli M (2008) Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. *Evolution* 62:2868–2883.
<https://doi.org/10.1111/J.1558-5646.2008.00482.X>
- Warren DL, Matzke NJ, Cardillo M, et al (2021) ENMTools 1.0: an R package for comparative ecological biogeography. *Ecography* 44:504–511.
<https://doi.org/10.1111/ecog.05485>
- Welch NT, Waldrop TA (2001) Restoring Table Mountain pine (*Pinus pungens* Lamb.) communities with prescribed fire: An overview of current research. *Castanea* 66:42–49.

- Williams JW, Jackson ST (2007) Novel climates, no-analog communities, and ecological surprises. *Front Ecol Environ* 5:475–482. <https://doi.org/10.1890/070037>
- Wright S (1943) Isolation by distance. *Genetics* 28:114–138.
- Wright JW (1959) Species hybridization in the white pines. *Forest Sci* 5:210–222.
- Yang R-C (1998) Estimating hierarchical *F*-statistics. *Evolution* 52:950–956. <https://doi.org/10.2307/2411227>
- Yang YX, Zhi LQ, Jia Y, et al (2020) Nucleotide diversity and demographic history of *Pinus bungeana*, an endangered conifer species endemic in China. *J Syst Evol* 58:282–294. <https://doi.org/10.1111/jse.12546>
- Zobel DB (1969) Factors affecting the distribution of *Pinus pungens*, an Appalachian endemic. *Ecol Monogr* 39:303–333.
- Zou J, Sun Y, Li L, et al (2013) Population genetic evidence for speciation pattern and gene flow between *Picea wilsonii*, *P. morrisonicola* and *P. neoveitchii*. *Ann Bot* 112:1829–1844. <https://doi.org/10.1093/AOB/MCT241>

Appendix 1

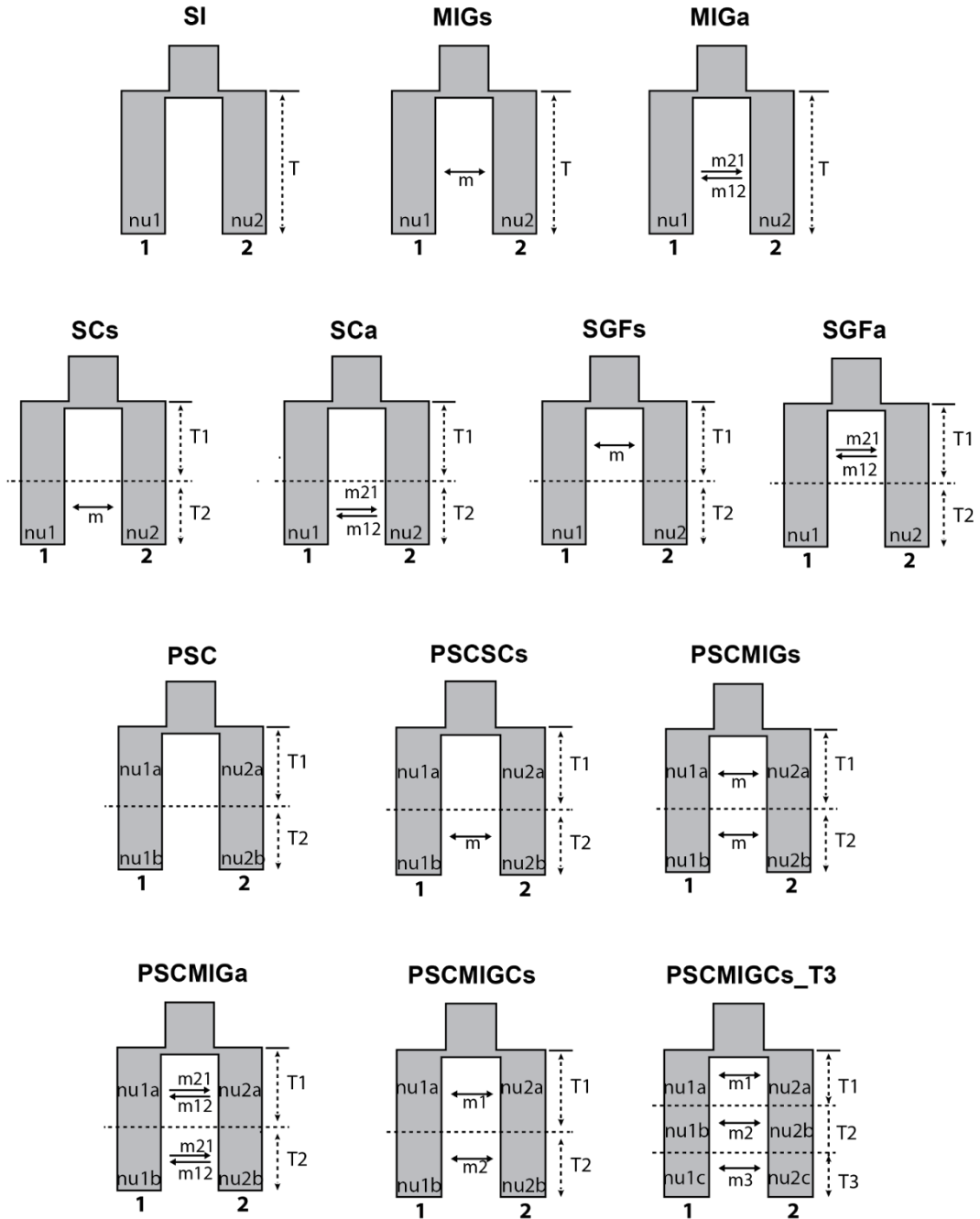


Figure 1.S1 The thirteen divergence scenarios tested within the program $\partial\alpha\partial i$.

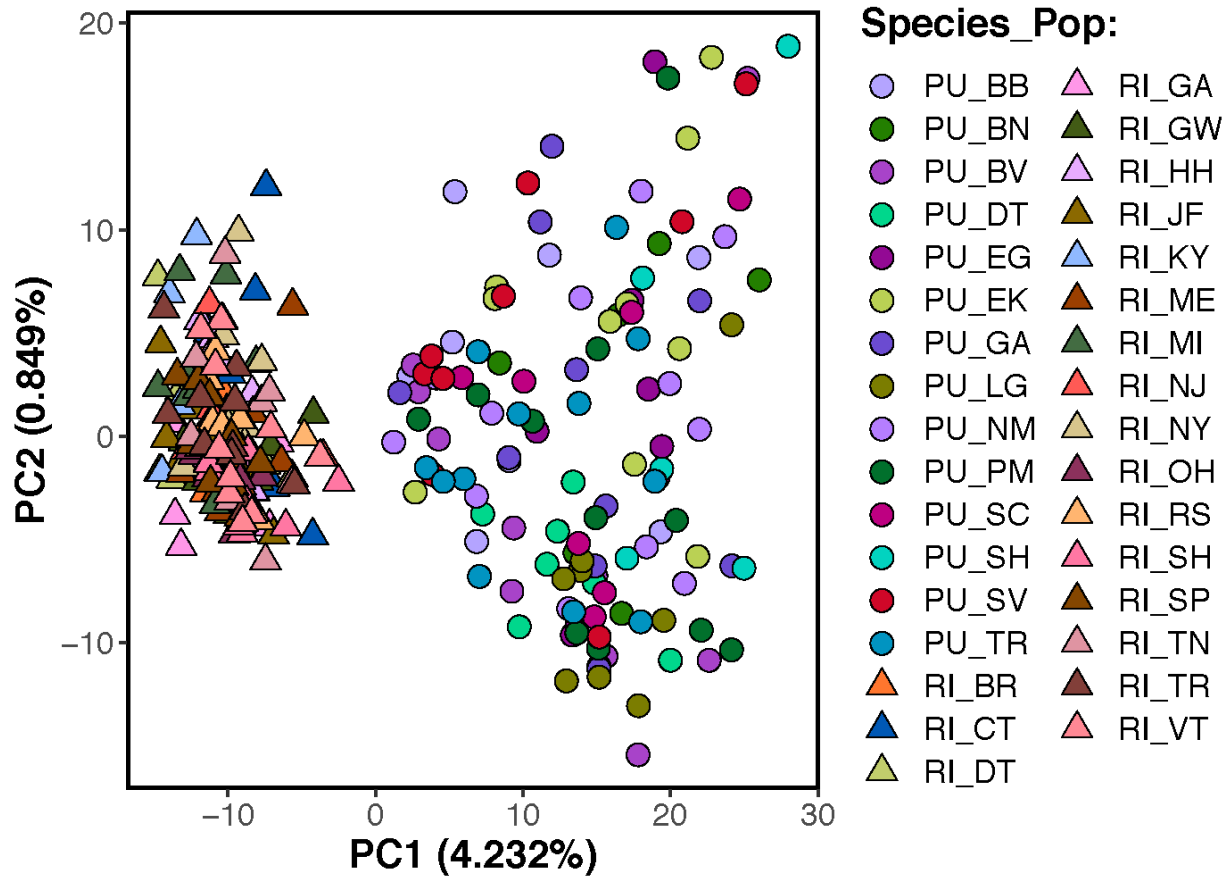


Figure 1.S2 Principal component analysis (PCA) of 300 *P. rigida* and *P. pungens* trees labeled by population assignment.

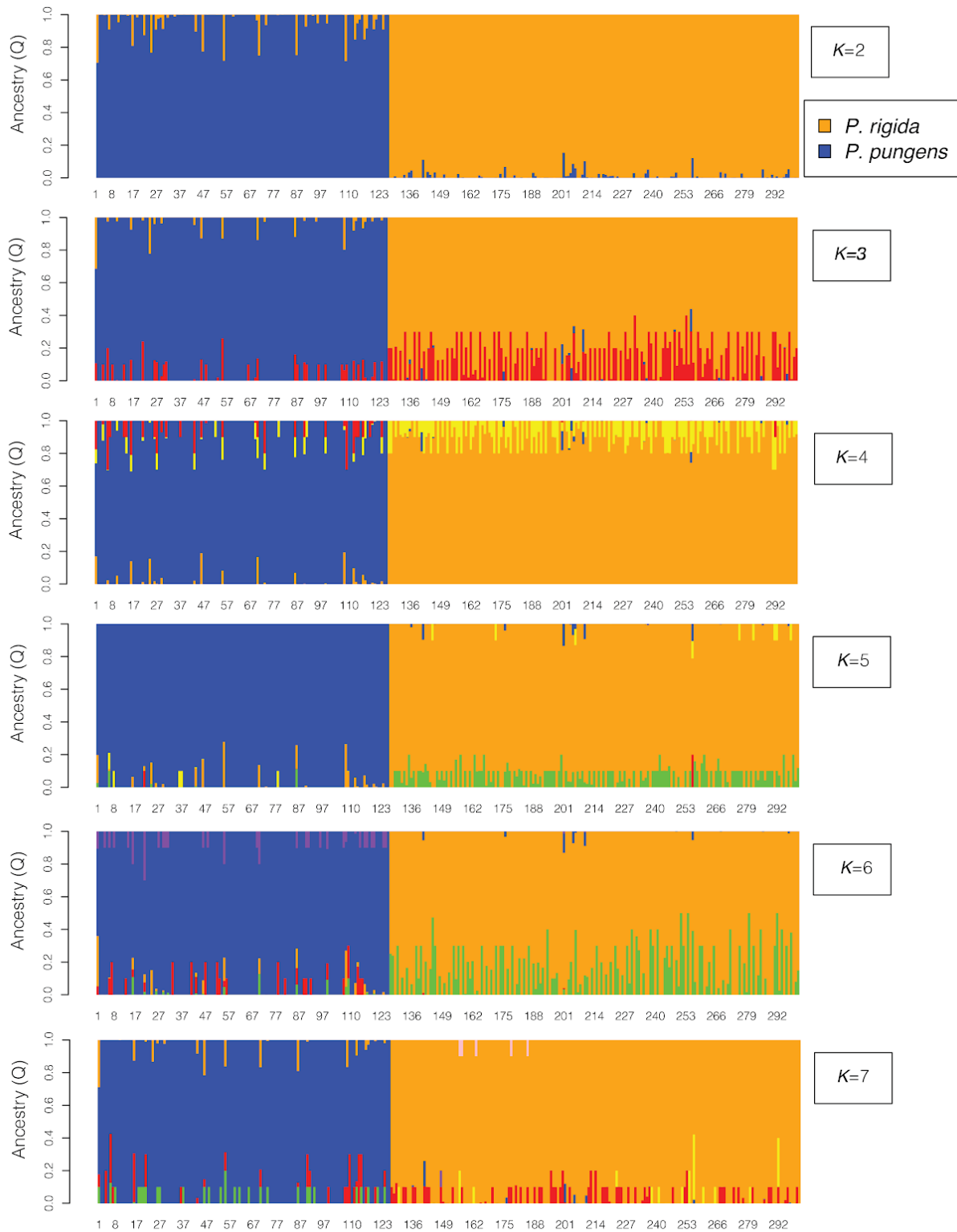


Figure 1.S3 Individual based assignments of admixture from analysis of *fastSTRUCTURE* for $K = 2$ through $K = 7$. The plot associated with each value of K represents the averages assignments for each individual across 10 replicate runs.

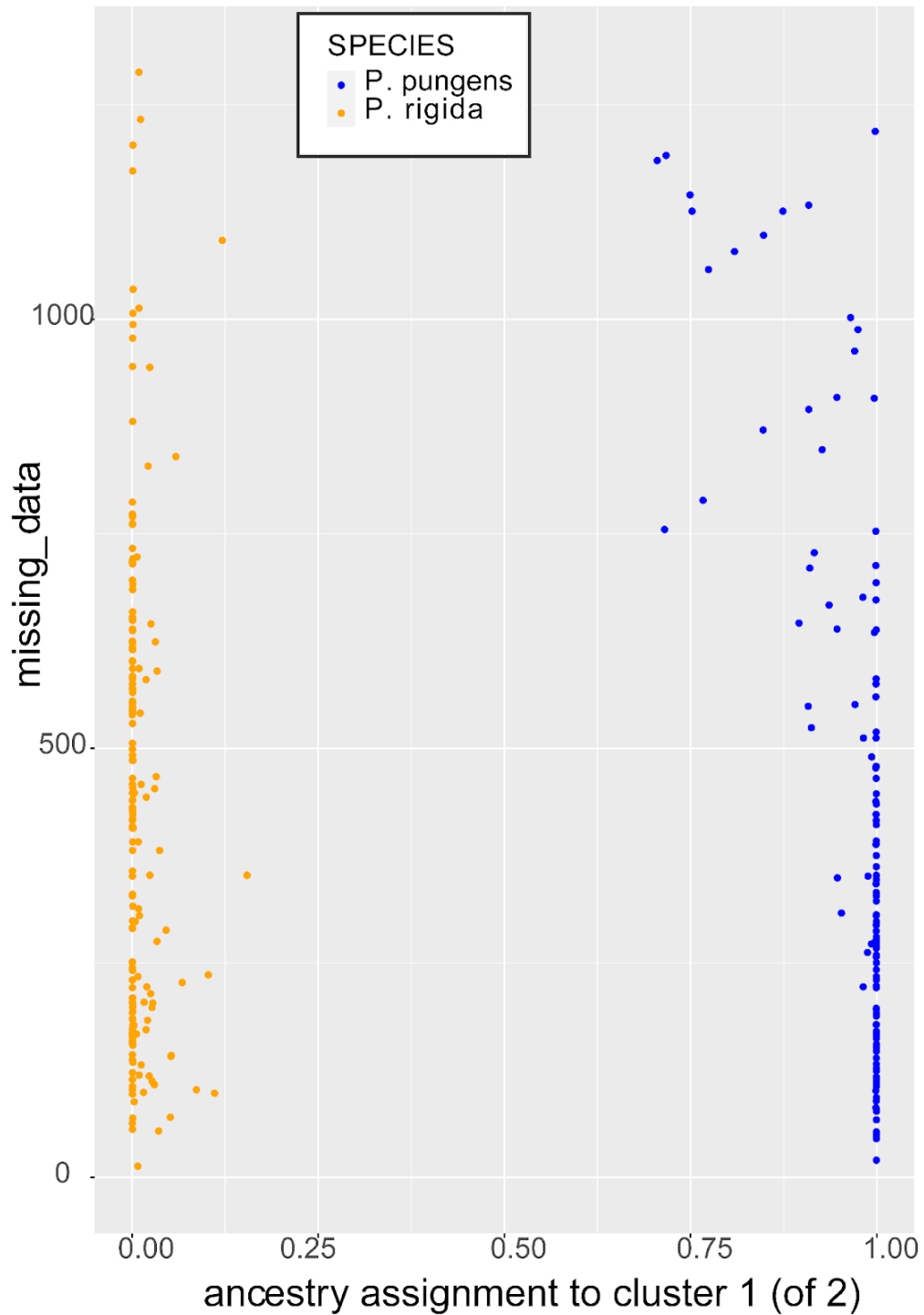


Figure 1.S4 Distribution of missing data across the sampled trees in relation to ancestral coefficients (from $K = 2$). Blue circles to the right are samples of *P. pungens*. Orange circles to the left are samples of *P. rigida*.

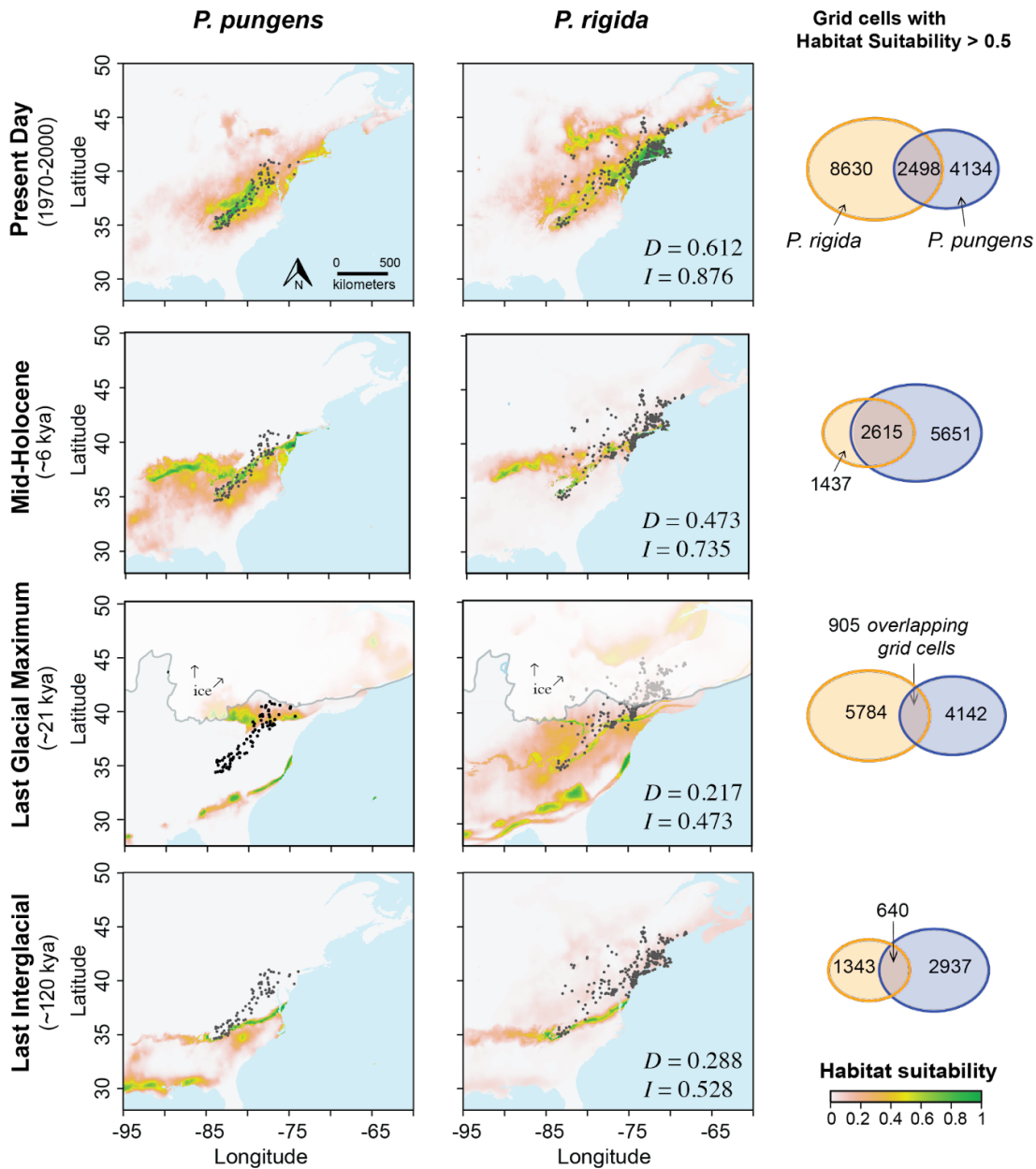


Figure 1.S5 Species distribution model (SDM) predictions across four time points for *P. pungens* and *P. rigida*. Measures of raster overlap in terms of Schoener's *D* and Warren's *I* index between the models of each species, and at each time point, are presented in the bottom right corner of the prediction plots for *P. rigida*. Venn diagrams illustrate the number of grid cells with moderate to high habitat suitability scores (> 0.5) for each SDM at a given time point, as well as the number of shared, or overlapping, grid cells. Blue Venn diagram ovals show grid cell counts from the *P. pungens* SDM, and orange Venn diagram ovals show grid cell counts from the *P. rigida* SDM for the aligning time point (denoted on the left side). Habitat suitability distributions for LGM and HOL depict ensemble predictions. Glacial extent data (labeled ice in LGM plots) for 18 kya was provided by Dyke (2003).

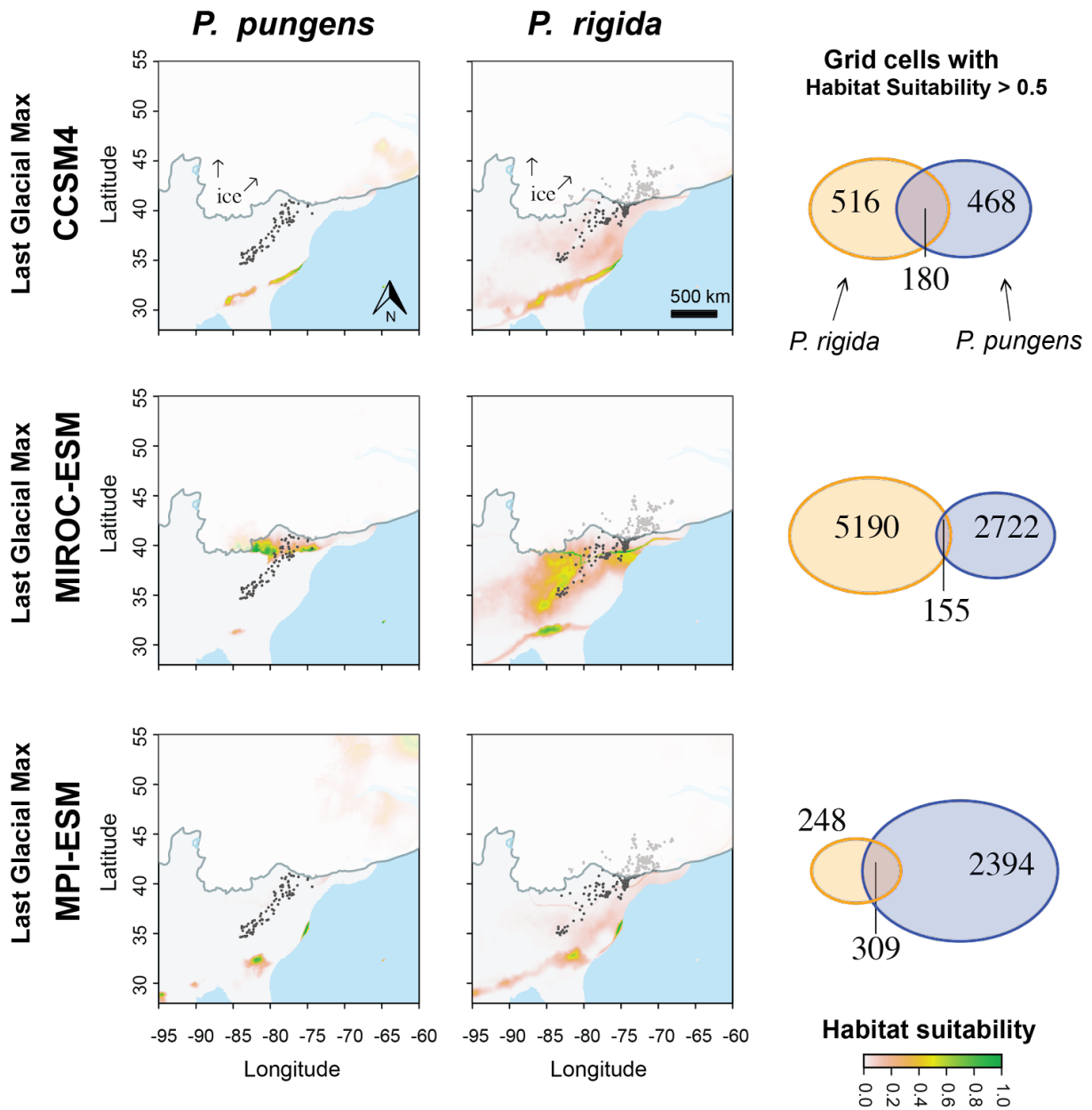


Figure 1.S6 Last Glacial Maximum (LGM, ~21 kya) model predictions from each GCM (CCSM4, MIROC-ESM, and MPI-ESM).

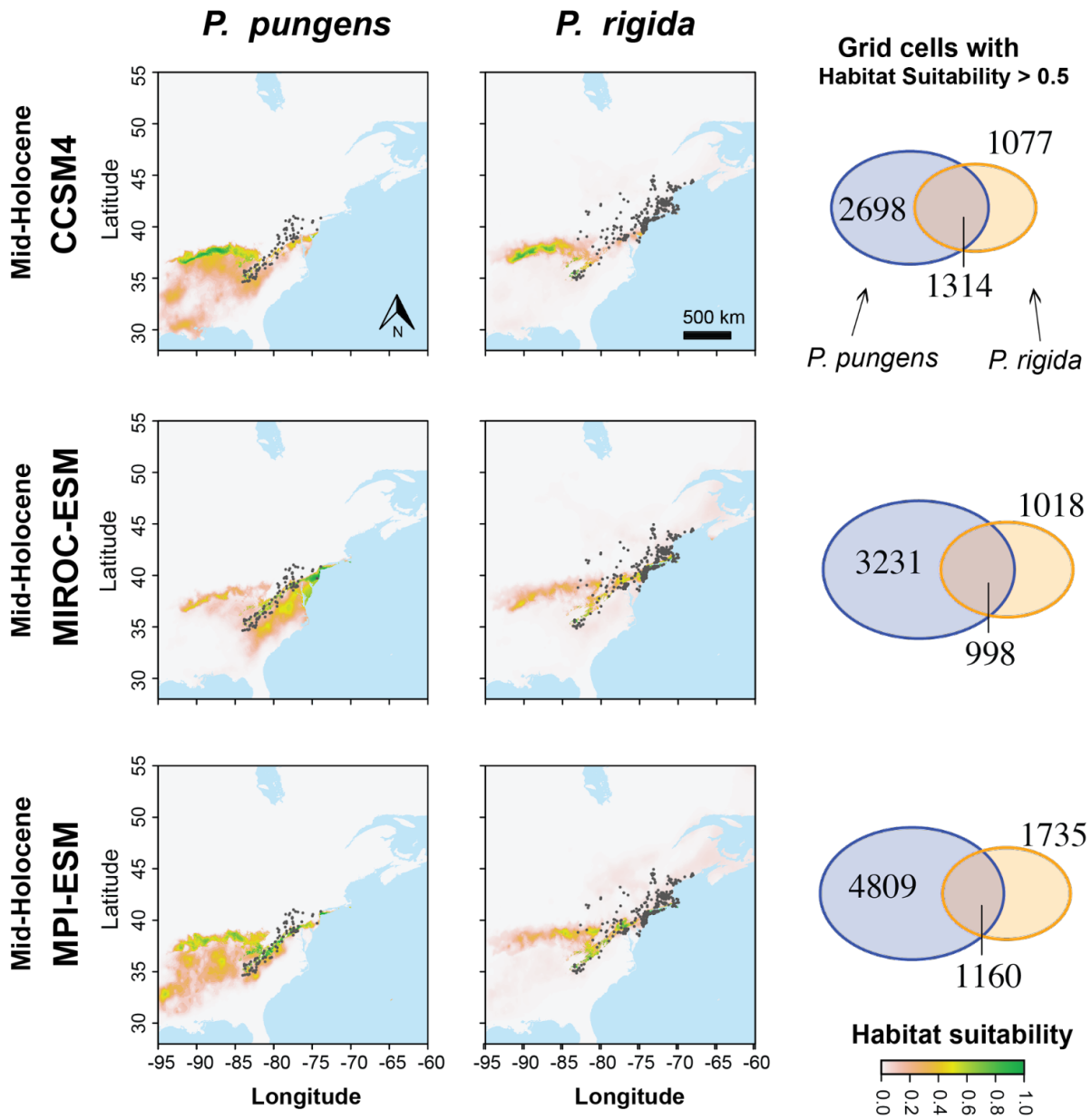


Figure 1.S7 Mid-Holocene (~6 kya) model predictions from each GCM (CCSM4, MIROC-ESM, and MPI-ESM).

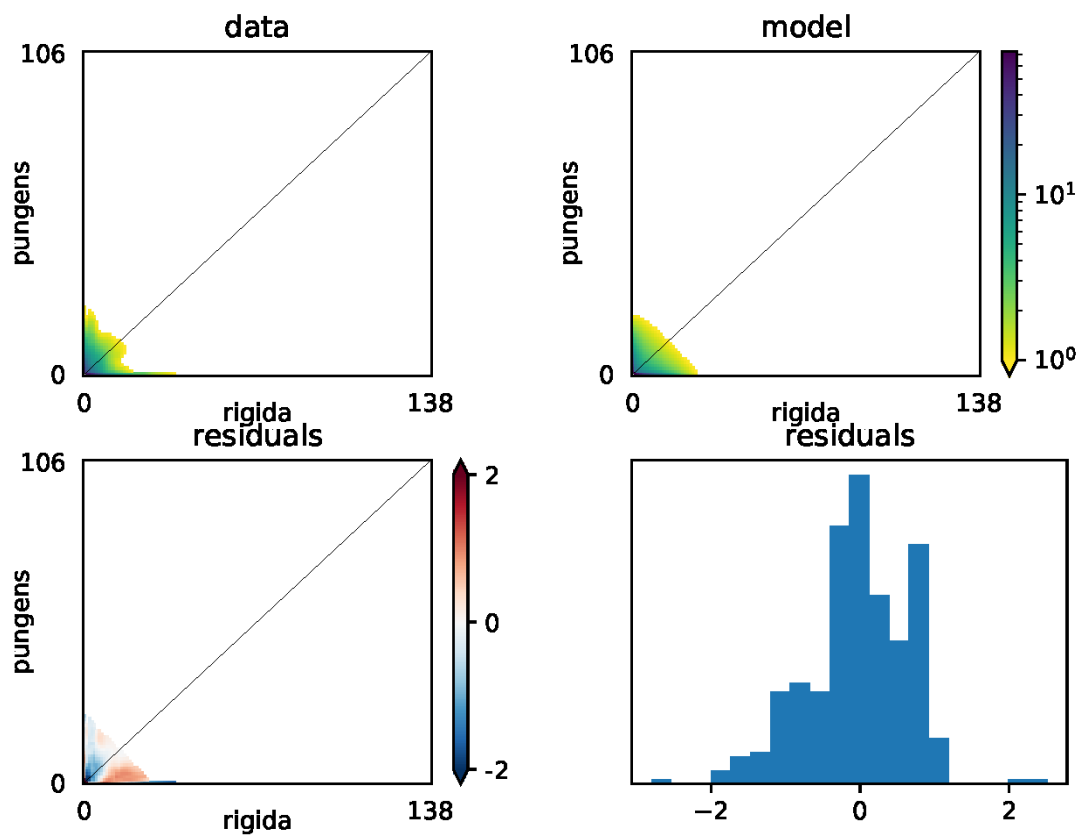


Figure 1.S8 Presentation of data-model fit to the PSCMIGCs model run with highest log likelihood.

Table 1.S1 Parameters and the estimates associated with the best run of each model type. The model with three time intervals (PSCMIGCsT3) is not included in this table. Those parameters are summarized in Table 1.S2.

Model	θ	T1	T2	NuP1	NuR1	nuP2	nuR2	mS1	mS2	mPR	mRP
Strict isolation (SI)	419.94	2.16 E-04	---	4.08 E-03	4.82E -03	---	---	---	---	---	---
Gene Flow- sym (MIGs)	423.44	5.41 E-03	---	7.11 E-02	8.68E -02	---	---	3.21E +01	---	---	---
Gene Flow- asym (MIGa)	431.96	8.21 E-03	---	9.96 E-02	9.33E -02	---	---	---	---	2.66E +01	3.13 E+01
Secondary Contact- sym (SCs)	403.07	6.92 E-03	4.78 E-03	1.46 E-01	1.74E -01	---	---	2.25E +01	---	---	---
Secondary Contact- asym (SCa)	414.49	5.90 E-04	2.73 E-04	1.47 E-02	2.24E -02	---	---	---	---	1.27E +01	5.56 E+01
Ancient GF-asym (SGFa)	400.32	2.28 E-02	5.88 E-04	2.69 E-01	1.84E -01	---	---	---	---	4.22E +00	2.20 E+01
Ancient GF-sym (SGFs)	402.43	7.89 E-03	2.10 E-03	1.44 E-01	1.54E -01	---	---	2.96E +01	---	---	---
Pop size change (PSC)	377.41	2.99 E-03	5.45 E-03	5.52 E-01	6.24E -01	1.61E -01	1.33E -01	---	---	---	---
Pop size change- symGF (PSCMIGs)	75.78	2.75 E+00	5.25 E-02	1.02 E+01	3.29E +01	2.83E -01	3.47E -01	1.26E +01	---	---	---
Pop size change- asymGF (PSCMIGa)	263.42	2.00 E-01	1.37 E-03	3.08 E+00	1.47E -01	2.62E -02	2.18E +00	---	---	5.62E -01	4.14 E+01
Pop size and symGF change (PSCMIGCs)	118.10	1.50 E+00	1.95 E-02	2.84 E+01	2.10E +01	9.54E -02	1.09E -01	4.86E +01	3.83E +01	---	---
Pop size change- SCsym (PSCSCs)	271.86	1.22 E-01	6.31 E-02	3.63 E+00	8.14E +01	4.07E -01	4.98E -01	5.47E +00	---	---	---

Table 1.S2 Parameter estimates from the best run (lowest AIC) for the model allowing 3 time intervals (PSCMIGCsT3).

Parameter	scaled	unscaled
nuP1	13.84	206,046.53
nuR1	16.39	243,895.25
T1	4.96	3,693,145.67
mS1	49.59	0.00167
nuP2	0.11	1,675.62
nuR2	0.18	2,671.88
T2	0.0077	5,756.79
mS2	15.76	0.00053
nuP3	3.88	57,810.81
nuR3	6.19	92,124.16
T3	0.0016	1,182.63
mS3	15.76	0.00053
θ (Nref)	48.64	(14,884.34)

Chapter 2

Potential drivers in the differential development of reproductive isolation for three cryptically related North American pine species (*Pinus pungens*, *P. rigida*, and *P. taeda*)

Abstract

Inferring divergence histories and drivers of reproductive isolation (RI) within clades of the genus *Pinus* requires a multidisciplinary approach as histories appear to be riddled with hybridization, periods of isolation, local adaptation, and effective population size changes that co-occurred with major shifts in climate. In this study, we performed historical species distribution modeling (SDM), population structure analysis, redundancy analysis, and demographic inference to help explain the differential development of RI across three eastern North American pine species (*Pinus pungens*, *P. rigida*, and *P. taeda*). The previous work done on these species helped construct a three-species demographic inference routine that sought to estimate when and to what extent gene flow occurred across ancestral and extant species boundaries. We found pairwise demographic inferences to be more informative than the seven three-species models we tested. Divergence occurred with gene flow for *P. pungens* and *P. rigida* as previously inferred for these two species. Unexpectedly, strict isolation was the best fit model of divergence for pairwise inferences with *P. taeda* even though hybridization between *P. rigida* and *P. taeda* is observable at present. Collectively we present strong support for a common ancestor between *P. pungens* and *P. rigida*, but placement of *P. taeda* relative to these other two was difficult to ascertain based on comparisons of model AIC scores and divergence time estimates. We further explored the relationships between and across

these three species by mapping our RADseq contigs to the annotated *P. taeda* genome. Pairwise analysis of F_{ST} for highly differentiated single nucleotide polymorphisms (SNPs) among contigs that associated with genic regions may help explain the less established RI between *P. rigida* and *P. taeda* and the stronger RI between *P. pungens* and the other two focal species. From the suite of analyses performed and literature reviewed, we concluded that geography, climate, gene flow, and ecological divergence have all contributed to standing levels of differentiation across these three pine species and that the challenges associated with delineation of species relationships from our study and past phylogenetic inferences may be linked to assumptions of tree bifurcation.

Introduction

The maintenance of species boundaries involves an array of mechanisms, requiring specific consideration of geography, climate, life history traits, and genetic architectures to adequately identify drivers of speciation. Investigations within model systems, such as *Arabidopsis*, *Mimulus*, and *Helianthus*, have helped elucidate the different genetic architectures associated with the development of pre and postzygotic reproductive isolation (RI), an important component to the process of speciation, in plants (Widmer, Lexer, and Cozzolino 2009; Rieseberg and Blackman 2010). A growing body of literature has associated the development of RI with adaptive evolution (e.g., Nosil and Feder 2012; Kremer and Hipp 2020). Emerging patterns suggest that adaptive traits are polygenic, genomic islands of divergence are small and spread throughout the genome, and species boundaries appear to be permeable with relatively few loci contributing to RI (Zukowska and Wachowiak 2016). *Populus trichocarpa* (Torr. and A.Gray ex. Hook.) was the first

sequenced forest tree genome (Tuskan et al. 2006) and the work done in *Populus* has initiated an understanding of how RI in long-lived trees may involve more complex genetic architectures (e.g., more traits that are polygenic in nature) than the architectures associated with RI in short-lived plant taxa (e.g., a simple inversion; Shang et al. 2020). In parallel, investigations into divergence among taxa of the genus *Quercus* has added depth to our comprehension of how genomes across closely related tree taxa are shaped by hybridization, ecology, and purifying selection (e.g., Cokus, Gugger and Sork 2015; Hipp et al. 2020). However, documentation and explanation of general evolutionary patterns related to the relative contribution of extrinsic and intrinsic barriers to RI, and how these barrier loci are distributed across the genomes of closely related tree taxa, remain in their infancy. Furthermore, it is unclear how the results of speciation studies in *Populus* and *Quercus* can be extrapolated to fit expectations for other tree taxa, especially those among conifers. Indeed, conifer genomes are substantially larger, have fewer chromosomes, lower levels of linkage disequilibrium, slower rates of genome evolution, and more transposable elements (Prunier et al. 2015). All these differences may contribute to contrasting expectations about the evolutionary tempo and mode for the development of RI.

Well-annotated genome sequences are useful to determine the distribution of barrier loci across the genome given that inversions, linkage groups, and functional groups of genes (e.g., disease resistance, drought tolerance, and phenology) have been previously described as contributors to RI (Rieseberg and Blackman 2010; Cokus, Gugger, and Sork 2015; Khodwekar and Gailing 2017). The large size and immense complexity of conifer

genomes (>15GB) have made sequencing and annotating them a challenge, but draft genomes with curated annotations are now available for *Pinus taeda* L. (Neale et al. 2014; Wegrzyn et al. 2014), *Picea abies* (L.) H. Karst (Nystedt et al. 2013), *Picea glauca* (Moench) Voss (Birol et al. 2013), and *Pinus lambertiana* Dougl. (Stevens et al. 2016), presenting opportunities to identify and functionally describe loci contributing to RI. Until population-level genomic resources for conifers become available, which will help clarify if islands of divergence or continents of divergence (Nosil and Feder 2012) can also describe RI in conifers, we can continue to utilize next generation sequencing and candidate gene approaches to infer demographic histories, identify environmental drivers of divergence, and assign biological function to highly differentiated loci that are within or near coding regions. As case studies that employ these methods accumulate, we suspect patterns related to tempo and mode of divergence will emerge among those that examine multiple closely related species of comparable genetic architecture, geography, and climate (Bolte and Eckert 2020). Given that interspecific gene flow is commonly observed in the divergence histories of forest trees, we anticipate patterns related to the contributory effects of gene flow to the development of RI to also emerge, such as the relative rates at which RI develops when reinforcement (hybrid fitness reduction) versus introgression (hybrid zones) is involved.

The genus *Pinus* is the most diverse group of conifers with over 110 species that inhabit an array of geographic and climatic gradients, providing an extensive resource for comparative investigation into conifer speciation and the development of RI (Zukowska and Wachowiak 2016; Jin et al. 2021), but even within the genus *Pinus*, hard and soft

pinus appear to be distinct in terms of artificial crossing success and diversification rate. Soft pines of sections *Quinquifoliae* and *Parrya* can be successfully crossed with one another, with the only exception being *P. lambertiana* Dougl., which suggests genetic incompatibilities are infrequent or weak in these groups (Critchfield 1967). In contrast, hard pines of sections *Trifoliae* and *Pinus* have more documented cases of reproductive incompatibilities among its members (Critchfield 1967), but why this is so has yet to be described. Most investigations into pine speciation, using two or more taxonomically established species, have taken a phylogeographic approach (e.g., Liu et al. 2014; Zhang et al. 2014; Zhou et al. 2017; Liu et al. 2019; Yang et al. 2020). Some have gone further to include evaluations of niche evolution to discern stabilizing selection or diversifying selection as drivers of divergence (e.g., Menon et al. 2018). Some have incorporated candidate loci for RI into their analyses to help genetically explain species-level boundaries (e.g., Gao et al. 2012; Wachowiak et al. 2018). Together, these efforts have laid essential groundwork for future investigations into the development of RI. Investigations relevant to hard pine speciation from molecular data are lacking though, especially in North America, which is surprising given the first conifer genome to be sequenced was *P. taeda*. Only two speciation studies have leveraged this genomic resource to identify biological functions among differentiated loci across defined species, but these studies involved hard pines clades of Europe and Asia (Gao et al. 2012; Wachowiak et al. 2018). Given that climatic drivers of divergence differ within and across continents (Jin et al. 2021), loci involved in RI may also differ regionally.

Here, we add to the body of pine speciation literature with an examination of three closely related eastern North American hard pine species: *P. pungens* Lamb., *P. rigida* Mill., and *P. taeda* L. (Gernant et al. 2018; Jin et al. 2021). The geographical distributions of these species differ (Figure 2.1) but have regions of overlap or are proximal enough to one another to dismiss geographical isolation as a contemporary boundary to gene flow. *P. pungens* and *P. rigida* have differences in pollen release timing that contribute to prezygotic isolation (Zobel 1969; Ladeau and Clark 2006), yet recurring interspecific gene flow characterizes their divergence history (Bolte et al. 2022), thus providing evidence of RI lability when phenological schedules are responsible, at least in part, for the maintenance of species boundaries (Vallejo-Marín and Hiscock 2016). Artificial crossing experiments have indicated though that hybrids of *P. pungens* and *P. rigida* have low yield of sound seeds, suggesting incompatibilities may also explain the lack of hybridization observed at present. For *P. rigida* and *P. taeda*, pollen release timing also differs by approximately four weeks (i.e., in North Carolina; Zobel 1969; Ladeau and Clark 2006) but these species appear to have remained genetically compatible throughout their divergence history (Hyun 1960; Critchfield 1963) and continue to hybridize in nature (Smouse and Saylor 1973). Moreover, hybrids bred between *P. rigida* and *P. taeda* are a valued source of fast-growing timber in cooler climates where natural populations of *P. taeda* cannot persist (Hyun and Ahn 1959; Knezick et al. 1985a). Describing the mode of RI between *P. pungens* and *P. taeda* is more cryptic. They have potentially overlapping pollen release dates (early to mid-April; Zobel 1969; Ladeau and Clark 2006), yet artificial crossing experiments did not produce sound seeds (Critchfield 1963), thus RI is reasonably stronger between *P. pungens* and *P. taeda*.

Inspired by the variable degrees of RI between these three species, we used a comprehensive analytical framework to address the following questions: 1) What are the relative contributions of geography and climate to genetic differences, species distributions, and patterns of niche evolution? 2) To what extent did gene flow occur across the ancestral populations and contemporary species boundaries of these three species? We found through our demographic inference routine, relying on both pairwise and three species models, confidence in the relationship between *P. pungens* and *P. rigida*. These two species shared a recent common ancestor, but placement of *P. taeda* in relation to these two species is less clear. We mapped 5050 RADseq contigs to the *P. taeda* annotated genome to characterize the distribution of our genome-wide data that was used in all genetic analyses. From the genic regions that associated with our data, we observed high differentiation in comparisons with *P. pungens* and low differentiation between *P. rigida* and *P. taeda* which may explain differences in genomic compatibility and relative strengths of RI.

Methods

Sampling

We obtained range-wide samples of needle tissue for 14 populations of *Pinus pungens*, 19 populations of *P. rigida*, and 25 populations of *P. taeda* (Fig. 2.1). Each population consisted of 2-12 trees with each sampled tree distanced by approximately 50 m from the

next to avoid potential kinship (Table 2.S1). Needle tissue was dried using silica beads, then 10 mg of tissue was cut and lysed for DNA extraction.

DNA sequence data

Genomic DNA was extracted from 606 trees using DNeasy Plant Kits (Qiagen) following the manufacturer's protocol. We then prepared ddRADseq libraries (Peterson et al. 2012), using the procedure from Parchman et al. (2012). EcoRI and MseI restriction enzymes were used to digest all four libraries before performing ligation of adaptors and barcodes. After PCR, agarose gel electrophoresis was used to separate then select DNA fragments between 300-500 bp in length. The pooled DNA was isolated using a QIAquick Gel Extraction Kit (Qiagen). Single-end sequencing was conducted on Illumina HiSeq 4000 platform by Novogene Corporation (Sacramento, CA). Raw fastq files were demultiplexed using GBSX (Herten et al. 2015) version 1.2, allowing two mismatches (-mb 2). The dDocent bioinformatics pipeline (Puritz et al. 2014) was used to generate a reference assembly and call variants. The reference assembly was optimized using shell scripts and documentation within dDocent (cutoffs: individual = 6, coverage = 6; clustering similarity: -c 0.92), utilizing cd-hit-est (Fu et al. 2012) for assembly. The initial variant calling produced 239,628 single nucleotide polymorphisms (SNPs) that were further filtered using *vcftools* (Danecek et al. 2011), version 0.1.15. We retained only biallelic SNPs with sequencing data for at least 50% of the samples, minor allele frequency (MAF) > 0.01, summed depth across samples > 100 and < 15000, and alternate allele call quality ≥ 50 . Sampled trees with excessive missing data ($\geq 50\%$) were removed from the data set leaving 515 trees. We further removed 75 samples of *P. taeda* (i.e., removed 5 - 7

samples from each population) to make population sample sizes more comparable across the three species. The remaining 440 samples (86 *P. pungens*, 122 *P. rigida*, 232 *P. taeda*) were used in all analyses.

To account for linkage disequilibrium before performing demographic inference (Gutenkunst et al. 2009), we thinned the dataset to one SNP per contig (--thin 100). Additionally, stringent filtering steps were taken to minimize the potential misassembly of paralogous genomic regions. Removing loci with excessive coverage and retaining only loci with two alleles present are expected to ameliorate the influence of misassembled paralogous loci in our data (Hapke and Thiele 2016; McKinney et al. 2018). Furthermore, we retained loci with $F_{IS} > -0.5$, as misassembly to paralogous genomic regions can lead to abnormal heterozygosity (Hohenlohe et al. 2013; McKinney et al. 2017). From the 5820 SNPs remaining, we identified 1397 SNPs that were fixed for the same allele in both *P. pungens* and *P. rigida*. To rectify the possibility that the de novo reference assembly process was biased toward *P. taeda* identity due to larger sample size (63% more trees than *P. pungens*, 48% more trees than *P. rigida*), we filtered out 55% (determined by averaging the aforementioned sample size discrepancies) of these 1397 SNPs by selecting 628 SNPs with the least amount of missing data. The final filtered data set for demographic inference was comprised of 5051 SNPs.

Population structure

Overall patterns of genetic structure for *P. pungens*, *P. rigida*, and *P. taeda* were investigated using principal component analysis (PCA), by following standardization

routines detailed in Patterson et al. (2006) and employing in the `prcomp` function of the *stats* version 4.0.4 package in R version 3.6.2 (R Development Core Team 2021) and following standardization routines detailed in Patterson et al. (2006). To further assess structure and presence of admixture across the 440 samples, an individual-based assignment test was conducted using *fastSTRUCTURE* (Raj et al. 2014) with cluster assignments ranging from $K = 3$ to $K = 7$. The cluster assignment with the highest average log-likelihood value across ten replicate runs was determined to be the best fit. Individual admixture assignments were then aligned and averaged across the 10 runs using the *pophelper* version 1.2.0 (Francis 2017) package in R.

Associations between genetic structure and environment

We tested the multivariate relationships among genotype, climate, and geography by conducting full and partial redundancy analyses (RDA) within the *vegan* version 2.5-7 package (Oksanen et al. 2020) in R version 4.0.4 (R Core Development Team 2021). Genotype data were coded as counts of the minor allele for each sample (i.e., 0, 1, or 2 copies) and then standardized following Patterson et al. (2006). Climate raster data (i.e., 19 bioclimatic variables at 30 arc second resolutions), as well as elevational raster data from WorldClim, were extracted from geographic coordinates for each sampled tree and then tested for correlation using Pearson's correlation coefficient (r) in R. Five bioclimatic variables that were not highly correlated ($r < |0.75|$) and known to influence diversification in the genus *Pinus* (Menon et al. 2018; Jin et al. 2021; Bolte et al. 2022) were retained for analysis: Bio 2 (mean diurnal range), Bio 4 (temperature seasonality), and Bio 9 (mean temperature of the driest quarter), Bio 12 (annual precipitation), and Bio 15 (precipitation

seasonality). The full explanatory data set included these five bioclimatic variables, latitude, longitude, and elevation. Statistical significance of all RDA models ($\alpha = 0.05$), as well as each axis within full models, was assessed using a permutation-based analysis of variance (ANOVA) procedure with 999 permutations (Legendre and Legendre 2012). The influence of predictor variables, as well as their confounded effects, in RDA were quantified using variance partitioning as employed in the *varpart* function of the *vegan* package in R. We used this same procedure to test the multivariate relationships between ancestral coefficients, climate, and geography.

Species distribution modeling and niche divergence

To help formulate a testable hypothesis in the inference of demography from genomic data, species distribution modeling (SDM) was performed for each species to identify areas of suitable habitat under current climate conditions and across three historical time periods (see Richards et al. 2007). These temporal inferences were then used to help identify plausible demographic responses. For example, if overlap in modeled habitat suitability changed over time, the hypothesis for demographic inference would include changes in gene flow parameters over time.

Occurrence records for *P. pungens* were downloaded from GBIF.org (18th December 2018; GBIF occurrence download (<https://doi.org/10.15468/dl.urehu0>) and combined with known occurrences published by Jetton et al. (2015). For *P. rigida* and *P. taeda*, all occurrence records were downloaded from GBIF.org (29th December 2015 and 18th December 2018; GBIF occurrence download (<http://doi.org/10.15468/dl.ak0weh>) and

<https://doi.org/10.15468/dl.kiknmo>). Records were examined for presence within or close to the known geographical range of each species (Little 1971), and any over 200 km outside the known geographic range were pruned. The remaining locations were then thinned to one occurrence per 10 km to reduce the effects of sampling bias using the *spThin* version 0.1.0.1 package (Aiello-Lammens et al. 2015) in R. The resulting occurrence dataset included 84 records for *P. pungens*, 252 records for *P. rigida*, and 361 for *P. taeda* (Online Resource 2). All subsequent analyses were performed in R version 3.6.2 (R Development Core Team 2021).

The same bioclimatic variables (Bio2, Bio4, Bio9, Bio12, Bio15) selected for RDA were used in species distribution modeling but were downloaded from WorldClim version 1.4 (Hijmans et al. 2005) at 2.5 arc minute resolution. Paleoclimate raster data for the LGM (~21 kya) and Holocene (HOL; ~6 kya) were based on three General Circulation Models (GCMs; CCSM4, MIROC-ESM, and MPI-ESM). Ensembles were built by averaging the grid cell values across the three GCMs for each time period, which were then used to predict species distributions and habitat suitability in the past. Paleoclimate data for the LIG (~120 kya; Otto-Bliesner et al. 2008) were only available at 30 arc second resolution, so we downscaled the raster files to 2.5 arc minute resolution to help facilitate comparative analyses across the four time points. Because only one GCM is available for the LIG, no ensemble was built.

We built species distribution models (SDMs) using MAXENT version 3.4.1 (Phillips et al. 2017) and determined the best-fit model for each of our focal species using the Akaike

information criterion (AIC) as implemented in the *ENMeval* version 2.0.0 R package (Kass et al. 2021). Raw raster predictions were standardized to have the sum of all grid cells equal the value of one using the *raster.standardize* function in the *ENMTools* version 1.0.5 (Warren et al. 2021) R package. We then transformed standardized rasters to cumulative raster predictions with habitat suitability scaled from 0 to 1, which allowed quantitative SDM comparisons across species and time. Next, SDM cumulative raster predictions were converted into coordinate points using the *sf* version 0.9-7 R package to calculate the number of points with habitat suitability values greater than 0.5 (i.e., moderate to high suitability areas). Overlap (i.e., shared points across species) in SDM predictions for each time period was measured using the *inner_join* function in the *dplyr* version 1.0.5 R package. The extent of modeled species distributional overlap was also quantified using the *raster.overlap* function in *ENMTools*, thus providing measures for Schoener's *D* (1968) and Warren's *I* (Warren et al. 2008). A background similarity test was also performed for each pairwise species comparison to describe niche evolution (conservatism vs. divergence) during speciation. The same five bioclimatic variables detailed above, along with the occurrence records from GBIF, were used in this analysis and executed within the *phyloclim* version 0.9.5 R package (Hiebl and Calenge 2018).

Demographic modeling

Demographic modeling was conducted using Diffusion Approximation for Demographic Inference (*∂α∂i* v.2.0.5; Gutenkunst et al. 2009). Among the seven complex models tested, we held certain relationships constant based on the results of previous studies. First, in each of these models, *P. pungens* and *P. rigida* maintained ongoing symmetrical

gene flow as was previously inferred (Bolte et al. 2022). Second, we dismissed investigating extant gene flow between *P. pungens* and *P. taeda* due to the results of experiments where artificial crosses were unable to produce seeds (Critchfield 1963). Finally, we assumed the topology *P. pungens* and *P. rigida* being more closely related and more recently diverged as reported from phylogenetic inference of Hernandez- Leon et al. (2013) and Saladin et al. (2017). Based on SDM predictions across four time points, we confirmed the findings in Bolte et al. (2022) that there was consistent overlap in suitable habitat between *P. pungens* and *P. rigida*, and we further hypothesized that the overlap between *P. rigida* and *P. taeda* was also consistent enough to allow interspecific gene flow. Given our research objectives here we focused on gene flow timing and directionality. While the results of Bolte et al. (2022) indicated recent and dramatic reductions in effective population sizes for both *P. pungens* and *P. rigida* during the last glacial period, working with three diverged lineages in a demographic inference framework is computationally taxing, so we omitted inference of population size changes. We instead fixed the ancestral size of *P. pungens* and *P. rigida* to be five times larger than the combined inferences for current effective population size to acknowledge this dynamic reported in Bolte et al. (2022).

Our null model considered the pure divergence between ancestral populations and strict isolation between *P. taeda* and *P. rigida*. The other six demographic models involved potential divergence scenarios for the ancestral populations and investigation into the gene flow dynamics between *P. rigida* and *P. taeda* (i.e., parameters shifting between two time intervals, symmetrical, and asymmetrical genetic exchange (Fig. 2.S2)). The two

models with the highest composite likelihood among the seven scenarios tested were then selected for parameter optimization. We performed five replicate runs of each model in *∂α∂i* with a 260 x 280 x 300 grid space and the nonlinear Broyden-Fletcher-Goldfarb-Shannon (BFGS) optimization routine. Model selection was conducted using AIC (Akaike 1974). Unscaled parameter estimates were obtained using a per lineage substitution rate of 7.28×10^{10} substitutions/site/year rate for *Pinaceae* (De La Torre et al. 2017) and a generation time of 25 years (Ma et al. 2006). Genome length was calculated as proposed in Bolte et al. (2022).

We also explored pairwise model (i.e., two species) inferences to determine level of accuracy in divergence time and gene flow estimates from our best-fit three population model. Model types included divergence with strict isolation, divergence with symmetrical gene flow, and divergence with asymmetrical gene flow for *P. pungens* and *P. rigida*, *P. pungens* and *P. taeda*, and *P. rigida* and *P. taeda*. AIC scores were used to assess goodness of fit across three replicate runs of each model type and pairwise species relationships. The best replicate run (lowest AIC) for each model was then used to calculate ΔAIC ($AIC_{\text{model } i} - AIC_{\text{best model}}$) scores (Burnham and Anderson 2002). From the best supported pairwise inferences, upper and lower 95% confidence intervals (CIs) for all parameters were obtained using the Fisher Information Matrix (FIM)-based uncertainty analysis.

Distribution of RADseq contigs across the Pinus taeda annotated genome

To determine the extent to which the 5051 SNPs in our analyses were identifiable within the *Pinus taeda* genome (Pita.2_01.fa; treegenomesdb.org) and associated with annotations, we mapped our RADseq contigs using blastn, version 2.5.0 (NCBI). Settings included e-values less than 10, word sizes greater than 4, and gaps penalized by 1. Under these settings, all but one contig successfully mapped to regions of the *P. taeda* genome. We kept the three best hits (i.e., lowest e-values) per contig. Each hit was matched with a scaffold identifier (i.e., seqid) from the *P. taeda* genome. We then further reduced the data to include only scaffold IDs that had annotations. Because the scaffold sizes can be long with multiple attributes (i.e., annotated regions), we compared the location of a given RADseq contig to locations of attributes along the respective scaffold. Attributes associated with the gene closest to or directly hit by the RADseq contig were retained for further analyses.

We calculated F -statistics for each of the 5051 SNPs in the *hierfstat* package (Goudet 2005) and outlier detection was performed in the R package, *OutFLANK*, version 0.2 (Whitlock and Lotterhos 2014). F_{CT} (species) values were then used to parse data into categories of species level differentiation (e.g., $F_{CT} < 0.3$, $F_{CT} \geq 0.3$, ≥ 0.75 , and ≥ 0.9) to report counts and observe trends. We measured the distance of SNPs in relation to genic regions and created three additional categories. We counted how many SNPs were outside 20k bp from a gene, within 20k bp from a gene, and within a gene. We subset our genetic data to include only SNPs having $F_{CT} \geq 0.3$ and then further subset those into the aforementioned distance categories to a gene. These three data sets were then subjected to pairwise estimates of F_{ST} for each species pair using the *hierfstat* package. This

analysis was performed to examine differences in the distribution of F_{ST} across the three distance categories to genes for each species pair (i.e., *P. pungens* - *P. rigida*, *P. pungens* - *P. taeda*, *P. rigida* - *P. taeda*). To see if EggNOG descriptions, provided with the *P. taeda* genome download (treegenomesdb.org), were enriched in our data at F_{CT} values ≥ 0.3 compared to counts with F_{CT} values < 0.3 , we performed Fisher's Exact tests for gene descriptions that had multiple records or close counts between the two F_{CT} categories.

Results

Population structure and genetic diversity

Principal component analysis (PCA) showed clear separation between *P. pungens*, *P. rigida*, and *P. taeda* across PC1 and PC2 (Figure 2.2a). The first PC axis explained 4.77% of the variation across the 5051 SNP x 440 tree data set, while PC2 explained 1.75%. Of the 5051 SNPs analyzed, 1876 SNPs were fixed in *P. pungens*, 1242 SNPs were fixed in *P. rigida*, and only 328 SNPs were fixed in *P. taeda*. Among those, *P. pungens* and *P. rigida* had 628 SNPs fixed for the same allele. Fewer SNPs were fixed for the same allele in comparing *P. taeda* to the other two species. Only 78 and 81 were shared among those of *P. pungens* and *P. rigida*, respectively. In the analysis of structure, $K = 3$ had the highest log-likelihood values (Figure 2.2b). We observed low levels of admixture (2-20%) in 14.0% of sampled *P. pungens* and 8.2% of sampled *P. rigida*. Most of this admixture was assigned to *P. taeda* ancestry. Among samples of *P. taeda*, several had low levels of admixture assigning to either *P. pungens* or *P. rigida*, but 14.6% of sampled *P. taeda* had

moderate to high levels of admixture (20 – 60%) with *P. rigida*. Most of this admixture was found in five of the twenty-five *P. taeda* populations (Table 2.S1), four of which are in regions over 400 km from where contemporary geographical distributions overlap (Figure 2.S3).

Associations between genetic structure and environment

The combined effects of climate and geography explained 4.31% (adj. r^2) to 6.05% (r^2) of the genetic variance across 5051 SNPs and 440 sampled trees. The first RDA axis accounted for the bulk of the explanatory variance (63.24%, Figure 2.3a) although RDA axes 2, 3 and 4 were also important in describing the genetic variation across *P. pungens*, *P. rigida* and *P. taeda* (p -values < 0.05). The combined variable loadings of RDA1 and RDA2 indicated elevation, latitude, and Bio4 (temperature seasonality) as the primary predictors of differentiation. With geography removed, Bio15 (precipitation seasonality) was the highest predictor of differentiation (Figure 2.3b), and with climate removed from the analysis, elevation and longitude were the highest predictors of differentiation.

The results of full and partial RDAs (Figure 2.3) are summarized in Table 2.1. The higher explanatory variance associated with the partial model for the independent effect of geography indicated that it, as opposed to climate alone, was the best predictor of genome-wide genetic variation across these three species (Figure 2.3c). Species level clustering was more diffuse among all partial RDAs conducted (Figure 2.3), however, suggesting both geography and climate are important to genetic differentiation across

species. We also observed that climate and geography were even stronger predictors of ancestry ($r^2 = 59.40$; Table 2.1; Figure 2.3d-f).

Partitioning the effects of each predictor set revealed that climate independently (i.e., conditioned on geography) accounted for 11.07% of the explained variance. Geography independently (i.e., conditioned on climate) accounted for 25.75% of the explained variance. The confounded effect, due to the correlations inherent to the chosen geographic and climatic predictor variables, was 63.17%.

Species distribution modeling

We used MAXENT to predict past geographical distributions during the LIG, LGM, and HOL and formed testable hypotheses within the demographic inference framework of *∂α∂i*, v.2.0.5. The best fit SDM for *P. pungens* used a linear and quadratic feature class with a 1.0 regularization multiplier, while the SDMs for both *P. rigida* and *P. taeda* used a linear, quadratic, and hinge feature class with a regularization multiplier of 3.0. All SDMs had AUC values over 0.85. Data inputs, outputs, and statistical results for model evaluation are available online (https://github.com/boltece/Species_boundaries_3pines). Bio15 (precipitation seasonality) was the most informative and contributive climate variable to the SDMs of *P. rigida* and *P. pungens*, and Bio9 (mean temperature of the driest quarter) was most important and contributive to the SDM of *P. taeda* (Figure 2.S2). Bio4 (temperature seasonality) was the second most important variable to the SDM predictions of all three species, and in the full RDA was the most important climate descriptor of genetic variation. Congruency between SDM and RDA variable importance was also observed in Bio 9, as the highest loadings along RDA axis 1 (Figure 2.3a) were in the

direction of *P. taeda* samples. Likewise, Bio15 was the most important variable in the partial RDA (with geography removed, Figure 2.3b).

Across the four time periods modeled, we observed fluctuations in the areas of moderate to high habitat suitability for all three species. The greatest differences observed were among the distributional overlap values (Venn Diagrams of Figure 2.4a) and raster overlap values (Schoener's *D*) associated with *P. pungens* and *P. taeda*, which increased over time (Figure 2.4b). Raster overlap between *P. pungens* and *P. rigida* was consistently high (0.529 – 0.599) relative to the other comparisons made (Figure 2.4b). The current model predictions, labeled NOW in Figure 2.4a, reflected current geographical distributions of each species, except for a few small disjunct regions deemed suitable for habitat. This likely resulted from using a data set reduced to five climatic variables (Figure 2.S2). Notably though, four of the five most admixed populations of *P. taeda* with *P. rigida* ancestry were from Louisiana and Mississippi (populations TA_LA, TA_LB, TA_MD, and TA_ME; Table 2.S1), a region that was predicted to also have suitable habitat during the LIG for *P. rigida* (Figure 2.4a), but at present is over 400 km away from natural *P. rigida* stands, based on distributional maps in Little (1971; Figure 2.S3).

The background similarity test yielded results of niche conservatism in all pairwise comparisons as measures of niche overlap were higher than the distributional ranges of background similarity values. The highest niche overlap was between *P. pungens* and *P. rigida* (Schoener's *D* = 0.570) with the distributions of asymmetrical background niche

similarity values far lower ($0.15 < \text{Schoener's } D < 0.3$) indicating relatively strong niche conservatism compared to the other pairwise species assessments (Figure 2.5). There were similar niche overlap values in the comparisons of *P. pungens* and *P. taeda* (Schoener's $D = 0.282$) as well as *P. rigida* and *P. taeda* (Schoener's $D = 0.295$), but the distributions of background niche similarity were more diverged between *P. pungens* and *P. taeda*.

Demographic modeling

Our workflow for demographic inference is summarized in Figure 2.6. The best-fit model from our first round of analyses described the two divergence events associated with T1 and T2 as occurring with symmetrical gene flow (Figure 2.6a). This model, as well as the other six variations tested, inferred an unreasonably shallow divergence time of approximately 7,310 years ago. Exceptionally high rates of gene flow during T2 were also consistently inferred across all models that had included those parameters. The best-fit model indicated 200 migrants per generation (gene flow rate; $m = 0.0022$) between *P. pungens* and *P. rigida* and 68 migrants per generation (gene flow rate; $m = 0.00076$) between *P. rigida* and *P. taeda*. Because divergence time estimates are sensitive to migration and effective population size estimates, we ran the best-fit model from the first round of inference under different lower and upper bounds (Figure 2.6b). This effort did not improve model fit. AIC scores were higher (Figure 2.6) than the best-fit model from the first round of inferences. From the three replicates that converged to provide an optimal value of θ , which is proportional to the ancestral effective population size ($\theta = 4N_e\mu$), divergence time estimates were larger but still unreasonable. Total divergence

time estimate ranged from 22,170 years ago. Rate of gene flow between *P. pungens* and *P. rigida* continued to be higher than inferences for *P. rigida* and *P. taeda*.

Given these results, we decided to examine the species topology assumed above where *P. rigida* and *P. pungens* were sister species using pairwise comparisons across two-population models Figure 2.6c-e. Strict isolation models had shallow divergence time as inferred in the three population models, but unexpectedly divergence time inferences that involved *P. taeda* were similar (~2,500 years ago) and more shallow than the divergence time inferred for *P. pungens* and *P. rigida* (20,535 years ago). The AIC scores were much higher, suggesting poor fit, for the models that involved *P. taeda* though (AIC = 8374 and 8943 versus 4159 in the model for *P. pungens* and *P. rigida*). Adding gene flow to the two-population models instantly alleviated shallow divergence time estimates (Figure 2.6d). Models with the lowest AIC indicated divergence between *P. pungens* and *P. rigida* to be approximately 1.11 mya with ongoing asymmetrical gene flow. Gene flow directionality was higher from *P. pungens* into *P. rigida* ($m_{21} = 1.2e-04$ versus $m_{12} = 8.5e-05$). The divergence time between *P. rigida* and *P. taeda* was deeper (~ 1.69 mya) and even deeper between *P. pungens* and *P. taeda* (~ 30.1 mya) when asymmetrical gene flow was allowed, but these models had higher AIC scores (9,086 and 9,191, respectively) than the strict isolation models (8,943 and 8,374; Figure 2.6). Calculations of 95% CIs for parameters estimated from the best-fit pairwise models (starred in Figure 2.6) were narrow and required an array of eps values (1.0E-02 - 1.0E-07; Table 2.S2). Small range in values around parameter inference should not be interpreted as well-fit.

Distribution of RADseq contigs across the Pinus taeda annotated genome

To further characterize our 5051 SNP data set, we mapped RADseq contigs to the *P. taeda* draft genome, version 2. All but one contig were successfully mapped. After filtering hits down to the best three scaffold IDs per contig, which was determined by the lowest e-values, 15,137 hits remained (Figure 2.7a), comprised of 13,249 unique scaffold IDs.

The e-values across the filtered hits ranged from $7.44e-34$ to 6.40. Associated scaffold IDs (i.e., seqid), percent match, number of base pairs included, gaps, location on each scaffold, and e-values are available online at (https://github.com/boltece/Species_boundaries_3pines). Of the 13,249 unique scaffold ID assigned to contigs, 16.21% matched with annotated attributes (i.e., curated annotations; PITA_x) of the *P. taeda* genome (Figure 2.7a). We used an arbitrary threshold of 20k bp to count the number of hits located close to genes. Of the 2444 unique contig-scaffold ID hits with annotations, 45.17% were over 20kbp from a gene, 38.75% were close to genes, and 16.08% were in genes (Figure 2.7a).

We then characterized our 2444 annotated hits by parsing them into categories respective to the F_{CT} values (species-level differentiation) associated with each RADseq contig/SNP. The higher the F_{CT} value, then the more differentiation there is across species at that SNP. Most SNPs had low F_{CT} values (Figure 2.7b). Likewise, most of the annotated hits were associated with SNPs that had an $F_{CT} < 0.3$. There were no outlier SNPs detected using OutFLANK (Figure 2.S4, all FDR q-values > 0.1). For SNPs with $F_{CT} \geq 0.3$, 57 were over 20k bp from a gene, 32 were close to a gene, and 15 were within genes (Figure 2.7

c-e). For those within genes, eleven were intronic and four were in coding regions. After performing Fisher's Exact Tests for seven EggNOG descriptions that had multiple occurrences within the category of $F_{CT} \geq 0.3$ or appeared to have similar counts between $F_{CT} \geq 0.3$ and $F_{CT} < 0.3$, we found two to be enriched. Heat shock protein and YT521-B-like domain had p -values < 0.05 .

Our collective observations from demographic modeling of both three-population and two-population configurations inspired pairwise analyses of F_{ST} across the SNPs that had $F_{CT} \geq 0.3$ in the three species comparisons. The distribution of F_{ST} values from each pairwise analysis at categorical levels of distance to a gene (i.e., outside 20k bp from a gene, inside 20k bp from a gene, and within a gene) are presented in Figure 2.7, panels f-h. Similar patterns in F_{ST} distributions were observed between SNPs outside and inside 20k bp of a gene. However, within genic regions, pairwise comparisons with *P. pungens* had higher F_{ST} (medians of 0.85 and 0.71) and comparisons between *P. rigida* and *P. taeda* were mostly below 0.3 F_{ST} (median = 0.06).

Discussion

We investigated the divergence history and drivers of differentiation for three North American pine species (*P. pungens*, *P. rigida*, and *P. taeda*) using a multidisciplinary approach that involved analyses of historical species distributions, niche evolution, genetic structure, RDA, demographic inference, and the distribution of RADseq contigs

along the annotated genome of *P. taeda*. Our demographic inference routine provided evidence that *P. pungens* and *P. rigida* shared a recent common ancestor, but placement of *P. taeda* in relation to these two species remains enigmatic. Gene flow between *P. pungens* and *P. rigida*, as inferred in this study and in Bolte et al. (2022), played an important role in the development of RI. Considering these two species have strikingly diverged traits yet conserved ancestral niches and distributional overlap, reinforcement (i.e., selection against hybrids or intermediate trait values) and character displacement are candidate causes towards the rapid development of pre and postzygotic isolation (Beans 2014). The development of RI between these two species may have carried over into different present-level compatibilities with *P. taeda*, without a history of gene flow being directly involved. Our pairwise demographic inferences support this notion. Divergence histories with *P. taeda* were best described through models of strict isolation. The similar trait values between *P. taeda* and *P. rigida* may indicate more similar or compatible genetic architectures for hybridization and introgression than those associated with traits of *P. pungens*.

Interestingly, geography explained more of the genetic differentiation across our three focal species compared to the five climate variables we selected for analysis. While climatic niches were statistically different from each other and genetic differentiation was strongly associated with precipitation seasonality, elevation, and latitude, our null model tests for niche evolution indicated ancestral niche conservatism. Thus, stabilizing selection may be stronger than diversifying selection along the niche axes we analyzed. This could be a product of historical gene flow homogenizing species level genetic

differentiation and therefore homogenizing niche differentiation and/or use of climatic variables describing the core aspects of pine niches (i.e., niche aspects shared by most pine species).

Climate and geography help contextualize differentiation

The SDM analysis we conducted was used to form a hypothesis for gene flow rates in demographic inference (see Richards, Carstens, and Knowles 2007). Gene flow rates corresponded to habitat suitability overlaps as hypothesized. More percent overlap of *P. pungens* within the distribution of *P. rigida* was observed relative to the overlap of habitat suitability between *P. rigida* and *P. taeda*. The rate of gene flow was highest between *P. pungens* and *P. rigida*. The SDMs for *P. pungens* and *P. rigida* were both mainly driven by Bio15 (precipitation seasonality). The SDMs for all three species were influenced by Bio4 (temperature seasonality), and the SDM for *P. taeda* was primarily driven by Bio9 (mean temperature of the driest quarter). While niches of all three species are relatively conserved based on the results of background similarity, niche identities defined by the five climatic variables we selected were statistically different with *P. pungens* and *P. rigida* being more similar than niche comparisons with *P. taeda*.

We found congruency between our SDM and RDA analyses. We observed the importance of precipitation seasonality to genetic differentiation across these three species, providing further support to the conclusions drawn in Jackson and Overpeck (2000) regarding adaptations to seasonality under Quaternary climate, in Jin et al. (2012) regarding drivers of divergence in eastern North American pines, and in Bolte et al. (2022)

regarding drivers of differentiation between two of our three focal species. Also observed from the RDA is the confounding nature of geography and climate. Drawing conclusions related to which climatic or geographic variables were driving forces to genetic differentiation should be done cautiously. What limits a niche or drives adaptation, could involve other climatic variables (e.g., aridity; Eckert et al. 2010) that were removed from analysis due to high correlation or variables that were not considered directly. Geographical factors of latitude, longitude, and elevation were able to explain more of the genetic differences across our focal species than the five climatic variables we included. Given the strikingly different trait values of *P. pungens* (e.g., cone serotiny, needle morphology, early reproductive age, seed size, etc.; Zobel 1969) against the more similar morphological characteristics shared between *P. rigida* and *P. taeda*, the importance of geography to genetic differentiation may be better explained by examining soil features (Scull et al. 2003), biotic interactions (e.g., mycorrhizae; Nunez, Horton and Simberloff 2009), and fire regimes (Kane et al. 2015), which have been associated with adaptive traits (Brady, Kruckeberg, and Bradshaw 2005; Keeley et al. 2011; Jin et al. 2021) and range limits (Pickles et al. 2015).

Interpretation of the SDMs any further than gene flow potential since the LIG should be done cautiously. Historical SDM predictions for the LGM vary greatly across GCMs. Climate variable selection (e.g., sensitivity to seasonality variables; Varela et al. 2015) and no-analog climate regimes (Veloz et al. 2012) have been attributed to variability across GCM predictions. The ensemble approach we employed for estimating HOL and LGM distributional overlap likely provided an over-prediction of habitat suitability, but

Bolte et al. (2022) showed that even with seasonality variables used in model predictions, overlap in habitat suitability for *P. pungens* and *P. rigida* was consistently observed and less variable than size of suitable habitat across each GCM as well as the ensemble. Considering these disclaimers, we report the SDM predictions of LGM dual refugia for *P. pungens* and *P. rigida*, east and west of the Appalachian Mountains. This geographic barrier has been responsible for restricting gene flow during glacial periods which resulted in genetic differences through both neutral and nonneutral processes (Soltis et al. 2006). The divergence histories for each of our focal taxa, regardless of phylogenetic topology, can potentially be explained as a dynamic interplay of mixing-isolation-mixing (MIM; see He et al. 2019), cycles of expansions and contractions, and natural selection (Wu et al. 2022). For instance, the four distant populations of *P. taeda* with high *P. rigida* ancestry could be artifacts of historical distributional overlap during the LIG (according to our SDM predictions), long distance dispersal into a favorable microclimate for first and second generational hybrid phenotypes, or unfortunately for students of phylogeography and speciation, human-mediated transplants of non-native populations. With consideration of only naturally occurring demographic processes and the hypothesis of He et al. (2019; the number of cycles of MIM is proportional to genetic differentiation), *P. rigida* and *P. taeda* may have experienced fewer cycles than *P. rigida* and *P. pungens*. It could also be the case that *P. rigida* and *P. pungens* had higher gene flow in more heterogeneous environments (e.g., mountains) for longer bouts of time which accelerated the development of RI through ecological character displacement (Cushman and Landguth 2016).

Three species demographic models require confidence in species relationships

The divergence histories of *P. pungens*, *P. rigida*, and *P. taeda* using a three-species inference framework, indicated unreasonably shallow divergence time estimates (e.g., thousands of years; Figure 2.6). Based on the wide and non-normally distributed residuals from data to model comparison (Figure 2.S4), the topology we used in the three-species models did not fit our data. The two-species models we examined provided needed context to the relationships among our focal species. AIC values were over 4,000 units lower for models of *P. pungens* and *P. rigida* suggesting well established demographic processes (i.e., sfs based models) better fit patterns in the site frequency spectrum for *P. pungens* and *P. rigida* than those from comparisons with *P. taeda*. Adding gene flow to the two-species models provided divergence time estimates more aligned with phylogenetic inferences (Hernandez-Leon et al. 2014; Saladin et al. 2017; Gernandt et al. 2018; Jin et al. 2021) suggesting gene flow was important to speciation, but AIC scores did not improve between gene flow models involving *P. taeda*. The best-fit, two-species model for *P. pungens* and *P. rigida* inferred a divergence time of 1.11 mya. Adding population effective size changes may have made our divergence time more comparable to the ~2.74 mya estimate reported in Bolte et al. (2022; see Momigliano, Florin, and Merilä 2021). Divergence ~2,500 years ago under strict isolation was the best fit demographic model for *P. rigida* and *P. taeda*, as well as for *P. pungens* and *P. taeda*, which is interesting given the amount of admixture present across our sampled trees (Figure 2.2). We simply interpret these results as evidence of complex divergence histories that can be more confidently inferred once species topology is resolved.

Past attempts to define relationships and infer divergence times for closely related hard pines of eastern North America (i.e., *P. pungens*, *P. rigida*, *P. serotina*, *P. taeda*, *P. echinata*) resulted in discordance (Gernandt et al. 2018), lower bootstrap confidence (Hernandez-Leon et al. 2013), and lower Bayesian posterior probabilities (Saladin et al. 2017) compared to most other clades belonging to the genus *Pinus*. We used a preferred data type for demographic inference (i.e., genome-wide nuclear data; Excoffier et al. 2013), but the topology we assumed was inferred from chloroplast DNA (cpDNA). Gernandt et al. (2018) found discordance between nuclear and cpDNA phylogenetic inferences. Nuclear DNA placed *P. taeda* in a separate clade with *P. echinata* which shared a once removed ancestor to the clade of *P. pungens*, *P. rigida*, and *P. serotina*. Given our results from two and three species demographic inferences from genome-wide nuclear DNA, it is possible that *P. taeda* did not share an exclusive common ancestor with *P. rigida* and *P. pungens* or that the relationships among these three species does not fit assumptions of bifurcation. Indeed, among the two other demographic inferences studies performed in hard pines, one described the hybrid speciation of *P. densata* (Gao et al. 2012; Wachowiak et al. 2018).

The extensive hybridization between *P. rigida* and *P. taeda*, as observed in our analysis of structure (Figure 2.2), likely challenged the topology we used in three species demographic inference. Moreover, *P. rigida* and *P. taeda* are part of a larger hybridizing complex involving *P. serotina* and *P. echinata* (Smouse and Saylor 1973), that may be obscuring species relationships through genomic homogenization. The species integrity of *P. echinata*, for example, has become a recent concern due to increased hybridization

with *P. taeda* over just the past few decades (Xu, Tauer, and Nelson 2008; Stewart et al. 2010; Stewart, Tauer, and Nelson 2012). Contrastingly, the *P. pungens* genome seems to be less vulnerable to homogenization at present. Among hard pines of eastern North America, successful crossings only occurred with *P. rigida* (Critchfield 1961) and even still, hybrid seed fill rates were low (< 20% of within species crosses). The best next step to understanding the differential development of RI in hard pines in eastern North America is to resolve the species relationships in a phylogenetic inference study that includes all naturally and extensively hybridizing species (*P. serotina*, *P. rigida*, *P. taeda*, *P. echinata*) and *P. pungens*. Demographic inferences can now be performed for 5 species in $\partial\alpha\partial i$ 2.1.0 using Graphics Processing Units (Gutenkunst 2021) opening the potential to gain insight into hybridizing complexes and species relationships among them.

RADseq data are not always anonymous and intergenic

We characterized the distribution of our SNPs in relation to the annotated genome of *P. taeda* to better understand the data used in demographic inference. The majority of our RADseq contigs (94%) were intergenic, but 299 contigs (6%) matched with 393 genic regions. These proportions of captured intergenic to genic regions align with general expectations for RADseq data (Parchman et al. 2018). Across the genic regions observed in our data, 11 contigs had 15 hits that were highly differentiated ($F_{CT} \geq 0.3$) across species, but pairwise estimates of F_{ST} across species provided unanticipated insight. We found that within gene measures of pairwise differentiation did not reflect the patterns of pairwise differentiation from intergenic regions. Contigs that associated within genic regions were highly differentiated between *P. taeda* and *P. pungens* and *P. rigida* and *P.*

pungens. In artificial crossing experiments for these two species pairs, hybrids were either unable to produce seeds or had low seed fertility suggesting reproductive isolation is quite strong. Contrastingly, there was low differentiation between *P. rigida* and *P. taeda* among 10 of the 11 contigs that mapped to genic regions (Figure 2.7). Again, this species pair successfully hybridizes both in nature and through breeding programs. We fully recognize that our genetic data type is not fit for fine scale genomic inferences related to RI, but the patterns of differentiation across species pairs, what is known about RI already from artificial crossing, and the differences between intergenic and genic regions is suggestive. We recommend a whole exome approach in future studies seeking to understand the genes involved in RI, especially between *P. pungens* and *P. rigida*, because they have a rich history of overlapping distributions, gene flow, and ecological divergence (Bolte et al. 2022).

The polygenic nature of adaptation and the limited genomic coverage that can be obtained from large genomes (*Pinus* > 20Gbp) with RADseq data does not allow powerful hypothesis tests regarding the loci contributing to RI (Lowry et al. 2017; McKinney et al. 2016). However, we note two annotations that were enriched in our $F_{CT} \geq 0.3$ data set, heat shock proteins and YT521-B-like-domains. We have provided a summary table of the *P. taeda* genome attributes associated with highly differentiated SNPs, e-values, and corresponding EggNOG descriptions (Table 2.S3) for which more robust future research endeavors can refer to for comparative purposes. Most intriguing are the heat shock proteins that were highly differentiated across species and enriched in our data. These proteins, as well as others that were highly differentiated (Table 2.S3; Pfam:DUF26,

cysteine-rich motifs, and phosphatase 2C), are associated with stress response (Fuchs et al. 2012; Delgado-Cerrone et al. 2018; Hussain et al. 2022), a well described driver of speciation (Lexer and Fay 2005; Cokus, Gugger, and Sork 2015).

Conclusions

There are several outcomes associated with hybridization ranging from genome-wide homogenization (Slatkin 1985) to rapid development of reproductive isolation through reinforcement of species boundaries (Howard 1993) which require methods past demographic inference to elucidate. Multidisciplinary approaches which incorporate SDMs, RDA, genome mapping, loci-specific differentiation in addition to demographic inference, provide a more cohesive understanding of the when, where, and how of lineage divergence. Yet, what will remain unknown for conifers are the relative rates at which RI strengthens in the presence or absence of gene flow, the genetic architectures of traits that promote or inhibit hybrid establishment and introgression, and the contribution of environmental complexity to mode and tempo of RI (Bolte and Eckert 2020) unless more clade specific investigations for conifers, such as the one we conducted here, lay groundwork for comparative analyses and even predictive modeling. Aptly predicting gains and losses to biodiversity under our rapidly changing climatic conditions will require an enhanced recognition of interspecific gene flow potentials, which only genomically-based comparative research can reveal.

Table 2.1. Summaries redundancy analyses with climate and geographic as predictors of genetic variation. Adjusted r^2 represents the individual contribution of the predictor with all others removed and the proportion of variance explained (PVE) represents the overall contribution without controlling for interactive effects among the predictors. An asterisk denotes model significance ($p < 0.01$).

Model	r^2 (%)	Adj. r^2 (%)	PVE (%), RDA1	$p < 0.05$ (by = axis)
Genetics ~ Climate + Geo*	6.05	4.31	63.24	RDA1-4
Genetics ~ Climate Geo*	1.56	0.477	38.42	NA
Genetics ~ Geo Climate*	1.75	1.11	63.78	NA
Ancestry ~ Climate + Geo*	59.92	59.17	81.91	RDA1-2
Ancestry ~ Climate Geo*	5.59	5.16	86.28	NA
Ancestry ~ Geo Climate*	13.85	13.73	86.47	NA

Table 2.2 Results from Fisher's Exact Tests for seven EggNOG descriptions associated with attributes of the *P. taeda* genome. Descriptions with p -values < 0.5 have an asterisk.

EggNOG Description	Counts $F_{CT} < 0.3$	Counts $F_{CT} \geq 0.3$	p-value	odds ratio	95% CI
Heat shock protein*	13	3	0.028	5.309	0.955 - 19.746
YT521-B-like domain*	5	2	0.033	9.135	0.869 - 56.624
agenet domain	10	2	0.090	4.563	0.480 - 21.818
mitogen-activated protein kinase	5	1	0.230	4.529	0.0949 - 41.000
pathogenesis-related protein	4	1	0.196	5.661	0.114 - 57.875
chaperone dnaJ	1	1	0.083	22.643	0.287 - 1756.883
pentatricopeptide repeat- containing protein	39	3	0.426	1.718	0.334 - 5.547

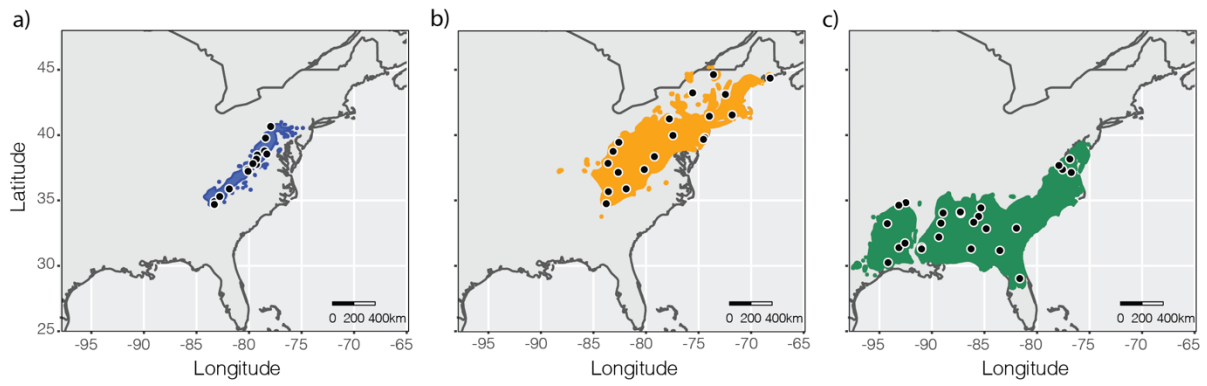


Figure 2.1 Known geographical distribution of focal species, a) *Pinus pungens*, b) *P. rigida*, c) *P. taeda* (Little 1971) in relation to populations sampled (black dots) for genetic analysis.

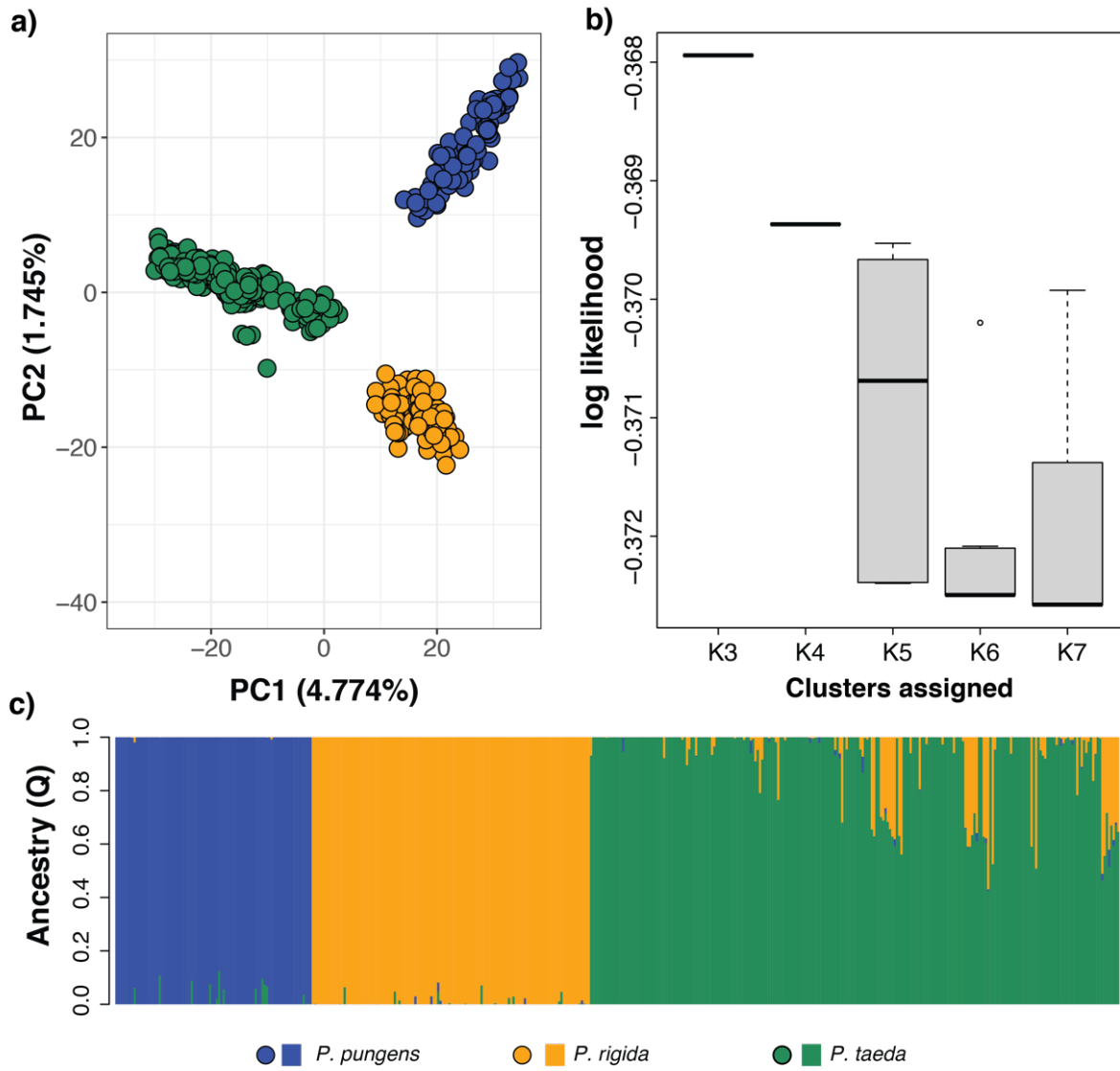


Figure 2.2 Measures of genetic differentiation and diversity among sampled trees of *P. pungens*, *P. rigida*, and *P. taeda*: a) Principal components analysis of 5051 genome-wide single nucleotide polymorphism (SNPs) for *Pinus pungens* (blue, right side of PC1), *P. rigida* (orange, right side of PC1), and *P. taeda* (green, left side of PC1); b) log-likelihood values across ten replicate runs in fastSTRUCTURE for $K = 3$ through $K = 7$; c) results of averaged $K = 3$ ancestry (Q) assignments for each sample arranged by population name in Table 2.S1.

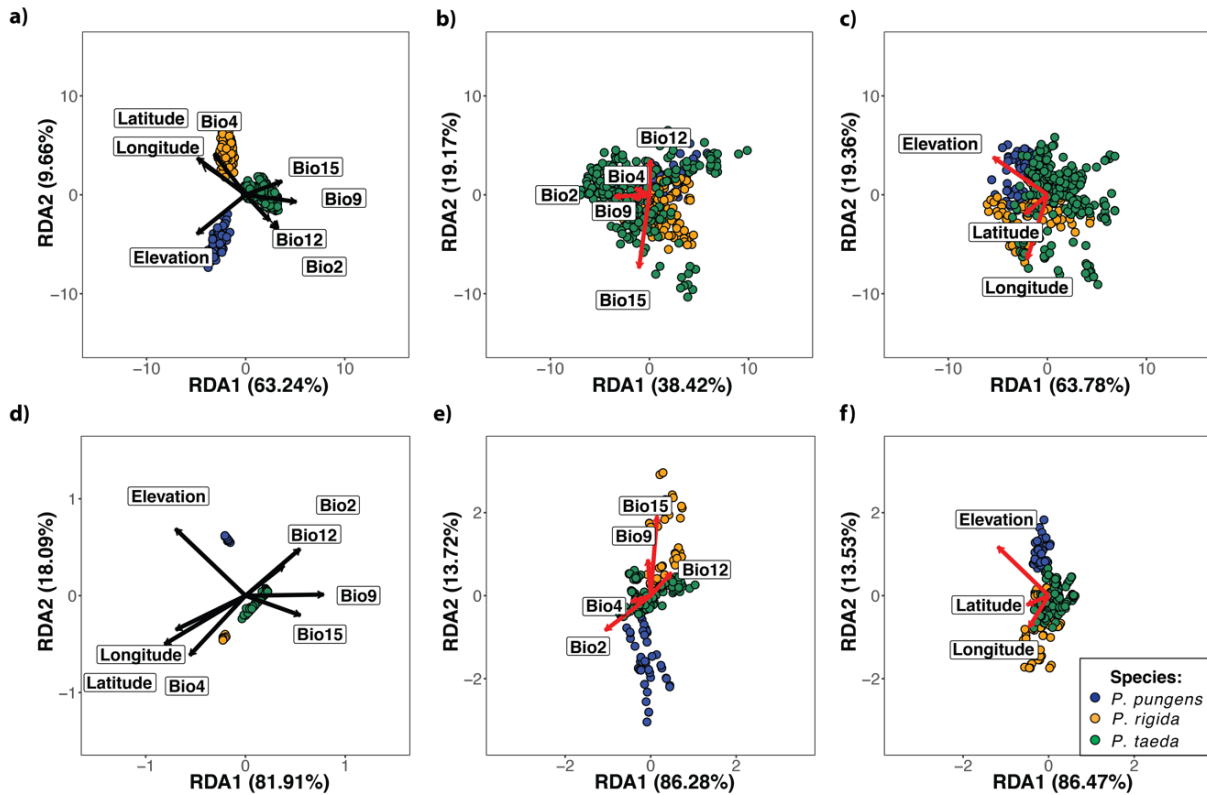


Figure 2.3 Redundancy analysis (RDA) of the multilocus genotypes for each tree with a) climate and geographic predictor variables (full model), b) climate predictor variables (geography removed), and c) geographic predictor variables (climate removed). Panels d-e present redundancy analysis of the ancestral coefficients from structure analysis ($K = 3$) for each tree with d) climate and geographic predictor variables (full model), e) climate predictor variables (geography removed), and f) geographic predictor variables (climate removed). Direction and length of arrows on each RDA plot correspond to the loadings of each variable.

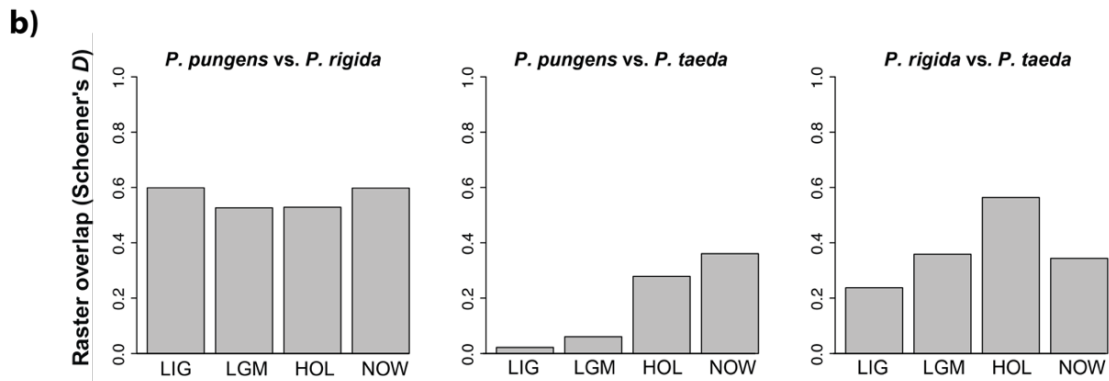
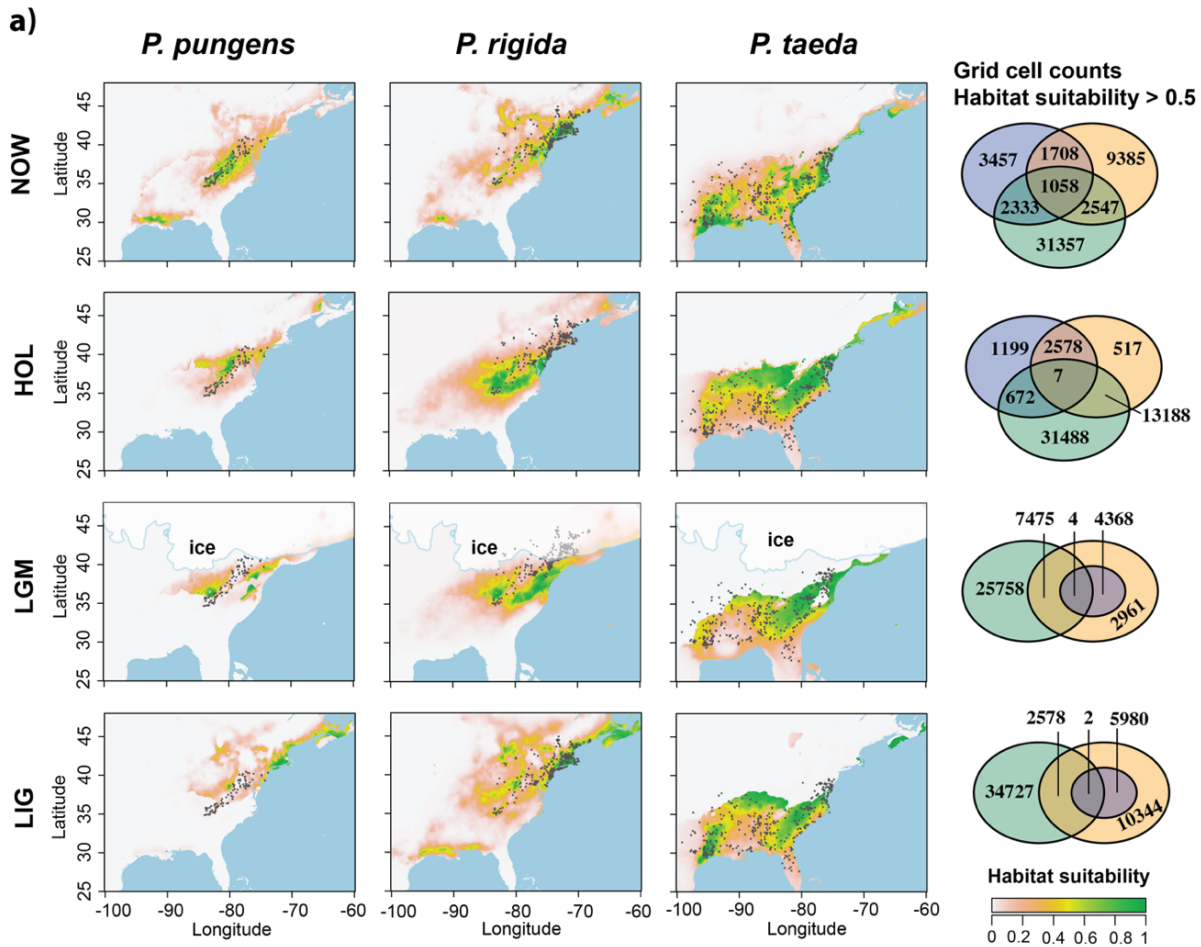


Figure 2.4 SDM predictions a) across four time points for *P. pungens*, *P. rigida*, and *P. taeda*. Occurrence records for each species (black dots) overlay habitat suitability predictions. Venn diagrams illustrate the number of grid cells with moderate to high habitat suitability scores (> 0.5) for each SDM at a given time point, as well as the number of overlapping grid cells. Blue ovals show counts for *P. pungens*, orange ovals show counts for *P. rigida*, and green ovals show counts for the *P. taeda* SDM predictions at each aligning time point. SDM Glacial extent data (labeled ice in LGM plots) for 18 kya was provided by Dyke (2003). Panel b illustrates pairwise comparisons of raster overlap across each time period.

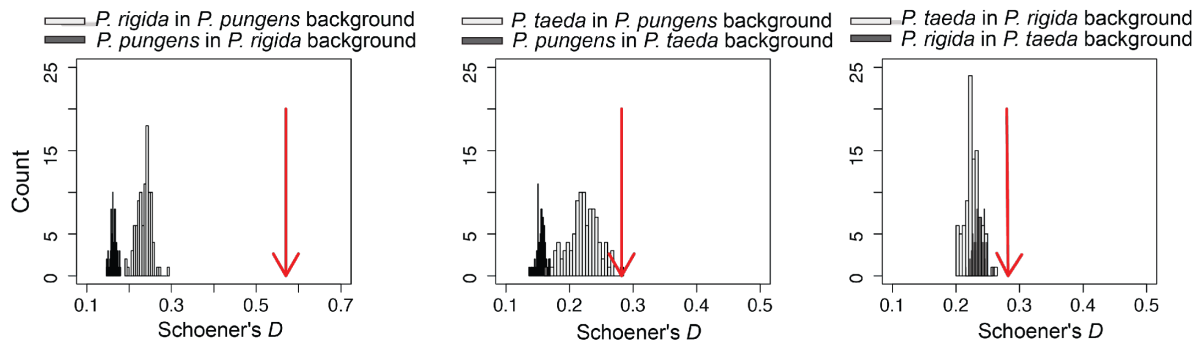


Figure 2.5 Relative distributions of asymmetrical background similarity tests (gray bars) to niche overlap (red arrow). Panels from left to right illustrate the niche relationships between *P. pungens* and *P. rigida*, *P. pungens* and *P. taeda*, and *P. rigida* and *P. taeda*, respectively. An arrow to the left of a background similarity distribution indicates niche divergence, while an arrow to the right indicates niche conservatism.

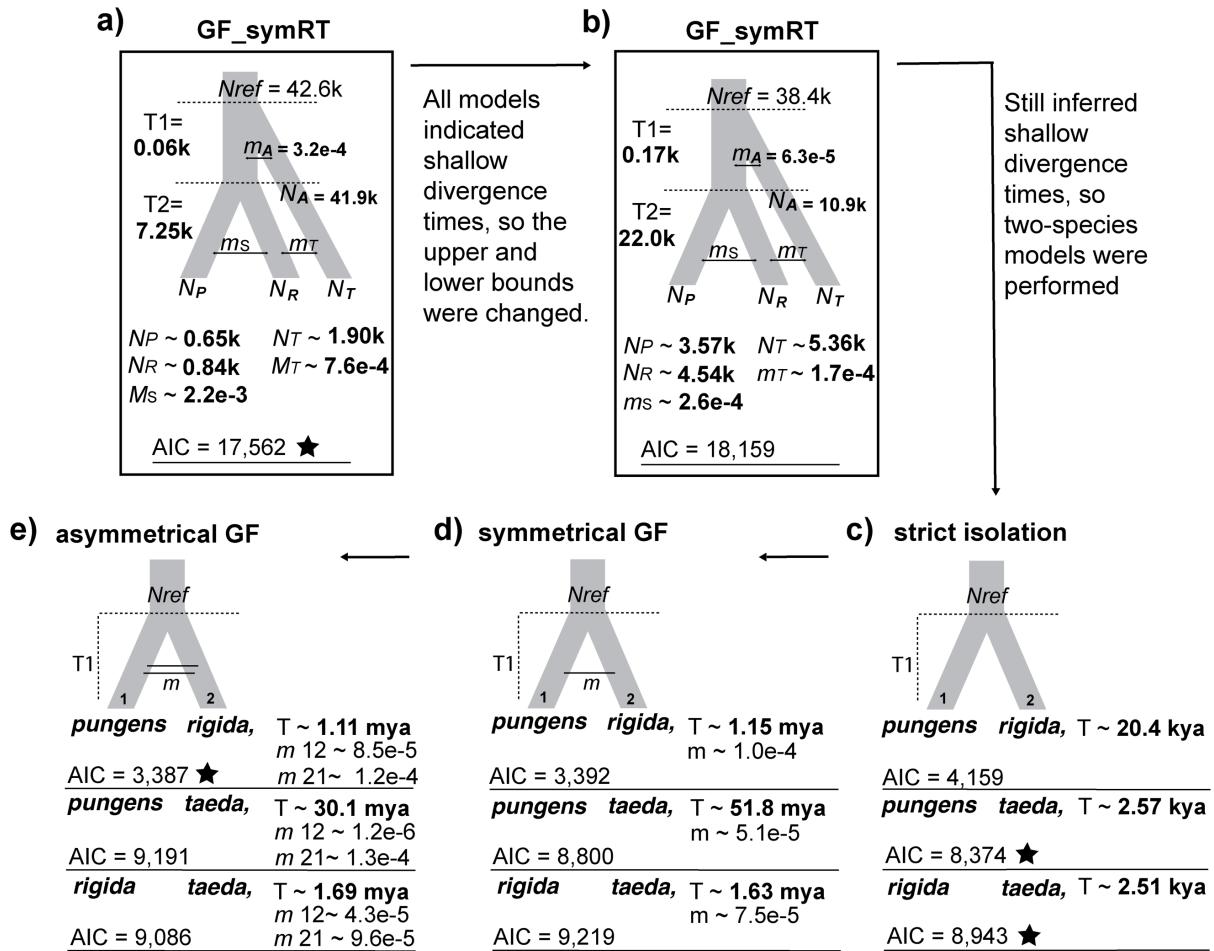


Figure 2.6 Demographic inference workflow where a) two models with the lowest AIC from the first round of inferences were used in b) to force deeper divergence time inferences through manipulation of lower and upper bounds of parameter space. Two population models to test species relationships and topology are presented in panels c - e. GF stands for gene flow. The acronyms symRT and asymRT stands for allowing symmetrical or asymmetrical gene flow between *P. rigida* and *P. taeda* during T2 (time interval 2). Respectively, N_A , N_P , N_R , and N_T are the effective population sizes of *P. taeda* at the end of T1, then *P. pungens*, *P. rigida*, and *P. taeda* at the end of T2.

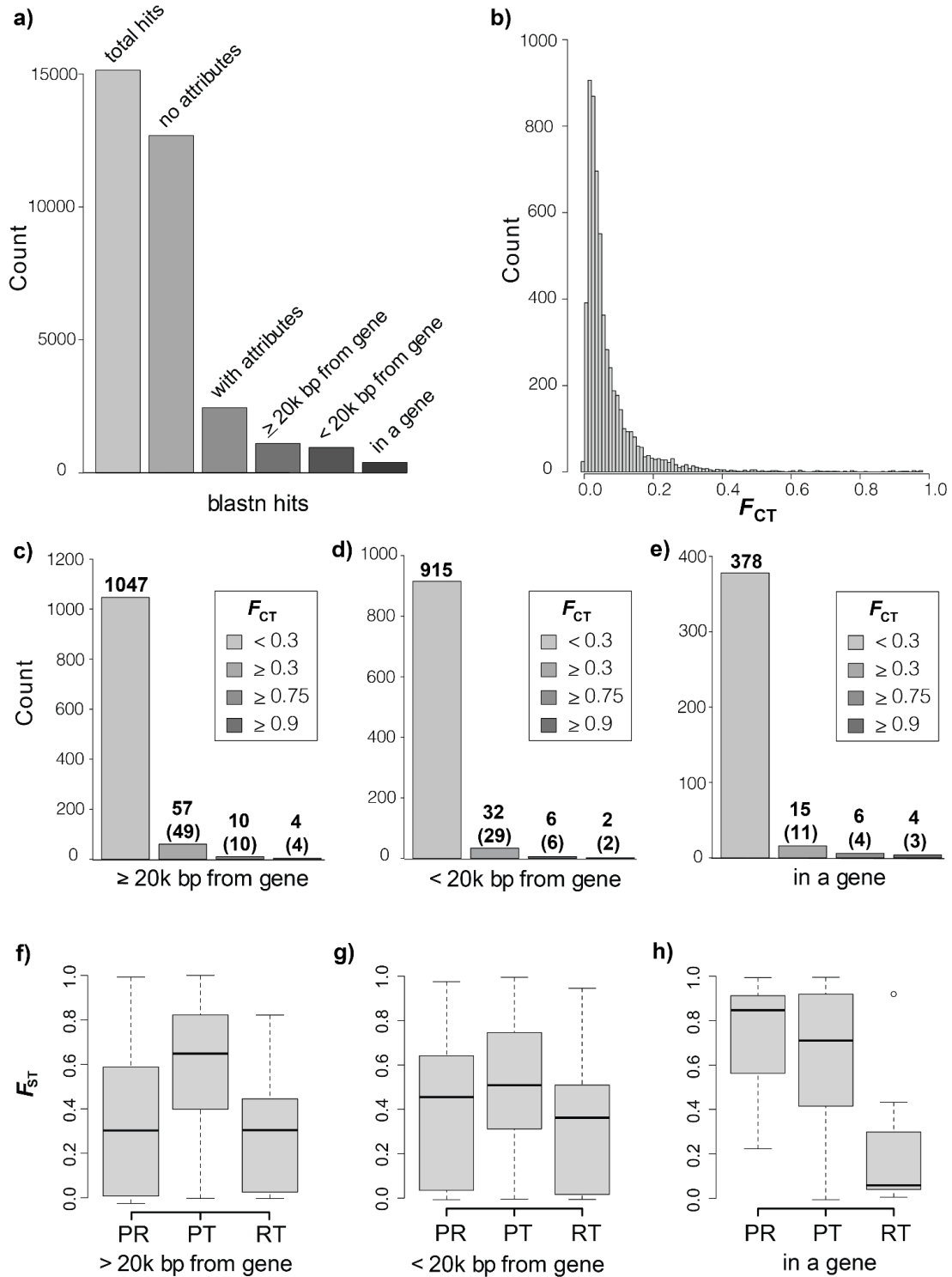


Figure 2.7 Description of blastn hits to the *P. taeda* draft genome. Panel a) shows the number of hits after data was filtered down to one RADseq contig per scaffold with max three unique scaffold IDs allowed per hit and how those relate to matched attributes (i.e. annotations) and locations to genes. Values associated with some bars are nested within bars to the left. Panel b) shows the distribution of F_{CT} values associated with our 5051

SNPs. The number of unique RADseq-scaffold hits and corresponding F_{CT} value ranges are shown in c) for those outside 20k bp of a gene, d) for those within 20k bp of a gene, and e) for those that hit within the gene. The third and fourth bars in panels c-e are nested components of the second bar. In parentheses are the number of unique RADtags (i.e., RADseq IDs) defining the number of hits. The distribution of F_{ST} values from pairwise species comparisons (PR, comparing variation between *P. pungens* and *P. rigida*; PT, between *P. pungens* and *P. taeda*; RT, between *P. rigida* and *P. taeda*) for SNPs that are f) relatively far from a gene, g) relatively close to a gene, and h) within a gene.

Literature Cited

- Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38(5), 541–545. doi: 10.1111/ECOG.01132
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716-723.
- Birol, I., Raymond, A., Jackman, S. D., Pleasance, S., Coope, R., Taylor, G. A., ... Jones, S. J. M. (2013). Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*, 29(12), 1492–1497. doi: 10.1093/bioinformatics/btt178
- Brady, K. U., Kruckeberg, A. R., & Bradshaw, H. D. (n.d.). Evolutionary Ecology Of Plant Adaptation To Serpentine Soils. *Annual Review of Ecology, Evolution, and Systematics*, 36, 243-266
- Bolte, C. E., & Eckert, A. J. (2020). Determining the when, where and how of conifer speciation: a challenge arising from the study 'Evolutionary history of a relict conifer *Pseudotsuga chienii*.' *Annals of Botany*, 125(1), v–vii. doi: 10.1093/AOB/MCZ201
- Bolte, C. E., Faske, T. M., & Friedline, C. J. (2022). Divergence amid recurring gene flow: complex demographic histories for *Pinus pungens* and *P. rigida* align with a growing expectation for forest trees. <https://doi.org/10.1101/2022.02.12.480138>
- Cokus, S. J., Gugger, P. F., & Sork, V. L. (2015). Evolutionary insights from de novo transcriptome assembly and SNP discovery in California white oaks. *BMC Genomics*, 16(1). doi: 10.1186/s12864-015-1761-4

- Cushman, S. A., & Landguth, E. L. (2016). Spatially heterogeneous environmental selection strengthens evolution of reproductively isolated populations in a dobzhansky-muller system of hybrid incompatibility. *Frontiers in Genetics*, 7(NOV). <https://doi.org/10.3389/fgene.2016.00209>
- Critchfield WB (1963) The Austrian x red pine hybrid. *Silvae Genetica* 12, 187-191
- Critchfield WB (1967) Crossability and relationships of the closed-cone pines. *Silvae Genetica*, 16, 89–97
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. doi: 10.1093/bioinformatics/btr330
- De La Torre, A. R., Li, Z., Van De Peer, Y., & Ingvarsson, P. K. (2017). Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Molecular Biology and Evolution*, 34(6), 1363–1377. doi: 10.1093/molbev/msx069
- Delgado-Cerrone, L., Alvarez, A., Mena, E., Ponce de León, I., & Montesano, M. (2018). Genome-wide analysis of the soybean CRK-family and transcriptional regulation by biotic stress signals triggering plant immunity. *PLoS ONE*, 13(11), 1–27. <https://doi.org/10.1371/journal.pone.0207438>
- Eckert, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González-Martínez, S. C., & Neale, D. B. (2010). Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*. <https://doi.org/10.1534/genetics.110.115543>
- Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, 9(10). doi: 10.1371/journal.pgen.1003905
- Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Molecular Ecology Resources*, 17(1), 27–32. doi: 10.1111/1755-0998.12509
- Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150–3152. doi: 10.1093/bioinformatics/bts565
- Fuchs, S., Grill, E., Meskiene, I., & Schweighofer, A. (2013). Type 2C protein phosphatases in plants. *FEBS Journal*, 280(2), 681–693. <https://doi.org/10.1111/j.1742-4658.2012.08670.x>
- Gao, J., Wang, B., Mao, J. F., Ingvarsson, P., Zeng, Q. Y., & Wang, X. R. (2012). Demography and speciation history of the homoploid hybrid pine *Pinus densata* on

- the Tibetan Plateau. *Molecular Ecology*, 21(19), 4811–4827. doi: 10.1111/j.1365-294X.2012.05712.x
- Gernandt, D. S., Aguirre Dugua, X., Vázquez-Lobo, A., Willyard, A., Moreno Letelier, A., Pérez de la Rosa, J. A., ... Liston, A. (2018). Multi-locus phylogenetics, lineage sorting, and reticulation in *Pinus* subsection Australes. *American Journal of Botany*, 105(4), 711–725. doi: 10.1002/AJB2.1052
- Goudet, J. (2005). hierfstat, a package for r to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, 5(1), 184–186. <https://doi.org/10.1111/J.1471-8286.2004.00828.X>
- Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10). doi: 10.1371/journal.pgen.1000695
- Gutenkunst, R. N. (2021). Dadi.CUDA: Accelerating Population Genetics Inference with Graphics Processing Units. *Molecular Biology and Evolution*, 38(5), 2177–2178. doi: 10.1093/MOLBEV/MSAA305
- Hapke, A., & Thiele, D. (2016). GIBPSs: a toolkit for fast and accurate analyses of genotyping-by-sequencing data without a reference genome. *Molecular Ecology Resources*, 16(4), 979–990. doi: 10.1111/1755-0998.12510
- He, Z., Li, X., Yang, M., Wang, X., Zhong, C., Duke, N. C., ... Shi, S. (2019). Speciation with gene flow via cycles of isolation and migration: Insights from multiple mangrove taxa. *National Science Review*, 6(2), 275–288. doi: 10.1093/nsr/nwy078
- Hernández-León, S., Gernandt, D. S., Pérez de la Rosa, J. a., & Jardón-Barbolla, L. (2013). Phylogenetic Relationships and Species Delimitation in *Pinus* Section Trifoliae Inferred from Plastid DNA. *PLoS ONE*, 8(7), 1–14. doi: 10.1371/journal.pone.0070501
- Herten, K., Hestand, M. S., Vermeesch, J. R., & Van Houdt, J. K. J. (2015). GBSX: A toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinformatics*, 16(1), 1–6. doi: 10.1186/s12859-015-0514-3
- Hiebl, C. & Challenge C. (2018). phyloclim: Integrating phylogenetics and climatic niche modeling. R package version 0.9.5. <https://CRAN.R-project.org/package=phyloclim>
- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15), 1965–1978. doi: 10.1002/JOC.1276
- Hohenlohe, P. A., Day, M. D., Amish, S. J., Miller, M. R., Kamps-Hughes, N., Boyer, M. C., ... Luikart, G. (2013). Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, 22(11), 3002–3013. doi: 10.1111/mec.12239

- Howard, D. J. (1993). Reinforcement: origin, dynamics, and fate of an evolutionary hypothesis. *Hybrid zones and the evolutionary process*, 46-69.
- Hussain, A., Asif, N., Pirzada, A. R., Noreen, A., Shaukat, J., Burhan, A., Zaynab, M., Ali, E., Imran, K., Ameen, A., Mahmood, M. A., Nazar, A., & Mukhtar, M. S. (2022). Genome wide study of cysteine rich receptor like proteins in *Gossypium* sp. *Scientific Reports*, 12(1), 1–18. <https://doi.org/10.1038/s41598-022-08943-1>
- Hyun, S. K. (1960). Mass production of control-pollinated seed of conifers. In *Proceedings of the Fifth World Forestry Congress held at the University of Washington, Seattle, August 29-September 10, 1960* (Vol. 2).
- Hyun, S. K., & Ahn, K. Y. (1959). Mass production of pitch-loblolly hybrid pine (x *Pinus rigitaeda*) seed. *Res. Rep. Ins. For. Genet*, (1), 11-24.
- Jackson, S. T., & Overpeck, J. T. (2000). *Responses of Plant Populations and Communities to Environmental Changes of the Late Quaternary*. 26(4), 194–220.
- Jetton, R. M., Crane, B. S., Whittier, W. A., & Dvorak, W. S. (2015). Genetic resource conservation of Table Mountain pine (*Pinus pungens*) in the central and southern Appalachian Mountains. *Tree Planters' Notes*, 58(1), 42–52.
- Jin, W. T., Gernandt, D. S., Wehenkel, C., Xia, X. M., Wei, X. X., & Wang, X. Q. (2021). Phylogenomic and ecological analyses reveal the spatiotemporal evolution of global pines. *Proceedings of the National Academy of Sciences of the United States of America*, 118(20). doi: 10.1073/PNAS.2022302118/-DCSUPPLEMENTAL
- Kane, V. R., Lutz, J. A., Alina Cansler, C., Povak, N. A., Churchill, D. J., Smith, D. F., ... North, M. P. (2015). Water balance and topography predict fire and forest structure patterns. *Forest Ecology and Management*, 338, 1–13. doi: 10.1016/j.foreco.2014.10.038
- Kass, J. M., Muscarella, R., Galante, P. J., Bohl, C. L., Pinilla-Buitrago, G. E., Boria, R. A., ... Anderson, R. P. (2021). ENMeval 2.0: Redesignated for customizable and reproducible modeling of species' niches and distributions. *Methods in Ecology and Evolution*, 12(9), 1602–1608. doi: 10.1111/2041-210X.13628
- Keeley, J. E., Pausas, J. G., Rundel, P. W., Bond, W. J., & Bradstock, R. A. (2011). Fire as an evolutionary pressure shaping plant traits. *Trends in Plant Science*, 16(8), 406–411. doi: 10.1016/j.tplants.2011.04.002
- Khodwekar, S., & Gailing, O. (2017). Evidence for environment-dependent introgression of adaptive genes between two red oak species with different drought adaptations. *American Journal of Botany*, 104(7), 1088–1098. doi: 10.3732/ajb.1700060
- Knezick, D. R., Kuser, J. E., & Garrett, P. W. (1985). Supplemental mass pollination of single clone orchards for the production of southern pine hybrids. In *Proc. Southern Forest Tree Improv. Conf*, 18, 187-193

- Ladeau, S. L., & Clark, J. S. (2006). Pollen production by *Pinus taeda* growing in elevated atmospheric CO₂. *Functional Ecology*, 20(3), 541–547. doi: 10.1111/j.1365-2435.2006.01133.x
- Legendre, P., & Legendre, L. (2012) Numerical Ecology. Third Edition. Elsevier.
- Lexer, C., & Fay, M. (2005) Adaptation to environmental stress: A rare or frequent driver of speciation? *Journal of Evolutionary Biology*, 18(4), 893-900.
- Little EL Jr. (1971) Atlas of United States trees, Vol. 1, conifers and important hardwoods: U.S. Department of Agriculture, 1146, 9, p200
- Liu, L., Hao, Z. Z., Liu, Y. Y., Wei, X. X., Cun, Y. Z., & Wang, X. Q. (2014). Phylogeography of *Pinus armandii* and its relatives: Heterogeneous contributions of geography and climate changes to the genetic differentiation and diversification of Chinese white pines. *PLoS ONE*, 9(1), 1–12. doi: 10.1371/journal.pone.0085920
- Liu, Y. Y., Jin, W. T., Wei, X. X., & Wang, X. Q. (2019). Cryptic speciation in the Chinese white pine (*Pinus armandii*): Implications for the high species diversity of conifers in the Hengduan Mountains, a global biodiversity hotspot. *Molecular Phylogenetics and Evolution*, 138(May), 114–125. doi: 10.1016/j.ympev.2019.05.015
- Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, 17(2), 142–152. <https://doi.org/10.1111/1755-0998.12635>
- Ma, X. F., Szmidt, A. E., & Wang, X. R. (2006). Genetic structure and evolutionary history of a diploid hybrid pine *Pinus densata* inferred from the nucleotide variation at seven gene loci. *Molecular Biology and Evolution*, 23(4), 807–816. doi: 10.1093/molbev/msj100
- Marchi, N., Schlichta, F., & Excoffier, L. (2021). Demographic inference. *Current Biology*, 31(6), R276–R279. doi: 10.1016/j.cub.2021.01.053
- McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al. (2016). In *Molecular Ecology Resources* (Vol. 17, Issue 3, pp. 356–361). Blackwell Publishing Ltd. <https://doi.org/10.1111/1755-0998.12649>
- McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, 17(4), 656–669. doi: 10.1111/1755-0998.12613

- McKinney, G. J., Waples, R. K., Pascal, C. E., Seeb, L. W., & Seeb, J. E. (2018). Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. *Molecular Ecology Resources*, *18*(3), 570–579. doi: 10.1111/1755-0998.12763
- Menon, M., Bagley, J. C., Friedline, C. J., Whipple, A. V., Schoettle, A. W., Leal-Sàenz, A., ... Eckert, A. J. (2018). The role of hybridization during ecological divergence of southwestern white pine (*Pinus strobiformis*) and limber pine (*P. flexilis*). *Molecular Ecology*, *27*(5), 1245–1260. doi: 10.1111/MEC.14505
- Momigliano, P., Florin, A. B., & Merilä, J. (2021). Biases in Demographic Modeling Affect Our Understanding of Recent Divergence. *Molecular Biology and Evolution*, *38*(7), 2967–2985. doi: 10.1093/molbev/msab047
- Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., ... Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, *15*(3), 1–13. doi: 10.1186/gb-2014-15-3-r59
- Nosil, P., & Feder, J. L. (2012). Widespread yet heterogeneous genomic divergence. *Molecular Ecology*, *21*(12), 2829–2832. doi: 10.1111/j.1365-294X.2012.05580.x
- Nunez, M. A., Horton, T. R., & Simberloff, D. (2009). Lack of belowground mutualisms hinders Pinaceae invasions. In *Ecology* (Vol. 90).
- Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., ... Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, *497*(7451), 579–584. doi: 10.1038/nature12211
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B., ... & Wagner, H. (2020). Vegan: Community ecology package. *R package version, 2.5-7*.
- Otto-bliesner, A. B. L., Marshall, S. J., Overpeck, J. T., Gifford, H., Otto-bliesner, B. L., Marshall, S., ... Miller, G. H. (2019). *Simulating Arctic Climate Warmth and Icefield Retreat in the Last Interglacial*
- Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C. A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, *21*(12), 2991–3005. doi: 10.1111/j.1365-294X.2012.05513.x
- Parchman, T. L., Jahner, J. P., Uckele, K. A., Galland, L. M., & Eckert, A. J. (2018). RADseq approaches and applications for forest tree genetics. *Tree Genetics and Genomes*, *14*(3). doi: 10.1007/S11295-018-1251-3
- Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and

- genotyping in model and non-model species. *PLoS ONE*, 7(5). doi: 10.1371/journal.pone.0037135
- Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: an open-source release of Maxent. *Ecography*, 40(7), 887–893. doi: 10.1111/ECOG.03049
- Pickles, B. J., Twieg, B. D., O'Neill, G. A., Mohn, W. W., & Simard, S. W. (2015). Local adaptation in migrated interior Douglas-fir seedlings is mediated by ectomycorrhizas and other soil factors. *New Phytologist*, 207(3), 858–871. doi: 10.1111/nph.13360
- Prunier, J., Verta, J. P., & Mackay, J. J. (2016). Conifer genomics and adaptation: At the crossroads of genetic diversity and genome function. *New Phytologist*, 209(1), 44–62. doi: 10.1111/nph.13565
- Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, 2014(1). doi: 10.7717/PEERJ.431
- Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, 197(2), 573–589. <https://doi.org/10.1534/genetics.114.164350>
- Richards, C. L., Carstens, B. C., & Lacey Knowles, L. (2007, November). Distribution modelling and statistical phylogeography: An integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography*, Vol. 34, pp. 1833–1845. doi: 10.1111/j.1365-2699.2007.01814.x
- Rieseberg, L. H., & Blackman, B. K. (2010). Speciation genes in plants. *Annals of Botany*, 106(3), 439–455. doi: 10.1093/aob/mcq126
- Saladin, B., Leslie, A. B., Wüest, R. O., Litsios, G., Conti, E., Salamin, N., & Zimmermann, N. E. (2017). Fossils matter: improved estimates of divergence times in *Pinus* reveal older diversification.
- Scull, P., Franklin, J., Chadwick, O. A., & McArthur, D. (2003). Predictive soil mapping: A review. *Progress in Physical Geography*, 27(2), 171–197. doi: 10.1191/0309133303pp366ra
- Shang, H., Hess, J., Pickup, M., Field, D. L., Ingvarsson, P. K., Liu, J., & Lexer, C. (2020). Evolution of strong reproductive isolation in plants: Broad-scale patterns and lessons from a perennial model group: Evolution of strong barriers in plants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1806). doi: 10.1098/rstb.2019.0544
- Slatkin, M. (1985). Gene flow in natural populations. *Annual Review of Ecology and Systematics*. Vol. 16, 393–430. doi: 10.1146/annurev.ecolsys.16.1.393

- Smouse, P. E. & Saylor, L. C. (1973). Studies of the *Pinus rigida*-*Serotina* Complex II. Natural hybridization among the *Pinus rigida*-*serotina* complex, *P. taeda* and *P. echinata*. *Annals of the Missouri Botanical Garden*, 60(2), 192-203.
- Soltis, D. E., Morris, A. B., McLachlan, J. S., Manos, P. S., & Soltis, P. S. (2006). Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, 15(14), 4261–4293. doi: 10.1111/j.1365-294X.2006.03061.x
- Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., ... Langley, C. H. (2016). Sequence of the sugar pine megagenome. *Genetics*, 204(4), 1613–1626. doi: 10.1534/genetics.116.193227
- Stewart, J. F., Tauer, C. G., & Nelson, C. D. (2012). Bidirectional introgression between loblolly pine (*Pinus taeda* L.) and shortleaf pine (*P. echinata* Mill.) has increased since the 1950s. *Tree Genetics and Genomes*, 8(4), 725–735. doi: 10.1007/s11295-011-0459-2
- Tuskan, G. A., DiFazio, S., Jansson, S., & Bohlmann, J. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313, 1596–1605.
- Vallejo-Marín, M., & Hiscock, S. J. (2016). Hybridization and hybrid speciation under global change. *The New Phytologist*, 211(4), 1170–1187. doi: 10.1111/nph.14004
- Varela, S., Lima-Ribeiro, M. S., & Terribile, L. C. (2015). A short guide to the climatic variables of the last glacial maximum for biogeographers. *PLoS ONE*, 10(6). doi: 10.1371/JOURNAL.PONE.0129037
- Veloz, S. D., Williams, J. W., Blois, J. L., He, F., Otto-Bliesner, B., & Liu, Z. (2012). No-analog climates and shifting realized niches during the late quaternary: implications for 21st-century predictions by species distribution models. *Global Change Biology*, 18(5), 1698–1713. doi: 10.1111/J.1365-2486.2011.02635.X
- Wachowiak, W., Zaborowska, J., Bartosz, Ł., Perry, A., Zucca, G. M., González-martínez, S. C., & Cavers, S. (2018). *Molecular signatures of divergence and selection in closely related pine taxa*.
- Warren, D. L., Glor, R. E., & Turelli, M. (2008). Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. *Evolution*, 62(11), 2868–2883. doi: 10.1111/J.1558-5646.2008.00482.X
- Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., Vasquez-Gross, H. A., ... Neale, D. B. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, 196(3), 891–909. doi: 10.1534/genetics.113.159996
- Whitlock, M. C. & Lotterhos, K. (2014). OutFLANK: Fst outliers with trimming. R package version 0.2

- Widmer, A., Lexer, C., & Cozzolino, S. (2009). Evolution of reproductive isolation in plants. *Heredity*, *102*(1), 31–38. doi: 10.1038/HDY.2008.69
- Wu, C. I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*(6), 851–865. doi: 10.1046/j.1420-9101.2001.00335.x
- Wu, S., Wang, Y., Wang, Z., Shrestha, N., & Liu, J. (2022). Species divergence with gene flow and hybrid speciation on the Qinghai–Tibet Plateau. *New Phytologist*, *392*–404. doi: 10.1111/nph.17956
- Yang, Y. X., Zhi, L. Q., Jia, Y., Zhong, Q. Y., Liu, Z. L., Yue, M., & Li, Z. H. (2020). Nucleotide diversity and demographic history of *Pinus bungeana*, an endangered conifer species endemic in China. *Journal of Systematics and Evolution*, *58*(3), 282–294. doi: 10.1111/jse.12546
- Zhang, D., Xia, T., Yan, M., Dai, X., Xu, J., Li, S., & Yin, T. (2014). Genetic introgression and species boundary of two geographically overlapping pine species revealed by molecular markers. *PLoS ONE*, *9*(6). doi: 10.1371/journal.pone.0101106
- Zhou, Y., Duvaux, L., Ren, G., Zhang, L., Savolainen, O., & Liu, J. (2017). Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. *Heredity*, *118*(3), 211–220. doi: 10.1038/hdy.2016.72
- Zobel, D. B. (1969). Factors Affecting the Distribution of *Pinus pungens*, an Appalachian Endemic. *Ecological Monographs*, *39*(3), 303–333. doi: 10.2307/1948548
- Zukowska, W. B., & Wachowiak, W. (2016). Utility of closely related taxa for genetic studies of adaptive variation and speciation: Current state and perspectives in plants with focus on forest tree species. *Journal of Systematics and Evolution*, *54*(1), 17–28. doi: 10.1111/jse.12177

Appendix 2

Table 2.S1 Location of sampled populations, number of trees (*n*) that were sampled for *Pinus pungens* (PU), *P. rigida* (RI), and *P. taeda* (TA) populations. Averaged ancestry assignments (with K=3) for each population are in the last three columns.

Species	Code	Location	Lat	Long	<i>n</i>	TA_anc	PU_anc	RI_anc
<i>P. pungens</i>	PU_BB	Briery Branch, VA	38.48	-79.22	3	0.000	1.000	0.000
<i>P. pungens</i>	PU_BN	Buchanan State Forest, PA	39.77	-78.43	6	0.010	0.986	0.004
<i>P. pungens</i>	PU_BV	Buena Vista, VA	37.76	-79.29	7	0.000	1.000	0.000
<i>P. pungens</i>	PU_DT	Dragon's Tooth, VA	37.37	-80.16	2	0.000	1.000	0.000
<i>P. pungens</i>	PU_EG	Edinburg Gap, VA	38.79	-78.53	8	0.013	0.987	0.000
<i>P. pungens</i>	PU_EK	Elliott Knob, VA	38.17	-79.30	7	0.000	1.000	0.000
<i>P. pungens</i>	PU_GA	Walnut Fork, GA	34.92	-83.28	7	0.013	0.987	0.000
<i>P. pungens</i>	PU_LG	Looking Glass Rock, NC	35.30	-82.79	8	0.034	0.966	0.000
<i>P. pungens</i>	PU_NM	North Mountain, VA	37.82	-79.63	8	0.000	1.000	0.000
<i>P. pungens</i>	PU_PM	Poor Mountain, VA	37.23	-80.09	9	0.017	0.983	0.000
<i>P. pungens</i>	PU_SC	Pine Mountain, VA	34.70	-83.30	6	0.023	0.975	0.002
<i>P. pungens</i>	PU_SH	Shenandoah NP, VA	38.55	-78.31	5	0.000	1.000	0.000
<i>P. pungens</i>	PU_SV	Stone Valley Forest, PA	40.66	-77.95	3	0.000	1.000	0.000
<i>P. pungens</i>	PU_TR	Table Rock Mountain, NC	35.89	-81.88	7	0.006	0.994	0.000
<i>P. rigida</i>	RI_BR	Bass River State Forest, NJ	39.80	-74.41	6	0.000	0.000	1.000
<i>P. rigida</i>	RI_CT	Pachaug State Forest, CT	41.54	-71.81	7	0.000	0.000	1.000
<i>P. rigida</i>	RI_DT	Dragon's Tooth, VA	37.37	-80.16	7	0.009	0.000	0.991
<i>P. rigida</i>	RI_GA	Chattahoochee NF, GA	34.75	-83.78	6	0.000	0.000	1.000
<i>P. rigida</i>	RI_GW	George Washington NF, VA	38.36	-79.20	7	0.000	0.000	1.000
<i>P. rigida</i>	RI_HH	Hudson Highlands State Park, NY	41.44	-73.97	6	0.010	0.000	0.989
<i>P. rigida</i>	RI_JF	Jefferson NF, VA	37.15	-82.64	6	0.000	0.000	1.000
<i>P. rigida</i>	RI_KY	Daniel Boone NF, KY	37.84	-83.62	7	0.000	0.004	0.996
<i>P. rigida</i>	RI_ME	Acadia NP, ME	44.36	-68.19	9	0.006	0.009	0.986
<i>P. rigida</i>	RI_MI	Michaux State Forest, PA	39.98	-77.44	9	0.000	0.000	1.000
<i>P. rigida</i>	RI_NJ	Wharton State Forest, NJ	39.68	-74.53	6	0.012	0.000	0.988
<i>P. rigida</i>	RI_NY	Macomb State Park, NY	44.63	-73.58	7	0.001	0.000	0.999
<i>P. rigida</i>	RI_OH	South Bloomingville, OH	39.45	-82.59	3	0.000	0.000	1.000
<i>P. rigida</i>	RI_RS	Rome Sand Plains, NY	43.23	-75.56	8	0.007	0.003	0.990
<i>P. rigida</i>	RI_SH	Shawnee State Park, OH	38.75	-83.13	3	0.000	0.000	1.000
<i>P. rigida</i>	RI_SP	Sproul State Forest, PA	41.24	-77.78	5	0.000	0.000	1.000
<i>P. rigida</i>	RI_TN	Great Smoky Mountains NP, TN	35.68	-83.58	5	0.000	0.000	1.000
<i>P. rigida</i>	RI_TR	Table Rock Mountain, NC	35.89	-81.89	8	0.007	0.000	0.993
<i>P. rigida</i>	RI_VT	Bellows Falls, VT	43.11	-72.44	7	0.001	0.002	0.997
<i>P. taeda</i>	TA_AA	Frank Jackson State Park, AL	31.30	-86.27	11	0.993	0.000	0.007
<i>P. taeda</i>	TA_AB	Clear Creek Rec. Area, AL	34.02	-87.27	10	0.994	0.006	0.000
<i>P. taeda</i>	TA_AC	Houston Rec. Area, AL	34.12	-87.29	10	1.000	0.000	0.000

Table 2.S1 continued

Species	Code	Location	Lat	Long	<i>n</i>	TA_anc	PU_anc	RI_anc
<i>P. taeda</i>	TA_AD	Coleman Lake, AL	33.78	-85.56	10	0.991	0.000	0.009
<i>P. taeda</i>	TA_AE	Talladega County, AL	33.34	-86.03	9	0.975	0.000	0.024
<i>P. taeda</i>	TA_AF	Jackson Township, AR	34.84	-92.48	9	0.988	0.000	0.011
<i>P. taeda</i>	TA_AG	Hot Springs Village, AR	34.64	-93.15	9	0.998	0.000	0.002
<i>P. taeda</i>	TA_FL	Pittman, FL	29.03	-81.64	15	0.948	0.004	0.047
<i>P. taeda</i>	TA_GA	Sloppy Floyd State Park, GA	34.43	-85.34	10	0.999	0.000	0.001
<i>P. taeda</i>	TA_GB	Pine Mountain, GA	32.84	-84.83	9	0.994	0.005	0.001
<i>P. taeda</i>	TA_GC	Ellenton, GA	31.18	-83.54	9	0.948	0.005	0.048
<i>P. taeda</i>	TA_GD	Jenkins County, GA	32.88	-81.96	12	0.980	0.005	0.015
<i>P. taeda</i>	TA_LA	Alco, LA	31.39	-93.14	7	0.770	0.003	0.228
<i>P. taeda</i>	TA_LB	Catahoula Nat. Wildlife Area, LA	31.74	-92.56	7	0.678	0.004	0.318
<i>P. taeda</i>	TA_MA	Choctaw Lake, MS	33.27	-89.14	8	0.967	0.001	0.032
<i>P. taeda</i>	TA_MB	Chickasaw County, MS	34.05	-88.94	10	0.995	0.000	0.005
<i>P. taeda</i>	TA_MC	Franklin County, MS	31.43	-90.99	9	0.992	0.000	0.008
<i>P. taeda</i>	TA_MD	Eunice, MS	31.29	-90.99	8	0.724	0.005	0.271
<i>P. taeda</i>	TA_ME	Montrose, MS	32.20	-89.34	5	0.636	0.005	0.359
<i>P. taeda</i>	TA_TA	Cass County, TX	33.23	-94.25	12	0.991	0.000	0.009
<i>P. taeda</i>	TA_TB	Village Creek State Park, TX	30.25	-94.17	8	0.887	0.000	0.113
<i>P. taeda</i>	TA_VA	Pocahontas State Park, VA	37.37	-77.58	9	0.975	0.000	0.025
<i>P. taeda</i>	TA_VB	Powhatan State Park, VA	37.68	-77.92	8	0.960	0.006	0.034
<i>P. taeda</i>	TA_VC	Chippokes Plant. State Park, VA	37.14	-76.74	10	0.956	0.001	0.043
<i>P. taeda</i>	TA_VD	Westmoreland State Park, VA	38.17	-76.87	8	0.604	0.015	0.381

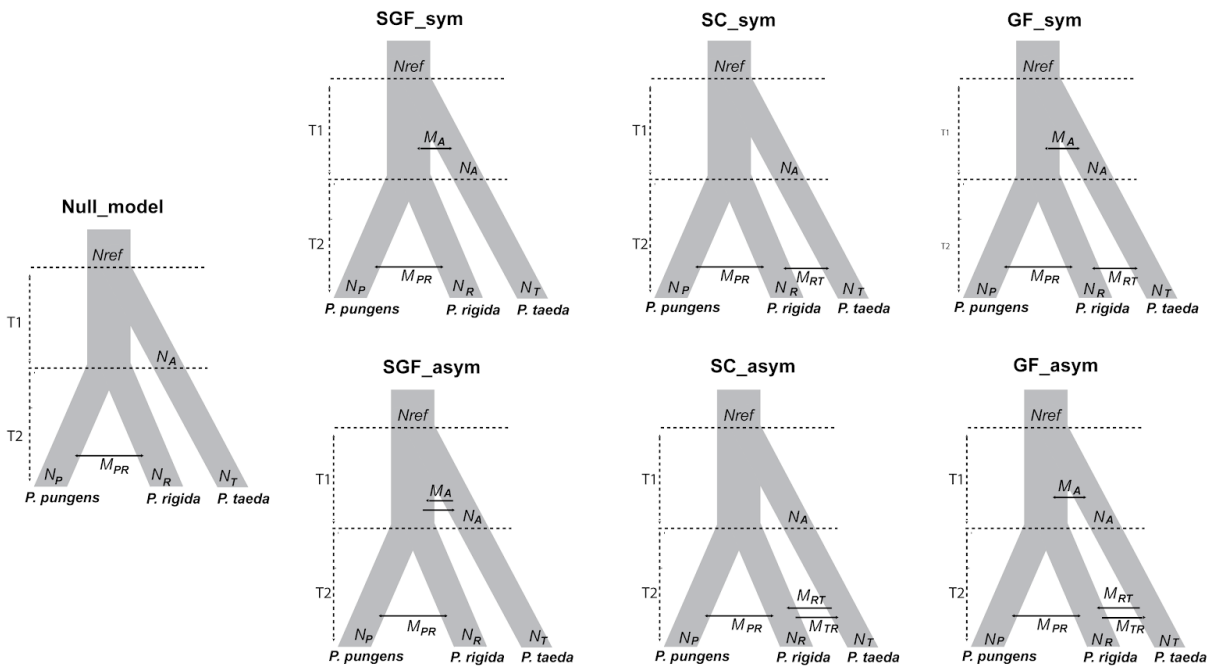


Figure 2.S1 Seven demographic models that were tested in the first round of model selection. SGF is speciation with gene flow. SC is secondary contact. GF allowed gene flow at T1 (first time interval) and T2 (second time interval). The acronym sym means the model inferred symmetrical gene flow. The acronym asym means the model inferred asymmetric gene flow.

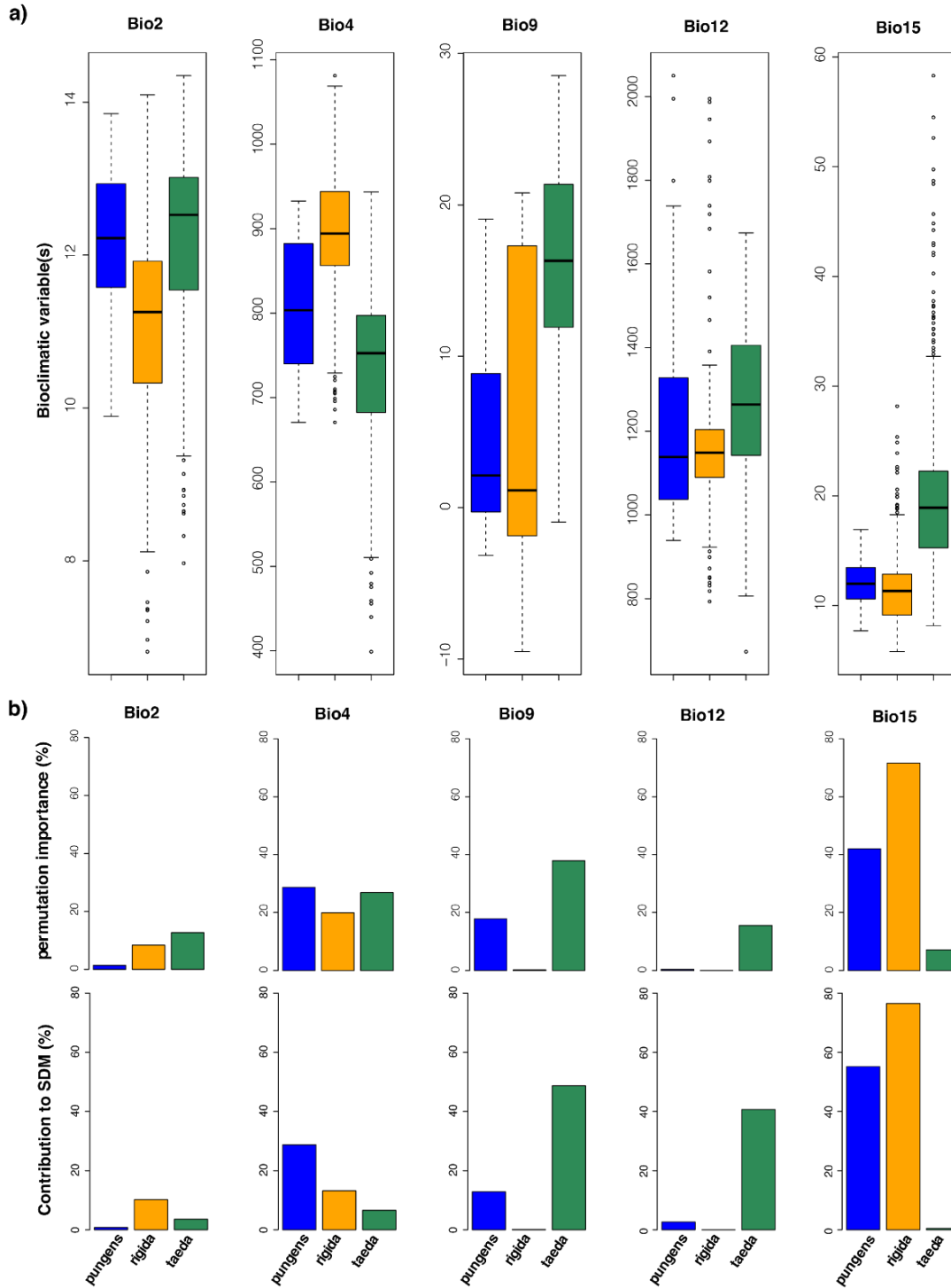


Figure 2.S2 Bioclimatic variable associations with a) occurrence data used in SDMs and b) SDM permutation importance and percent contribution to each model. Blue bars correspond to *P. pungens*. Orange bars correspond to *P. rigida*. Green bars correspond to *P. taeda*.

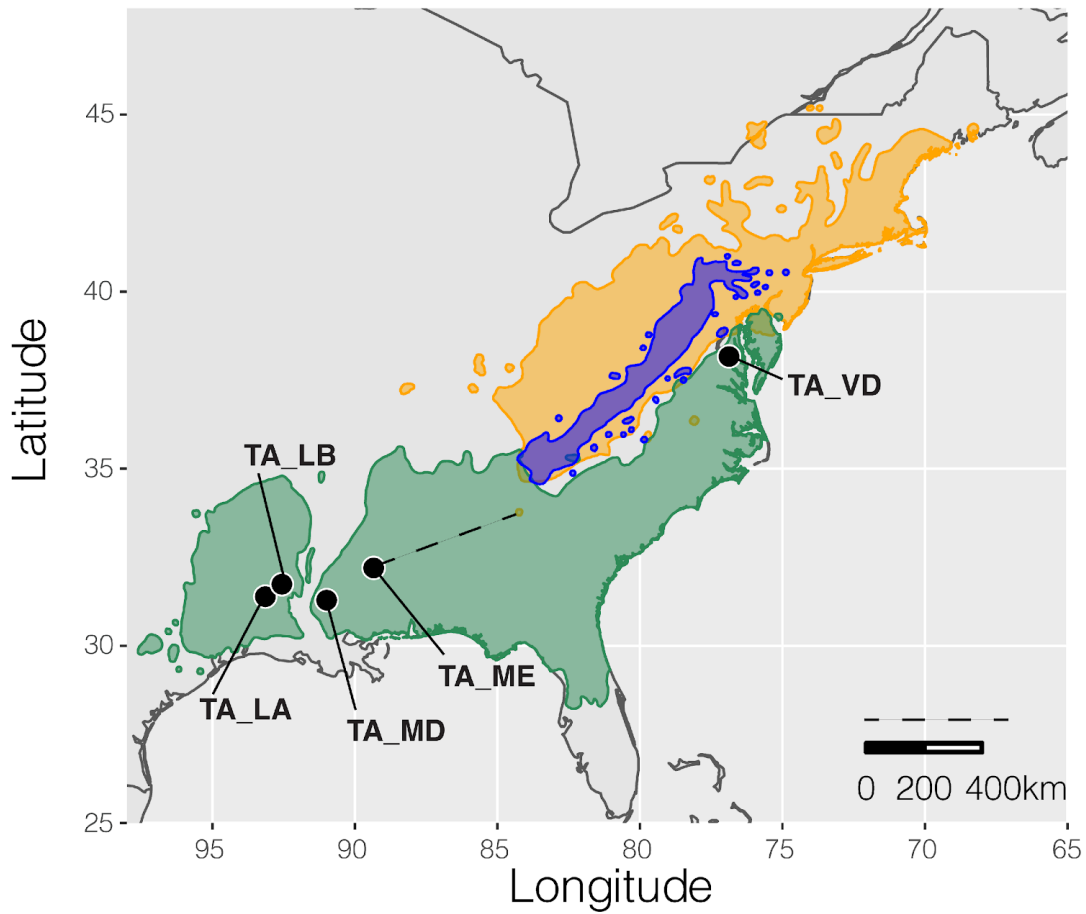


Figure 2.S3 Geographical distributions of *P. pungens* (blue), *P. rigida* (orange), *P. taeda* (green), as described in Little (1971). Five populations with the most admixture present between *P. taeda* and *P. rigida* are plotted (black dots) and labeled. The dashed line illustrates distance between the closest region of geographical overlap between natural stands of *P. taeda* in Louisiana and Mississippi in relation to suitable habitat of *P. rigida*.

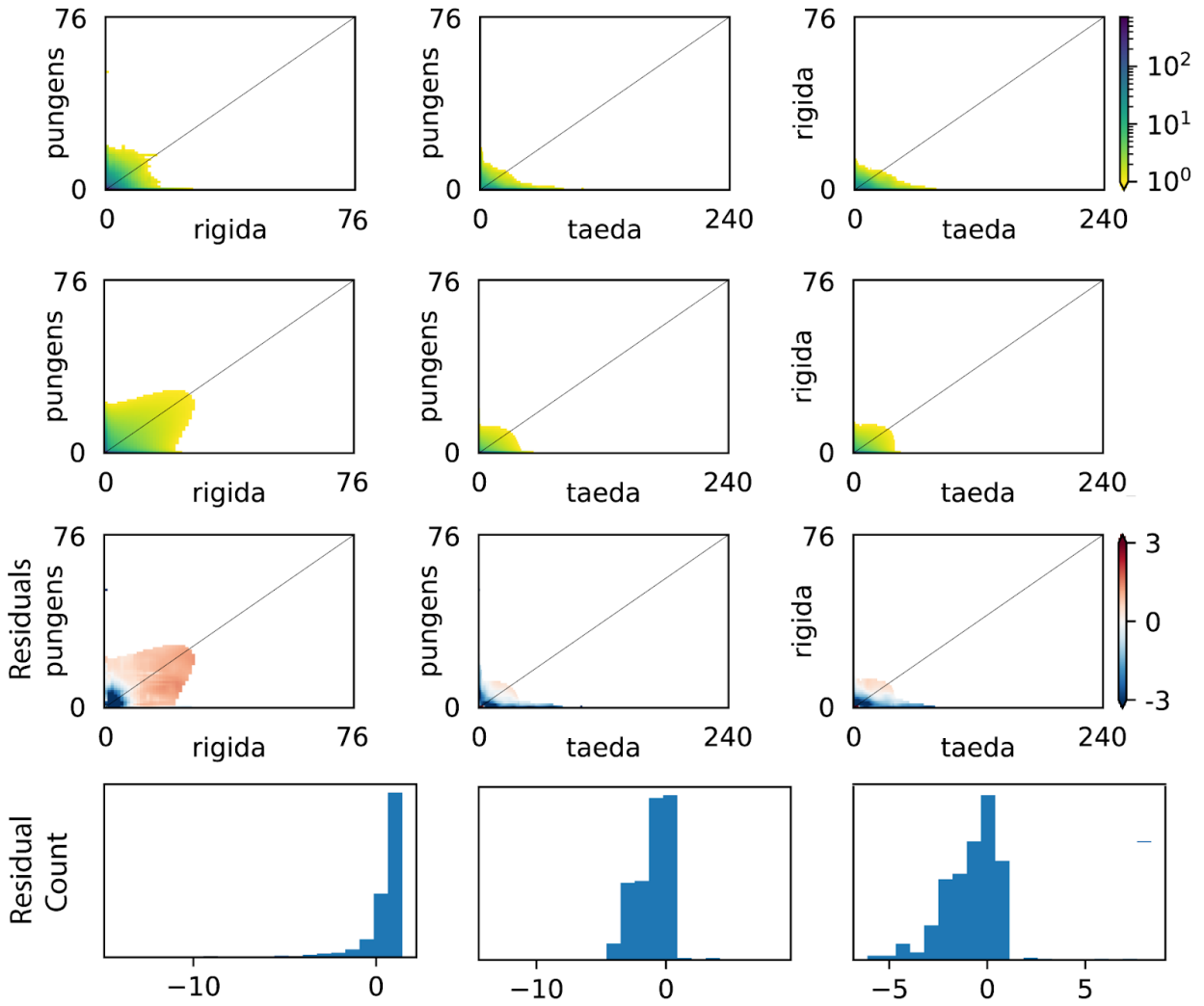


Figure 2.S4 Folded site frequency spectrum for the data (top row) and symmetrical gene flow model (second row). Residuals are plotted in the last two rows and correspond to the three-species model run with the lowest AIC.

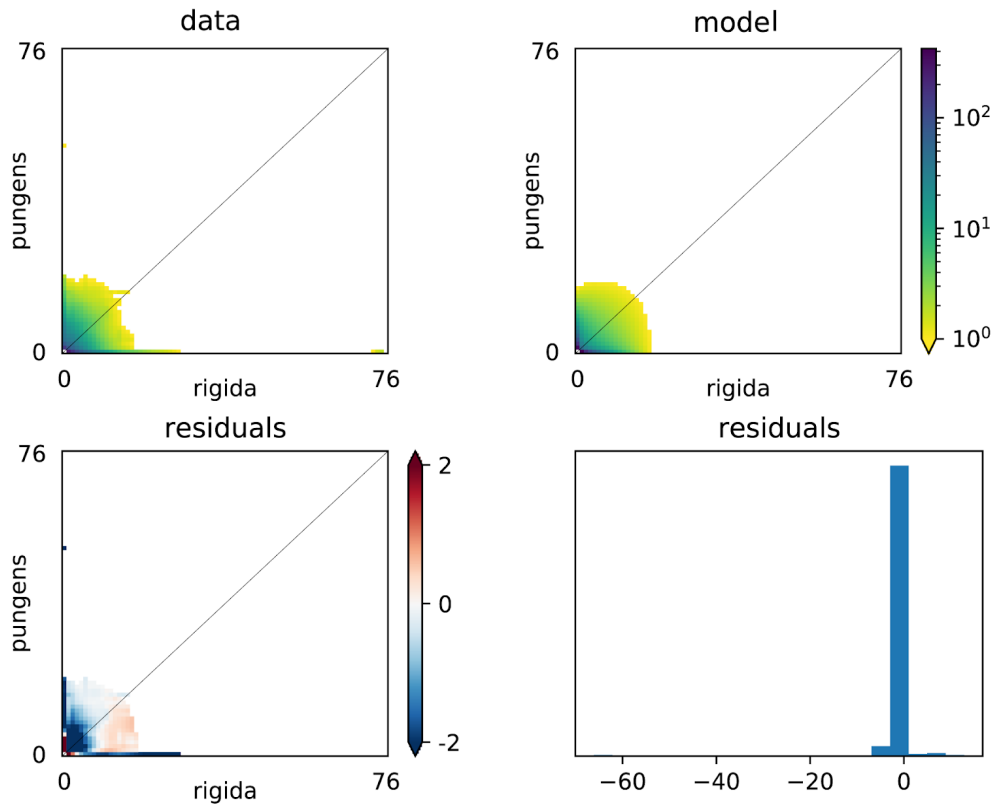


Figure 2.S5 Folded site frequency spectrum for the data and asymmetrical gene flow model for the *P. pungens* and *P. rigida* two-species model. Residuals are plotted in the bottom row and correspond to the two-species model run with the lowest AIC.

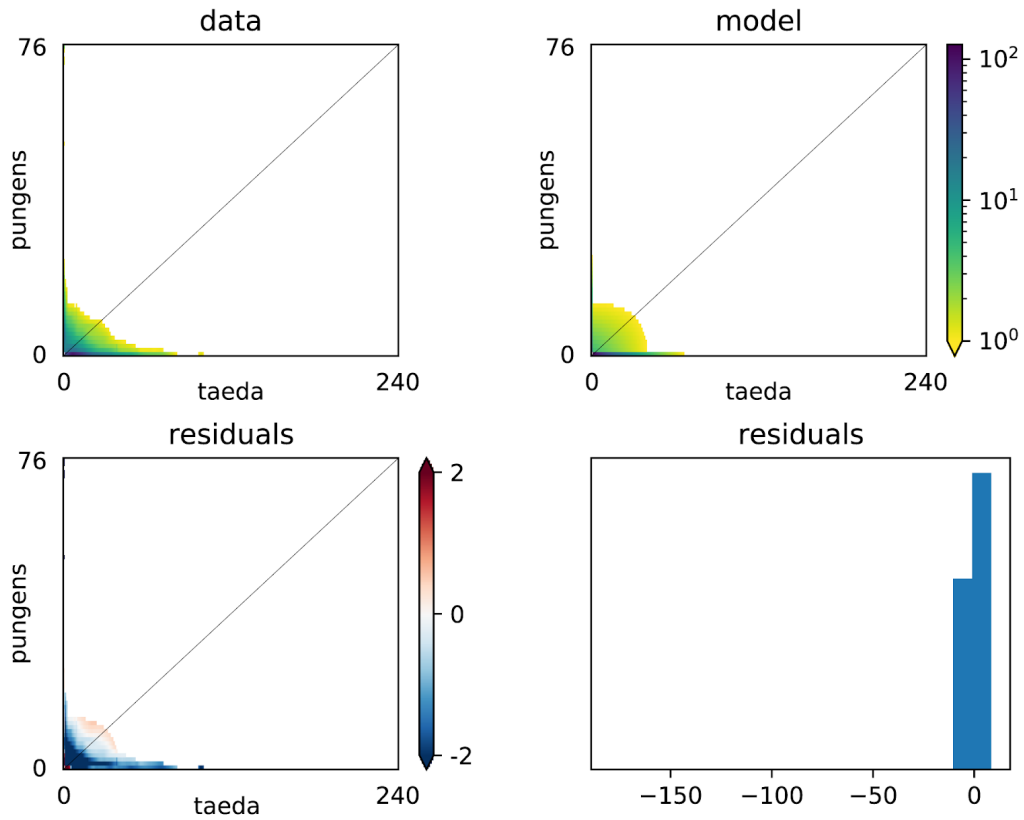


Figure 2.S6 Folded site frequency spectrum for the data and strict isolation model for the *P. pungens* and *P. taeda* two-species model. Residuals are plotted in the bottom row and correspond to the two-species model run with the lowest AIC.

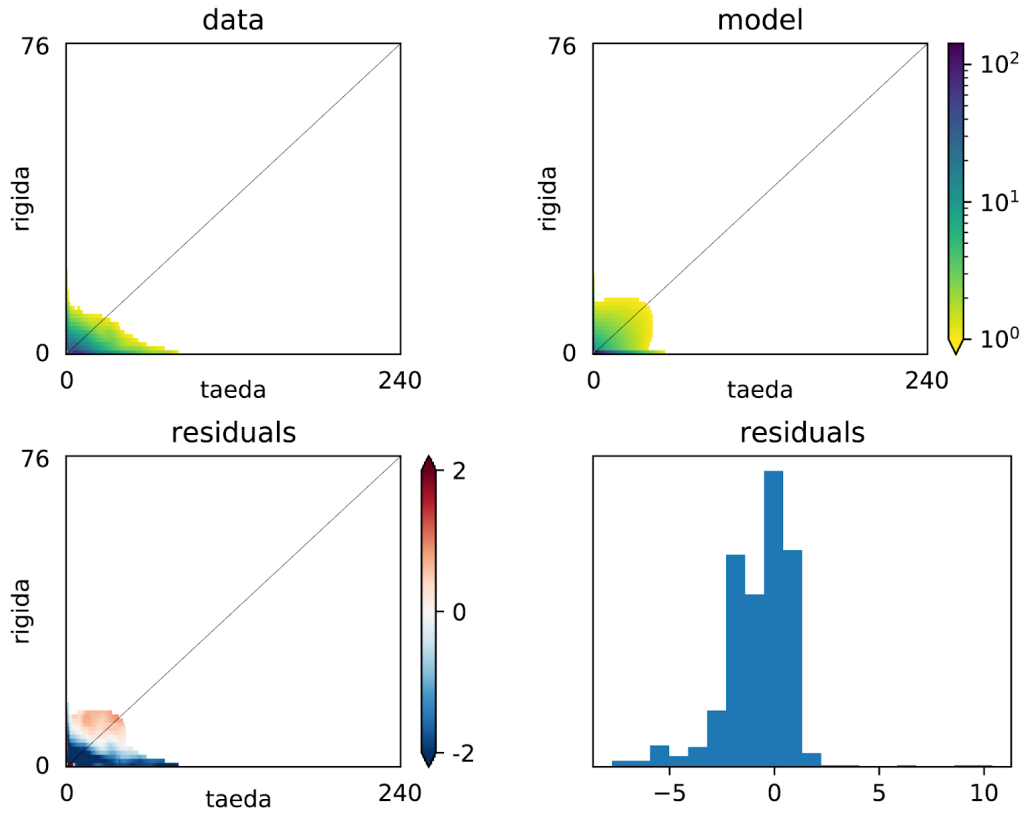


Figure 2.S7 Folded site frequency spectrum for the data and strict isolation model for the *P. rigida* and *P. taeda* two-species model. Residuals are plotted in the bottom row and correspond to the two-species model run with the lowest AIC.

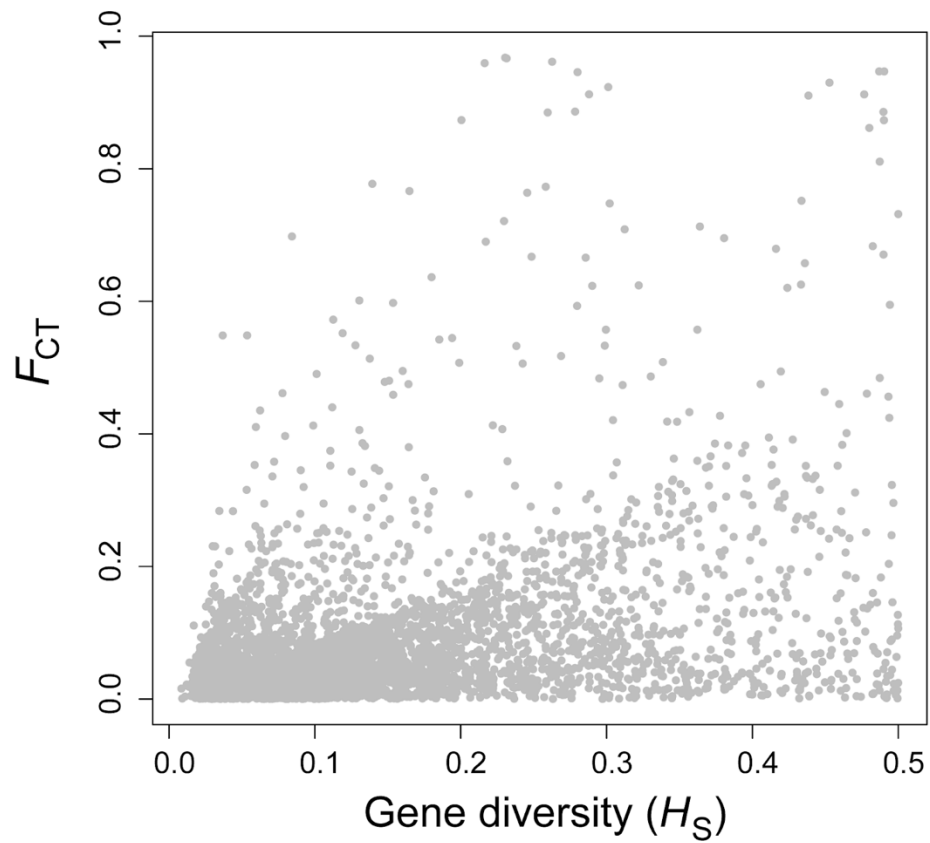


Figure 2.S8 Output from OutFLANK showing the distribution of 5051 SNPs according to measure of gene diversity and F_{CT} .

Table 2.S2 Parameter estimates and 95% Confidence Intervals (CI) for the two-species models with the lowest AIC scores for each pairwise species inference. Values are unscaled. The eps value in the FIM uncertainty test is the relative step size used when taking numerical derivatives.

a) Strict isolation *P. rigida* and *P. taeda*

Parameters	dadi estimate	FIM (± SD)	95% CI	lower CI	upper CI	eps value
NuR	581.43	3.62	7.10	574.33	588.53	1.00E-02
NuT	1346.25	8.33	16.33	1,329.92	1,362.57	1.00E-02
T1	2,509.88	2.254980	4.420	2,505.46	2,514.30	1.00E-03
θ (Nref)	48782.81	742.80	1455.88	47,326.93	50,238.69	1.00E-02

b) Strict isolation *P. pungens* and *P. taeda*

Parameters	dadi estimate	FIM (± SD)	95% CI	lower CI	upper CI	eps value
NuP	380.93	113.38	222.23	158.70	603.15	1.00E-03
NuT	1162.82	23.08	45.24	1,117.58	1,208.06	1.00E-02
T1	2,571.52	852.74	1671.38	900.14	4,242.90	1.00E-03
θ (Nref)	48427.88	797.41	1562.93	46,864.95	49,990.81	1.00E-02

c) Asymmetrical migrations, *P. pungens* and *P. rigida*

Parameters	dadi estimate	FIM (± SD)	95% CI	lower CI	upper CI	eps value
NuP	20594.43	6.21	12.18	20,582.24	20,606.61	1.00E-07
NuR	22763.12	510.35	1000.28	21,762.85	23,763.40	1.00E-04
T1	1,106,344.41	15,576.09	30,529.13	1,075,815.28	1,136,873.54	1.00E-03
m12	8.48E-05	2.38E-08	4.66E-08	8.47E-05	8.48E-05	1.00E-06
m21	1.16E-04	3.26E-08	6.40E-08	1.16E-04	1.16E-04	1.00E-07
θ (Nref)	1782.47	31.77	62.28	1,720.19	1,844.74	1.00E-07

Table 2.S3 Summary results from pairwise analysis of F_{ST} across SNPs that hit within genic regions and information extracted from the *P. taeda* annotated genome files (Query Sequence and EggNOG Description) for each match. The lower the blastn e-value the better the match.

RADtag_ID	F_{ST_RT}	F_{ST_PT}	F_{ST_PR}	Fixed for same allele	blastn e-value	Query Sequence	EggNOG Description
Contig_28882	0.15	0.59	0.22	NA	5.18E-18	PITA_26761	aspartic proteinase-like protein
Contig_38296	NA	0.48	NA	NA rigida-no data	2.12E-29	PITA_41641	pectinesterase
Contig_38794	0.30	0.31	NA	rigida pungens	3.02E-17	PITA_50952	Essential component of the PAM complex, a complex required for the translocation of transit peptide-containing proteins from the inner membrane into the mitochondrial matrix in an ATP-dependent manner (By similarity)
Contig_38922	0.01	0.71	0.47	taeda	3.46E-23	PITA_21651	Cysteine-rich receptor-like protein kinase
Contig_38922	0.01	0.71	0.47	taeda	3.46E-23	PITA_11424	cysteine-rich repeat secretory protein
Contig_38922	0.01	0.71	0.47	taeda	3.46E-23	PITA_36166	Pfam:DUF26
Contig_44880	NA	0.87	0.80	rigida taeda	9.87E-19	PITA_05170	signal peptide peptidase-like
Contig_46405	0.43	0.35	NA	rigida pungens	4.52E-12	PITA_22637	Branched-chain-amino-acid aminotransferase-like protein 3
Contig_46405	0.43	0.35	NA	rigida pungens	1.48E-13	PITA_26461	receptor-like protein kinase At1g80640-like
Contig_50030	0.92	-0.01	0.89	pungens	3.95E-06	PITA_22771	NA
Contig_50030	0.92	-0.01	0.89	pungens	3.95E-06	PITA_32008	NA
Contig_58047	0.04	0.83	0.65	taeda	1.81E-22	PITA_03098	Heat Shock Protein
Contig_65343	0.01	0.97	0.99	pungens	1.58E-16	PITA_18237	phosphatase 2C
Contig_69929	0.06	0.99	0.92	pungens	2.46E-08	PITA_24066	26S proteasome non-atpase regulatory subunit
Contig_73630	0.06	0.99	0.91	taeda	5.67E-22	PITA_44592	NA

Chapter 3

The extent of genetic diversity and hybridization within sympatric stands of two closely related pine species (*Pinus pungens* and *P. rigida*) in the southern Appalachian Mountains

Abstract

Climate change affects species distributions, population connectivity, and reproductive phenology, thus influences the rate of gene flow across populations and species boundaries. Intraspecific and interspecific gene flow increases genetic diversity, and with this increase, comes greater adaptive potential. Preparing for climate change will require predictions of adaptive potential which is dependent on assessments of hybridization and standing levels of population genetic diversity. Forest trees have long generation times and low migratory potential. Therefore, under rapidly changing environmental conditions, adaptational lags, population fragmentation, and genetic diversity reductions are generally expected. Increase in interspecific gene flow (i.e., hybridization) has been observed in plant species under warming climatic conditions, though, and linked to a breakdown of reproductive phenological barriers and increases in species distributional overlap. We focus here on two pine species, *Pinus pungens* and *P. rigida*, with overlapping species distributions along the southern Appalachian Mountains, attempting to assess the current extent of hybridization and genetic diversity at three sympatric forest stands. Even though these species have had recurring gene flow throughout their divergence history, our genome-wide nuclear data indicate that interspecific boundaries are strongly maintained in sympatry, as highly differentiating single nucleotide polymorphisms (SNPs) are consistently identified across the three stands. Additionally,

intraspecific population structure was observed across the three stands indicating potential roles of population fragmentation (i.e., localized drift and low connectivity) as well as local adaptation in structuring allele frequencies across sampled stands. Given the results of past studies and those presented here, ecological character displacement, coupled with disruptive selection, has probably been involved in the development of reproductive isolation (RI). Evidence from previous work on *P. pungens* and *P. rigida* suggests distributional ranges have cyclically or consistently overlapped, but we suspect some populations of our focal species have not interacted and may be less genetically isolated.

Introduction

Defining the extent and role of hybridization between two or more species is foundational to studies of ecology and evolution. The consequences of hybridization have implications for conservation management and contribute to our general understanding of reproductive isolation (RI) in relation to the speciation continuum (Seehausen et al. 2014). Hybridization was once narrowly viewed as a process that reinforced species boundaries. We now recognize that hybridization can also lead to increased biodiversity through hybrid speciation (Abbott et al. 2013), increased genetic diversity through adaptive introgression (Rieseberg and Wendel 1993), or even reduce interspecific biodiversity through lineage fusion and/or species displacement (Grant and Grant 2014). Recent

observations of hybridization between once prezygotically isolated species suggests that phenological barriers, such as timing of pollen release and flowering in plants, may not be permanently established and may break down under warming climatic conditions (Vallejo-Marín and Hiscock 2016). This suggests the importance of re-evaluating the extent of hybridization as climate changes.

Extrinsic barriers to reproduction and phenological prezygotic isolation are commonly reported between closely related plant species (Lowry et al. 2008; Baack et al. 2015). For tree taxa, long generation times and low migratory potential are threats to population persistence under rapidly changing climate conditions (Petit and Hampe 2006). If prezygotic isolation through mainly phenological schedules are labile in one or both closely related species, however, secondary contact may occur and promote an increase in genetic diversity (Abbott 2017). Indeed, populations with high genetic diversity hold a greater capacity for adaptive evolution (Seehausen 2004; Gompert et al. 2017). Quantifying the drivers of standing levels of genetic diversity through evolutionary processes including hybridization and introgression can lead to better forest management outcomes (Janes and Hamilton 2017). Some tree species with a rich history of interspecific gene flow may not have hybrids with intermediate morphologies (e.g., transgressive phenotypes; Stelkens and Seehausen 2009) or be actively hybridizing with closely related taxa under the current climate conditions (Linan et al. 2021). Regardless of whether a management plan seeks to promote hybridization or restrict it, studies that consider both ecological and genetic data are likely to provide the most accurate depiction of the present, the past, and thus the future.

The instability of climate during the Quaternary Period has left imprints across the genomes of many temperate and boreal tree species, revealing changes in effective population size and extent of gene flow (e.g., Levensen, Tiffin, and Olson 2012; Li et al. 2013; Yang et al. 2020). For instance, a history of recurring gene flow describes the divergence of *Pinus pungens* and *P. rigida* (Bolte et al. 2022), despite rare observations of hybrids in nature (Zobel 1969; Brown 2021) and reduced fertility in artificial crossing experiments (5-14% of seeds were filled; Critchfield 1963). While range-wide estimates of genetic diversity are now available for both *P. pungens* and *P. rigida* (Bolte et al. 2022), a more well-resolved estimate of hybridization and genetic differentiation between these two species is important for forest management planning. Fire suppression practices during the 20th century compromised population persistence of the fire-adapted *P. pungens* and *P. rigida* (Brose and Waldrop 2006). Because they are foundational species to a unique montane ecosystem of southern Appalachia, management efforts have been made to restore stands of these two species through prescribed burning and, specifically for *P. pungens*, seed banking for assisted migration (Jetton et al. 2015). Trait differences across populations of *P. rigida* were quantified in a common garden study (Ledig et al. 2015) and indicated three genetic groupings arranged latitudinally and two outlier populations along the northeast coastline.

The existence of genetic groupings and outlier populations may be related to geographically separated refugia during the last glacial maxima (LGM), traits that confer post-glacial expansion, and present-day population fragmentation and its influence on

intraspecific gene flow (Govindaraju 1989; Ledig et al. 2015). While these neutral processes have almost certainly played a role in genetic differentiation, local adaptation to differing niche optima may also explain some of the population level differences, as evidenced by the strong trait differentiation across populations but low genome-wide estimates of population structure. As the climate continues to warm, forest management plans will be most effective if populations are fully characterized, especially those at the southern, rear edge of a species distribution (Hampe and Petit 2005). These populations have higher risk of extirpation yet may carry adaptive alleles conferring tolerance to higher temperature and drought than more northerly distributed populations (Rehm et al. 2015; Issac-Renton et al. 2018) making them potentially well-fit candidates for assisted migration to projected warmer and drier climate regimes.

In this study, we estimated the extent of hybridization between *P. pungens* and *P. rigida* within three sympatric stands and compared genetic diversity estimates for each species at each stand using 6343 genome-wide single nucleotide polymorphisms (SNPs). We concluded that active hybridization in sympatric stands under current climate conditions is indeed rare, with only one advanced generational hybrid observed in our data. Many of the SNPs associated with high species-level genetic differentiation ($F_{ST} \geq 0.8$) at each stand were also shared across all three stands (~77%). Contrastingly, SNPs with low levels of genetic differentiation ($0.3 > F_{ST} > 0.1$) at each stand were not as commonly shared (~26%). This provides evidence that species level boundaries, at least in sympatry, involve the same genomic regions. We also present evidence of population structure within both species. From our estimates of genetic diversity, trailing edge

populations of both species may be experiencing inbreeding and/or population contraction. The only population that had similar values between observed and expected heterozygosity was that of *P. pungens* at Brown Mountain suggesting relative decreases in inbreeding. This stand had the highest levels of genetic diversity and its more central location within the geographical distribution suggests higher intraspecific gene flow may be occurring with other nearby populations. From these results we have gained a greater understanding of the strength of species boundaries between *P. pungens* and *P. rigida* in sympatric stands and provide population genetic information that can guide forest conservation and management planning.

Methods

Sampling of sympatric stands

Leaf tissue from *P. pungens* and *P. rigida* were collected from three forest stands along the Appalachian Mountains where sympatry occurs (Figure 3.1a). One of these stands was on Brown Mountain of Shenandoah National Park (coordinates: 38.30 N, -78.67 W), the most northern population we sampled (Figure 3.1b). This stand is part of a wilderness area with a mix of established trees and post-wildfire regenerating stands. The second stand was Laurel Falls (coordinates: 35.67 N, -83.59 W), a rear edge population within Great Smoky Mountains National Park of Tennessee. This stand represented the most southern and western sympatric stand sampled (Figure 3.1c). The third stand was at the junction of the Dragon Tooth and Appalachian Trail within Jefferson National Forest

(Figure 3.1d; coordinates: 37.37 N, -80.17 W). Our sampling scheme involved sampling all trees resembling either parental species that occurred within 20 meters of the marked trail and were perceivably safe to obtain (i.e., some trees growing on the sides of steep cliffs were not collected). The number of trees sampled ranged from 26 to 37 across sites (Figure 3.1). All samples of needle tissue were dried using silica beads, followed by cutting and lysing of 10 mg of tissue for DNA extraction.

DNA sequence data and SNP calling

Total genomic DNA was extracted from 205 trees using DNeasy Plant Kits (Qiagen) following the manufacturer's protocol and subsequently used in a reduced-representation workflow to produce DNA sequencing libraries using the procedures outlined in Parchman et al. (2012). We sized-selected DNA fragments from 350 to 450 bp in length using the PippinPrep quantitative gel electrophoresis unit (Sage Science, Beverley, MA) at the University of Texas Genome and Sequencing Analysis Center in Austin, TX. Fragments were then sequenced using the Illumina NovaSeq platform with S2 chemistry.

Contaminants (e.g., PhiX and *E. coli*) and Illumina sequencing oligos were then filtered from the sequencing data using bowtie_db2 (langmead12) and a pipeline of Perl and bash scripts (<http://github.com/ncgr/tapioca>). To demultiplex reads by sample, we corrected 1-2 bp errors in barcode sequences, removed restriction site-associated bases, and then matched each sampled tree to its corresponding DNA barcode sequence. This process was accomplished using a custom Perl script that ultimately produced individual fastq files of sequence data for each tree that was sampled.

We used the dDocent bioinformatics pipeline (Puritz et al. 2014) to generate a reference assembly and call single nucleotide polymorphisms (SNPs). The reference assembly was optimized using shell scripts and documentation within dDocent (cutoffs: individual = 6, coverage = 6; clustering similarity: -c 0.92) and cd-hit-est (Fu et al. 2012) for assembly. The initial SNP call produced 199,897 variant sites. These were further filtered using *vcftools* (Danecek et al. 2011), version 0.1.15, to retain only biallelic SNPs with sequencing data for at least 60% of the samples, minor allele frequency (MAF) > 0.02, summed depth across samples > 50 and < 5000, and alternate allele call quality ≥ 50 . Additionally, due to issues in genotype bias leading to mis-assembly of paralogous genomic regions, we reduced the probability of variant calling in these regions by only retaining biallelic SNPs and removing loci with abnormal heterozygosity ($F_{IS} > -0.5$; Hapke and Thiele 2016; Hohenlohe et al. 2013; McKinney et al. 2017; 2018). Sampled trees with excessive missing data ($\geq 50\%$) were removed from the data set leaving 194 trees in our analysis (*P. pungens*: $n = 97$, *P. rigida*: $n = 97$).

Genetic structure across species and population

To incorporate genotype uncertainty stemming from sequencing and alignment error, as well as low and variable sequencing depth across individuals and loci, we used a hierarchical Bayesian model (*ENTROPY*; Gompert et al. 2014; Shastry et al. 2021) to estimate genotype probabilities for each tree at each locus, infer number of populations (k), and estimate ancestry coefficients (q). This model is similar to that of *STRUCTURE* (Pritchard et al. 2000) but uses allele frequency priors and genotype likelihoods calculated

in *samtools* with linear discriminant analysis following *k-means* clustering for starting values of ancestry coefficients (q). Seven total models were assessed ($k = 2-8$) across 4 chains each based on 60,000 MCMC iterations with a burn-in of 10,000 and thinned to every 10th step. The best model was for $k = 2$, assessed with the deviance information criterion (DIC) value, and our *a priori* assumption given there are two species. Genotype probabilities and ancestry coefficients (q) were averaged across all chains and summarized DICs for each population are reported in Table 3.S1.

Genetic structure across the samples of *P. pungens* and *P. rigida* at each sympatric stand was further visualized using principal component analysis (PCA), following standardization routines detailed in Patterson et al. (2006). For PCA, we employed the `prcomp` function of the *stats* version 4.0.4 package in R version 3.6.2 (R Development Core Team 2021). We estimated genetic diversity of each population per species in terms of observed and expected heterozygosity using the same custom script employed in Bolte et al. (2022). Calculations of F_{ST} for species-level differentiation at each stand was performed in *hierfstat*, version 0.5-10, in R (Goudet 2005) using the 'varcomp.glob' function. Confidence intervals (95%) were calculated using the 'boot.vc' function with 1000 replications. Additionally, pairwise F_{ST} was estimated between populations within each species and was assessed for statistical significance through a permutation-based analysis ($n = 1000$ permutations of population identifiers across samples).

Genomic differentiation across species

To estimate species-level differentiation at each SNP and the amount of shared genetic differences across stands, we used the same three parsed (i.e., stand-specific) genetic data frames and methods (i.e., *hierfstat*) that were used for global estimates of F_{ST} as detailed in the previous section. The corresponding values of F_{ST} for each SNP were reduced to two categories for analysis, those with moderately low F_{ST} ($0.3 > F_{ST} > 0.1$) and those with exceptionally high F_{ST} (≥ 0.8). The threshold of 0.3 for the moderately low F_{ST} category was determined from the average species-level F_{ST} value estimated for each stand (0.29 - 0.30, Table 3.1). For simplicity, these categories will be referred to as low and high moving forward. Next, we counted the number of SNPs at low and high F_{ST} associated with species-level differentiation at each sympatric stand and compared how many SNPs in each of these categories were commonly shared across the sympatric stands. This provided a proxy for how random or uniform genetic differentiation was across species using our sampled sites as replicates.

Finally, RADtag sequences associated with each SNP ID in the category of high F_{ST} were mapped to the *Pinus taeda* L. annotated genome (version 2; Wegrzyn et al. 2014) using BLAST, version 2.5.0, to characterize the genomic distribution of variation in relation to coding and non-coding regions. We used a word size of 15 and penalized e-value scores by 5 for each open gap and 2 for each gap extension. E-values less than 10 were retained. We then filtered hits based on the best three e-value scores for each RADtag sequence using a custom python script (https://github.com/boltece/filter_blast). Scaffold identifiers from the *P. taeda* genome that associated with RADtag sequences were then matched

with gene attributes. We determined the distance of each hit on a *P. taeda* scaffold in relation to the gene annotations (i.e., attributes) using a custom python script. We kept the attribute that was closest to the RADtag read for summaries. Three additional categories were also made as done in Chapter 2 (Bolte 2022): RADtag sequences that hit within a gene, those $\leq 20\text{kbp}$ from a gene, and those $> 20\text{kbp}$ from a gene.

Results

Genetic structure across species and populations

The distribution of all samples across the three sympatric stands in PCA space show clear separation according to species along PC1, which explained 77.63% of the genetic variance in our 6,343 SNP data set (Figure 3.2b). The only tree with admixture was a *P. pungens* sample (11% assignment to *P. rigida* ancestry) from Dragon Tooth. This sample (PU_DT_22) is separated from the others along PC1 (Figure 3.2b, c). Field notes indicated that PU_DT_22 (coordinates: 37.366 N, -80.168 W, 809.244 meters elevation) was a young tree with no cones. Based on the amount of occupied PC space as well as estimates of observed (H_O) and expected (H_E) heterozygosity (Table 3.1), there is less genetic variation in *P. pungens* (H_O : 0.119 - 0.143) at each stand than the genetic variation associated with *P. rigida* (H_O : 0.132 - 0.154).

Genetic diversity estimates across the three populations of *P. rigida* had lower observed heterozygosity than expected. This was also the case for *P. pungens* at Dragon Tooth and Laurel Falls. While Brown Mountain had the highest genetic diversity estimates for both species across the three sites, the *P. pungens* population at Brown Mountain was the only sampled population that had higher observed heterozygosity than expected (Table 3.1). Independent PCA for each species were also used to visualize population genetic variation differences across sympatric stands (Figure 3.3a). Examination of sample distributions along PC1 and PC2 for both independent plots of *P. pungens* and *P. rigida* samples revealed that the Brown Mountain (BM) stand had the greatest genetic diversity as it occupied more PC space (see Figure 3.S1 as well for *P. pungens* populations in PC space with the hybrid individual removed). This aligned with the higher estimates of observed heterozygosity at Brown Mountain. Laurel Falls (LF), one of the most southern and western regions where the two species have distributional overlap, had the lowest genetic diversity according to estimates of observed heterozygosity. Trees were more genetically similar between Dragon Tooth (DT) and Laurel Falls for both species than those sampled at Brown Mountain (Figure 3.3b). The two most distant populations of each other, Brown Mountain and Laurel Falls, were the most dissimilar.

Genomic differentiation across species level boundaries

To observe species-level genomic differentiation we categorized SNPs into two categories: those with low F_{ST} ($0.3 > F_{ST} > 0.1$) and those with high F_{ST} (≥ 0.8) for each sampled stand. More SNPs had low F_{ST} compared to high F_{ST} (328 versus 162, respectively). Within each category, counts were similar across the stands, but the

proportion of shared SNPs across the stands in each F_{ST} category differed substantially. For SNPs in the category of low F_{ST} , 26.4 - 27.4% were commonly shared across all three sympatric stands. In contrast, 73.6 - 79.4% of the SNPs categorized as having high F_{ST} were commonly shared (Figure 3.4). We mapped shared RADseq contigs of both F_{ST} categories to the *P. taeda* draft genome, version 2, to further characterize genomic differentiation captured in our data. All contigs were successfully mapped. After filtering hits down to keep the best three scaffold IDs per contig, determined by the lowest e-values, 486 hits for the high F_{ST} and 1043 hits for low F_{ST} were retained for summaries. The e-values across the filtered hits ranged from 1.47e-38 to 8.80. Of the 328 shared SNPs with low F_{ST} , 55.2% were > 20kbp to a gene, 30.2% were \leq 20kbp to a gene, and 14.6% were within a gene (Table 3.2). Of the 162 shared SNPs with high F_{ST} , 62.3% were > 20kbp to a gene, 23.5% were \leq 20kbp to a gene, and 14.2% were within a gene (Table 3.2).

Of the 486 blast hits for the high F_{ST} category, 150 matched with scaffolds that had annotated gene attributes. Among those, 68.0% were over 20kbp from a gene, 13.3% were close to genes, and 18.7% were in genes. Hits within genes were mostly intronic but two matched coding DNA sequences (CDS; Table 3.S2). Related EggNOG descriptions and GO terms of *P. taeda* attributes for each contig-scaffold ID hit are summarized in Table 3.S2 - Table 3.S4. Seven of the 150 *P. taeda* attributes listed in these tables did not have EggNOG descriptions, and 45 did not have EggNOG GO terms. Among those with descriptions, some terms often cited in literature had multiple occurrences: 4 zinc finger, 4 retrotransposon, 4 dnaJ chaperone, and 3 Fbox proteins.

Discussion

It is well-established that intraspecific and interspecific gene flow dynamics affect the rate of speciation, population structure, and overall measures genetic and biological diversity (Savolainen et al. 2007; Petit and Excoffier 2009; Wang and Bradburd 2014), but the resulting directionality and intensity of these effects are dependent on many factors such as life history traits, environmental complexity, and genetic architecture (Abbott 2017; Bolte and Eckert 2020; Kulmuni et al. 2020; Wu et al. 2022). As case studies accumulate, patterns will emerge to help us better understand the development of RI in conifers. In this study, we added to a growing base of speciation literature for *P. pungens* and *P. rigida* by examining the extent of hybridization, genetic diversity, and genetic differentiation across three sympatric stands along the Appalachian Mountains. We present evidence of species boundaries being strongly maintained while in sympatry and explain how ecological and reproductive character displacement were potentially driven through reinforcement (i.e., reduced hybrid fitness and selection towards diverged trait optima). Only 1 out of 194 sampled trees had admixture and the admixture that was present was in low proportion (11% *P. rigida* ancestry in a sampled tree of *P. pungens* at Dragon Tooth). This lack of hybridization observed from genetic data is consistent with the morphological observations of sympatric stands (Zobel 1969; Brown 2021).

Across the three sympatric stands more SNPs with high F_{ST} were shared (~76%) compared to SNPs with low F_{ST} (~27% in common). This suggests species level genetic differences are driven by the same genomic regions across sites. For the ~24% of SNPs

that were not shared across the three stands but had high F_{ST} , other evolutionary forces such as genetic drift or local adaptation may have driven differentiation. Within species, population structure was observed across the three stands. Greater genetic diversity was estimated for populations at Brown Mountain compared to Laurel Falls and Dragon Tooth. This could be due to its more central (i.e., core) location in relation to geographical distributions. Laurel Falls, a rear edge population, had the least genetic diversity, a pattern found in other rear edge populations of species due to migratory and adaptational lags under a rapidly changing climate (Bridle and Vines 2007; Zhu, Woodall, and Clark 2011).

Character displacement through reinforcement may explain co-existence

Past work on *P. pungens* and *P. rigida* reported differences in reproductive phenological schedules such as timing of pollen release to be partially responsible for RI (Zobel 1969). Trait differences related to seed size (i.e., dispersal capability), rate of seedling establishment, and serotiny (i.e., differential adaptations to fire frequency and intensity), needle morphology, and soil mycorrhizae associations (Zobel 1969) may also contribute to RI. Some of these traits have been defined as quantitative (Caignard et al. 2019), highly heritable (e.g., seed mass; Harper et al. 1970), polygenic and widely distributed across the genome (e.g., growth; Lind et al. 2018), or associated with only a few larger effect loci (e.g., serotiny; Parchman et al. 2012). While ecological divergence has been linked to strengthening of RI through the development of both pre- and postzygotic isolating mechanisms (Baack et al. 2015), it is also possible that divergent selection acting on pre-mating traits (e.g., pollen release timing) tandemly drove divergence in ecological traits (Widmer, Lexer, and Cozzolino 2009). Disentangling the epistatic or pleiotropic

interactions between reproductive phenological traits and ecological traits is exceptionally challenging when genomic resources are as limited as they are for pines (Lind et al. 2018). Nonetheless, ecological divergence plays a role in the maintenance of species boundaries for *P. pungens* and *P. rigida* (Bolte et al. 2022) and the complex genomic architectures of diverged traits may explain low crossability observed in artificial crossing experiments of Critchfield (1963). Indeed, there is a link between ecological divergence and intrinsic barriers to gene flow in other plant taxa (Widmer, Lexer, and Cozzolino 2009). In light of the demographic inference, species distribution modeling, and association analyses reported in Bolte et al. (2022), which detailed divergence as occurring with gene flow, consistent overlap in species distributions over the past 120,000 years, and the importance of seasonality to genetic differentiation, ecological character displacement through reinforcement (Levin 2006) likely drove RI between these two species when in sympatry.

Deciphering between the relative contributions of allopatry and sympatry to the evolution of RI is challenging when working with conifers due to long generation times, long distance pollen dispersal, and a limited fossil record (Betancourt et al. 1991), but demographic inferences from genetic data and historical species distribution modeling have provided some indication of when, where, and how species and populations have diverged (Richards et al. 2007). For *P. pungens* and *P. rigida*, initial divergence aligns in timing with the start of the Quaternary period (~2.7 mya; Bolte et al. 2022). In studies of other plant taxa, the extreme climatic oscillations of the Quaternary period appear to have caused changes in effective population sizes (i.e., contraction - expansion cycles), gene

flow dynamics, and adaptations to seasonality (Soltis et al. 2006; Jackson and Overpeck 2000). To further elucidate the relative contribution of demographic and adaptive processes to the development of RI in *P. pungens* and *P. rigida*, we recommend three future study designs.

First, we need to determine if trees within allopatrically distributed stands are as prezygotically isolated and ecologically diverged as are trees in sympatry by following an experimental design such as the one first proposed in Lack (1947). This design compares traits values between sympatric and allopatric stands that have similar abiotic and biotic factors (i.e., eliminate variation due to environment) so genetically based differences in trait values can be observed (Calabrese and Pfenning 2020). A previous study on hybridization across four hard pines of eastern North America (*P. rigida*, *P. serotina*, *P. taeda*, and *P. echinata*) found that species integrities were upheld in sympatry, but hybrids based on intermediate trait values were observed in allopatric or parapatric populations (Smouse and Saylor 1973). To date pollen release timing for *P. pungens* and *P. rigida* has only been measured and compared within sympatric stands (Zobel 1969), but the frequency of cone serotiny, for example, does vary between sympatric and allopatric stands of *P. rigida*. In sympatric stands, *P. rigida* has solely non-serotinous cones and *P. pungens* has solely serotinous cones. However, in the northeastern coastal region of the *P. rigida* geographic distribution, far from any extant stand of *P. pungens*, there are two outlier populations (Pine Plains in New Jersey and Acadia National Park in Maine) that exhibit serotiny, faster seedling establishment, and shorter stature than more southern and western populations (Ledig et al. 2015). These two populations are suspected to have

resided in refugia just south of the last glacial extent on what is now the continental shelf. At present, a cline of mixed serotiny is observable along 300 km transects from these outlier populations which has been attributed to spatially varying selection pressures at migration-selection equilibrium (Ledig and Fryer 1972). In contrast, given the conclusions from the common garden study of Ledig et al. (2015) and those presented here, we reinterpret the cline for serotiny as a result of secondary contact between two refugial populations. Indeed, clinal trends observed in regions where two refugial populations have reconnected (a suture zone) have been observed in *P. ponderosa* (Johansen and Latta 2002). If the southern refugia was shared between *P. pungens* and *P. rigida* or at least proximal enough to have recurring contact over the course of climate oscillations (mixing-isolation-mixing model; He et al. 2019), it could explain the promotion of ecological and reproductive character displacement. If this dynamic was absent in the northeastern refugia, it could explain less diverged trait values between the two outlier populations of *P. rigida* to *P. pungens*.

The other study designs we suggest involve comparisons between *P. pungens* and outlier *P. rigida* populations. A simple first step would involve the same methods we employed here and compare the number of high F_{ST} SNPs shared between outlier *P. rigida* and *P. pungens* populations to the counts we reported in this study. Fewer counts of shared SNPs may be an indication of less evolved RI. Another route for investigation could include a crossing experiment to provide a test for the relative contributions of gene flow to intrinsic postzygotic barriers. This study may find higher hybrid fertility than the 5-13% hybrid seed fill reported in Critchfield (1963). Coupling this effort with an assessment of

hybrid fitness from experimental or common garden approaches may elucidate the relative contribution of extrinsic postzygotic barriers to RI through genotype-environment interactions. Indeed, the amount and type of introgressed variants in a population are determinants of where a population can establish, persist, and contribute to adaptation in parental taxa (Hamilton and Miller 2016; Janes and Hamilton 2017; Menon et al. 2018).

Implications for forest conservation and management planning

Climate affects species distributions and thus genetic diversity. The three stands we intensively sampled had genetic diversity estimates that correlated with latitude and size of stand. The low genetic diversity at Laurels Fall, a trailing edge population for both *P. pungens* and *P. rigida*, fits theory and empirical evidence echoed in a host of literature (e.g. Lawton 1993; Vucetich and Waite 2003; Bridle and Vines 2007; Zhu et al. 2012) such as trees have difficulty tracking niche optima so populations at trailing edges contract and genetic diversity reduces. While low genetic diversity limits adaptation potential, trailing edge populations may have specific adaptations that confer population persistence at higher latitudes as climate warms (Hampe and Petit 2005; Jump and Peñuelas 2005; Rehm et al. 2015), making them prime candidates for assisted migration (Aitken et al. 2008). Brown Mountain, on the other hand, has more centrally located populations and is a managed wilderness area, unlike the other two stands in our study, and had the highest genetic diversity estimates. It is hard to discern though whether place within the geographic distribution, management strategies, fire activity, or a combination of all three has promoted greater genetic diversity.

Conservation and management strategies are often resource intensive making it important to comprehensively consider available information and weigh the benefits and risks associated with management options such as prescribed burning, assisted migration, and facilitated introgression. Past studies have provided information to help guide prescribed burning practices to restore populations of both species (Welch and Waldrop 2001). Information related to habitat fragmentation and dwindling population sizes for *P. pungens* helped initiate seed banking for assisted migration (Jetton et al. 2015). Here, we provide valuable information related to genetic diversity for trailing edge populations and hybridization potential within naturally shared stands of *P. pungens* and *P. rigida*. We also present important considerations and directions for future research. If hybrids are more often found in allopatric stands or if genomic compatibility is indeed higher between outlier populations of *P. rigida* to those of *P. pungens*, then population seed source determines outcomes of management plans in terms of hybridization. Avenues for facilitated introgression may even arise as possible management strategies (e.g., American and Chinese chestnuts for disease resistance; Newhouse and Powell 2020). We provided in this study a foundational base of characterized genomic differentiation by mapping our RADtag sequences to the *P. taeda* genome, to which future genomic research can use as a reference. More importantly, we presented an efficient way to compare species level differentiation across populations that provides insight into the development of RI and hybridization potential. Populations with lower levels of genetic differentiation may imply less RI and higher hybridization potential.

Table 3.1 Genetic diversity estimates expected heterozygosity (H_E) and observed heterozygosity (H_O), for each species at each sympatric stand. Estimates of genetic differentiation across species (F_{ST}) at each sympatric stand are also provided.

	Brown Mountain	Dragon Tooth	Laurel Falls
<i>P. pungens</i>			
H_E (SD)	0.139 (0.157)	0.138 (0.157)	0.136 (0.159)
H_O (SD)	0.143 (0.189)	0.129 (0.157)	0.119 (0.161)
<i>P. rigida</i>			
H_E (SD)	0.171 (0.151)	0.169 (0.152)	0.168 (0.154)
H_O (SD)	0.154 (0.165)	0.135 (0.148)	0.132 (0.145)
F_{ST} (95% CI)	0.290 (0.276- 0.304)	0.293 (0.279 - 0.308)	0.299 (0.285 - 0.313)

Table 3.2 Counts of RADtag sequences (i.e., contigs) and how they mapped to the *P. taeda* genome for each F_{ST} and distance category.

F_{ST} category	Contigs mapped	>20kbp from a gene	≤ 20kbp from a gene	Within a gene
Shared low	328	181	99	48
Shared high	162	101	38	23

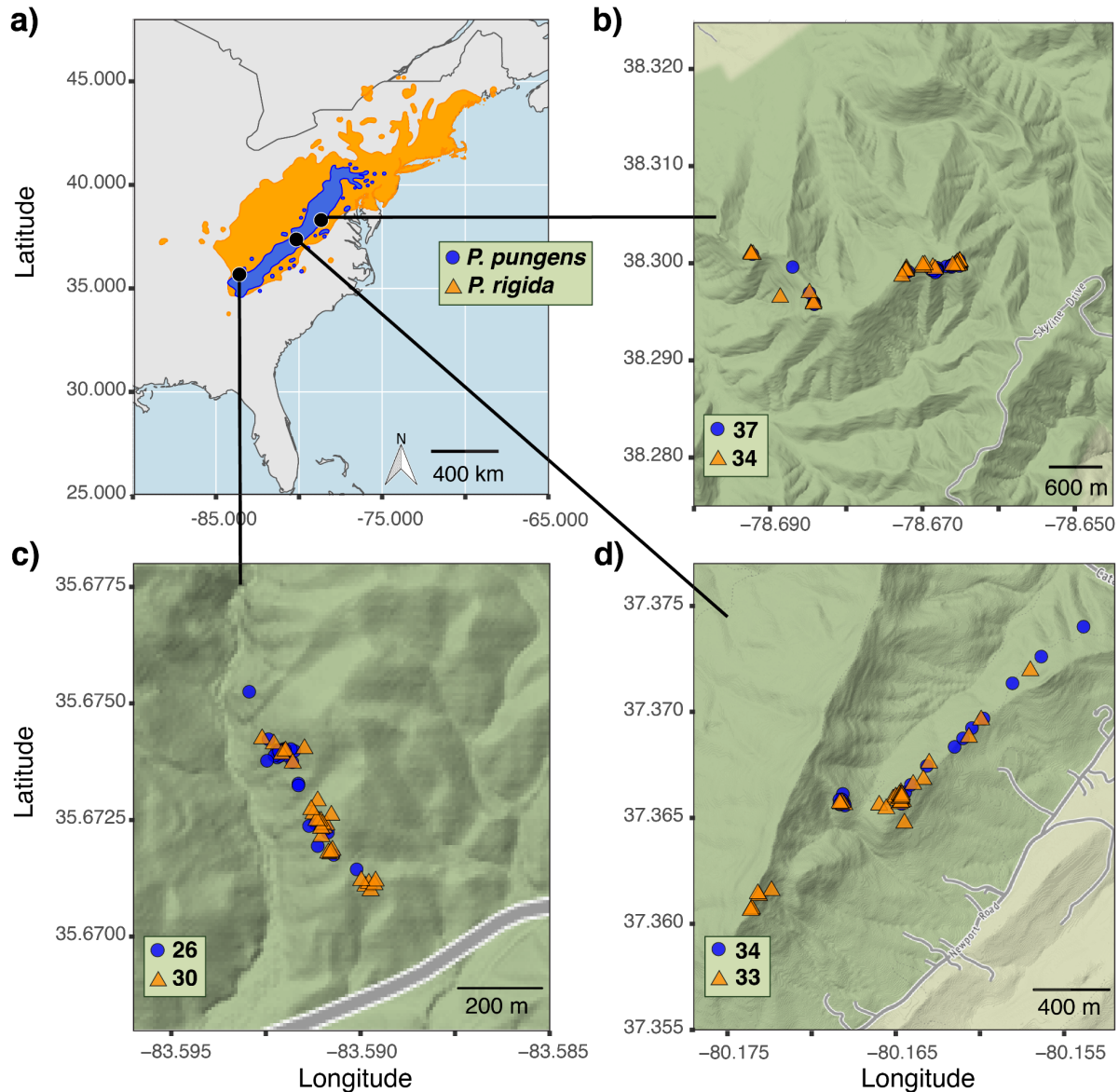


Figure 3.1 Distribution of sampled sympatric populations a) in relation to each other geographically and across the described geographic range of each species in Little (1975). The trees sampled within each population are shown in for b) Brown Mountain of Shenandoah National Park, c) Laurel Falls of Great Smoky Mountains National Park, and d) Dragon Tooth of Jefferson National Forest. Blue circles indicate samples morphologically identified as *P. pungens*. Orange triangles are samples indicative of *P. rigida*.

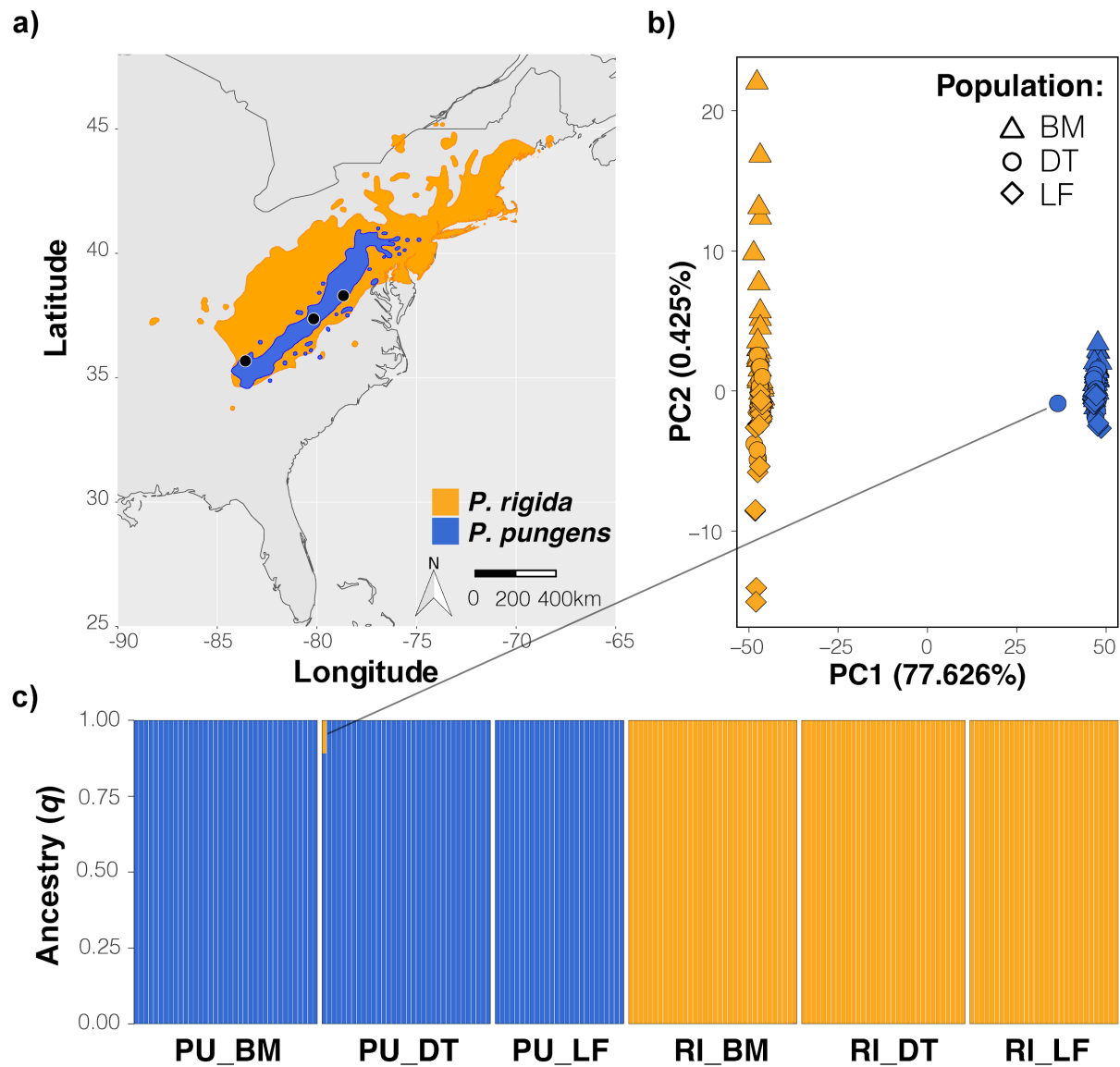


Figure 3.2 Species level genetic differentiation for 194 sampled trees across three sympatric stands (map, panel a). Principal component analysis results based on multilocus genotypes across 6343 SNPs are provided in panel b. Inference of structure from ($k = 2$) is provided in panel c.

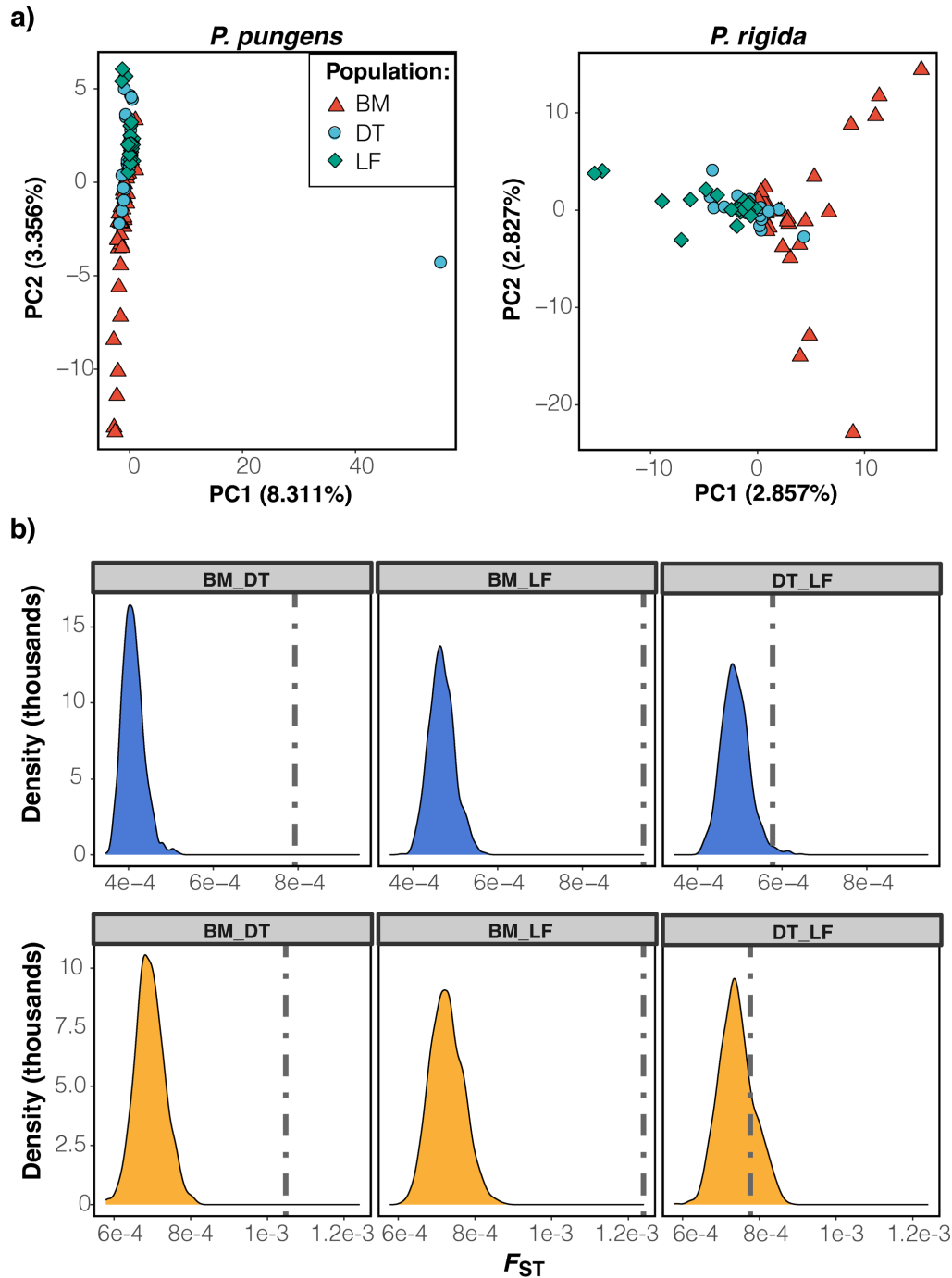


Figure 3.3 Population-level genetic differentiation across 6343 SNPs illustrated in a) principal component analysis (PCA) for *P. pungens* and *P. rigida* sampled trees, and b) Pairwise population level comparisons for *P. pungens* (top row, blue) and *P. rigida* (bottom row, orange) where BM is Brown Mountain, DT is Dragon Tooth, and LF is Laurel Falls. Dashed line is the realized pairwise F_{ST} in each plot. Distributions are permutations of F_{ST} based on random selection of individuals. If the dashed line is to the right of the distribution, then populations are more different than expected by random chance.

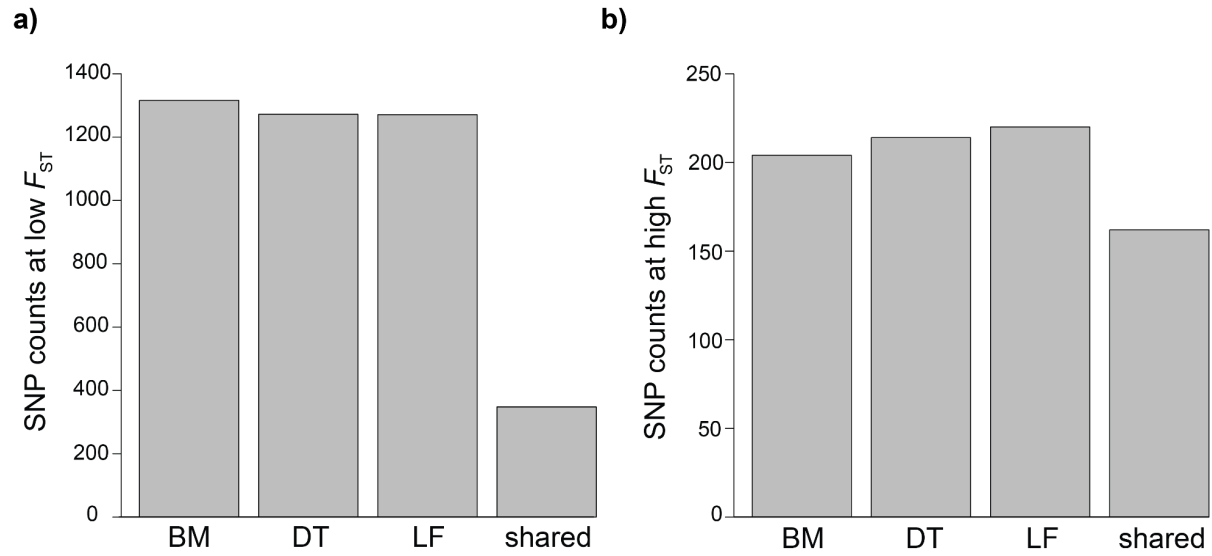


Figure 3.4 Counts of SNPs based on two categories of F_{ST} , low ($0.3 > F_{ST} > 0.1$; panel a) and high ($F_{ST} \geq 0.8$; panel b) for Brown Mountain (BM), Dragon Tooth (DT), and Laurel Falls (LF). The last bar in each plot represents the number of SNP IDs (dDocent contigs) that were shared between BM, DT, and LF.

Literature Cited

- Abbott R., Albach D., Ansell S., Arntzen J.W., Baird S.J.E., et al. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, 26, 229–246.
- Abbott, R. J. (2017). Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. *Journal of Systematics and Evolution*, 55(4), 238–258. doi: 10.1111/JSE.12267
- Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T., & Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*, 1(1), 95–111. <https://doi.org/10.1111/J.1752-4571.2007.00013.X>
- Baack, E., Melo, M. C., Rieseberg, L. H., & Ortiz-Barrientos, D. (2015). The origins of reproductive isolation in plants. *New Phytologist*, 207(4), 968–984. doi: 10.1111/NPH.13424
- Beans, C. M. (2014). The case for character displacement in plants. *Ecology and Evolution*, 4(6), 862–875. doi: 10.1002/ece3.978
- Betancourt, J. L., Schuster, W. S., Mitton, J. B., & Anderson, R. S. (1991). Fossil and genetic history of a pinyon pine (*Pinus edulis*) isolate. *Ecology*, 72(5), 1685–1697. doi: 10.2307/1940968
- Bolte, C. E., & Eckert, A. J. (2020). Determining the when, where and how of conifer speciation: a challenge arising from the study 'Evolutionary history of a relict conifer *Pseudotsuga chienii*.' *Annals of Botany*, 125(1), v–vii. doi: 10.1093/AOB/MCZ201
- Bolte, C. E., Faske, T. M., Friedline, C. J., & Eckert, A. J. (2022). Divergence amid recurring gene flow: complex demographic processes during speciation are the growing expectation for forest trees. *bioRxiv*.
- Bridle, J. R., & Vines, T. H. (2007). Limits to evolution at range margins: when and why does adaptation fail? *Trends in Ecology and Evolution*, 22(3), 140–147. doi: 10.1016/j.tree.2006.11.002
- Brose, P. H., & Waldrop, T. A. (2006). Fire and the origin of Table Mountain pine pitch pine communities in the southern Appalachian Mountains, USA. *Canadian Journal of Forest Research*, 36(3), 710–718. <https://doi.org/10.1139/x05-281>
- Brown, A. L. (2021). Phenotypic characterization of Table Mountain (*Pinus pungens*) and pitch pine (*Pinus rigida*) hybrids along an elevational gradient in the Blue Ridge Mountains , Virginia. Virginia Commonwealth University.

- Brown, W. L., Jr., Wilson, E. O. (1956). Character displacement. *Systematic Zoology*, 5, 49–64.
- Caignard, T., Delzon, S., Bodénès, C., Dencausse, B., & Kremer, A. (2019). Heritability and genetic architecture of reproduction-related traits in a temperate oak species. *Tree Genetics and Genomes*, 15(1). doi: 10.1007/s11295-018-1309-2
- Calabrese, G. M., & Pfennig, K. S. (2020). Reinforcement and the Proliferation of Species. *Journal of Heredity*, 111(1), 138–146. doi: 10.1093/jhered/esz073
- Critchfield, W. B. (1963). Hybridization of the southern pines in California. Pages 40-48 in *Southern Forest Tree Improvement Conference*. Publ. 22.
- Darwin, C. (1859). The origin of species by means of natural selection.
- Gompert, Z., Lucas, L. K., Buerkle, C. A., Forister, M. L., Fordyce, J. A., & Nice, C. C. (2014). Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology*, 23(18), 4555–4573. doi: 10.1111/mec.12811
- Gompert, Z., Mandeville, E. G., & Buerkle, C. A. (2017). Analysis of Population Genomic Data from Hybrid Zones. *Annual Review of Ecology, Evolution, and Systematics*, 48, 207–229. doi: 10.1146/annurev-ecolsys-110316-022652
- Goudet, J. (2005). hierfstat, a package for r to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, 5(1), 184–186. doi: 10.1111/J.1471-8286.2004.00828.X
- Govindaraju, D. R. (1989). Estimates of gene flow in forest trees. *Biological Journal of the Linnean Society*, 37(4), 345–357. doi: 10.1111/j.1095-8312.1989.tb01910.x
- Grant P.R., Grant B.R. (2014). Evolutionary biology: speciation undone. *Nature*, 507, 178–179.
- Hamilton, J. A., & Miller, J. M. (2016). Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conservation Biology*, 30(1), 33–41. doi: 10.1111/cobi.12574
- Hampe, A., & Petit, R. J. (2005). Conserving biodiversity under climate change: The rear edge matters. *Ecology Letters*, 8(5), 461–467. doi: 10.1111/j.1461-0248.2005.00739.x
- Hapke, A., & Thiele, D. (2016). GIBPSs: a toolkit for fast and accurate analyses of genotyping-by-sequencing data without a reference genome. *Molecular Ecology Resources*, 16(4), 979–990. doi: 10.1111/1755-0998.12510
- Hohenlohe, P. A., Day, M. D., Amish, S. J., Miller, M. R., Kamps-Hughes, N., Boyer, M. C., ... Luikart, G. (2013). Genomic patterns of introgression in rainbow and

- westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, 22(11), 3002–3013. doi: 10.1111/mec.12239
- Isaac-Renton, M., Montwé, D., Hamann, A., Spiecker, H., Cherubini, P., & Treydte, K. (2018). Northern forest tree populations are physiologically maladapted to drought. *Nature Communications*, 9(1), 1–9. doi: 10.1038/s41467-018-07701-0
- Jackson, S. T., & Overpeck, J. T. (2000). *Responses of Plant Populations and Communities to Environmental Changes of the Late Quaternary*. 26(4), 194–220.
- Janes, J. K., & Hamilton, J. A. (2017, July 4). Mixing it up: The role of hybridization in forest management and conservation under climate change. *Forests*, Vol. 8. MDPI AG. doi: 10.3390/f8070237
- Jetton, R. M., Crane, B. S., Whittier, W. A., & Dvorak, W. S. (n.d.). *Genetic Resource Conservation of Table Mountain Pine (Pinus pungens) in the Central and Southern Appalachian Mountains*.
- Johansen, & Latta, R. G. (2003). Mitochondrial haplotype distribution, seed dispersal and patterns of postglacial expansion of ponderosa pine. *Molecular Ecology*, 12(1), 293–298. <https://doi.org/10.1046/j.1365-294X.2003.01723.x>
- Jump, A. S., & Peñuelas, J. (2005). Running to stand still: Adaptation and the response of plants to rapid climate change. *Ecology Letters*, 8(9), 1010–1020. doi: 10.1111/j.1461-0248.2005.00796.x
- Kulmuni, J., Butlin, R. K., Lucek, K., Savolainen, V., & Westram, A. M. (2020). Towards the completion of speciation: The evolution of reproductive isolation beyond the first barriers: Progress towards complete speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1806), 1–9. doi: 10.1098/rstb.2019.0528
- Lack, D. (1947). *Darwin's finches*. Cambridge University Press, Cambridge.
- Lawton, J. H. (1993). Range, population abundance and conservation. *Trends in Ecology and Evolution*, 8(11), 409–413. doi: 10.1016/0169-5347(93)90043-O
- Ledig, F. T., & Fryer, J. H. (1972). A pocket of variability in *Pinus rigida*. *Evolution*, 26(2), 259-266.
- Ledig, F. T., Smouse, P. E., & Hom, J. L. (2015). Postglacial migration and adaptation for dispersal in pitch pine (Pinaceae). *American Journal of Botany*, 102(12), 2074–2091. doi: 10.3732/AJB.1500009
- Levin, D. A. (2006). Flowering phenology in relation to adaptive radiation. *Systematic Botany*, 31(2), 239-246.

- Levensen, N. D., Tiffin, P., & Olson, M. S. (2012). Pleistocene speciation in the genus *populus* (salicaceae). *Systematic Biology*, 61(3), 401–412. doi: 10.1093/sysbio/syr120
- Li, L., Abbott, R. J., Liu, B., Sun, Y., Li, L., Zou, J., ... Liu, J. (2013). Pliocene intraspecific divergence and Plio-Pleistocene range expansions within *Picea likiangensis* (Lijiang spruce), a dominant forest tree of the Qinghai-Tibet Plateau. *Molecular Ecology*, 22(20), 5237–5255. doi: 10.1111/MEC.12466
- Linan, A. G., Lowry, P. P., Miller, A. J., Schatz, G. E., Sevathian, J. C., & Edwards, C. E. (2021). RAD-sequencing reveals patterns of diversification and hybridization, and the accumulation of reproductive isolation in a clade of partially sympatric, tropical island trees. *Molecular Ecology*, 30(18), 4520–4537. doi: 10.1111/mec.15736
- Lind, B. M., Menon, M., Bolte, C. E., Faske, T. M., & Eckert, A. J. (2018). The genomics of local adaptation in trees: are we out of the woods yet? *Tree Genetics and Genomes*, 14(2). doi: 10.1007/s11295-017-1224-y
- Little, E. L. (1975). *Rare and local conifers in the United States* (No. 19). US Department of Agriculture, Forest Service.
- Lowry, D. B., Modliszewski, J. L., Wright, K. M., Wu, C. A., & Willis, J. H. (2008). The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1506), 3009–3021.
- McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, 17(4), 656–669. doi: 10.1111/1755-0998.12613
- McKinney, G. J., Waples, R. K., Pascal, C. E., Seeb, L. W., & Seeb, J. E. (2018). Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. *Molecular Ecology Resources*, 18(3), 570–579. doi: 10.1111/1755-0998.12763
- Newhouse, A. E., & Powell, W. A. (2021). Intentional introgression of a blight tolerance transgene to rescue the remnant population of American chestnut. *Conservation Science and Practice*, 3(4), 1–8. doi: 10.1111/csp2.348
- Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C. A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, 21(12), 2991–3005. doi: 10.1111/j.1365-294X.2012.05513.x
- Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12), 2074–2093. <https://doi.org/10.1371/journal.pgen.0020190>

- Petit, R. J., & Excoffier, L. (2009). Gene flow and species delimitation. *Trends in Ecology and Evolution*, 24(7), 386–393. doi: 10.1016/j.tree.2009.02.011
- Pfennig, D. W., & Pfennig, K. S. (2010). Character displacement and the origins of diversity. *American Naturalist*, 176(SUPPL. 1). doi: 10.1086/657056
- Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, 2014(1). doi: 10.7717/peerj.431
- Rehm, E. M., Olivas, P., Stroud, J., & Feeley, K. J. (2015). Losing your edge: Climate change and the conservation value of range-edge populations. *Ecology and Evolution*, 5(19), 4315–4326. doi: 10.1002/ece3.1645
- Richards, C. L., Carstens, B. C., & Lacey Knowles, L. (2007, November). Distribution modelling and statistical phylogeography: An integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography*, Vol. 34, pp. 1833–1845. doi: 10.1111/j.1365-2699.2007.01814.x
- Rieseberg LH, Wendel JF. 1993. Introgression and its consequences in plants. In: Harrison R, ed. Hybrid zones and the evolutionary process. New York, NY, USA: Oxford University Press, 70–109.
- Savolainen, O., Pyhäjärvi, T., & Knürr, T. (2007). Gene Flow and Local Adaptation in Trees. *Annual Review of Ecology, Evolution, and Systematics*. doi: 10.1146/annurev.ecolsys.38.091206.095646
- Schluter, D. (2001). Ecology and the origin of species. *Trends in Ecology & Evolution*, 16(7), 372–380.
- Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., ... Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, 15(3), 176–192. doi: 10.1038/nrg3644
- Shastri, V., Adams, P. E., Lindtke, D., Mandeville, E. G., Parchman, T. L., Gompert, Z., & Buerkle, C. A. (2021). Model-based genotype and ancestry estimation for potential hybrids with mixed-ploidy. *Molecular Ecology Resources*, 21(5), 1434–1451. <https://doi.org/10.1111/1755-0998.13330>
- Smouse, P. E., & Saylor, L. C. (1973). Studies of the *Pinus rigida*-*Serotina* Complex II. Natural Hybridization Among the *Pinus rigida*-*Serotina* Complex, *P. taeda* and *P. echinata*. *Annals of the Missouri Botanical Garden*, Vol. 60, p. 192. St. Louis, Missouri Botanical Garden Press, 1914-. doi: 10.2307/2395085
- Soltis, D. E., Morris, A. B., McLachlan, J. S., Manos, P. S., & Soltis, P. S. (2006). Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, 15(14), 4261–4293. doi: 10.1111/j.1365-294X.2006.03061.x

- Stelkens, R., & Seehausen, O. (2009). Genetic distance between species predicts novel trait expression in their hybrids. *Evolution*, 63(4), 884–897. doi: 10.1111/j.1558-5646.2008.00599.x
- Vallejo-Marín, M., & Hiscock, S. J. (2016). Hybridization and hybrid speciation under global change. *The New Phytologist*, 211(4), 1170–1187. doi: 10.1111/nph.14004
- Vucetich, J. A., & Waite, T. A. (2003). Spatial patterns of demography and genetic processes across the species' range: Null hypotheses for landscape conservation genetics. *Conservation Genetics*, 4(5), 639–645. doi: 10.1023/A:1025671831349
- Wang, I. J., & Bradburd, G. S. (2014). Isolation by environment. *Molecular Ecology*, 23(23), 5649–5662. doi: 10.1111/mec.12938
- Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., Vasquez-Gross, H. A., ... Neale, D. B. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, 196(3), 891–909. doi: 10.1534/genetics.113.159996
- Welch, N. T., & Waldrop, T. A. (2001). Restoring table mountain pine (*Pinus pungens* Lamb.) communities with prescribed fire: An overview of current research. *Castanea*, 66(1), 42–49.
- Widmer, A., Lexer, C., & Cozzolino, S. (2009). Evolution of reproductive isolation in plants. *Heredity*, 102(1), 31–38. doi: 10.1038/HDY.2008.69
- Wu, S., Wang, Y., Wang, Z., Shrestha, N., & Liu, J. (2022). Species divergence with gene flow and hybrid speciation on the Qinghai–Tibet Plateau. *New Phytologist*, 392–404. doi: 10.1111/nph.17956
- Yang, Y. X., Zhi, L. Q., Jia, Y., Zhong, Q. Y., Liu, Z. L., Yue, M., & Li, Z. H. (2020). Nucleotide diversity and demographic history of *Pinus bungeana*, an endangered conifer species endemic in China. *Journal of Systematics and Evolution*, 58(3), 282–294. doi: 10.1111/jse.12546
- Zhu, K., Woodall, C. W., & Clark, J. S. (2012). Failure to migrate: Lack of tree range expansion in response to climate change. *Global Change Biology*, 18(3), 1042–1052. doi: 10.1111/j.1365-2486.2011.02571.x
- Zobel, D. B. (1969). Factors Affecting the Distribution of *Pinus pungens*, an Appalachian Endemic. *Ecological Monographs*, 39(3), 303–333. doi: 10.2307/1948548

Appendix 3

Table 3.S1 DIC scores from analysis of structure across four replicate runs (# of chains) of each cluster assignment (k).

k	# of chains	mean	min	max
2	4	7759042.83	7662333.02	7907748.68
3	4	738152283.2	31634296.32	1142885883
4	4	770237112.4	7594790.78	2452668406
5	4	449283354.6	56484031.51	827382799.8
6	4	1040226154	56605947.68	2501654986
7	4	156167843.4	23235437.52	240757035.8
8	4	571312240.7	272083887.3	1095240329

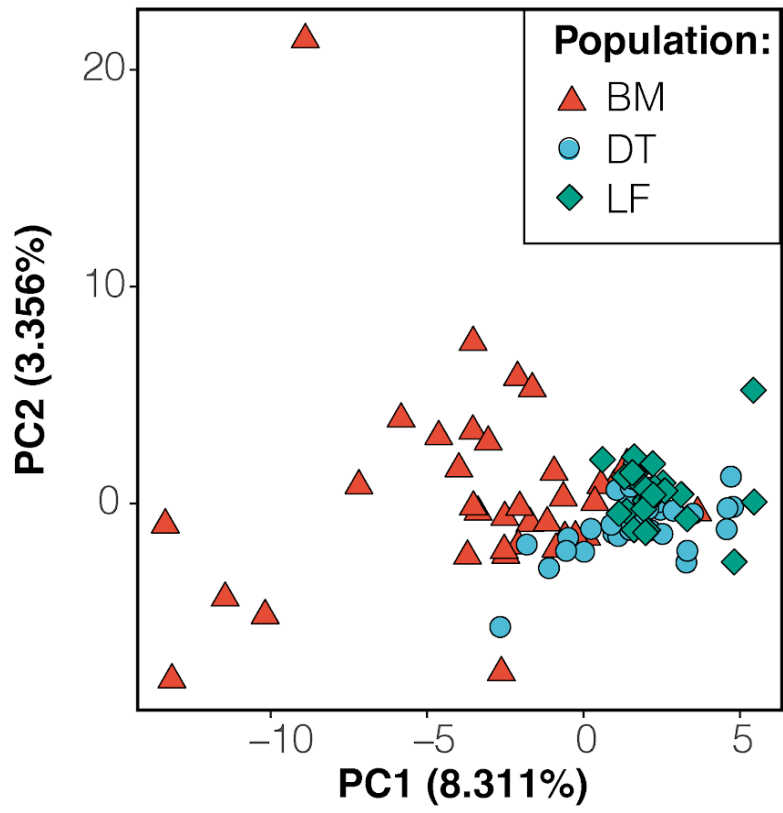


Figure 3.S2 PCA of *P. pungens* populations with the hybrid sample removed.

Table 3.S2 Summary of BLAST results for RADtag_ID matches to the *Pinus taeda* (PITA) genome that were within a gene. Annotations based on EggNOG descriptions and GO terms were sourced directly from the annotation file that accompanies the genome on treegenomesdb.org.

RADtag_ID	seqid	Blastn e-value	Query Sequence	type	EggNOG Description	EggNOG.GO.Biological
Contig_27087	super3404	8.01E-37	PITA_02305	intron	Pfam:DUF1630	
Contig_35847	super1143	3.02E-25	PITA_02984	intron	Phenazine biosynthesis-like protein	GO:0008152-metabolic process(L=1),GO:0050896-response to stimulus(L=1)
Contig_43608	super3003	1.00E-06	PITA_04633	intron	inositol transporter	GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0044699-single-organism process(L=1),GO:0051179-localization(L=1)
Contig_52087	super513	1.47E-38	PITA_05034	intron	Leo1-like protein	GO:0008152-metabolic process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_72286	scaffold8269	3.02E-25	PITA_05457	intron	inositol hexakisphosphate and diphosphoinositol-pentakisphosphate	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_38481	scaffold61163	1.46E-27	PITA_10355	intron	ribosomal protein	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)

Contig_71144	scaffold220265	1.47E-38	PITA_15238	intron	tHO complex	GO:000003-reproduction(L=1),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0040007-growth(L=1),GO:0040011-locomotion(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1),GO:0051704-multi-organism process(L=1),GO:0065007-biological regulation(L=1)
Contig_66403	super22	1.82E-19	PITA_17774	intron	glycine-rich protein	
Contig_38878	scaffold181463	3.40E-21	PITA_21711	intron	DUF4206	GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1)
Contig_45936	super1179	1.47E-38	PITA_28156	intron	Mitotic checkpoint protein	GO:0007094-mitotic spindle assembly checkpoint(L=10),GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0065007-biological regulation(L=1)
Contig_53151	super4483	3.76E-17	PITA_28781	intron	DNA ligase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1)
Contig_76277	scaffold67599	4.36E-35	PITA_31323	intron	RNA helicase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0050896-response to stimulus(L=1),GO:0071840-cellular component organization or biogenesis(L=1)

Contig_54496	scaffold187632	4.33E-24	PITA_32602	intron	interconversion of serine and glycine (By similarity)	GO:0002376-immune system process(L=1),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1)
Contig_67171	scaffold69450	2.38E-33	PITA_33597	CDS	acetolactate synthase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0050896-response to stimulus(L=1)
Contig_29660	scaffold226780	5.46E-05	PITA_34199	intron	Endoplasmic reticulum metalloproteinase	GO:0008152-metabolic process(L=1)
Contig_25443	scaffold3476	5.46E-05	PITA_34268	intron	CONTAINS InterPro DOMAINs Galactose oxidase kelch, beta-propeller (InterPro IPR011043), Kelch repeat type 1 (InterPro IPR006652), Kelch repeat type 2 (InterPro IPR011498), Kelch-type beta propeller (InterPro IPR015915)	
Contig_29660	super1865	4.36E-35	PITA_34828	intron	chaperone protein DnaJ	GO:0000003-reproduction(L=1),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051704-multi-organism process(L=1),GO:0071840-cellular component organization or biogenesis(L=1)

Contig_34583	super654	1.47E-38	PITA_35434	intron	protein kinase kinase kinase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051704-multi-organism process(L=1),GO:0065007-biological regulation(L=1)
Contig_36192	scaffold91612	2.66E-18	PITA_37231	intron	4-coumarate--CoA ligase-like	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0050896-response to stimulus(L=1)
Contig_38481	super1333	2.10E-26	PITA_37604	intron	protease	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0065007-biological regulation(L=1)
Contig_39650	super1882	2.10E-26	PITA_37913	intron	ATP-dependent RNA helicase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_26987	scaffold153489	2.99E-14	PITA_38445	CDS	DnaJ homolog subfamily B member	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0050896-response to stimulus(L=1),GO:0051704-multi-organism process(L=1),GO:0065007-biological regulation(L=1)
Contig_45936	super4300	0.16	PITA_40827	intron	B3 domain- containing protein	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_46581	scaffold67986	6.13E-23	PITA_42479	intron	Lipid-A- disaccharide	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)

Contig_54333	scaffold127918	6.23E-23	PITA_42634	intron	phosphatidylinositol-4-phosphate 5-kinase	GO:0000003-reproduction(L=1),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0040007-growth(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1),GO:0051704-multi-organism cellular component organization or biogenesis(L=1)
Contig_35847	super4047	3.02E-25	PITA_44268	intron	Histone deacetylase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0065007-biological regulation(L=1),GO:0071840-cellular component organization or biogenesis(L=1)
Contig_71795	super3043	1.47E-38	PITA_46678	intron	response regulator	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_53151	scaffold38373	2.05E-15	PITA_48888	intron	diaminopimelate decarboxylase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)

Table 3.S3 Summary of BLAST results for RADtag_ID matches to the *Pinus taeda* (PITA) genome and within 20kbp of a gene. Annotations based on EggNOG descriptions and GO terms were sourced directly from the annotation file that accompanies the genome on treegenomesdb.org.

RADtag_ID	PITA_seqid	Blastn e-value	Query Sequence	Type	EggNOG Description	EggNOG.GO.Biological
Contig_52360	super2600	2.10E-26	PITA_00766	CDS	DnaJ homolog subfamily C	GO:0000003-reproduction(L=1),GO:0009987-cellular process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1),GO:0071840-cellular component organization or biogenesis(L=1)
Contig_47103	scaffold84156	2.33E-22	PITA_03370	CDS		
Contig_44277	scaffold180381	7.01E-19	PITA_04889	CDS	Protein of unknown function (DUF1664)	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_33280	scaffold33454	2.64E-29	PITA_05145	CDS	NA	
Contig_45936	scaffold69035	0.61	PITA_10646	CDS	Homeobox-leucine zipper protein	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0065007-biological regulation(L=1)
Contig_36552	scaffold33327	3.34E-21	PITA_14229	CDS	to conserved	
Contig_28907	scaffold93211	7.01E-19	PITA_14689	CDS	phospholipase C	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1)

Contig_38843	scaffold135214	8.8	PITA_17270	CDS	UDP-N-acetylglucosamine--peptide N-acetylglucosaminyltransferase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1)
Contig_46822	scaffold39564	8.01E-37	PITA_18521	CDS	F-box domain	
Contig_23415	scaffold146298	4.36E-35	PITA_19161	CDS	Cysteine-rich receptor-like protein kinase	GO:0002376-immune system process(L=1),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1),GO:0051704-multi-organism process(L=1)
Contig_60840	scaffold179606	2.38E-33	PITA_21499	CDS	YT521-B-like domain	
Contig_57733	scaffold2134	3.85E-28	PITA_22868	CDS	Inherit from euNOG: Endonuclease Exonuclease Phosphatase	
Contig_28454	C5160949	2.3	PITA_29710	CDS	receptor-like kinase protein	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_54333	scaffold182088	6.23E-23	PITA_31618	CDS	mitochondrial ubiquitin ligase activator of nfkb	GO:0008152-metabolic process(L=1)
Contig_73393	scaffold17543	2.36E-22	PITA_32532	CDS	Polyamine oxidase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_60723	scaffold77078	8.01E-37	PITA_34873	CDS	stem 28 kDa	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)

Contig_38319	super1793	8.01E-37	PITA_37919	CDS	acetyl-coa carboxylase	GO:0000003-reproduction(L=1),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1)
Contig_24989	scaffold206382	0.000207	PITA_40681	CDS	protein BREVIS RADIX-like	
Contig_50456	scaffold218490	1.47E-38	PITA_49009	CDS	NA	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0040007-growth(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1),GO:0065007-biological regulation(L=1)
Contig_20979	scaffold111511	3.02E-25	PITA_49754	CDS	Zinc finger, C3HC4 type (RING finger)	

Table 3.S4 Summary of BLAST results for RADtag_ID matches to the *Pinus taeda* (PITA) genome and over 20kbp from a gene. Annotations based on EggNOG descriptions and GO terms were sourced directly from the annotation file that accompanies the genome on treegenomesdb.org.

RADtag_ID	PITA seqid	Blastn e-value	Query Sequence	type	EggNOG Description	EggNOG.GO.Biological
Contig_36014	scaffold81562	4.33E-24	PITA_00410	CDS	histone H3	GO:0000003-reproduction(L=1),GO:0007094-mitotic spindle assembly checkpoint(L=10),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0022610-biological adhesion(L=1),GO:0023052-signaling(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0040007-growth(L=1),GO:0040011-locomotion(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1),GO:0065007-biological regulation(L=1),GO:0071840-cellular component organization or biogenesis(L=1)
Contig_38823	scaffold98725	3.40E-21	PITA_00612	CDS	shikimate quinate	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0065007-biological regulation(L=1)
Contig_23440	scaffold82932	1.01E-17	PITA_00850	CDS	Tetraspanin family	
Contig_51130	scaffold33886	1.63E-12	PITA_01338	CDS	proton pump interactor	GO:0065007-biological regulation(L=1)
Contig_38709	scaffold105914	4.85E-09	PITA_02307	CDS	Putative methyltransferase	
Contig_22404	super2964	3.85E-28	PITA_02432	CDS	ParB	GO:0008152-metabolic process(L=1),GO:0050896-response to stimulus(L=1)

Contig_23415	scaffold14908	3.02E-25	PITA_03648	CDS	Mitogen-activated protein kinase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_38558	scaffold140954	6.19E-12	PITA_03866	CDS	reductase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0048511-rhythmic process(L=1),GO:0050896-response to stimulus(L=1)
Contig_61095	scaffold108859	4.36E-35	PITA_04003	CDS	wound stress protein	
Contig_38801	scaffold225025	1.83E-30	PITA_05080	CDS	transferase activity, transferring hexosyl groups	GO:0008152-metabolic process(L=1)
Contig_59507	C5161785	0.16	PITA_05501	CDS	ATP-dependent DNA helicase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_71596	scaffold57237	1.30E-31	PITA_05721	CDS	May be involved in pre-mRNA splicing (By similarity)	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0065007-biological regulation(L=1)
Contig_26987	scaffold59649	1.14E-24	PITA_05818	CDS	Inherit from euNOG: Protein of unknown function (DUF 659)	
Contig_23835	super2953	7.90E-15	PITA_06105	CDS	Rubredoxin	
Contig_68879	scaffold187397	7.06E-30	PITA_06555	CDS	transcription factor	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)

Contig_32448	super3943	0.011	PITA_07094	CDS	ribosome biogenesis regulatory protein	GO:0071840-cellular component organization or biogenesis(L=1)
Contig_20833	scaffold130938	3.04E-36	PITA_07390	CDS	F-box kelch-repeat protein	
Contig_23453	scaffold75992	1.01E-17	PITA_08100	CDS	Small nuclear ribonucleoprotein	
Contig_22781	scaffold52038	3.82E-17	PITA_09433	CDS	protein ethylene insensitive	GO:0008152-metabolic process(L=1),GO:0009987- cellular process(L=1),GO:0023052- signaling(L=1),GO:0044699- single-organism process(L=1),GO:0050896- response to stimulus(L=1),GO:0051704- multi-organism process(L=1),GO:0065007- biological regulation(L=1)
Contig_26746	super3932	0.61	PITA_10969	CDS	ribosomal protein S6	GO:0008152-metabolic process(L=1),GO:0009987- cellular process(L=1)
Contig_69994	super3883	2.08E-15	PITA_11172	CDS	Inherit from euNOG: Endonuclease Exonuclease Phosphatase	
Contig_62818	scaffold126221	2.38E-33	PITA_11348	CDS	SpoU rRNA Methylase family	GO:0008152-metabolic process(L=1),GO:0009987- cellular process(L=1)
Contig_30273	scaffold80810	7.06E-30	PITA_11790	CDS		
Contig_27442	scaffold25008	2.38E-33	PITA_12332	CDS	exocyst complex component	GO:0009987-cellular process(L=1),GO:0044699- single-organism process(L=1),GO:0051179- localization(L=1)
Contig_52360	super1309	4.36E-35	PITA_13459	CDS	Mediator complex subunit MED14	GO:0008152-metabolic process(L=1),GO:0009987- cellular process(L=1),GO:0032501- multicellular organismal process(L=1),GO:0032502- developmental process(L=1),GO:0044699- single-organism process(L=1),GO:0065007- biological regulation(L=1)

Contig_28454	super4416	1.47E-38	PITA_14189	CDS	glycine-rich protein	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0051179-localization(L=1),GO:0071840-cellular component organization or biogenesis(L=1)
Contig_53342	scaffold65104	1.63E-34	PITA_14366	CDS	membrane-associated kinase regulator	
Contig_20833	scaffold227224	1.46E-27	PITA_14429	CDS	Methyltransferase domain	GO:0008152-metabolic process(L=1)
Contig_21309	scaffold58735	8.01E-37	PITA_14553	CDS	Homeobox-leucine zipper protein	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_38558	scaffold111589	1.14E-13	PITA_15078	CDS	CBL-interacting protein kinase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)

Contig_60142	scaffold93589	4.33E-24	PITA_15241	CDS	S-phase kinase-- associated protein	GO:0000003- reproduction(L=1),GO:0007 035-vacuolar acidification(L=11),GO:0007 610- behavior(L=1),GO:0008152- metabolic process(L=1),GO:0009987- cellular process(L=1),GO:0023052- signaling(L=1),GO:0031145- anaphase-promoting complex-dependent catabolic process(L=10),GO:0031146- SCF-dependent proteasomal ubiquitin-dependent protein catabolic process(L=10),GO:0032501- multicellular organismal process(L=1),GO:0032502- developmental process(L=1),GO:0040007- growth(L=1),GO:0040011- locomotion(L=1),GO:004469 9-single-organism process(L=1),GO:0050896- response to stimulus(L=1),GO:0051179- localization(L=1),GO:00514 37-positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition(L=11),GO:005143 9-regulation of ubiquitin- protein ligase activity involved in mitotic cell cycle(L=10),GO:0051452- intracellular pH reduction(L=10),GO:005170 4-multi-organism process(L=1),GO:0065007- biological regulation(L=1),GO:0071840 -cellular component organization or biogenesis(L=1),GO:200005 8-regulation of protein ubiquitination involved in ubiquitin-dependent protein catabolic process(L=10),GO:2000060- positive regulation of protein ubiquitination involved in ubiquitin-dependent protein catabolic process(L=10)
---------------------	---------------	----------	------------	-----	--	---

Contig_49266	scaffold164519	1.14E-24	PITA_16790	CDS	Zinc finger, C3HC4 type (RING finger)
---------------------	----------------	----------	------------	-----	---

Contig_23440	scaffold127477	1.85E-19	PITA_16855	CDS	Inherit from euNOG: Protein of unknown function (DUF 659)	
Contig_74841	super4142	3.34E-21	PITA_17814	CDS	RING	
Contig_26665	scaffold210482	4.36E-35	PITA_18115	CDS	Pyruvate kinase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1)
Contig_52827	scaffold145065	2.35E-11	PITA_18139	CDS	BTB POZ domain-containing protein	GO:0000003-reproduction(L=1),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_61084	super996	0.61	PITA_18315	CDS	agenet domain-containing protein	
Contig_44282	super3365	1.85E-19	PITA_18370	CDS	MYSc	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_50400	scaffold27804	6.97E-08	PITA_18754	CDS	Myb-like DNA-binding domain	
Contig_28907	scaffold196538	1.85E-19	PITA_19216	CDS	senescence-associated protein	GO:0050896-response to stimulus(L=1)

Contig_40612	scaffold227774	3.85E-28	PITA_19387	CDS	Serine threonine-protein kinase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0065007-biological regulation(L=1)
Contig_54333	super222	1.47E-38	PITA_19506	CDS	CBL-interacting protein kinase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_25141	scaffold58227	2.38E-33	PITA_19615	CDS	glutamate synthase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0022610-biological adhesion(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0040007-growth(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_69511	super276	7.06E-30	PITA_20523	CDS	UBX domain-containing protein	
Contig_61095	C5091181	3.40E-21	PITA_21653	CDS	Protein of unknown function (DUF1399)	
Contig_40435	super3974	6.23E-23	PITA_22100	CDS	Zinc ion binding	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_69082	scaffold92839	1.02E-28	PITA_22289	CDS	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)

Contig_34816	scaffold223532	1.29E-20	PITA_22683	CDS	phosphatase 2C	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1)
Contig_72286	scaffold21954	1.30E-31	PITA_23100	CDS	dsRNA-binding protein	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_75244	super2162	2.11E-37	PITA_23672	CDS	PXA domain	GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1)

Contig_43291	super3510	2.10E-26	PITA_23792	CDS	<p>Component of the FACT complex, a general chromatin factor that acts to reorganize nucleosomes. The FACT complex is involved in multiple processes that require DNA as a template such as mRNA elongation, DNA replication and DNA repair. During transcription the FACT complex acts as a histone chaperone that both destabilizes and restores nucleosomal structure. It facilitates the passage of RNA polymerase II and transcription by promoting the dissociation of one histone H2A-H2B dimer from the nucleosome, then subsequently promotes the reestablishment of the nucleosome following the passage of RNA polymerase II</p>	<p>GO:0000003-reproduction(L=1),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)</p>
Contig_66409	super709	4.91E-31	PITA_24604	CDS	Inherit from KOG: Retrotransposon protein	
Contig_28824	scaffold42968	2.10E-26	PITA_24822	CDS	vesicle-associated membrane protein	<p>GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1),GO:0071840-cellular component organization or biogenesis(L=1)</p>
Contig_75244	super419	1.14E-24	PITA_24919	CDS	NA	

Contig_38709	scaffold108429	3.37E-10	PITA_27242	CDS	Lipid-A-disaccharide	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_40432	scaffold19135	2.08E-15	PITA_28229	CDS	protein TRANSPARENT TESTA	GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1)
Contig_26665	super3179	3.85E-28	PITA_30787	CDS	tocopherol cyclase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1),GO:0065007-biological regulation(L=1)
Contig_33213	scaffold42137	7.95E-26	PITA_31144	CDS	subtilisin-like protease-like	GO:0008152-metabolic process(L=1),GO:0065007-biological regulation(L=1)
Contig_54843	super4204	1.47E-38	PITA_31311	CDS	F-box kelch-repeat protein	
Contig_43213	scaffold207871	7.06E-30	PITA_31566	CDS	Inherit from euNOG: expressed protein	
Contig_68036	scaffold106098	1.83E-30	PITA_32182	CDS	O-methyltransferase	GO:0008152-metabolic process(L=1)
Contig_37388	scaffold85011	4.36E-35	PITA_32221	CDS	glyoxal or galactose oxidase	
Contig_66403	scaffold85011	1.63E-34	PITA_32221	CDS	glyoxal or galactose oxidase	
Contig_18941	super1289	1.47E-38	PITA_33233	CDS	Retrotransposon protein	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)

Contig_38558	super3839	1.47E-38	PITA_33295	CDS	Alpha-amylase C-terminal beta-sheet domain	GO:0008152-metabolic process(L=1)
Contig_57465	super1249	1.64E-23	PITA_33924	CDS	Inherit from KOG: Retrotransposon protein	
Contig_33266	scaffold144802	3.85E-28	PITA_34661	CDS	reductase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0048511-rhythmic process(L=1),GO:0050896-response to stimulus(L=1)
Contig_51130	super4260	1.14E-13	PITA_36327	CDS	COP9 signalosome complex subunit	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_32625	scaffold4420	2.68E-29	PITA_36640	CDS	allene cyclase	oxide GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0050896-response to stimulus(L=1),GO:0051704-multi-organism process(L=1)
Contig_57465	scaffold98160	6.23E-23	PITA_36650	CDS	Pfam:DUF231	GO:0050896-response to stimulus(L=1),GO:0051704-multi-organism process(L=1)
Contig_53151	super80	2.99E-36	PITA_36960	CDS)-oxidoreductase	GO:0008152-metabolic process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0044699-single-organism process(L=1)

Contig_40432	scaffold196624	2.99E-14	PITA_37066	CDS	T-complex protein 1 subunit	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0050896-response to stimulus(L=1)
Contig_25199	super3811	1.85E-19	PITA_38103	CDS	DSBA-like thioredoxin domain	GO:0002376-immune system process(L=1),GO:0008152-metabolic process(L=1),GO:0050896-response to stimulus(L=1),GO:0051704-multi-organism process(L=1)
Contig_33266	scaffold133555	3.85E-28	PITA_38218	CDS	26S protease regulatory subunit 6B	GO:0008152-metabolic process(L=1)
Contig_50456	scaffold20831	2.99E-14	PITA_39205	CDS	leucine-rich repeat receptor-like serine threonine-protein kinase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_27551	super1638	7.01E-19	PITA_39622	CDS	NA	
Contig_57733	scaffold52254	3.85E-28	PITA_39630	CDS	Inherit from KOG: Retrotransposon protein	
Contig_22779	C5117729	6.23E-23	PITA_39659	CDS		
Contig_46822	C5117729	6.23E-23	PITA_39659	CDS		
Contig_33213	C5117729	3.02E-25	PITA_39659	CDS		
Contig_74841	scaffold105593	0.16	PITA_39844	CDS	repeat-containing protein	
Contig_20986	scaffold103990	6.19E-12	PITA_40915	CDS	UDP-Glycosyltransferase	GO:0008152-metabolic process(L=1)
Contig_44282	C5125621	7.01E-19	PITA_41876	CDS	Cysteine-rich receptor-like protein kinase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_52822	super723	1.45E-16	PITA_41903	CDS	zinc finger	

Contig_54333	scaffold127918	6.23E-23	PITA_42634	CDS	phosphatidylinositol-4-phosphate 5-kinase	GO:0000003-reproduction(L=1),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0032501-multicellular organismal process(L=1),GO:0032502-developmental process(L=1),GO:0040007-growth(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1),GO:0051704-multi-organism process(L=1),GO:0071840-cellular component organization or biogenesis(L=1)
Contig_52827	scaffold117547	2.38E-33	PITA_43562	CDS	expressed protein	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_68036	scaffold227647	1.81E-08	PITA_43694	CDS	wall-associated receptor kinase-like	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_50852	scaffold88724	1.14E-13	PITA_44865	CDS	transcription	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0044699-single-organism process(L=1),GO:0065007-biological regulation(L=1)
Contig_40005	scaffold43788	8.89E-11	PITA_44971	CDS	Inherit from euNOG: Transcription factor	
Contig_41424	scaffold123907	1.00E-28	PITA_45129	CDS	quinone-oxidoreductase homolog, chloroplastic-like	GO:0008152-metabolic process(L=1)
Contig_28824	scaffold72308	2.10E-26	PITA_45227	CDS	RabGAP TBC domain-containing protein	GO:0065007-biological regulation(L=1)
Contig_75680	scaffold207298	5.53E-27	PITA_46565	CDS	Dienelactone hydrolase family	GO:0008152-metabolic process(L=1),GO:0050896-response to stimulus(L=1)

Contig_31511	super906	4.36E-35	PITA_46723	CDS	nucleoside diphosphate kinase	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0023052-signaling(L=1),GO:0044699-single-organism process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_38843	scaffold17966	1.47E-38	PITA_46749	CDS		
Contig_34816	scaffold108578	1.29E-20	PITA_47880	CDS	glucose-6-phosphate isomerase	GO:0002376-immune system process(L=1),GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0050896-response to stimulus(L=1),GO:0065007-biological regulation(L=1)
Contig_73361	scaffold9408	1.85E-19	PITA_48225	CDS	ribosomal protein	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_71596	C5146975	4.30E-13	PITA_48564	CDS	flavanone 3-hydroxylase	GO:0008152-metabolic process(L=1),GO:0050896-response to stimulus(L=1)
Contig_31511	scaffold65544	1.84E-08	PITA_48737	CDS	Protein of unknown function, DUF604	GO:0008152-metabolic process(L=1)
Contig_28804	scaffold126409	2.10E-26	PITA_48788	CDS	nuclear transport factor 2 (NTF2) family protein RNA recognition motif (RRM)-containing protein	GO:0050896-response to stimulus(L=1),GO:0051179-localization(L=1)
Contig_59903	scaffold10302	6.23E-23	PITA_49712	CDS		
Contig_38823	super2641	4.36E-35	PITA_50046	CDS	FGGY carbohydrate kinase domain-containing	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1)
Contig_59507	scaffold72084	0.043	PITA_50648	CDS	Chaperone protein dnaJ 8	GO:0008152-metabolic process(L=1),GO:0009987-cellular process(L=1),GO:0050896-response to stimulus(L=1)