Exome Resequencing Reveals Evolutionary History, Genomic Diversity, and Targets of Selection in the Conifers *Pinus taeda* and *Pinus elliottii*

Juan J. Acosta[1,6,†], Annette M. Fahrenkrog[1,2,†], Leandro G. Neves[1,2,7], Márcio F.R. Resende[3], Christopher Dervinis[1], John M. Davis[1], Jason A. Holliday[4], and Matias Kirst[1,2,5,*]

[1]School of Forest Resources and Conservation, University of Florida

[2]Plant Molecular and Cellular Biology Graduate Program, University of Florida

[3]Horticultural Sciences Department, University of Florida

[4]Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University

[5]University of Florida Genetics Institute, University of Florida

[6]Present address: Camcore, Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC

[7]Present address: RAPiD Genomics, Gainesville, FL

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: mkirst@ufl.edu.

Accepted: January 22, 2019

Abstract

Loblolly pine (*Pinus taeda*) and slash pine (*Pinus elliottii*) are ecologically and economically important pine species that dominate many forest ecosystems in the southern United States, but like all conifers, the study of their genetic diversity and demographic history has been hampered by their large genome size. A small number of studies mainly based on candidate-gene sequencing have been reported for *P. taeda* to date, whereas none are available for *P. elliottii*. Targeted exome resequencing has recently enabled population genomics studies for conifers, approach used here to assess genomic diversity, signatures of selection, population structure, and demographic history of *P. elliottii* and *P. taeda*. Extensive similarities were revealed between these species: both species feature rapid linkage disequilibrium decay and high levels of genetic diversity. Moreover, genome-wide positive correlations for measures of genetic diversity between the species were also observed, likely due to shared structural genomic constraints. Also, positive selection appears to be targeting a common set of genes in both pines. Demographic history differs between both species, with only *P. taeda* being affected by a dramatic bottleneck during the last glacial period. The ability of *P. taeda* to recover from a dramatic reduction in population size while still retaining high levels of genetic diversity shows promise for other pines facing environmental stressors associated with climate change, indicating that these too may be able to adapt successfully to new future conditions even after a drastic population size contraction.

Key words: *Pinus taeda*, *Pinus elliottii*, demography, genetic diversity, natural selection, exome.

Introduction

Loblolly pine (*Pinus taeda*) and slash pine (*Pinus elliottii*) dominate many forest ecosystems in the southern United States, spanning a wide variety of geographical terrains, climates, and environmental conditions (Nelson et al. 2013). Significant past demographic events are likely to have contributed to the extant genetic composition of these species. Existing *P. taeda* populations are likely derived from two refugia present during the Pleistocene glaciation: south Texas/northeast Mexico and south Florida/Caribbean, whereas *P. elliottii* was apparently present only in the latter refugium (Wells et al. 1991; Schmidtling and Hipkins 1998; Schmidtling 2003; Soltis et al. 2006). The Pleistocene glaciation probably caused a significant reduction in the size of southern pine populations, but the effect of this bottleneck (BN) on contemporary genetic diversity and the timing, magnitude, and the extent of

recovery of the population size have remained largely unknown. Such a BN may parallel current events that are dramatically reducing the size of several forest tree populations, such as that of lodgepole pine (*Pinus contorta*) in Western North America (Kurz et al. 2008). Thus, understanding the consequences of past genetic BNs for conifer forests may inform the future evolutionary impact of today's events.

A limiting factor for population genomics studies of pines and other conifers is their exceedingly large and complex genomes that contain abundant repetitive sequences. The genome size of loblolly and slash pines is estimated to be >21 Gbp (Ahuja and Neale 2005; Neale et al. 2014). Although the introduction of next-generation sequencing created the opportunity for development of the first draft of a pine genome sequence (*P. taeda*, Neale et al. 2014), efforts to extensively characterize DNA sequence variation by resequencing the entire genomes of large numbers of individuals are not yet feasible. Previous population genetics studies (Brown et al. 2004) and studies that have used coalescent simulation to uncover the evolutionary history of pine species (Heuertz et al. 2006; Pyhäjärvi et al. 2007; Wachowiak et al. 2011) were based on genotyping few nuclear loci. In recent years, resequencing the coding and regulatory regions of the large pine genomes has become a viable genotyping alternative (Neves 2013; Lu et al. 2016). Coding regions are often of primary interest for comparative genomics, population genetics, and molecular breeding. Targeted resequencing allows genotyping individuals from natural populations at a large number of markers across the coding part of the genome, offering the opportunity to uncover past demographic events using many loci, and to identify genes implicated in adaptation. A better understanding of the genetic basis of adaptation may be critical for survival of forest tree species in a scenario of climate change and increasing pressure from pathogens (Aitken et al. 2008).

Recently, exome genotyping was used to explore linkage disequilibrium (LD) and population structure in an association mapping population of *P. taeda* (Lu et al. 2016). This study found fast LD decay ($r^2 < 0.22$ within 55 bp) and the presence of two main genetic clusters across the species range separated by the Mississippi River, confirming findings from previous studies (Al-Rabab'Ah and Williams 2002; Soltis et al. 2006; Eckert, Liechty, et al. 2010; Eckert, van Heerwaarden, et al. 2010). Although the subdivision into two genetic clusters located west and east of the Mississippi River is the strongest genetic structure pattern in *P. taeda*, the eastern cluster is likely divisible into further clusters along a west to east axis (Eckert, Liechty, et al. 2010; Eckert, van Heerwaarden, et al. 2010; Lu et al. 2016). Unlike for *P. taeda*, to the best of our knowledge, there are no equivalent studies available for *P. elliottii*. An assessment of the feasibility of sequence capture in *P. elliottii* using oligonucleotide probes designed for *P. taeda* showed a high degree of transferability of probes between these species, with highly consistent capture efficiency of individual probes across samples (Neves et al. 2013). Interspecific capture was also applied successfully in an exome resequencing study in other pine and spruce species (Suren et al. 2016). Here, we explored the extensive transferability of sequence capture methods across related species to conduct a genome-wide survey of DNA polymorphisms in *P. taeda* and *P. elliottii*. We sequenced >10,000 genes for which most of the variation is explored (including low-frequency alleles), avoiding biases that are associated with preselection of polymorphisms for analysis. The main objectives of this study were to uncover the demographic events that shaped the evolutionary history of *P. taeda* and *P. elliottii*, to assess whether population structure is present in *P. elliottii*, to analyze the genetic diversity of these pine species and to identify genes putatively under positive selection (henceforth referred to as "candidate selection genes") for their possible role in adaptation.

Materials and Methods

Sample Collection, Library Preparation, and Sequence Capture

Seeds of loblolly pine (*P. taeda* L.) and slash pine (*P. elliottii* Engelm.) were collected from trees growing in natural stands distributed across the range of the species. Haploid genomic DNA was extracted from the megagametophyte tissue from 24 seeds from each species using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA). Each seed was obtained from trees located in distinct populations (fig. 1 and supplementary table S1, Supplementary Material online). Genomic DNA libraries were prepared following Bentley et al. (2008) with modifications. Genomic DNA was sheared using a Covaris E210 focused ultrasonicator (Covaris, Woburn, MA) to an average size of 250 bp. Sheared DNA was end-repaired using the End-It DNA End-Repair kit (Epicentre Biotechnologies, Madison, WI) producing blunt-end fragments and a single adenine was added to the 3′ end of each molecule using 15-U Klenow fragment (3′ to 5′ exo-; New England Biolabs Inc., Ipswich, MA) and 0.2-mM deoxyadenosine triphosphate (dATP) (Promega, Madison, WI). To allow multiplexing before sequence capture, the libraries were ligated to barcoded adapters described previously (Neves et al. 2013) using the Fast-Link DNA Ligation kit (Epicentre Biotechnologies). Following ligation, the libraries were amplified by eight cycles of polymerase chain reaction (PCR) using Illumina paired-end primers and Phusion High-Fidelity DNA Polymerase (New England Biolabs Inc.). The samples were purified using the Agencourt AMPure XP bead purification kit (Beckman Coulter, Brea, CA) after each step of the protocol. After PCR enrichment, the libraries were quantified with the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Waltham, MA). Individually barcoded libraries derived from eight megagametophytes were pooled, as described previously (Neves et al. 2013), for a total of three pools per species. A set of 54,773 120-mer RNA probes

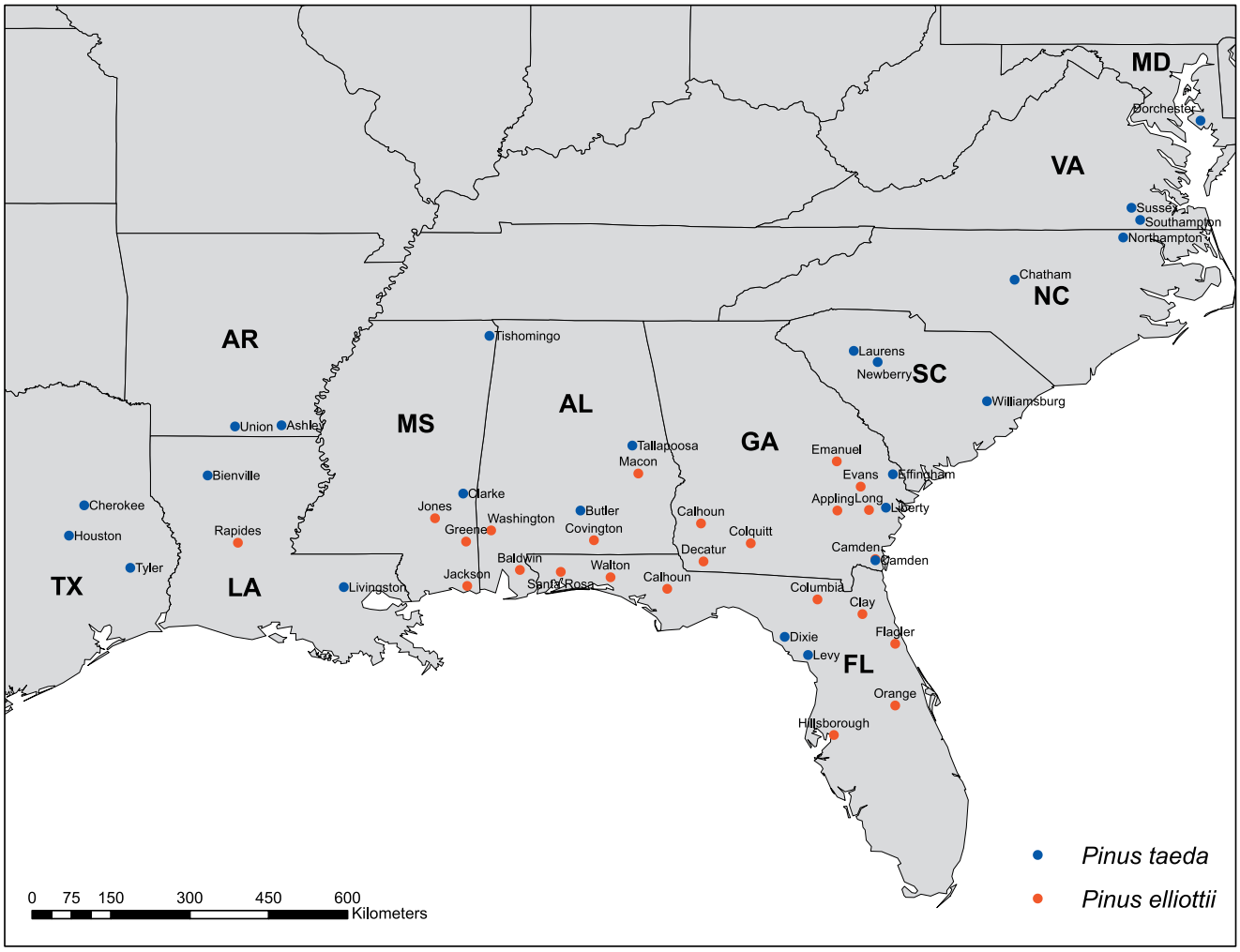

Fig. 1.—Seed collection sites for natural populations of *Pinus elliottii* (slash pine) and *Pinus taeda* (loblolly pine) in the southeastern United States.

previously designed based on an assembly of *P. taeda* expressed sequence tag (EST) sequences (Neves et al. 2013) was used to capture 14,729 genes, for a total target size of 6.57 Mbp. Multiplexed exome capture was performed following Agilent's Sure Select Target Enrichment (Agilent Technologies, Santa Clara, CA) protocol (Gnirke et al. 2009). Briefly, biotinylated probes and pooled libraries were mixed and hybridized at 65 °C for 24 h. The probe-DNA hybrids were then separated from nontarget DNA with magnetic streptavidin coated beads (Dynabeads, Invitrogen), eluted from the beads, and amplified by eight cycles of PCR prior to paired-end sequencing in individual lanes of a HiSeq2000 DNA Sequencing System (Illumina, San Diego, CA). Raw sequence data were submitted previously to the Short Read Achieve under study accession SRP018726 (Neves et al. 2013).

Identification of Genetic Variants

The sequencing reads obtained for each lane were processed using custom Perl and Python scripts. First, using FASTX-Toolkit version 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html; last accessed December 31, 2018), the reads from each individual sample were separated based on their barcode sequences, and assigned to the corresponding genotype. After barcode sequences were removed, reads were processed based on sequence quality, excluding bases from the 3′ end that had Phred scores below 20. In addition, reads shorter than 50 bases were discarded and only reads with a Phred score above 20 for 90% or more of the bases were maintained. Quality filtered reads were aligned to the reference genome using MOSAIK 2.2 (Lee et al. 2014) allowing 5% mismatch. The genome reference used was a combination of the original EST/UniGene-based reference and the de novo assembly for *P. taeda* created by Neves et al. (2013), which includes introns and 5′ and 3′ expansions for 11,386 genes. Single-nucleotide polymorphisms (SNPs) were discovered using FREEBAYES version 0.9.6 (Garrison and Marth 2012), with the following parameter settings: –pvar (reports sites if the probability that there is a polymorphism at the site is greater than specified) 0.90; –theta (expected mutation rate

or pairwise nucleotide diversity among the population under analysis) 0.01; –ploidy (ploidy for the analysis) 1; –min-alternate-fraction (fraction of observations supporting an alternative allele within a single individual in order to evaluate the position) 0.8; –min-alternate-count (count of observations supporting an alternative allele within a single individual in order to evaluate the position) 8; –min-coverage (required coverage to process a site) 10; –no-indels (ignore insertion and deletion alleles); –no-mnps (ignore multinucleotide polymorphisms). The SNPs identified above were used for assessment of LD, population structure, genetic diversity, and identification of genes under positive selection.

To investigate the extent of shared polymorphism and estimate demography, reads aligning to 1,000 genes present in the EST-based reference were retrieved from the alignment files from both species and realigned to the *P. taeda* v1.01 reference using MOSAIK 2.2. Variants were identified with FREEBAYES version 1.0.2 as described above. Shared polymorphism between species was determined by comparison of the polymorphic positions identified in both species. A subset of silent sites was generated for estimation of demography by annotating SNPs with VEP version 92 (Variant Effect Predictor; McLaren et al. 2016) and selecting variants annotated as intergenic, synonymous, upstream or downstream of a gene, and intronic variants.

LD and Population Structure

For each species, pairwise LD (r^2) between SNPs was calculated with PLINK 1.9 (Purcell et al. 2007) for each gene that contained two or more markers, using 65,325 SNPs (out of 67,071 SNPs) and 55,263 SNPs (out of 56,144 SNPs) for *P. taeda* and *P. elliottii*, respectively (supplementary table S3, Supplementary Material online). LD decay with physical distance within genes was estimated as described elsewhere (Remington et al. 2001; Marroni et al. 2011). Population structure was assessed using principal component analysis (PCA) implemented in the R package SnpRelate (Zheng et al. 2012). SNPs in LD were pruned, removing one SNP of a pair showing an r^2 above 0.2. This procedure resulted in the selection of 26,577 and 24,524 SNPs for population structure analysis in *P. taeda* and *P. elliottii*, respectively (supplementary table S3, Supplementary Material online).

Genetic Diversity

The program DNASAM version 20100621 (Eckert, Liechty, et al. 2010) was used to estimate genetic diversity (number of segregating sites, S; Watterson estimator, θ_W; and nucleotide diversity, π) and the summary statistics Tajima's D (Tajima 1989), Fay and Wu's H (Fay and Wu 2000), and Wall's B (Wall 1999). To generate the multiple sequence alignments required by DNASAM as input files, the "genome sequence" for each individual of the two species was generated by modifying the reference genome sequence replacing the reference alleles by the variants detected in a given sample. These sequences were then used to generate one file per contig containing the sequences from all samples from a given species. Outlier loci for D, H, and B were obtained selecting the genes in the top 1% of the empirical distribution for these test statistics (negative tail of the distribution for D and H; positive tail of the distribution for B). A subset of genes could be placed on the 12 main linkage groups corresponding to the *P. taeda* chromosomes using a gene map available for this species (Neves et al. 2014). Nucleotide diversity (π and θ_W) and summary statistics (D, H, and B) were plotted along linkage groups for these genes using the software CIRCOS version 0.69 (Krzywinski et al. 2009).

Functional Annotation of Genes

Gene annotation was carried out with BLAST2GO version 4.1.1 (Götz et al. 2008) for the *P. taeda* EST/UniGene-based reference used for alignment and SNP calling. The steps performed by the BLAST2GO program were BLAST against the Viridiplantae nr database using an E-value threshold of 1×10^{-3}, mapping to retrieve gene ontology (GO) terms for the BLAST hits, and annotation to select GO terms from the GO term pool collected for a sequence during the mapping step. This process retrieved annotations for 3,424 contigs corresponding to 2,856 genes. Sequences still lacking annotation were aligned to the *P. taeda* transcriptome (Pita.IU.TranscriptomeMainsV1.102013.fasta file downloaded in August 2017 from https://treegenesdb.org/FTP/Genomes/Pita/v1.01/transcriptome/) using BLAST version 2.6.0 (Altschul et al. 1990) with an E-value threshold of 1×10^{-3}, and annotations were retrieved from the functional annotation file available for the transcriptome (pita.-transcriptome.functional_annotation.tsv downloaded in August 2017 from https://treegenesdb.org/FTP/Genomes/Pita/v1.01/transcriptome/).

Simulation of Demographic Events and Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) analyses were applied to evaluate the consistency of the empirical estimates of genetic diversity and summary statistics under several demographic models (Beaumont et al. 2002; Beaumont 2010; Lopes and Beaumont 2010; Baragatti and Pudlo 2014). The summary statistics θ_W, π, Tajima's D, Wall's B, Fay and Wu's H, and Rozas' ZA were obtained for both species with the R package PopGenome (Pfeifer et al. 2014) using silent SNPs (supplementary table S2, Supplementary Material online). For each species, sequences of the same length as the contigs harboring the silent sites (359 and 290 contigs in *P. taeda* and *P. elliottii*, respectively) were simulated in 24 haploid genomes under three demographic models: neutral equilibrium (NE), BN, and population growth (GR), following the methods of Holliday et al. (2010). For each of these models,

expected mutation rate (θ_W) and recombination parameters (ρ) were drawn from uniform prior distributions with range [0, 0.01]. The NE model included only these parameters, whereas the BN model contained three additional parameters: 1) timing of the BN, 2) reduction in effective population size (BN severity), and 3) durations of the BN, which were drawn from log-uniform priors with ranges [0.0001, 0.1]. The GR model also included two additional parameters in comparison to the NE model: 1) the population was assumed to have begun growing at a time in the past drawn for log-uniform prior distribution with range [0.0001, 0.1] and 2) with an ancestral population size drawn for log-uniform prior distribution with range [0.0001, 0.5]. Coalescent simulations were performed using MLCOALSIM v2 (Ramos-Onsins and Mitchell-Olds 2007) executing a total of 50,000 simulations for each demographic model. Subsequent ABC analyses were conducted using the following R scripts provided by Dr Mark Beaumont (Beaumont et al. 2002): "calmod.R" for model selection; "make_2pd.R" for ABC; and "loc2plot_d.R" for estimating posterior modes and highest posterior density intervals (Holliday et al. 2010). A tolerance level of $\delta = 0.05$, corresponding to the 2,500 closest points, was used for the ABC step. Summary statistics used for ABC included the means for θ_W, π, Tajima's *D*, Wall's *B*, Fay and Wu's *H*, and Rozas' ZA, as well as their standard deviations.

Results

Sequence Capture of Haploid Genomic DNA from Pine Megagametophytes

The genetic diversity of loblolly and slash pine was characterized by sequence capture, using a set of probes that were shown previously to be suitable for analysis of their megagenomes (Neves et al. 2013). A significant challenge in the analysis of genetic polymorphisms from next-generation sequencing data is the detection of heterozygous loci and determination of gametic phase. To avoid these difficulties, we sequenced genomic DNA from haploid tissue obtained from seed megagametophytes. On average, we obtained 18.3- and 21.0-M reads for each megagametophyte of *P. taeda* and *P. elliottii*, respectively, of which 34% had similarity to the probes used for sequence capture. This resulted in the capture of 11,695 and 11,591 genes in *P. taeda* and *P. elliottii*, respectively (supplementary fig. S1, Supplementary Material online). Sequencing depth by gene was highly correlated in both species (Pearson correlation coefficient = 0.997). A set of 474 genes with mean sequencing depth above 100× in both species was removed from analysis because they most likely correspond to repetitive regions. This resulted in a final set of 11,221 and 11,117 genes surveyed in *P. taeda* and *P. elliottii*, respectively. The mean (median) sequencing depth per gene was 13.74× (9.19×) in *P. taeda* and 14.95× (9.98×) in *P. elliottii*. Despite the high genetic diversity expected for these species (Brown et al. 2004) and the small fraction of the genome being captured by probes (6.6 Mb/21,700 Mb = 0.03%), the reasonable capture efficiency suggests that this approach is suitable for analysis of very large, complex genomes.

LD Decay and Population Structure

To identify SNPs in the 48 haploid pine samples, captured sequences were aligned to the reference genome for the corresponding species generated previously (Neves et al. 2013). A total of 67,071 and 56,144 biallelic SNPs were identified in *P. taeda* and *P. elliottii*, respectively (supplementary table S3, Supplementary Material online), with about one quarter of them (25.2% and 24.7% in *P. taeda* and *P. elliottii*, respectively) being polymorphisms with minor allele frequency below 0.05.

LD decay with physical distance was estimated by gene in both pine species. The maximum distance between markers was given by the length of the contigs conforming the reference genomes, with a mean (maximum) distance of 195 bp (1,506 bp) between markers for *P. elliottii* and 193 bp (1,303 bp) for *P. taeda*. In *P. taeda*, LD decayed below an r^2 threshold of 0.2 after 123 bp, slower than reported previously for this species when using a much larger population and greater number of SNPs (55–79 bp; Lu et al. 2016). Very similar results were obtained for *P. elliottii*, where LD decayed below an r^2 threshold of 0.2 after 134 bp.

In *P. taeda*, population structure has been shown to be driven by the Mississippi River discontinuity (Al-Rabab'Ah and Williams 2002; Eckert, van Heerwaarden, et al. 2010; Lu et al. 2016). Here, we found the same pattern, with eigenvector 1 separating the samples collected west of the Mississippi River from those collected east of the river (fig. 2*A*). We believe that the limited sample size of our study did not allow resolving finer patterns of genetic structure, but it was sufficient to reveal a major division such as the genetic structure caused by the expansion of *P. taeda* from two different refugia after the last glaciation. In contrast, *P. elliottii* likely remained in one refugium located in south Florida and the Caribbean during the last glaciation, and its native range does not extend west of the Mississippi River (Schmidtling 2003). Most of the samples analyzed in this study were collected east of the Mississippi River, with only one sample collected west of the river in the state of Louisiana (LA). The LA collection site that falls outside the native range of the species, but *P. elliottii* now reproduces naturally in this region after being introduced there in commercial plantations (https://www.srs.fs.usda.gov/pubs/misc/ag_654/volume_1/pinus/elliottii.htm; last accessed December 31, 2018). PCA revealed this sample to be an outlier with respect to the remaining population (fig. 2*B*). The reason why this sample appears as an outlier is unknown, but we can speculate that either the material planted in LA came from a genetically distinct population not sampled in our study or that this sample shows a

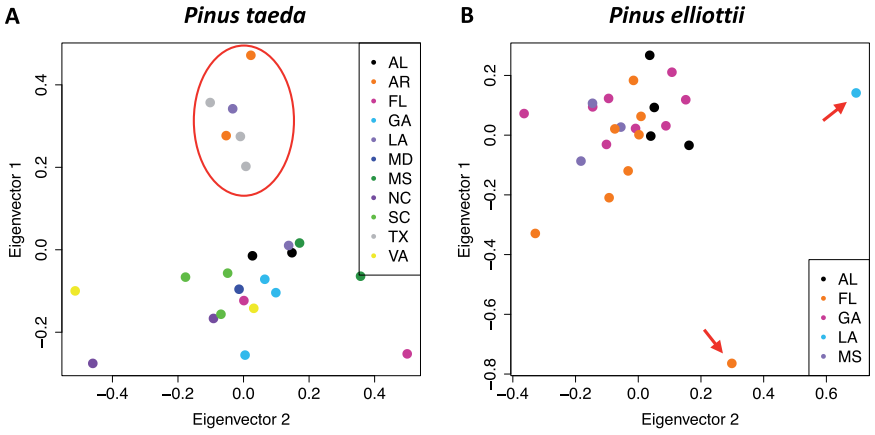

Fig. 2.—Population structure of two southern pine species assessed with PCA. Each point on the plots represents one sample colored by the state where it was collected. (*A*) *Pinus taeda* (loblolly pine). A cluster of samples collected west of the Mississippi River is marked with a red oval. (*B*) *Pinus elliottii* (slash pine). Two outlier samples are marked with red arrows.

certain degree of hybridization with other species. A second outlier was found among the samples collected in the state of Florida. This corresponds to the most southern sample included in this study, collected in the Tampa Bay area. This sample could correspond to a *P. elliottii* var. *densa* tree, a variety that grows in south Florida (https://www.srs.fs.usda.gov/pubs/misc/ag_654/volume_1/pinus/elliottii.htm; last accessed December 31, 2018). The remaining samples most likely belong to *P. elliottii* var. *elliottii*, the variety growing from Central Florida towards the north (https://www.srs.fs.usda.gov/pubs/misc/ag_654/volume_1/pinus/elliottii.htm; last accessed December 31, 2018). Further studies including a larger number of samples from south Florida are needed to clearly elucidate the genetic structure present across *P. elliottii*'s native range.

Genetic Diversity and Signatures of Selection

Genetic diversity estimates provide a foundation for the identification of loci that may be under natural positive selection and, therefore, relevant for novel adaptions of *P. elliottii and P. taeda*. Although in this study the number of individuals sampled across geographic locations was limited, it provides an initial assessment of which loci display patterns of diversity that suggest the action of positive selection. Furthermore, the availability of data for two distinct species allows for a genome-wide comparative analysis by correlating estimates of population genetic parameters across loci. The mean nucleotide diversity (π) observed in *P. elliottii and P. taeda* was 0.00132 and 0.00136, respectively (table 1 and fig. 3). Watterson's estimator of nucleotide diversity (θ_W) was 0.00136 for *P. elliottii* and 0.00140 for *P. taeda* (table 1 and fig. 3). Estimates of θ_W were slightly higher than values reported for nonsynonymous sites in a previous study assessing genetic diversity *P. taeda* using SNPs in 19 nuclear loci ($\pi = 0.00114$, $\theta_W = 0.00108$; Brown et al. 2004). The Spearman's rank correlation coefficient of nucleotide diversity by gene between species was positive and highly significant for both diversity measures (π: $\rho = 0.36$, *P* value $< 2.2e\text{-}16$; θ_W: $\rho = 0.38$, *P* value $< 2.2e\text{-}16$) for a subset of 6,798 genes with these statistics available in both species (supplementary fig. S2, Supplementary Material online).

Three summary statistics were estimated for each species to identify genes putatively under selection. The first statistic, Tajima's *D* (henceforth referred to as *D*), is a normalized version of the difference between π, based on the average number of nucleotide differences between sequences, and θ_W, based on the total number of segregating sites (Tajima 1989). It is most influenced by low-frequency variants. Loci carrying deleterious mutations that are kept at low frequency in the population by purifying selection and loci where new mutations appear after an allele has been fixed or its frequency increased drastically by a selective sweep will show a negative value for *D* (Tajima 1989; Nielsen 2005; Hartl and Clark 2007). *D* can also reflect demographic processes. When population size is expanding, for example after a BN, low-frequency mutations will increase relative to neutral expectation generating negative *D* values (Tajima 1989). On the other hand, recent admixture between populations, balancing selection or diversifying selection will decrease low and high-frequency mutations, generating positive *D* values (Hartl and Clark 2007). The mean values obtained for *D* were −0.084 for *P. elliottii* and −0.083 for *P. taeda* (fig. 3 and table 1). Spearman's rank correlation coefficient was estimated for *D* between species. Although not as high as for the diversity metrics, the correlation was also positive and highly significant (6,798 genes; $\rho = 0.19$, *P* value $< 2.2e\text{-}16$, supplementary fig. S2, Supplementary Material online). For a more specific comparison between species, we selected D outlier loci (top 1% negative values of the empirical distribution

Fig. 3.—Genome-wide distribution of genetic diversity and summary statistics for *Pinus elliottii* (slash pine) and *Pinus taeda* (loblolly pine). Before plotting, a loess smoothing function was applied to the data for a subset of genes placed on a *P. taeda* genetic map. Black lines represent the karyotype composed of 12 linkage groups (LG) and the length of each LG is shown in centimorgan (cM). A total of 2,252 genes was included to plot nucleotide diversity (π, innermost or first track), Watterson's estimator of nucleotide diversity (θ_W, second track), Tajima's *D* (third track, placed outside the karyotype), and Fay and Wu's *H* (fourth track). The Wall's *B* plot (outermost or fifth track) includes data from 1,774 genes. The gray line shown in tracks three and four represents a value of zero.

corresponding to 95 and 87 genes in *P. elliottii* and *P. taeda*, respectively) and searched for overlapping genes likely under positive selection. Three genes in common between the species were identified, a greater overlap than expected by chance (*P* value = 0.0125; supplementary table S4, Supplementary Material online).

The second summary statistic calculated, Fay and Wu's *H* (henceforth referred to as *H*), is based on the difference between π and θ_H, where θ_H is an estimator of θ weighted by the homozygosity of the derived (nonancestral) variants (Fay and Wu 2000). *H* is greatly influenced by high-frequency variants, which increase with positive selection due to hitchhiking. An excess of high-frequency derived variants will cause negative values of *H*, indicating positive selection (Fay and Wu 2000; Innan and Kim 2004; Wright and Gaut 2005). *H* can also be affected by demography, showing negative values under a scenario of population shrinkage (Zeng et al. 2006). Compared with *D*, both statistics (*D* and *H*) will reject neutrality in the presence of an excess of both low- and high-frequency variants, a scenario that occurs shortly before and

Table 1
Genome-Wide Observed Population Genetic Parameters in *Pinus taeda* and *Pinus elliottii*

Parameter		*Pinus taeda*	*Pinus elliottii*
π	μ	0.00136	0.00132
	σ	0.00120	0.00119
θ_W	μ	0.00140	0.00136
	σ	0.00110	0.00109
Tajima's *D*	μ	−0.08316	−0.08444
	σ	1.02132	1.01389
Fay and Wu's *H*	μ	−0.23320	−0.3195
	σ	1.34153	1.53563
Wall's *B*	μ	0.12340	0.12200
	σ	0.24418	0.23546

after a linked beneficial mutation reaches fixation during a selective sweep (Fay and Wu 2000; Zeng et al. 2006). When only low-frequency variants are in excess, which happens during the phase when genetic variation is recovered after fixation of a selected allele, *D* but not *H* will reject neutrality. Conversely, when only high-frequency variants are in excess, for example, after population shrinkage, neutrality will be rejected by *H* but not *D* (Fay and Wu 2000; Zeng et al. 2006). The mean values obtained for *H* were −0.3195 and −0.2332 for *P. elliottii* and *P. taeda*, respectively (fig. 3 and table 1). This statistic also showed a positive and highly significant correlation between species (6,798 genes; Spearman's $\rho = 0.38$, *P* value < 2.2e-16, supplementary fig. S2, Supplementary Material online). Twelve outlier loci likely under positive selection were identified in common in both pines based on this statistic (*P* value = 0.0125; supplementary table S4, Supplementary Material online).

The third summary statistic analyzed, Wall's *B* (henceforth referred to as *B*), is a measure of LD among segregating sites. It ranges between 0 and 1, with high values indicating greater LD between sites and departure from the standard neutral model caused by balancing selection or population subdivision (Wall 1999). *B* was included in this study as a measure of LD within the genes analyzed. For this statistic, mean values of 0.1220 and 0.1234 were obtained for *P. elliottii* and *P. taeda*, respectively (fig. 3 and table 1). The correlation of *B* between species was also significantly positive (4,729 genes; Spearman's $\rho = 0.15$, *P* value < 2.2e-16, supplementary fig. S2, Supplementary Material online).

For a given species, the genes identified as outliers by two neutrality tests may be stronger candidates for positive selection/adaptation than those identified by a single test. The statistics *D* and *H* utilized here are complementary because they are both able to detect positive selection while being affected by different demographic processes. For example, *D* is most sensitive to GR and less sensitive to population shrinkage, whereas the opposite is true for *H* (Fay and Wu 2000; Zeng et al. 2006). Thus, the overlap between *D* and *H* within species was assessed (supplementary table S5, Supplementary Material online), revealing 10 and 13 overlapping genes for *P. taeda* and *P. elliottii*, respectively. Finally, the overlap between outliers for different statistics between species was investigated, revealing three additional candidate genes (supplementary table S5, Supplementary Material online). All possible overlaps between the summary statistics *D* and *H* within and between species revealed a set of 16 candidate genes probably under positive selection in parallel in *P. taeda* and *P elliottii* (supplementary table S6, Supplementary Material online). Six out of these 16 candidate selection genes were identified by more than one overlap between statistics. We were able to obtain functional annotations for three of these six genes, with the first gene (0_17165) predicted to encode CUT1, an enzyme involved in cuticular wax production and pollen fertility (Millar et al. 1999). The second gene (2_10367) encodes a probable chloroplastic anion transporter and the third gene (2_5963) encodes a probable BAH1-like E3 ubiquitin-ligase involved in the accumulation of salicylic acid and immune responses in plants (Yaeno and Iba 2008).

Additionally, a set of nine genes showed signatures of positive selection in *P. elliottii* only (supplementary table S7, Supplementary Material online), whereas six genes were identified as selection candidates in *P. taeda* only (supplementary table S7, Supplementary Material online). Each of these genes was identified by a single overlap between statistics.

Simulation of Demographic Events and ABC

Coalescent simulations under three different demographic models—BN, GR, and NE—were evaluated. Observed and estimated population genetics parameters were used to select the most likely demographic model. Estimated posterior probabilities for each demographic scenario were for BN: 0.045 and 0.721 for *P. elliottii and P. taeda*, respectively; GR: 0.950 and 0.270; and NE 0.005 and 0.009. Consequently, the demographic scenario that fits best the observed data for *P. taeda* is the BN model, whereas the best fit for *P. elliottii* is the GR model. For *P. taeda*, posterior distributions of the BN model parameters "timing of BN" and "reduction in effective population size (N_e)" are unimodal and bell shaped (supplementary fig. S3, Supplementary Material online), and their expected values (posterior mode for "timing of BN" = 0.012; "reduction in N_e" = 0.001) indicate that this species experienced a drastic reduction in N_e during the BN event that occurred $0.012 \times 4N_e$ generations ago, where only 0.1% of the ancestral N_e survived. The time in the past when the BN affected *P. taeda* was estimated using the posterior mode of the simulated parameter "timing of BN" (0.012), and expressed in the time at which it occurred in terms of $4N_e$ generations (Holliday et al. 2010). As $\theta_W = 4N_e\mu$, we first estimated an N_e of ~17,000 using the posterior mode of θ_W for *P. taeda* (0.0023), a per year mutation rate of 1.3×10^{-9} (Willyard et al. 2007), and a generation time of 25 years ($0.0023/4 \times 1.3 \times 10^{-9} \times 25 = 17{,}692$). Using this estimate

of N_e, we dated the BN event that drastically reduced the US southeastern population of *P. taeda* to ~21,000 years before present (BP; $0.012 \times 4 \times 17{,}692 \times 25 = 21{,}230$).

For *P. elliottii*, posterior distributions of the growth parameters "ancestral N_e" and "timing that growth started" are also unimodal and bell-shaped (supplementary fig. S3, Supplementary Material online), with their expected values (posterior mode for "ancestral N_e" = 0.12; "timing that growth started" = 0.034) indicating that growth started $0.034 \times 4N_e$ generations ago from a population whose N_e was 12% the N_e of the current population. The time in the past when GR started was estimated using the posterior mode of the parameter "timing that growth started" (0.034), the posterior mode of θ_W (0.0011) and an estimate of N_e of ~8,000 ($0.0011/4 \times 1.3 \times 10^{-9} \times 25 = 8{,}462$). Assuming the same per year mutation rate and generation time as for *P. taeda*, it was estimated that growth started ~28,000 years BP ($0.034 \times 4 \times 8{,}462 \times 25 = 28{,}771$).

Discussion

We used targeted resequencing to genotype samples from two southern pine species, *P. taeda* and *P. elliottii*, with the objectives of understanding their evolutionary history and analyzing their genetic diversity. We previously showed that interspecific analysis of *P. elliottii*, including the identification of polymorphic sites, could be achieved using sequence capture probes designed for *P. taeda* (Neves et al. 2013). These probes were used here for targeted resequencing of the coding fraction of the genome in both species, allowing the identification of genetic polymorphisms in ~11,000 genes in *P. taeda* and *P. elliottii*. Given the success of capture in *P. elliottii*, these probes may also be suitable for analysis of more distantly related species.

One of the main challenges of probe design for conifers is the abundance of repetitive regions (Neves et al. 2013), estimated to correspond to 82% of the *P. taeda* genome (Neale et al. 2014). As shown by analysis of sequencing depth, the probes used here mainly retrieved low-copy genomic regions, with only 3.2% of the captured genes being excluded from downstream analyses for showing depth $\geq 100\times$ in both species. Repetitive regions and paralogs also pose a challenge to the identification of genetic polymorphisms from sequencing data (Kumar et al. 2012). In diploid samples, reads from paralogous regions aligning to the same gene in the reference sequence will generate erroneous heterozygous calls. This risk was avoided in this study by using haploid genomic DNA extracted from seed megagametophytes.

A great advantage of targeted resequencing-based genotyping methods over genotyping with SNP arrays is the lack of ascertainment bias in marker selection (Gilissen et al. 2011). Ascertainment bias in SNP genotyping is caused by the identification of markers in a small population during SNP array development, which may not be representative of the genetic diversity and allele frequency spectrum present in the species as a whole. Subsequent genotyping of a different population leads to biased allele discovery, usually favoring higher-frequency alleles (Albrechtsen et al. 2010). With few exceptions (Jaramillo-Correa et al. 2015; Lu et al. 2016), most of the population genetics studies reported to date in pine species or other conifers have been based on the analysis of very few loci (Brown et al. 2004; Pyhäjärvi et al. 2007; Soto et al. 2010; Wachowiak et al. 2011; Y. Zhou et al. 2014) or on genotypes obtained with SNP arrays (Prunier et al. 2011; Mosca et al. 2014, 2016). The ascertainment bias likely present in the latter data sets could distort estimated population genetics parameters and, therefore, the conclusions drawn from them (Albrechtsen et al. 2010). Our approach minimized this drawback.

In this study, exome resequencing allowed the identification of a large number of SNPs in *P. taeda* and *P. elliottii* (67 and 56 K, respectively). These SNPs were used to assess LD decay and population structure in both species, revealing patterns in agreement with previous reports for *P. taeda* (Al-Rabab'Ah and Williams 2002; Soltis et al. 2006; Eckert, Liechty, et al. 2010; Eckert, van Heerwaarden, et al. 2010; Lu et al. 2016). It should be noted, however, that estimating LD on the basis of relatively short resequenced genes, rather than full genome data, may overestimate the rate of LD decay. Although early studies of LD in forest trees suggested that it decayed very rapidly, at the level of individual genes (e.g., Brown et al. 2004), whole-genome resequencing in *Populus trichocarpa* revealed LD over much greater distances than was previously believed to be the case for temperate trees (Slavov et al. 2012).

Characterization of genetic diversity in the coding fraction of the *P. taeda* and *P. elliottii* genomes revealed that the level of diversity measured by π and θ_W is similar in both species and falls within the lower range previously reported for genes in outcrossing, perennial species of the genera *Populus*, *Pinus*, and *Amborella* (Brown et al. 2004; Keller et al. 2010; Olson et al. 2010; Amborella Genome Project 2013; Wachowiak et al. 2013; L. Zhou et al. 2014). The observed negative Tajima's *D* values in genic regions in both species indicate an excess of low-frequency polymorphisms relative to neutral expectation, suggesting population size expansion after a BN, positive selection and/or purifying selection. Overall negative Tajima's *D* in genic regions has been reported in other tree species like poplar (L. Zhou et al. 2014), consistent with purifying selection acting against mutations causing alterations in protein sequence or silent mutations affecting codon usage. Because SNPs were identified using relatively stringent criteria, true variants may have been excluded from analysis. Consequently, the results presented here may underestimate nucleotide diversity. However, this underestimation does not affect the general conclusions. Instead, it supports the hypothesis of recent GR,

positive selection or of purifying selection given that including more low-frequency polymorphisms would make the estimation of Tajima's *D* more negative.

In addition to studying genetic diversity and signatures of selection, we investigated the demographic history of *P. taeda* and *P. elliottii* using coalescent simulations and ABC. Three potential scenarios were evaluated (NE, GR, and BN) identifying differences in the demographic histories of both species: signatures of a BN were detected in *P. taeda*, whereas *P. elliottii* showed only evidence of growth. It was estimated that the BN event that drastically reduced the US southeastern population of *P. taeda* occurred ~21,000 BP. This estimate is coherent with the occurrence of the late Wisconsin glacial interval 30,000–11,000 years BP, with the last glacial maximum dated between 19,000 and 23,000 years BP (Fisher et al. 2010; Schmittner et al. 2011). Although some previous studies of temperate trees dated BNs to much earlier timeframes, on the order of ~200,000 years BP (e.g., Heuertz et al. 2006; Pyhäjärvi et al. 2007; Ingvarsson 2008; Holliday et al. 2010), our results agree with other studies in conifers that detected a BN event occurring during the last glacial period (Grivet et al. 2009; Wachowiak et al. 2011). For example, in the Mediterranean conifer Aleppo pine, demography estimates identified the signatures of a BN that drastically reduced the population ~18,000 years before present (Grivet et al. 2009). The agreement of this estimate to ours is of particular interest because, just like Aleppo pine, *P. taeda* is adapted to a warm climate. The discrepancy between studies may be related to the variation in how climatic processes affected individual species at particular times, and the signature those processes left in the genome, as well as our ability to account for signatures of these events through the particular genomic and population sampling strategy employed (e.g., resequencing of a small number of gene fragments vs. exome capture). Nevertheless, our results confirm the demographic history suggested by genome-wide negative Tajima's *D* values and reveal that *P. taeda*'s population is expanding after surviving a drastic BN that reduced its population to about 0.1% of its N_e. This result also agrees with previous studies in *P. taeda* based on a small number of nuclear loci, where lower than expected genetic diversity given its life history traits, and demographic history shaped by at least one BN were reported (Brown et al. 2004; Ersoz et al. 2010). The existence of a severe BN in the demographic history of *P. taeda* reveals a similar demographic phenomenon to what was previously observed in Iberian pine species adapted to warm climates, which underwent strong demographic fluctuations during glacial cycles as a consequence of their maladaptation to cold (Soto et al. 2010).

For *P. elliottii*, it was estimated that GR started ~28,000 years BP. The time estimates presented here are only approximate because they are based on an estimate of N_e that is prone to errors for depending on assumptions like the mutation rate value (Pyhäjärvi et al. 2007). Our estimates indicate that *P. elliottii*'s population started growing during the last glaciation, although it is more likely that growth began after the last glacial maximum. The absence of evidence supporting a BN in *P. elliottii* suggests that this species was abundant in south Florida and the Caribbean prior to the last glaciation, where it remained in a sufficiently large population during glacial periods to presently show only evidence of growth.

Summary statistics reflect the demographic history of a species, as seen above, but they can also reveal genes targeted by positive selection (Biswas and Akey 2006). Genes showing extreme values in the distribution of a summary statistic are more likely to be outliers with possible roles in adaptation (Savolainen et al. 2013). Here, we identified candidate genes putatively under positive selection based on the empirical distributions for Tajima's *D*, Fay and Wu's *H*, and Wall's *B*. Especially, interesting selection candidates are 47 genes showing signatures of positive selection in both pine species or identified as outliers by more than one method within a species (supplementary table S4, Supplementary Material online). Many of these genes are putatively involved in response to the environment, including biotic and abiotic stress response, immune response, hormone response, protein degradation, intracellular transport, and signaling, among others. Molecular characterization of these candidate genes is needed to confirm their function and gain insight into the mechanisms by which they may favor adaptation of *P. elliottii* and *P. taeda* to their environment.

Although a reduced number of candidate genes for selection (29 genes, supplementary table S4, Supplementary Material online) were identified in common in both southern pines, significantly positive genome-wide correlations of nucleotide diversity and summary statistics were detected between species. A similar finding was recently reported for three poplar species, where the correlation of nucleotide diversity was attributed to conserved mutation rates across the genome and/or shared selective constraints among species (Wang et al. 2016). Similarly, a recent study conducted on phylogenetically distant bird species reported the correlation of genome-wide patterns of genetic diversity between species to be driven by local levels of recombination rate conserved in syntenic regions and associated with certain genomic features of low recombination (e.g., centromeric regions; Vijay et al. 2017). These genomic regions of low recombination may exhibit linked selection (reduction of neutral genetic variation linked to loci under positive or negative selection) not related to adaptation (Vijay et al. 2017). Other possible explanations for a positive correlation of nucleotide diversity and summary statistics between species are a similar demographic history (Wang et al. 2016) and shared ancestral genetic variation (Vijay et al. 2017). All of the above are plausible explanations for the correlation of genetic diversity and summary statistics among pine species observed here. Due to the lack of an assembled reference genome for pines, we were not able to assess if the correlation of summary statistics

between species is driven by specific genomic features. Also, based on their divergence time, it is possible for *P. elliottii* and *P. taeda* to share polymorphisms. The ancestors of these pine species diverged in the early Pliocene, ~5 Myr BP (Hernández-León et al. 2013). Assuming a generation time of 25 years (Brown et al. 2004), divergence occurred about 200,000 generations ago (5,000,000 years/25 years/generation = 200,000 generations) or 11.76N_e generations ago (calculated for *P. taeda*, N_e = 17,000). This divergence time is more recent than the time required to prevent species from sharing ancestral genetic variation (9–-12N_e generations; Hudson and Coyne 2002; Vijay et al. 2017). To assess the extent to which both species share polymorphisms, we realigned reads extracted from the alignments to 1,000 genes present in our EST-based reference to the *P. taeda* v1.01 genome and, after calling SNPs, identified a set of shared variants corresponding to 15.9% of the SNPs identified in *P. elliottii* (305/1,915) and 13.3% of *P. taeda* SNPs (305/2,299). Thus, both species share a considerable proportion of genetic variation, explaining in part the significantly positive genome-wide correlations of nucleotide diversity and summary statistics observed.

In summary, this study revealed extensive similarities between *P. elliottii* and *P. taeda*, two southern pine species with largely overlapping native ranges. Both species feature similar LD decay and similar levels of genetic diversity but differ in their demographic history. Their similarities and their need to adapt to a shared environment have resulted in genome-wide positive correlations of measures of genetic diversity and summary statistics between species. Also, positive selection appears to be targeting a common set of genes in both species, either through selection of shared standing genetic variants (parallelism at the nucleotide level) or through selection acting upon the same set of genes without sharing variation (parallelism at the gene or pathway level, Yeaman et al. 2016). This study, combining population genomics with an assessment of demographic history, revealed that one of this southern pine species, *P. taeda*, was able to recover from a dramatic reduction in N_e during the last glacial period and still harbors high levels of genetic diversity. This is promising for other pine species facing challenges like climate change and pathogen attack today, indicating that they may be able to overcome these challenges as well and adapt successfully to novel future conditions.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

The Cooperative Forest Genetics Research Program (CFGRP) at the University of Florida collected the seeds for *P. elliottii*, whereas *P. taeda* seeds were provided by ARBORGEN Inc. We acknowledge financial support from the US Department of Agriculture National Institute of Food and Agriculture (USDA/NIFA) awards 2013-67009-21200 and 2013-67013-21159 to M.K. We also acknowledge all the staff and students from the Forest Genomics Laboratory at the University of Florida for the help with the data collection.

Author Contributions

J.J.A. performed the research, data collection, data analysis, and interpretation and wrote the article; A.M.F. performed the research, data analysis, and interpretation and wrote the article; L.G.N. performed the research, data collection, data analysis, and interpretation; M.F.R.R. performed data analysis and interpretation; C.D. designed and performed the research and data collection; J.A.H. planned and designed the research and performed data analysis and interpretation; M.K. planned and designed the research, performed data analysis and interpretation, and wrote the article.

Literature Cited

Ahuja M, Neale D. 2005. Evolution of genome size in conifers. Silvae Genet. 54(1–6):126–137.

Aitken SN, Yeaman S, Holliday JA, Wang T, Curtis-McLane S. 2008. Adaptation, migration or extirpation: climate change outcomes for tree populations. Evol Appl. 1(1):95–111.

Al-Rabab'Ah MA, Williams CG. 2002. Population dynamics of *Pinus taeda* L. based on nuclear microsatellites. For Ecol Manage. 163:263–271.

Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol. 27(11):2534–2547.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Amborella Genome Project. 2013. The amborella genome and the evolution of flowering plants. Science 342:1241089.

Baragatti M, Pudlo P. 2014. An overview on approximate Bayesian computation. ESAIM Proc. 44:291–299.

Beaumont MA. 2010. Approximate Bayesian computation in evolution and ecology. Annu Rev Ecol Evol Syst. 41(1):379–406.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162(4):2025–2035.

Bentley DR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218):53–59.

Biswas S, Akey JM. 2006. Genomic insights into positive selection. Trends Genet. 22(8):437–446.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci U S A. 101(42):15255–15260.

Eckert AJ, Liechty JD, Tearse BR, Pande B, Neale DB. 2010. DnaSAM: software to perform neutrality testing for large datasets with complex null models. Mol Ecol Resour. 10(3):542–545.

Eckert AJ, et al. 2010. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics 185(3):969–982.

Ersoz ES, Wright MH, González-Martínez SC, Langley CH, Neale DB. 2010. Evolution of disease response genes in loblolly pine: insights from candidate genes. PLoS One 5(12):e14234.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155(3):1405–1413.
Fisher T, Ager TA, Carrara PE, McGeehin JP. 2010. Ecosystem development in the Girdwood area, south-central Alaska, following late Wisconsin glaciation. Can J Earth Sci. 47:971–985.
Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv Prepr. arXiv:1207.3907.
Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2011. Unlocking Mendelian disease using exome sequencing. Genome Biol. 12(9):228.
Gnirke A, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 27(2):182–189.
Götz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36(10):3420–3435.
Grivet D, Sebastiani F, González-Martínez SC, Vendramin GG. 2009. Patterns of polymorphism resulting from long-range colonization in the Mediterranean conifer Aleppo pine. New Phytol. 184(4):1016–1028.
Hartl DL, Clark AG. 2007. Principles of population genetics. 4th ed. Sunderland (MA): Sinauer Associates, Inc.
Hernández-León S, Gernandt DS, Pérez de la Rosa JA, Jardón-Barbolla L. 2013. Phylogenetic relationships and species delimitation in Pinus section trifoliae inferrred from plastid DNA. PLoS One 8(7):e70501.
Heuertz M, et al. 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. Genetics 174(4):2095–2105.
Holliday JA, Yuen M, Ritland K, Aitken SN. 2010. Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. Mol Ecol. 19(18):3857–3864.
Hudson RR, Coyne JA. 2002. Mathematical consequences of the genealogical species concept. Evolution (N Y). 56:1557–1565.
Ingvarsson PK. 2008. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. Genetics 180(1):329–340.
Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci U S A. 101(29):10667–10672.
Jaramillo-Correa J-P, et al. 2015. Molecular proxies for climate maladaptation in a long-lived tree (*Pinus pinaster* Aiton, Pinaceae). Genetics 199(3):793–807.
Keller SR, Olson MS, Silim S, Schroeder W, Tiffin P. 2010. Genomic diversity, population structure, and migration following rapid range expansion in the Balsam poplar, *Populus balsamifera*. Mol Ecol. 19(6):1212–1226.
Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19(9):1639–1645.
Kumar S, Banks TW, Cloutier S. 2012. SNP discovery through next-generation sequencing and its applications. Int J Plant Genomics. 2012:831460.
Kurz W, et al. 2008. Mountain pine beetle and forest carbon feedback to climate change. Nature 452(7190):987–990.
Lee W-P, et al. 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One 9(3):e90581.
Lopes JS, Beaumont MA. 2010. ABC: a useful Bayesian tool for the analysis of population data. Infect Genet Evol. 10(6):826–833.
Lu M, et al. 2016. Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). BMC Genomics. 17:730.
Marroni F, et al. 2011. Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (CAD4) gene. Tree Genet Genomes 7(5):1011–1023.
McLaren W, et al. 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17:1–14.
Millar AA, et al. 1999. CUT1, an Arabidopsis gene required for cuticular wax biosynthesis and pollen fertility, encodes a very-long-chain fatty acid condensing enzyme. Plant Cell 11(5):825–838.
Mosca E, González-Martínez SC, Neale DB. 2014. Environmental versus geographical determinants of genetic structure in two subalpine conifers. New Phytol. 201(1):180–192.
Mosca E, Gugerli F, Eckert AJ, Neale DB. 2016. Signatures of natural selection on *Pinus cembra* and *P. mugo* along elevational gradients in the Alps. Tree Genet Genomes 12:9.
Neale DB, et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15(3):R59.
Nelson CD, et al. 2013. Pines. In: Singh, BP, editor. Biofuel crops: production, physiology and genetics. Oxfordshire, UK: CABI. p. 427–459.
Neves LG. 2013. Exome sequencing for high-throughput genomic analysis of trees. Gainesville (FL): University of Florida.
Neves LG, Davis JM, Barbazuk WB, Kirst M. 2013. Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J. 75(1):146–156.
Neves LG, Davis JM, Barbazuk WB, Kirst M. 2014. A high-density gene map of loblolly pine (*Pinus taeda* L.) based on exome sequence capture genotyping. G3 4:29–37.
Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39:197–218.
Olson MS, et al. 2010. Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*). New Phytol. 186(2):526–536.
Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol. 31(7):1929–1936.
Prunier J, Laroche J, Beaulieu J, Bousquet J. 2011. Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. Mol Ecol. 20(8):1702–1716.
Purcell S, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81(3):559–575.
Pyhäjärvi T, et al. 2007. Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics 177(3):1713–1724.
Ramos-Onsins S, Mitchell-Olds T. 2007. Mlcoalsim: multilocus coalescent simulations. Evol Bioinf. 3:41–44.
Remington DL, et al. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci U S A. 98(20):11479–11484.
Savolainen O, Lascoux M, Merilä J. 2013. Ecological genomics of local adaptation. Nat Rev Genet. 14(11):807–820.
Schmidtling RC. 2003. The southern pines during the Pleistocene. Acta Hortic. (615):203–209.
Schmidtling RC, Hipkins V. 1998. Genetic diversity in longleaf pine (*Pinus palustris*): influence of historical and prehistorical events. Can J For Res. 28(8):1135–1145.
Schmittner A, et al. 2011. Climate sensitivity estimated from temperature reconstructions of the last glacial maximum. Science 334(6061):1385–1389.
Slavov GT, et al. 2012. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. New Phytol. 196(3):713–725.
Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS. 2006. Comparative phylogeography of unglaciated eastern North America. Mol Ecol. 15(14):4261–4293.
Soto A, Robledo-Arnuncio JJ, González-Martínez SC, Smouse PE, Alía MR. 2010. Climatic niche and neutral genetic diversity of the six Iberian pine species: a retrospective and prospective view. Mol Ecol. 19(7):1396–1409.
Suren H, et al. 2016. Exome capture from the spruce and pine giga-genomes. Mol Ecol Resour. 16(5):1136–1146.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595.

Vijay N, et al. 2017. Genomewide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa. Mol Ecol. 26(16):4284–4295.
Wachowiak W, Boratynska K, Cavers S. 2013. Geographical patterns of nucleotide diversity and population differentiation in three closely related European pine species in the *Pinus mugo* complex. Bot J Linn Soc. 172(2):225–238.
Wachowiak W, Salmela MJ, Ennos RA, Iason G, Cavers S. 2011. High genetic diversity at the extreme range edge: nucleotide variation at nuclear loci in Scots pine (*Pinus sylvestris* L.) in Scotland. Heredity (Edinb). 106(5):775–787.
Wall JD. 1999. Recombination and the power of statistical tests of neutrality. Genet Res. 74(1):65–79.
Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016. Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. Genetics 202(3):1185–1200.
Wells O, Switzer G, Schmidtling R. 1991. Geographic variation in Mississippi loblolly pine and sweetgum. Silvae Genet. 40:105–119.
Willyard A, et al. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Mol Biol Evol. 24(1):90–101.,
Wright SI, Gaut BS. 2005. Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol. 22(3):506–519.
Yaeno T, Iba K. 2008. BAH1/NLA, a RING-type ubiquitin E3 ligase, regulates the accumulation of salicylic acid and immune responses to *Pseudomonas syringae* DC3000. Plant Physiol. 148(2):1032–1041.
Yeaman S, et al. 2016. Convergent local adaptation to climate in distantly related conifers. Science 353(6306):1431–1433.
Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174(3):1431–1439.
Zheng X, et al. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28(24):3326–3328.
Zhou L, Bawa R, Holliday JA. 2014. Exome resequencing reveals signatures of demographic and adaptive processes across the genome and range of black cottonwood (*Populus trichocarpa*). Mol Ecol. 23(10):2486–2499.
Zhou Y, Zhang L, Liu J, Wu G, Savolainen O. 2014. Climatic adaptation and ecological divergence between two closely related pine species in Southeast China. Mol Ecol. 23(14):3504–3522.

Associate editor: Ellen Pritham