Genomic signature of reproductive isolation between the last two remnant populations of Torrey pine (*Pinus torreyana* Parry)

Lionel N. Di Santo[1,2,*], Alayna Mead[3,*], Jessica W. Wright[4], and Jill A. Hamilton[1,3,5]

[1] North Dakota State University, Department of Biological Sciences, Fargo, ND.

[2] University of Geneva, Department of Genetics and Evolution, Geneva, Switzerland.

[3] Pennsylvania State University, Department of Ecosystem Science and Management, State College, PA.

[4] Pacific Southwest Research Station, USDA-Forest Service, Placerville, CA.

[5] Corresponding author: jvh6349@psu.edu

[*] Contributed equally to the study

Abstract

Understanding the genomic mechanisms contributing to speciation requires studies of taxa that are in the early stages of divergence, before complete reproductive isolation has evolved. One of the rarest pines in the world, Torrey Pine (*Pinus torreyana* Parry), persists naturally across one island and one mainland population in southern California, and is an ideal system for assessing the evolution of divergence as they are morphologically and genetically differentiated, but experience some connectivity. Using ddRADseq and GWAS for a common garden experiment of island, mainland, and F1 genotypes, we identified divergent loci, their function, and their relationship to variation in fitness. Simulating neutral evolution and admixture, loci were identified that were more differentiated than expected and exhibited reduced admixture in the F1s. SNPs with reduced admixture were enriched for reproductive-related functions, such as growth and pollination. Additionally, one locus strongly associated with differential fitness exhibited reduced admixture, suggesting it could be involved in promoting reproductive isolation. With persistent genomic signatures of the evolution of reproductive isolation, Torrey pine may be at the early stages of speciation. As Torrey pine is a candidate for interpopulation genetic rescue, caution is warranted where longer-term gene flow between diverged populations may result in reduced fitness if barriers have evolved.

Key words: conifers, forest trees, genetic rescue, reproductive isolation, Torrey Pine

Background

Forest trees are well-known for maintaining signatures of local adaptation among populations even when experiencing high levels of homogenizing gene flow [1–3]. This apparent paradox is explained by strong competition and selection at early life stages: deleterious non-adaptive alleles introduced by migrant individuals or pollen are likely to be removed from the gene pool [1]. In some cases, population differentiation may lead to reproductive isolation and eventually speciation [4–9]. However, the evolutionary and genomic mechanisms underlying the shift from local adaptation with gene flow to the evolution of reproductive isolation (RI) is poorly documented in trees, particularly conifers [10]. The prevalence of local adaptation even under high levels of gene flow makes trees a good system for better understanding how and when divergent selection may drive the evolution of reproductive isolation, a prerequisite for speciation.

Speciation requires the evolution of genetic divergence, which can arise in response to environmental differences, as a result of isolation, or a combination of both [4,8,11]. Allopatry may cause incompatibilities to evolve between populations, resulting in reproductive isolation if they come into secondary contact. However, reproductive isolation can evolve even when gene flow is occurring if selection against migrants and hybrids is strong enough. If non-local alleles are deleterious, selection against them may favor the evolution of reproductive barriers.

Reproductive isolation can arise through many mechanisms, including both prezygotic and postzygotic barriers, and identifying the genes underlying these reproductive barriers may help us better understand the process of speciation at the genomic level [12–17]. One way to identify mechanisms underlying the evolution of reproductive isolation is through signatures left in the genome [5,11,14,16]. When reproductive isolation is developing amidst ongoing gene flow, as in the early stages of ecological speciation where species diverge as a consequence of natural selection among contrasting environments, loci that are under selection

will become more differentiated between two populations than other genomic regions [16]. As speciation proceeds, variants linked to those regions under selection diverge, eventually leading to differentiation throughout the genome. The functions of differentiated genomic regions can be further understood using phenotypic data for both populations grown in a common environment. Using genotype-phenotype associations, differentiated loci can be linked to reproductive and fitness-associated traits. As reproductive-aged common gardens are uncommon for long-lived species, our understanding of reproductive barriers and their genomic underpinnings in trees is limited.

In this study, we took advantage of a 20-year common garden with Torrey Pine (*Pinus torreyana* Parry), established by the USDA-Forest Service Pacific Southwest Research Station. Torrey pine is one of the rarest pines in the world with only two native populations remaining. Separated by approximately 280 km (170 miles), the island population native to Santa Rosa Island, one of the Channel Islands, experiences cooler temperatures and greater precipitation on average than the mainland population located at the Torrey Pine State Natural Reserve in La Jolla, California (Appendix S1). When grown in a common environment, the two populations exhibit phenotypic differences consistent with adaptation to their native environments [18], but also exhibit relatively low genetic differentiation (F_{ST} = 0.0129) and some evidence of ongoing gene flow [19]. In an open-pollinated common garden including mainland and island trees, only F1s from island maternal trees fertilized by mainland pollen were produced, with no F1s observed between mainland maternal trees fertilized by island pollen [18]. Asymmetric crossing barriers can be evidence of cytonuclear or maternal incompatibility [20–22], and may be a sign that reproductive barriers between the two populations have already begun to evolve [23]. Thus, Torrey pine is an ideal system for investigating the early stages of reproductive isolation between two geographically separated populations occurring in contrasting environments. To evaluate the genomic and fitness

consequences of inter-population gene flow we leverage a common garden experiment of reproductive-aged trees of both parental populations and F1 hybrids.

Understanding the nature of the divergence between Torrey pine populations will be critical to inform species management. Both populations exhibit exceedingly low genetic diversity, particularly for a conifer [19,24] and F1 hybrids between the two populations appear to have higher fitness proxies than the parental populations, suggesting genetic rescue may be beneficial [18]. However, increased F1 fitness may be the result of heterosis, in which excess homozygosity is reduced [25–27]. Given this, it is unclear whether increased hybrid fitness would persist in future generations of hybrids backcrossed with either parental population. If the parental populations have evolved partial reproductive isolation as a result of ecological speciation, F2s and backcrossed individuals may ultimately have lower fitness despite heterosis in F1s, limiting the value of intra-population genetic rescue [28,29].

In this paper, we investigate whether there is evidence for incipient ecological speciation between Torrey pine populations by testing whether highly differentiated loci are associated with fitness-related phenotypic differences. We address two main questions: 1) Do genomic and phenotypic differences observed between Torrey pine populations contribute to their reproductive isolation? 2) What are the implications for conservation of this endangered species – would genetic rescue be beneficial or should these two populations be managed as separate taxonomic groups?

Methods

Multigenerational common garden

Prior to 1960, two plots of 20 mainland and 20 island Torrey pine trees were established adjacent to each other at the USDA Horticultural Field Station (now the Scripps Institute) in La Jolla, CA [18,30]. Between 2004 and 2006, open-pollinated cones from 10 mainland and

16 island trees were sampled, yielding a total of 643 seeds which were subsequently planted and grown in a greenhouse at the USDA-Forest Service Pacific Southwest Station, Institute of Forest Genetics, Placerville, CA. As seedlings emerged, one cotyledon was clipped from each plant for genetic analyses and identification of seedlings' ancestry (island, mainland, or F1 hybrid) using two population-specific allozyme markers [31]. Of the 643 seeds planted, 128 were genetically confirmed as being of pure mainland ancestry, 134 were identified as being of pure island ancestry, and 381 were identified as mixed ancestry (F1 hybrids) [18,31]. These F1 hybrids solely reflected crossings of island females with mainland pollen donors, indicating hybrids were the product of unidirectional gene flow between parental populations.

Following emergence, a subset of 360 seedlings representing each ancestry (n=120 each of island, mainland, and F1 hybrids) were planted following a randomized complete block design with six replicate blocks at the common garden site in Montecito, CA (34.42963, -119.56396). Each of the 360 trees planted were measured for four quantitative traits, including tree height in cm (measured between 2008 and 2018), the number of conelets or first-year cones produced (measured between 2013 and 2018), the number of immature cones or second-year cones produced (measured between 2015 and 2018), and the number of mature or third-year cones produced (measured between 2015 and 2018). These traits were selected as they provide either proxies (height) or empirical estimates of fitness needed to examine the impact of interpopulation gene flow on Torrey pine fitness (see below). Given that selection likely acts across each reproductive stage and there was limited correlation among them over time (see below), we analyzed each cone production trait separately using all available years. For full details about the common garden experiment see [18].

DNA extraction and reduced representation sequencing

During the summer of 2016, needle tissue was collected from the 262 individuals surviving of the 360 seedlings originally planted within the common garden [18], including 89 pure

mainland, 88 pure island, and 85 F1 hybrid trees. Following collection, needles were dried and maintained on silica gel until genomic DNA could be extracted using approximately 20 to 35 mg of dried needle tissue based on a modified version of the CTAB protocol [32]. In this modified version, rough mixing (e.g., vortex) was replaced with slow manual shaking of samples to reduce DNA shearing. The concentration and purity of extracted DNA was evaluated for each sample separately using a NanoDrop 1000 Spectrophotometer (Thermo Scientific). Overall, purity ratios averaged 1.86 and 2.21 for 260/280 and 260/230 respectively, with concentrations of DNA ranging from 153.78 to 4,463.31 ng/µL (average = 1,138.41 ng/µL). Following extraction, DNA concentration was standardized to 85 ng/µL for all 262 samples and genomic libraries were prepared using the protocol described in [33]. Once constructed, all genomic libraries were pooled and sent to the Genomic Sequencing and Analysis Facility (GSAF; Austin, TX) for single-end sequencing (1x 100 bp) on 5 lanes of an Illumina HiSeq 2500. Prior to sequencing, the pooled library was selected for fragments within the range of 450-500 bp.

Reference-guided assembly and calling of genetic variants

Raw sequenced libraries were demultiplexed using *ipyrad* version 0.9.12 [34]. To reduce the probability of assigning a read to the wrong individual, *ipyrad* was parameterized to allow only a single mismatch in the barcode sequence. Raw reads were subsequently filtered for quality using *dDocent* version 2.8.13 pipeline [35,36]. Using *Trimmomatic* [37], base pairs with PHRED score below 20 at the beginning and end of reads and Illumina adapter sequences were removed, and reads were trimmed once the average PHRED score within a 5 bp window dropped below 10. Cleaned reads were mapped to the *Pinus taeda* (loblolly pine) draft genome version 2.0 (GeneBank accession: GCA_000404065.3) using the *BWA-MEM* algorithm [38] with default values, except for the mismatch penalty (-B, default: 4) and the gap open penalty (-O, default: 6), which were relaxed to 3 and 5, respectively, to account for additional genomic

differences that may have evolved between Torrey pine and closely related loblolly pine. Genetic variants were called from sequence alignments using the *SAMtools* and *BCFtools* pipeline [39]. Specifically, we used the multiallelic and rare-variant calling model (-m) with no prior expectation on the substitution rate (--prior 0). In total, the pipeline identified 44,493,992 genetic variants, including single-nucleotide (SNP) and insertion/deletion (INDEL) polymorphisms, that were subjected to quality filtering. First, variants with a genotype quality (GQ) < 20 and a genotype depth (DP) < 3 were marked as missing. Next, variants with PHRED scores ≤ 20 (QUAL), minor allele counts (MAC) < 3, minor allele frequencies (MAF) < 0.01, genotyping rates across individuals < 0.8, average depth across individuals > 23 (based on [40]'s equation), and inbreeding coefficients (F_{IS}) < -0.5 or > 0.5 were filtered out of the raw data. Of the initial 262 genotyped individuals, 53 were removed from subsequent analyses as they exhibited greater than 50% missing values. In total, following the removal of INDELs and SNPs with 3 or more alleles, 11,379 biallelic SNPs across 209 individuals, including 68 pure island, 75 pure mainland, and 66 F1 hybrid individuals were identified and used for analysis.

Admixture simulations

To evaluate whether barriers to gene flow may have evolved between Torrey pine populations, we quantified genome-wide admixture in F1 hybrids, measured as locus-specific observed heterozygosity (H_O), specifically searching for low-diversity SNPs. Using a customized R script relying on R packages *adegenet* [41,42] and *hierfstat* [43], for each of the 11,379 SNPs, we simulated the distribution of H_O values expected in F1 hybrids under the null hypothesis that H_O at any hybrid locus would be based on the frequency of alleles in the island and mainland populations. To do so, we used the function *hybridize()* that simulates hybridization between two populations using a 2-step approach. First, allele frequencies specific to each parental population are derived from genotypes of their individuals. Then, hybrid genotypes are generated by sampling gametes in each of these populations using a multinomial probability

distribution. SNP-specific null distributions were produced by repeating this process 1000 times and estimating locus-specific H_O each time from simulated hybrid genotypes using the function *basic.stats()*. To mirror the empirical SNP data set, we simulated 11,379 genotypes for 66 F1 hybrids within each iteration, using allele frequencies derived from 68 and 75 pure island and pure mainland individuals, respectively.

To identify SNPs that may exhibit reduced admixture relative to expectation, we compared simulated null H_O distributions (expected H_O) with H_O values estimated from empirical F1 hybrid genotypes (observed H_O). For each locus, we computed a one-sided p-value, defined as the probability that an expected H_O value is equal or lower than the observed H_O value at this locus. All p-values were corrected for multiple testing using Benjamini and Hochberg's [44] False Discovery Rate procedure implemented within the R function *p.adjust()*. SNPs that exhibited significantly reduced admixture relative to expectation based on parental allele frequencies (FDR < 0.1) were classified as potential candidate regions of the genome associated with the evolution of barriers to gene flow between island and mainland populations of Torrey pine.

Functional plant ontology (PO) and gene ontology (GO) annotation

SNPs exhibiting lower admixture relative to expectations were functionally annotated to determine whether their functions, processes, or anatomical and temporal expressions may be important to Torrey pine fitness using *blastx* 2.9.0+ [45–47] and a combination of customized R scripts. First, a region of 5,000 bps (5 kbps) from *Pinus taeda* draft genome version 2.0 (GenBank accession: GCA_000404065.3) was extracted 2,500 bps before and after each target locus using *samtools* [48]. If a SNP was within the first 2,500 bp of a scaffold, the first 5,000 bps were extracted. If a scaffold harboring a SNP was shorter than 5 kbps, the whole scaffold length was used instead. Extracted sequences were then blasted against the TAIR10 peptide blastset (TAIR10_pep_20101214_updated). Homology between query and blastset sequences

was assessed using *blastx* default parameters, an expectation value threshold of 10^{-3} (-evalue 0.001), and a maximum number of database hits of 5 (-max_target_seqs 5). Lastly, mapping of gene ontology (GO) and plant ontology (PO) terms onto annotated sequences was performed in R using a customized script and GO (ATH_GO_GOSLIM.txt.gz) as well as PO (grow.txt.gz) annotations available online. All databases mentioned throughout this section are available from https://www.arabidopsis.org/download/. Using the TAIR database, while neither conifer- nor pine-specific, provides an exhaustive set of GO and PO annotations to associate with reduced admixture. This database remains the most complete with respect to plant gene functions and shares orthologous and homologous sequences with pine that will aid in determining functional categories critical to plant fitness [e.g., 49]. Functional gene annotation is limited for conifers, particularly for genes involved in reproduction and development [50]. Furthermore, some genes involved in flowering in angiosperms pre-date their divergence with gymnosperms [50–52]. However, because of the evolutionary distance between *Arabidopsis* and Torrey pine, the annotations will be best understood within broad functional categorizations, rather than species-specific contexts.

Following the same procedure as above, we also annotated 5 kb regions around 500 randomly selected SNPs of the total 11,379 called to investigate whether the set of candidate loci for reproductive isolation was enriched for specific functions or processes. Identifying which functional annotations (GO or PO) may be enriched in loci with reduced admixture relative to expectation was performed in R using Pearson's chi-squared test as implemented in the function *prop.test()*. First, we determined which annotations were shared between the candidate and the random set of loci. Then, we estimated the proportion (frequency) of each shared annotation within both sets of loci separately and tested the null hypothesis of equal proportions (frequencies) between data sets, correcting all p-values for multiple testing using Benjamini and Hochberg's [44] False Discovery Rate procedure implemented within the R

function *p.adjust()*. Proportions of a given annotation were assumed significantly different between random loci and loci candidate for reproductive isolation when the false discovery rate was less than 10% (FDR < 0.1).

Correlation analysis between measured phenotypes

To limit redundancy across traits and ensure each phenotype largely reflects unique inter-individual variation, we conducted a repeated measures correlation analysis in R based on temporal measurements taken on all 209 samples using the package *rmcorr* [53]. Significance of estimated correlation coefficients was evaluated considering a threshold $\alpha = 0.05$. A correlogram with correlations between all pairs of phenotypes was generated using the R package *corrplot* [54] and is shown in Appendix S2. While some correlation coefficients were significant, no phenotypes were highly correlated ($r \geq 0.6$). Consequently, all four temporally assessed traits (tree height in cm, the number of conelets (first-year cones), the number of immature cones (second-year cones), and the number of mature cones (third-year cones) produced) were kept for subsequent analyses.

Genome-wide association analysis (GWAS)

To assess whether genotypic variation at select SNPs may be associated with variability in temporally measured phenotypes, we conducted a repeated-measures genome-wide association analysis based on all 209 samples and 11,379 genetic variants using the R package *RepeatABEL* [55]. For each trait, we first used the function *preFitModel()* to fit a linear mixed model to the data without SNP effects to estimate variance components for the trait, including a random polygenic (based on a genetic relationship matrix) and permanent environmental effect to account for population structure and the repeated nature (temporal sampling) of the data. Both of these effects are computed and assessed internally within the function. In addition, we considered "block" (spatial position of a sample within the common garden) and "year" as additional random variables within the model. Following this, we used the function *rGLS()*,

257 fitting a generalized least square model to each genetic marker given a covariance matrix
258 (estimated using *preFitModel*), to test for associations between SNPs (considered as fixed
259 effects) and phenotypes while correcting for the effect of random variables. Lastly, SNP-
260 specific p-values were corrected for multiple testing using Benjamini and Hochberg's [44]
261 False Discovery Rate procedure implemented within the R function *p.adjust()*. Of the 11,379
262 SNPs tested for association with each of the four phenotypic traits, only those with FDR < 0.1
263 were considered potential candidate markers underlying genotype-phenotype association.

264 ***Evaluating fitness consequences following interpopulation genetic mixing***

265 Comparing SNP lists, one locus was identified that exhibited both lower admixture than
266 expected under a neutral model and was statistically associated with one or more of the four
267 fitness or fitness-related phenotypes (hereafter referred to as the shared locus, see Results). To
268 identify whether this locus was a potential candidate associated with reproductive isolation
269 between Torrey pine populations, we required that (1) genotypes frequencies for the locus
270 would be biased by population, with each population favoring a different homozygous
271 genotype, and (2) fitness of individuals with non-local and admixed genotypes would be
272 reduced relative to individuals with the more local genotype. As the common garden was
273 planted in Montecito near Santa Barbara, CA, we considered mainland trees as local and island
274 trees as non-local. To determine whether the number of homozygous genotypes statistically
275 differed between parental ancestries, we counted genotype occurrences across pure island and
276 pure mainland individuals and conducted a chi-square test on the resulting contingency table.

277 To understand the effects of climate on fitness, we compared climatic variables from
278 the two native Torrey Pine populations and the common garden site, extracting data from
279 climate rasters from BioClim version 2.1 for the years 1970-2000 at 30 second resolution [56]
280 using the R package terra, version 1.7.71 [57] in R 4.3.2 [58]. We considered the common
281 garden site more similar to the native mainland site because the island site generally

experiences more moderate conditions; for example, having lower temperatures and more precipitation during the warmest quarter (Appendix S1). While the common garden site is a novel environment compared to either native site, and is more similar to the island site along PC1, the severity of the summer drought is the primary stressor in Mediterranean climates and is likely to be a more important factor in adaptation [59–62].

A linear mixed model and expected marginal means approach was used to evaluate whether average phenotypes statistically differed among genotypes by leveraging R packages *lme4* [63] and *emmeans* [64]. For each trait, we built a model where phenotypic variation was explained by the fixed effect of individuals' genotypes at the shared locus, the fixed effect of among-individual genetic relationships, the random effect of blocks assigned to individuals within the common garden, the random effect of years phenotypes were measured, and the random effect of individuals themselves (to account for the repeated nature of response variables). Genetic relationships among common garden individuals were estimated using a principal component analysis implemented within the *dudi.pca()* function from the *adegenet* R package. This analysis was performed on allele frequencies of the full genomic data set (11,379 SNPs across 209 individuals), with missing values replaced by the average allele frequencies across individuals. The first two principal components were used in all four linear mixed models to summarize genetic relationships among individuals, as these components provide good proxies for among-ancestry (PC1) and within-ancestry genetic structure (PC2) (Appendix S3).

The *emmeans()* function, computing expected marginal means, was used for the final assessment of phenotype average differences among genotypes. Briefly, the function computes genotype-specific trait averages and standard errors corrected for all effects included within linear mixed models. The function also computes adjusted p-values estimated from t-ratios calculated between all possible pairwise genotype comparisons. Ultimately, we used this

307 statistic to infer statistical differences in average phenotypes for all traits measured between
308 genotypes.
309
310 **Results**

311 ***SNPs exhibiting lower-than-expected admixture and their functional relevance***
312 Simulation of first-generation progenies from island and mainland individuals using genotypes
313 at 11,379 SNPs revealed some loci in F1 hybrids that may exhibit a higher degree of fixation
314 relative to expectations under the hypothesis that genotypes are solely determined by allele
315 frequencies in each parental population. A comparison between observed and simulated H_O
316 estimates at each locus identified 185 SNPs (FDR < 0.1) with reduced admixture. Of the 184
317 5kb-long sequences containing these SNPs (two SNPs were located on the same segment), 75
318 (41%) could be annotated with the TAIR 10 peptide blastset, with hits in 59 described *A.*
319 *thaliana* genes. Plant ontology terms could be retrieved for 37 (63%) of the 59 gene hits, while
320 gene ontology terms were retrieved for all 59 gene hits.
321 Various plant developmental stages are expressed in those regions of the genome that
322 exhibited less admixture than expected (Figure 1). This includes embryo development stages
323 (e.g., C globular stage, D bilateral stage, E expanded cotyledon stage, or F mature embryo
324 stage), vegetative development stages (e.g., 2-, 4-, 6-, 8-, 10-, 12-leaf stages, or seedling
325 development stage), and reproductive development stages (e.g., 4 anthesis stage, petal
326 differentiation and expansion stage, M germinated pollen stage, or L mature pollen stage).
327 Functionally, the vast majority of these loci are located or active in the nucleus with molecular
328 functions including protein binding, catalytic activity, RNA binding, as well as kinase and
329 transferase activities (Figure 2a, b). Biological processes associated with these molecular
330 functions covered responses to various stimuli and stresses, and plant development (Figure 2c).
331 Finally, enrichment analyses demonstrated that while no PO terms could preferentially be

associated with SNPs exhibiting reduced admixture relative to expectation, some GO terms were significantly enriched in the latter SNP set (Table 1). Interestingly, the products of genes associated with these loci appeared to often be located in the plasma membrane, involved in important developmental and reproductive processes, including cell differentiation, growth, and pollination, and have molecular functions including kinase or nuclease activity.

SNPs exhibiting statistical association with fitness traits and their proxies

Genome-wide association analysis indicated alleles at several loci may play an important role in contributing to fitness variation in Torrey pine. In total, 12 SNPs were significantly associated (FDR < 0.1) with variation in phenotypes (Table 2). While 8 (67%) of the 12 SNPs were associated with a single trait, the remaining 4 were associated with two or more traits. Phenotypic value of these traits all increased with the number of derived alleles at associated SNPs except for one, locus_8778 [APFE031129984.1:16200] (Table 2), where it decreased. Interestingly, comparing SNPs associated with fitness and fitness-related traits with SNPs exhibiting low-level admixture in F1 hybrids identified one common locus (locus_4218 [APFE030529380.1:392010]). This locus, in addition to exhibiting a reduced degree of admixture, was also the only locus that significantly explained phenotypic variation across all four traits measured (Table 2). Unfortunately, functional annotations for the 5 kbps region surrounding this locus were limited. While this region may be located in the mitochondrion (GO:0005739), nothing is known about its molecular function or the biological process it might be involved in.

Fitness consequences of hybridization at a locus exhibiting reduced admixture and significant genotype-phenotype association

The chi-squared test performed on homozygous genotype counts for locus_4218 [APFE030529380.1:392010] demonstrated unequal distribution of these genotypes between pure parental ancestries ($\chi^2_1 = 104.1$, $P < 0.001$). While homozygotes for the reference allele

(0/0) were dominant among pure island individuals (94% of all genotypes), homozygotes for the derived allele (1/1) formed the core of pure mainland individuals (97% of all genotypes). Given this result, we considered the reference allele (0) as the island allele, and the derived allele (1) as the mainland allele.

The evaluation of expected marginal means (phenotype averages corrected for within- and among-ancestry genetic structure, blocks within the common garden, year of trait measurements, and sample ID) revealed unequal fitness across reference, derived, and heterozygous genotypes. For three out of the four phenotypic traits, individuals homozygous for the island allele exhibited reduced fitness, while individuals homozygous for the mainland allele exhibited greater fitness, with heterozygous genotypes intermediate to the two (Figure 3; Appendix S4). These traits included tree height (cm), the number of conelets produced, and the number of mature cones produced. For the number of immature cones produced, the distribution of fitness among genotypes was slightly different. While homozygous individuals for the island allele still exhibited the lowest fitness, fitness of homozygous individuals for the mainland allele and heterozygous individuals were the highest, with no significant differences between genotypes. We considered the mainland allele (1) as the local, and the island allele (0) as non-local because both the common garden site and the mainland experiences warmer and drier summers. Our results thus indicate that carrying at least one copy of the local allele increases tree fitness, with a direct additive relationship between the dosage of mainland allele and fitness for tree height (cm), the number of conelets produced, and the number of mature cones produced (Figure 3).

Discussion

Local adaptation is common in tree species [65–68], but it is unclear when and how population differentiation may drive the evolution of reproductive isolation. In Torrey pine, previous work

has shown that the island and mainland populations are genetically and phenotypically different despite some recurrent gene flow, consistent with adaptation to differing environments [18,19]. Furthermore, asymmetric gene flow between populations when grown together provides some evidence for the evolution of reproductive isolation [18]. Together, this suggests that the two Torrey pine populations could be at an early stage of speciation. Here, we tested for further evidence for the evolution of reproductive isolation by identifying loci that appear highly differentiated between the populations, exhibit less recombination than expected, and are associated with differential fitness effects. One highly diverged locus was associated with fitness differences in the common garden, and could be a candidate locus for reproductively isolating island and mainland Torrey pine populations on an ecological scale. Highly differentiated loci were enriched for functions important in reproduction, such as growth, pollination, and cell differentiation, suggesting these could be the mechanisms underlying the evolution of reproductive isolation between the two populations.

While demographic modeling has provided evidence for historic speciation with gene flow in pines [69,70], the specific molecular mechanisms underlying reproductive isolation in conifers are not well understood [10]. We identified loci that may be involved in the evolution of reproductive isolation between two populations, including pollination-related genes and one allele associated with fitness differences expected when hypothesizing local adaptation, suggesting that both intrinsic and extrinsic barriers could be contributing to reproductive isolation in this species. These results thus identify potential molecular mechanisms underlying the process of reproductive isolation in a critically endangered pine and suggest that differential adaptation could be a driving force, which is important information for establishing optimal conservation management strategies for this species.

One locus shows patterns suggestive of extrinsic postzygotic reproductive isolation

We identified a locus that is divergent between the two populations and is also associated with fitness differences (locus_4218 [APFE030529380.1:392010]). Individuals with the "mainland" variant of this locus have higher growth and reproduction in the mainland common garden, consistent with local adaptation to the hotter, drier environment shared by both the mainland and garden environment (Appendix S1). The differentiation between populations at this allele could result from differential selection if heterozygous hybrids have lower fitness in both natural populations relative to individuals that are homozygous for the local variant. If this locus experiences differential selection between the two Torrey pine populations, it could contribute to reproductive isolation. While our results suggest that this locus may be beneficial in the hotter, drier summers experienced by mainland environments relative to the island population (Appendix S1), a reciprocal transplant of the two populations and their F1 hybrids would still be required to definitively show that the non-local variant decreases fitness in both natural populations.

While there is limited functional annotation for genomic regions proximal to this locus, it does have the annotation "mitochondrion" (GO:0005739) and may be encoded in the mitochondrion. Alternatively, it may represent ancient paralogy between nuclear and mitochondrial genomes. The locus could also be part of a nuclear mitochondrial gene present in the nucleus following mito-nuclear gene transfer, with its product interacting with mitochondria. Consequently, the locus could be involved in cytoplasmic incompatibility between the maternally-inherited mitochondrion and nuclear paternal genes (such as those involved in pollen tube growth) [13,21], or with the chloroplast, which is paternally inherited in pines [71,72].

Local adaptation is generally polygenic in forest trees [73,74], suggesting ecological speciation may also have a broad genetic basis [15,75]. However, we only find one locus

showing evidence consistent with ecological speciation. Because we used ddRADseq data and only sampled a fraction of the genome of Torrey pine, it is likely that other loci underlying reproductive isolation exist, but were not captured in this study. Some of the loci we identified as highly diverged may be linked to loci under differential selection, resulting in divergence among populations without an actual association with fitness. The fact that we identified a locus linked to both reproductive isolation and fitness despite this limitation suggests that the pattern may be more prevalent throughout the genome and could be better assessed using whole-genome sequence data.

Functions of differentiated genes

If loci that are highly differentiated between island and mainland populations are "speciation genes" underlying hybrid incompatibilities, differential adaptation, or both (or linked to involved genes), they should be enriched for related functions [13,15,28,76]. Intrinsic barriers are likely to be related to pollination or reproduction, and extrinsic barriers may be involved in traits under divergent selection, such as abiotic or biotic conditions. We found that highly differentiated loci were located near genes that were enriched for a wide range of functions (Table 1). Similar to previous studies, enriched functions included those related to pollination and development (“growth”, “pollination”, “cell differentiation”) [13,77]. This suggests that these differentiated regions may be involved in intrinsic reproductive barriers, such as pollination incompatibilities or failed development post-fertilization [78,79]. Other enriched functions include broad categories such as “plasma membrane”, “kinase activity”, and “biosynthetic process”, so it’s difficult to determine whether these genes could be involved in adaptation to contrasting environments, particularly as only one differentiated gene was associated with fitness differences in the common garden. Previously, island seedlings were found to germinate later and have a lower growth rate than mainland seedlings [18], and the enrichment of loci with the “growth” GO term may be related to these differential growth

455 rates. It is also possible that the enriched functions are associated with fitness in traits or
456 environmental conditions that were not measured in the common garden.

457 ***Conservation implications***

458 F1 hybrids between mainland and island populations exhibit heterosis, with a higher growth
459 rate and fecundity than either parental population [80]. Alone, this result would suggest that
460 the natural populations might benefit from genetic rescue, in which hybridization between the
461 two populations could introduce new genetic variation and mask deleterious homozygous
462 alleles. However, our results here warrant caution before introducing non-local or hybrid trees
463 into the wild. If the natural populations have already developed reproductive isolation,
464 introgression may result in decreased fitness that only presents in F2s and later generation
465 hybrids. Our data suggests that the island and mainland populations be managed separately at
466 present, but that future work should use crossing experiments and reciprocal transplants to
467 determine whether genetic rescue could have long-term benefits to the small populations of
468 this endangered tree species.

469

470 **Conclusions**

471 Taken together, the evidence of asymmetric barriers to gene flow, highly differentiated
472 genomic regions related to varying reproductive functions, and one highly differentiated locus
473 that associates with reduced fitness in individuals carrying one or two non-mainland alleles for
474 all four measured fitness metrics suggest that Torrey Pine populations are beginning to evolve
475 reproductive isolation, and this may be partly driven by adaptation to contrasting island and
476 mainland environments. In the future, whole-genome sequencing combined with development
477 of experimental crosses including F2s and backcrosses will be poised to identify the genes that
478 may be under divergent selection and contributing to the evolution of extrinsic reproductive

479 isolation, and determine whether the two processes occur simultaneously through evolution in
480 the same genes, or independently through evolution in different genes [15].
481
482 **Data accessibility**
483 Genomic and phenotypic data generated for this study, as well as scripts used for analysis are
484 available from Dryad (DOI): 10.5061/dryad.pc866t1xg.
485
486 **Author's contributions**
487 ***L.N.D.S.*** designed the study, contributed to data collection, generated, processed, and analyzed
488 genomic data, interpreted the data, and led the writing of the manuscript. ***A.M.*** contributed to
489 data collection, analyzed the climate data, interpreted the results, and led the writing of the
490 manuscript. ***J.W.W.*** contributed to data collection and edited the manuscript. ***J.A.H.*** designed
491 the study, contributed to data collection, interpreted the results, and wrote the manuscript.
492 ***L.N.D.S.*** and ***A.M.*** contributed equally to the study.
493
494 **Funding**
495 This work was funded by USDA Forest Service Health Protection Gene Conservation
496 Programme and Western Wildlands Environmental Threat Assessment Center to JAH and
497 JWW, Morton Arboretum Center for Tree Science Fellowship to JAH, a new faculty award
498 from the office of North Dakota Experimental Programme to Stimulate Competitive Research
499 (ND-EPSCoR NSFIIA-1355466), the Schatz Center for Tree Molecular Genetics, and the
500 NDSU Environmental and Conservation Sciences Graduate Programme to LNDS and an NSF-
501 PRFB (Award #2209410) to AM.
502

Acknowledgements

Tom Ledig, USDA-Forest Service, PSW, had the foresight to establish the common garden study at the end of his career. We thank him for sharing this project with us. For their help collecting needle tissue and measuring phenotypes within the common garden, the authors would like to thank Alexis Pearson, Tyler Stadel, Courtney Canning, and Madison Snider. The authors also would like to thank John Gabbert for allowing the scientific crew to access the common garden study. Finally, the authors would like to acknowledge the Michumash and the Kumeyaay people as the traditional caretakers of the Torrey pine ecosystems. Any use of product names is for informational purposes only and does not imply endorsement by the US Government. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy. This work was supported in part by the U.S. Department of Agriculture, Forest Service.

References

1. Petit RJ, Hampe A. 2006 Some evolutionary consequences of being a tree. *Annu. Rev. Ecol. Evol. Syst.* **37**, 187–214.
2. Janes JK, Hamilton JA. 2017 Mixing It Up: The Role of Hybridization in Forest Management and Conservation under Climate Change. *Forests* **8**, 237. (doi:10.3390/f8070237)
3. Cannon CH, Petit RJ. 2020 The oak syngameon: more than the sum of its parts. *New Phytol.* **226**, 978–983. (doi:10.1111/nph.16091)
4. Schluter D. 2001 Ecology and the origin of species. *Trends Ecol. Evol.* **16**, 372–380. (doi:10.1016/S0169-5347(01)02198-X)
5. Schluter D. 2009 Evidence for Ecological Speciation and Its Alternative. *Science* **323**, 737–741. (doi:10.1126/science.1160006)
6. Hendry AP, Nosil P, Rieseberg LH. 2007 The Speed of Ecological Speciation. *Funct. Ecol.* **21**, 455–464.
7. Hendry AP. 2009 Ecological speciation! Or the lack thereof? *Can. J. Fish. Aquat. Sci.* **66**, 1383–1398. (doi:10.1139/F09-074)
8. Nosil P, Harmon LJ, Seehausen O. 2009 Ecological explanations for (incomplete) speciation. *Trends Ecol. Evol.* **24**, 145–156. (doi:10.1016/j.tree.2008.10.011)
9. Andrew RL, Rieseberg LH. 2013 Divergence is focused on few genomic regions early in speciation: Incipient speciation of sunflower ecotypes. *Evolution* **67**, 2468–2482. (doi:10.1111/evo.12106)
10. 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 Pseudotaxus

chienii'. *Ann. Bot.* **125**, v–vii. (doi:10.1093/aob/mcz201)
11. Nosil P, Funk DJ, Ortiz-Barrientos D. 2009 Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* **18**, 375–402. (doi:10.1111/j.1365-294X.2008.03946.x)
12. Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH. 2008 The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 3009–3021. (doi:10.1098/rstb.2008.0064)
13. Rieseberg LH, Blackman BK. 2010 Speciation genes in plants. *Ann. Bot.* **106**, 439–455. (doi:10.1093/aob/mcq126)
14. Strasburg JL, Sherman NA, Wright KM, Moyle LC, Willis JH, Rieseberg LH. 2012 What can patterns of differentiation across plant genomes tell us about adaptation and speciation ? , 364–373. (doi:10.1098/rstb.2011.0199)
15. Schluter D, Rieseberg LH. 2022 Three problems in the genetics of speciation by selection. *Proc. Natl. Acad. Sci.* **119**, e2122153119. (doi:10.1073/pnas.2122153119)
16. Feder JL, Egan SP, Nosil P. 2012 The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350. (doi:10.1016/j.tig.2012.03.009)
17. Choi JY, Purugganan M, Stacy EA. 2020 Divergent Selection and Primary Gene Flow Shape Incipient Speciation of a Riparian Tree on Hawaii Island. *Mol. Biol. Evol.* **37**, 695–710. (doi:10.1093/molbev/msz259)
18. Hamilton JA, Royauté R, Wright JW, Hodgskiss P, Ledig TF. 2017 Genetic conservation and management of the California endemic, Torrey pine (Pinus torreyana Parry): Implications of genetic rescue in a genetically depauperate species. *Ecol. Evol.* **7**, 7370–7381. (doi:10.1002/ece3.3306)
19. Di Santo LN, Hoban S, Parchman TL, Wright JW, Hamilton JA. 2022 Reduced representation sequencing to understand the evolutionary history of Torrey pine (*Pinus torreyana* parry) with implications for rare species conservation. *Mol. Ecol.* **31**, 4622–4639. (doi:10.1111/mec.16615)
20. Tiffin P, Olson S, Moyle LC. 2001 Asymmetrical crossing barriers in angiosperms. *Proc. R. Soc. Lond. B Biol. Sci.* **268**, 861–867. (doi:10.1098/rspb.2000.1578)
21. Turelli M, Moyle LC. 2007 Asymmetric Postmating Isolation: Darwin's Corollary to Haldane's Rule. *Genetics* **176**, 1059–1088. (doi:10.1534/genetics.106.065979)
22. Case AL, Finseth FR, Barr CM, Fishman L. 2016 Selfish evolution of cytonuclear hybrid incompatibility in Mimulus. *Proc. R. Soc. B Biol. Sci.* **283**, 20161493. (doi:10.1098/rspb.2016.1493)
23. Barnard-Kubow KB, So N, Galloway LF. 2016 Cytonuclear incompatibility contributes to the early stages of speciation. *Evolution* **70**, 2752–2766. (doi:10.1111/evo.13075)
24. Farjon A. 2013 IUCN Red List of Threatened Species: *Pinus torreyana*.
25. Williams CG, Savolainen O. 1996 Inbreeding Depression in Conifers: Implications for Breeding Strategy. *For. Sci.* **42**, 102–117. (doi:10.1093/forestscience/42.1.102)
26. Tallmon DA, Luikart G, Waples RS. 2004 The alluring simplicity and complex reality of genetic rescue. *Trends Ecol. Evol.* **19**, 489–496. (doi:10.1016/j.tree.2004.07.003)
27. Johnson WE *et al.* 2010 Genetic Restoration of the Florida Panther. *Science* **329**, 1641–1645. (doi:10.1126/science.1192891)
28. Walter GM, Richards TJ, Wilkinson MJ, Blows MW, Aguirre JD, Ortiz-Barrientos D. 2020 Loss of ecologically important genetic variation in late generation hybrids reveals links between adaptation and speciation. *Evol. Lett.* **4**, 302–316. (doi:10.1002/evl3.187)
29. Christie K, Fraser LS, Lowry DB. 2022 The strength of reproductive isolating barriers in seed plants: Insights from studies quantifying premating and postmating reproductive barriers over the past 15 years. *Evolution* **76**, 2228–2243. (doi:10.1111/evo.14565)
30. Haller JR. 1967 *Proceedings of the Symposium on the Biology of the California Islands*. Santa Barbara Botanic Garden.

31. Ledig TF, Conkle TM. 1983 Gene diversity and genetic structure in a narrow endemic, Torrey Pine (Pinus torreyana Parry ex Carr.). *Evolution* **37**, 79–85. (doi:10.2307/2408176)
32. Doyle JJ, Doyle JL. 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15.
33. Di Santo LN, Hoban S, Parchman TL, Wright JW, Hamilton JA. 2022 Reduced representation sequencing to understand the evolutionary history of Torrey pine (Pinus torreyana parry) with implications for rare species conservation. *Mol. Ecol.* **31**, 4622–4639. (doi:10.1111/mec.16615)
34. Eaton DAR, Overcast I. 2020 ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics* **36**, 2592–2594. (doi:10.1093/bioinformatics/btz966)
35. Puritz JB, Hollenbeck CM, Gold JR. 2014 dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* **2**, e431–e431. (doi:10.7717/peerj.431)
36. Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE. 2014 Demystifying the RAD fad. *Mol. Ecol.* **23**, 5937–5942. (doi:10.1111/mec.12965)
37. Bolger AM, Lohse M, Usadel B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. (doi:10.1093/bioinformatics/btu170)
38. Li H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (doi:10.48550/arXiv.1303.3997)
39. Li H. 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993. (doi:10.1093/bioinformatics/btr509)
40. Li H. 2014 Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851. (doi:10.1093/bioinformatics/btu356)
41. Jombart T. 2008 adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405. (doi:10.1093/bioinformatics/btn129)
42. Jombart T, Ahmed I. 2011 adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071. (doi:10.1093/bioinformatics/btr521)
43. Goudet J, Jombart T. 2021 hierfstat: Estimation and Tests of Hierarchical F-Statistics.
44. Benjamini Y, Hochberg Y. 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300. (doi:10.1111/j.2517-6161.1995.tb02031.x)
45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410. (doi:10.1016/S0022-2836(05)80360-2)
46. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)
47. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009 BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421. (doi:10.1186/1471-2105-10-421)
48. Li H *et al.* 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)
49. Eckert AJ, Van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González-Martínez SC, Neale DB. 2010 Patterns of Population Structure and Environmental Associations to Aridity Across the Range of Loblolly Pine (Pinus taeda L., Pinaceae). *Genetics* **185**, 969–982. (doi:10.1534/genetics.110.115543)
50. De La Torre AR, Piot A, Liu B, Wilhite B, Weiss M, Porth I. 2020 Functional and morphological evolution in gymnosperms: A portrait of implicated gene families. *Evol. Appl.* **13**, 210–227. (doi:10.1111/eva.12839)

51. Moyroud E, Monniaux M, Thévenon E, Dumas R, Scutt CP, Frohlich MW, Parcy F. 2017 A link between LEAFY and B-gene homologues in Welwitschia mirabilis sheds light on ancestral mechanisms prefiguring floral development. *New Phytol.* **216**, 469–481. (doi:10.1111/nph.14483)
52. Liu Z-X, Xiong H-Y, Li L-Y, Fei Y-J. 2018 Functional Conservation of an AGAMOUS Orthologous Gene Controlling Reproductive Organ Development in the Gymnosperm Species Taxus chinensis var. mairei. *J. Plant Biol.* **61**, 50–59. (doi:10.1007/s12374-017-0154-4)
53. Bakdash JZ, Marusich LR. 2022 rmcorr: Repeated Measures Correlation.
54. Wei T, Simko V. 2021 R package 'corrplot': Visualization of Correlation Matrix.
55. Rönnegård L, McFarlane SE, Husby A, Kawakami T, Ellegren H, Qvarnström A. 2016 Increasing the power of genome wide association studies in natural populations using repeated measures – evaluation and implementation. *Methods Ecol. Evol.* **7**, 792–799. (doi:10.1111/2041-210X.12535)
56. Fick SE, Hijmans RJ. 2017 WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315. (doi:10.1002/joc.5086)
57. Hijmans RJ. 2024 *terra: Spatial Data Analysis*. See https://CRAN.R-project.org/package=terra.
58. R Core Team. 2023 *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. See https://www.R-project.org/.
59. Riordan EC, Gugger PF, Ortego J, Smith C, Gaddis K, Thompson P, Sork VL. 2016 Association of genetic and phenotypic variability with geography and climate in three southern California oaks. *Am. J. Bot.* **103**, 73–85. (doi:10.3732/ajb.1500135)
60. Granda E, Scoffoni C, Rubio-Casal AE, Sack L, Valladares F. 2014 Leaf and stem physiological responses to summer and winter extremes of woody species across temperate ecosystems. *Oikos* **123**, 1281–1290. (doi:10.1111/oik.01526)
61. Nardini A, Lo Gullo MA, Trifilò P, Salleo S. 2014 The challenge of the Mediterranean climate to plant hydraulics: Responses and adaptations. *Environ. Exp. Bot.* **103**, 68–79. (doi:10.1016/j.envexpbot.2013.09.018)
62. DeSilva R, Dodd RS. 2020 Association of genetic and climatic variability in giant sequoia, Sequoiadendron giganteum, reveals signatures of local adaptation along moisture-related gradients. *Ecol. Evol.* **10**, 10619–10632. (doi:10.1002/ece3.6716)
63. Bates D, Mächler M, Bolker B, Walker S. 2015 Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48. (doi:10.18637/jss.v067.i01)
64. Lenth RV. 2023 emmeans: Estimated Marginal Means, aka Least-Squares Means.
65. Savolainen O, Pyhäjärvi T, Knürr T. 2007 Gene Flow and Local Adaptation in Trees. *Annu. Rev. Ecol. Evol. Syst.* **38**, 595–619. (doi:10.1146/annurev.ecolsys.38.091206.095646)
66. Lind BM, Menon M, Bolte CE, Faske TM, Eckert AJ. 2018 The genomics of local adaptation in trees: are we out of the woods yet? *Tree Genet. Genomes* **14**, 29. (doi:10.1007/s11295-017-1224-y)
67. Gugger PF, Fitz-Gibbon ST, Albarrán-Lara A, Wright JW, Sork VL. 2021 Landscape genomics of *Quercus lobata* reveals genes involved in local climate adaptation at multiple spatial scales. *Mol. Ecol.* **30**, 406–423. (doi:10.1111/mec.15731)
68. Lind BM, Friedline CJ, Wegrzyn JL, Maloney PE, Vogler DR, Neale DB, Eckert AJ. 2017 Water availability drives signatures of local adaptation in whitebark pine (Pinus albicaulis Engelm.) across fine spatial scales of the Lake Tahoe Basin, USA. *Mol. Ecol.* **26**, 3168–3185. (doi:10.1111/mec.14106)
69. Bolte CE, Faske TM, Friedline CJ, Eckert AJ. 2022 Divergence amid recurring gene flow: complex demographic histories for two North American pine species (Pinus

689 pungens and P. rigida) fit growing expectations among forest trees. *Tree Genet. Genomes*
690 **18**, 35. (doi:10.1007/s11295-022-01565-8)
691 70. Menon M *et al.* 2018 The role of hybridization during ecological divergence of
692 southwestern white pine (Pinus strobiformis) and limber pine (P. flexilis). *Mol. Ecol.* **27**,
693 1245–1260. (doi:10.1111/mec.14505)
694 71. Neale DB, Sederoff RR. 1989 Paternal inheritance of chloroplast DNA and maternal
695 inheritance of mitochondrial DNA in loblolly pine. *Theor. Appl. Genet.* **77**, 212–216.
696 (doi:10.1007/BF00266189)
697 72. Mogensen HL. 1996 INVITED SPECIAL PAPER: The hows and whys of cytoplasmic
698 inheritance in seed plants. *Am. J. Bot.* **83**, 383–404. (doi:10.1002/j.1537-
699 2197.1996.tb12718.x)
700 73. Yeaman S *et al.* 2016 Convergent local adaptation to climate in distantly related conifers.
701 *Science* **353**, 1431–1433. (doi:10.1126/science.aaf7812)
702 74. MacLachlan IR, McDonald TK, Lind BM, Rieseberg LH, Yeaman S, Aitken SN. 2021
703 Genome-wide shifts in climate-related variation underpin responses to selective breeding
704 in a widespread conifer. *Proc. Natl. Acad. Sci.* **118**. (doi:10.1073/pnas.2016900118)
705 75. Rose NH, Bay RA, Morikawa MK, Palumbi SR. 2018 Polygenic evolution drives species
706 divergence and climate adaptation in corals. *Evolution* **72**, 82–94.
707 (doi:10.1111/evo.13385)
708 76. Wright KM, Lloyd D, Lowry DB, Macnair MR, Willis JH. 2013 Indirect Evolution of
709 Hybrid Lethality Due to Linkage with Selected Locus in Mimulus guttatus. *PLOS Biol.*
710 **11**, e1001497. (doi:10.1371/journal.pbio.1001497)
711 77. Leroy T *et al.* 2020 Massive postglacial gene flow between European white oaks
712 uncovered genes underlying species barriers. *New Phytol.* **226**, 1183–1197.
713 (doi:10.1111/nph.16039)
714 78. McWilliam JR. 1959 Interspecific Incompatibility in Pinus. *Am. J. Bot.* **46**, 425–433.
715 (doi:10.1002/j.1537-2197.1959.tb07033.x)
716 79. Fernando DD, Long SM, Sniezko RA. 2005 Sexual reproduction and crossing barriers in
717 white pines: the case between Pinus lambertiana (sugar pine) and P. monticola (western
718 white pine). *Tree Genet. Genomes* **1**, 143–150. (doi:10.1007/s11295-005-0015-z)
719 80. Hamilton JA, Royauté R, Wright JW, Hodgskiss P, Ledig FT. 2017 Genetic conservation
720 and management of the California endemic, Torrey pine (*Pinus torreyana* Parry):
721 Implications of genetic rescue in a genetically depauperate species. *Ecol. Evol.* **7**, 7370–
722 7381. (doi:10.1002/ece3.3306)

Tables

Table 1. GO annotations enrichment analysis. Listed are GO categories, GO terms, term proportions for RI candidates (Cand), term proportions for randomly selected SNPs (Rand), and the estimated FDR.

GO Category*	GO Term	Cand	Rand	FDR
CC	plasma membrane	0.121	0.024	<0.001
MF	Unknown molecular functions	0.136	0.082	0.048
	kinase activity	0.103	0.056	0.048
	nuclease activity	0.049	0.021	0.055
	other molecular functions	0.043	0.012	<0.001
BP	protein metabolic process	0.035	0.01	<0.001
	growth	0.026	0.002	<0.001
	pollination	0.023	0.001	<0.001
	secondary metabolic process	0.02	0.007	0.054
	cell differentiation	0.012	0.003	0.08
	biosynthetic process	0.009	0.055	<0.001

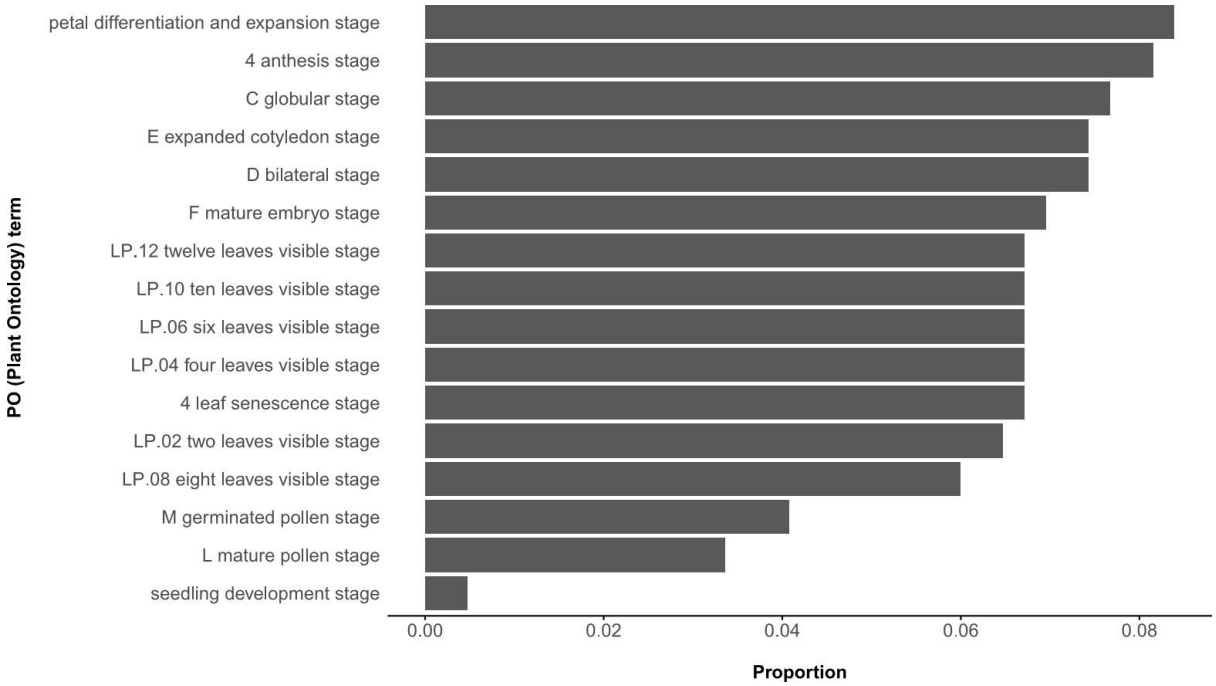
* CC: Cellular Component, MF: Molecular Function, and BP: Biological Process.

Table 2. SNPs significantly associated with one or more phenotypes (FDR < 0.1), including tree height (cm), number of conelets, number of immature cones, and number of mature cones. Listed are generic SNP IDs given for this study, SNP IDs retrieved from *Pinus taeda* draft genome (scaffold ID:position on the scaffold), the trait(s) each SNP is associated with, the effect of increasing the number of derived alleles on the associated phenotype(s), and the FDR of the genotype-phenotype association.

Generic SNP ID	SNP ID within *P. taeda* genome	Phenotypic trait association	Phenotypic effect	FDR
locus_4218	APFE030529380.1:392010	Tree height (cm)	+	0.047
		Nb conelets	+	<0.001
		Nb immature cones	+	0.085
		Nb mature cones	+	0.037
locus_12498	APFE031619559.1:67487	Tree height (cm)	+	0.047
locus_433	APFE030047412.1:104020	Nb conelets	+	0.078
		Nb mature cones	+	0.057
locus_3004	APFE030371338.1:302521	Nb conelets	+	0.078
locus_3676	APFE030458498.1:93076	Nb conelets	+	0.024
locus_9630	APFE031249435.1:61968	Nb conelets	+	0.011
		Nb immature cones	+	0.085
		Nb mature cones	+	0.027
locus_13093	APFE031688147.1:172170	Nb conelets	+	0.072
		Nb mature cones	+	0.027
locus_1164	APFE030138697.1:151998	Nb mature cones	+	0.032
locus_6624	APFE030828725.1:2769	Nb mature cones	+	0.027
locus_8019	APFE031020621.1:50406	Nb mature cones	+	0.053
locus_8778	APFE031129984.1:16200	Nb mature cones	-	0.057
locus_10668	APFE031380592.1:30882	Nb mature cones	+	0.053

737 **Figures**

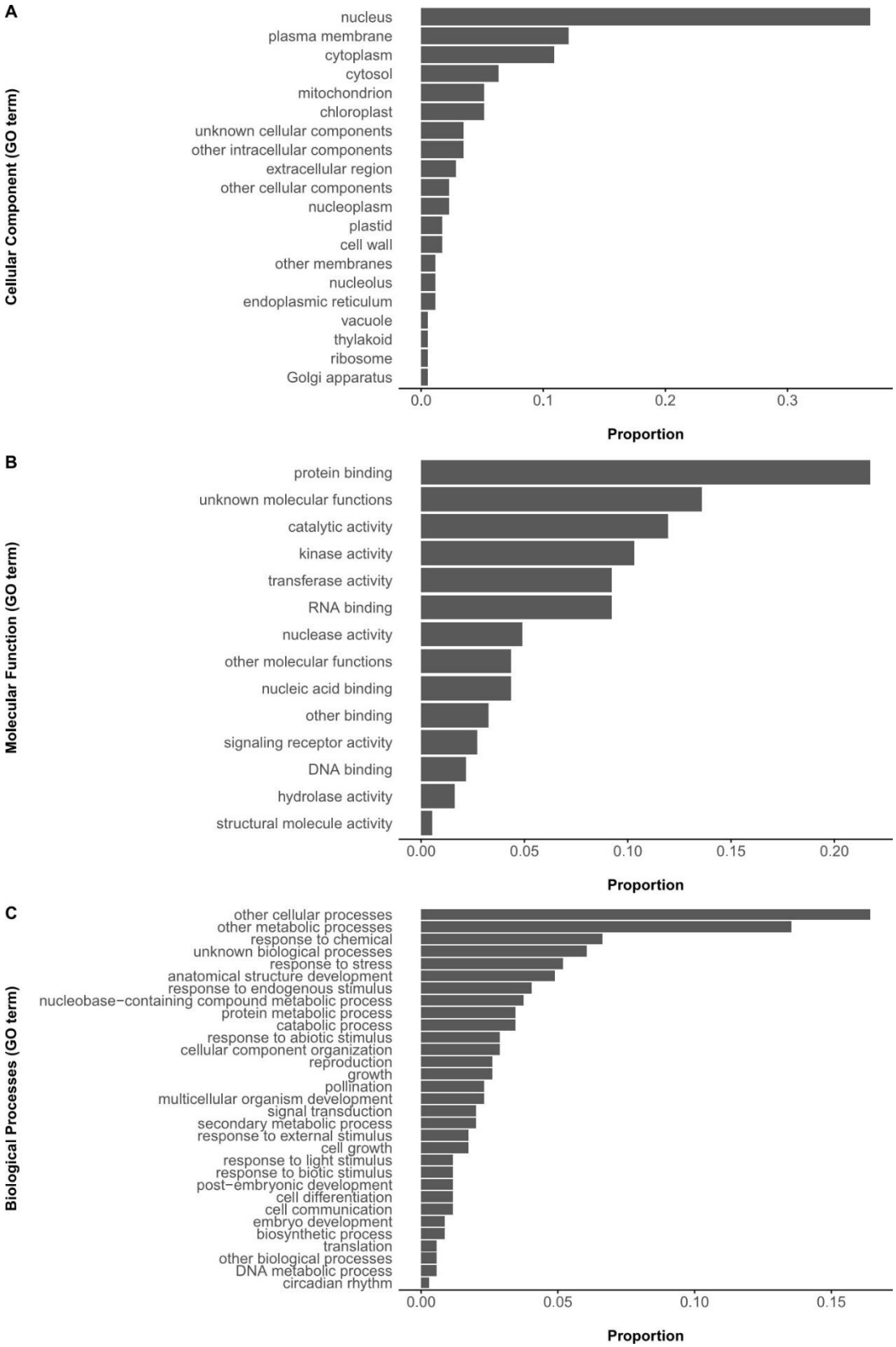
738



739

740 **Figure 1.** Proportion (x axis) of each plant ontology (PO) term (y axis) associated with all 75
741 annotated (out of 184) RI candidate segments (5kb-long sequence flanking candidate SNPs).

742



743

744 **Figure 2.** Proportion (x axis) of each gene ontology (GO) term (y axis) associated with all 75
745 annotated (out of 184) RI candidate segments (5kb-long sequence flanking candidate SNPs).
746 **(a)** Cellular Component (CC) terms. **(b)** Molecular Function (MF) terms. **(c)** Biological Process
747 (BP) terms.

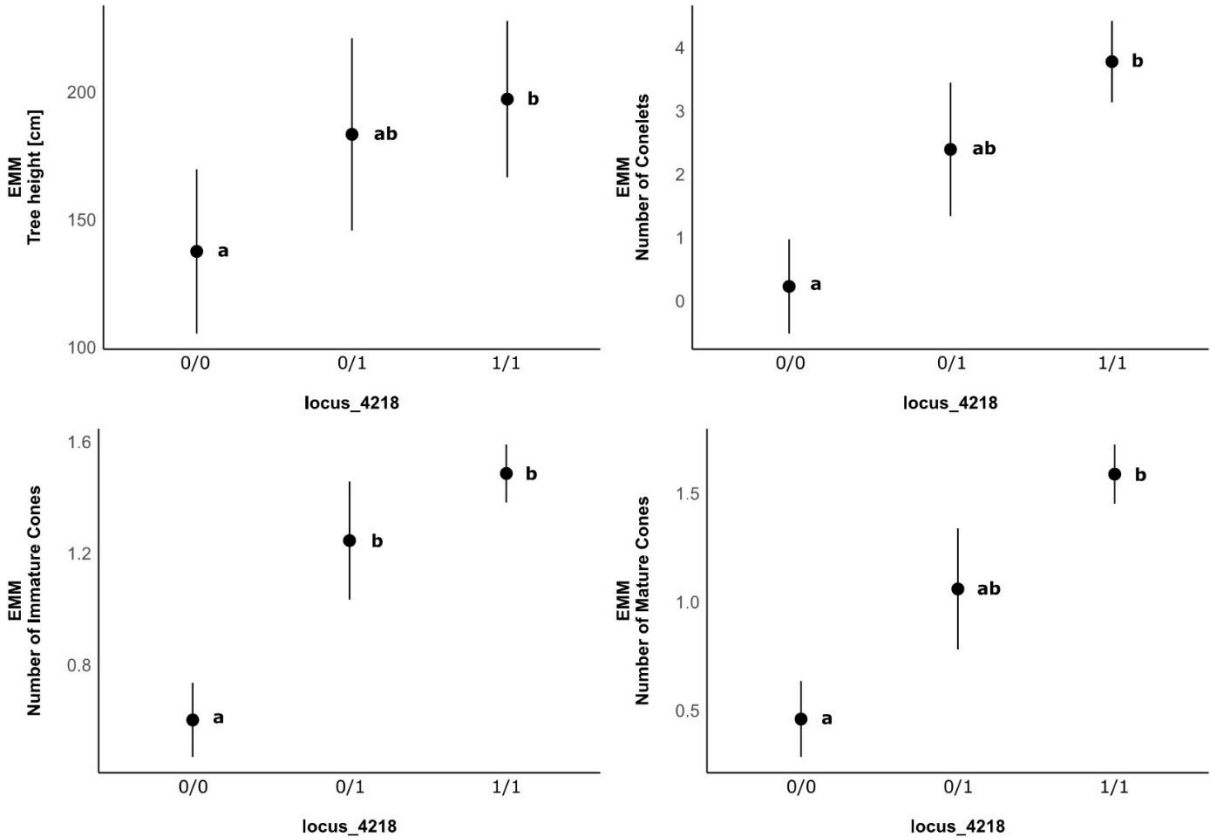


Figure 3. Expected Marginal Means (EMM) ± standard error of all four phenotypes investigated (y axis) given the genotype (x axis) of island, mainland, and hybrid individuals at locus_4218. Phenotypes measured include tree height (cm), number of conelets, number of immature cones, and number of mature cones. Distinct bolded lowercase letters indicate significant differences in phenotypes (adjusted $P < 0.05$) among genotypes (see Appendix S4 for details). Based on genotype frequencies (see Results), 0 was defined as the island allele, and 1 as the mainland allele.