AoB PLANTS
The open-access journal for plant sciences

Research Article

Conflicting genomic signals affect phylogenetic inference in four species of North American pines

Tomasz E. Koralewski[1*], Mariana Mateos[2] and Konstantin V. Krutovsky[1,3,4,5]

[1] Department of Ecosystem Science and Management, Texas A&M University, 2138 TAMU, College Station, TX 77843-2138, USA
[2] Department of Wildlife and Fisheries Sciences, Texas A&M University, 2258 TAMU, College Station, TX 77843-2258, USA
[3] Department of Forest Genetics and Forest Tree Breeding, Büsgen-Institute, Georg-August University of Göttingen, Büsgenweg 2, D-37077 Göttingen, Germany
[4] N.I. Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow 119333, Russia
[5] Genome Research and Education Center, Siberian Federal University, 50a/2 Akademgorodok, Krasnoyarsk 660036, Russia

Received: 17 September 2015; **Accepted:** 19 March 2016; **Published:** 8 April 2016

Associate Editor: Chelsea D. Specht

Citation: Koralewski TE, Mateos M, Krutovsky KV. 2016. Conflicting genomic signals affect phylogenetic inference in four species of North American pines. *AoB PLANTS* **8**: plw019; doi:10.1093/aobpla/plw019

Abstract. Adaptive evolutionary processes in plants may be accompanied by episodes of introgression, parallel evolution and incomplete lineage sorting that pose challenges in untangling species evolutionary history. Genus *Pinus* (pines) is one of the most abundant and most studied groups among gymnosperms, and a good example of a lineage where these phenomena have been observed. Pines are among the most ecologically and economically important plant species. Some, such as the pines of the southeastern USA (southern pines in subsection *Australes*), are subjects of intensive breeding programmes. Despite numerous published studies, the evolutionary history of *Australes* remains ambiguous and often controversial. We studied the phylogeny of four major southern pine species: shortleaf (*Pinus echinata*), slash (*P. elliottii*), longleaf (*P. palustris*) and loblolly (*P. taeda*), using sequences from 11 nuclear loci and maximum likelihood and Bayesian methods. Our analysis encountered resolution difficulties similar to earlier published studies. Although incomplete lineage sorting and introgression are two phenomena presumptively underlying our results, the phylogenetic inferences seem to be also influenced by the genes examined, with certain topologies supported by sets of genes sharing common putative functionalities. For example, genes involved in wood formation supported the clade *echinata*–*taeda*, genes linked to plant defence supported the clade *echinata*–*elliottii* and genes linked to water management properties supported the clade *echinata*–*palustris*. The support for these clades was very high and consistent across methods. We discuss the potential factors that could underlie these observations, including incomplete lineage sorting, hybridization and parallel or adaptive evolution. Our results likely reflect the relatively short evolutionary history of the subsection that is thought to have begun during the middle Miocene and has been influenced by climate fluctuations.

Keywords: *Australes*; drought tolerance; parallel evolution; phylogeny; plant defence; southern pines; wood formation.

Introduction

Plant genome structure and evolution are subjects of intensive investigations with recent milestone advances spanning whole-genome sequencing of some of the largest plant genomes, such as those of pines and spruces (Birol *et al.* 2013; Nystedt *et al.* 2013; Krutovsky *et al.* 2014; Neale *et al.* 2014; Zimin *et al.* 2014). These advancements bring opportunities to ask more focussed

* Corresponding author's e-mail address: tkoral@tamu.edu

questions about the dynamic processes that have contributed to plant genome evolution and influenced phylogenetic relationships among plant species. Duplications, from single gene to whole genome, are an intrinsic process that appears to be the major force driving genome evolution in plants (Freeling 2009; Li *et al.* 2015). Other processes, such as introgression, incomplete lineage sorting and parallel evolution, primarily influenced by extrinsic factors, additionally contribute to plant genome complexity and evolution (Martinsen *et al.* 2001; Wood *et al.* 2005; Willyard *et al.* 2009). The latter processes may also shape phenotypic and genotypic similarities among species becoming a challenge in phylogenetic studies.

Pines (genus *Pinus*, family Pinaceae) comprise a group of coniferous tree species occurring almost exclusively in the Northern Hemisphere, with populations often dominant across areas of the North Temperate Zone (Critchfield and Little 1966). They are keystone species of the vast boreal forest ecosystem. They provide pulp and timber products as primary commodities (Lowe *et al.* 1999; Borders and Bailey 2001; Alexander *et al.* 2002; Wear and Greis 2002; Croitoru 2007), but some additional pine forest products may include, for example, pine nuts, berries, herbs and mushrooms (Alexander *et al.* 2002; Bürgi *et al.* 2013; Nagasaka 2013). They provide a pivotal habitat for wildlife species (Brockerhoff *et al.* 2008), and greatly contribute to carbon storage and other ecosystem services (Goodale *et al.* 2002; McKinley *et al.* 2011). Due to their aesthetic value, they add recreational and ornamental dimensions in urban and suburban areas (Tyrväinen and Väänänen 1998; Knoth *et al.* 2002). These qualities extend to southeastern forest ecosystems of the USA, where pines may grow also in mixed forests with hardwoods (Sharitz *et al.* 1992), and where they additionally play an important role in ecosystem recovery from natural disturbances. For example, longleaf pine has developed complex adaptations to fire that allow fast stand regeneration (Outcalt 2000), and loblolly pine helps minimize soil erosion and provides watershed protection due to its fast growth (Schultz 1997).

Four major pines of the southeastern USA were investigated in this study: shortleaf (*Pinus echinata*), slash (*P. elliottii*), longleaf (*P. palustris*) and loblolly (*P. taeda*) (four of the 'southern pines'; subsection *Australes*, section *Trifoliae*, genus *Pinus*). They are widely cultivated, greatly dominating the southern US forest inventory (Sternitzke and Nelson 1970), and are, therefore, a subject of breeding programmes in the region (Wear and Greis 2002; McKeand *et al.* 2003; Fox *et al.* 2007). The traditional classification (Little and Critchfield 1969) considered 11 species in this subsection with habitat stretching cumulatively from the southeastern USA, through Mexico, to the Caribbean and Central America: slash (*P. elliottii*), spruce (*P. glabra*), longleaf (*P. palustris*), pond (*P. serotina*) and loblolly (*P. taeda*) pines in the southeastern USA; shortleaf (*P. echinata*), Table Mountain (*P. pungens*) and pitch (*P. rigida*) pines in the eastern USA; Cuban pine (*P. cubensis*) in Cuba; West Indian pine (*P. occidentalis*) in the West Indies; and Caribbean pine (*P. caribaea*) in both the West Indies and adjacent Central America. Attempts to refine the phylogeny of *Australes* are ongoing, but typically have been approached in the broader context of the Pinaceae family or along with other subsections.

Several previous investigations have provided insights to the phylogenetic relationships of the four species on which we focus in our study (i.e. *P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda*; Fig. 1). Adams and Jackson (1997) found a very close relationship between *P. palustris* and *P. taeda* based on 21 morphological characters, but their relationship to *P. echinata* and *P. elliottii* remained unresolved, and no statistical support was provided. Later, using RAPD markers and the neighbour-joining method, Dvorak *et al.* (2000) suggested a close relationship among *P. echinata*, *P. palustris* and *P. taeda*, in which *P. echinata* and *P. taeda* were sister lineages, and *P. elliottii* was sister to *P. caribaea*. Using a supertree approach and previously published phylogenies based on both morphological and molecular data, Grotkopp *et al.* (2004) also inferred a close relationship among *P. echinata*, *P. palustris* and *P. taeda*, but found that *P. palustris* was sister to *P. taeda*; *P. elliottii* was, again, sister to *P. caribaea*. Gernandt *et al.* (2005) used

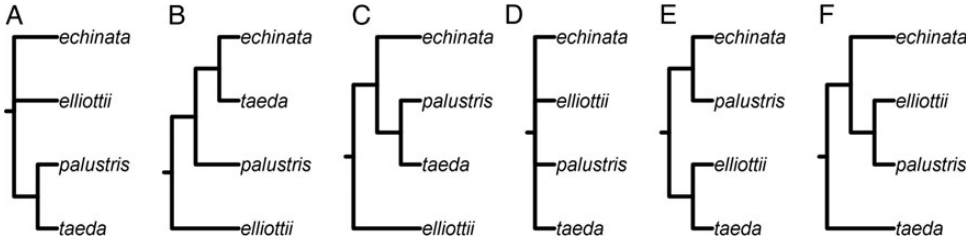

Figure 1. Cladograms for the four *Australes* species—*P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda*—from published studies. (A) Adams and Jackson (1997). (B) Dvorak *et al.* (2000). (C) Grotkopp *et al.* (2004). (D) Gernandt *et al.* (2005). (E) Eckert and Hall (2006). (F) Hernández-León *et al.* (2013). The original studies presented the four species in a broader context along with other pines.

two chloroplast genes in 101 pine species, but relationships among the four *Australes* species remained unresolved. Eckert and Hall (2006) used four chloroplast genes in 83 pine species. In their study, *P. echinata* and *P. palustris* were placed in one clade with low support, and *P. elliottii* and *P. taeda* were placed in another. Most recently, based on five chloroplast DNA (cpDNA) markers, Hernández-León *et al.* (2013) placed *P. echinata*, *P. elliottii* and *P. palustris* within the same clade, where *P. elliottii* and *P. palustris* were sisters, and *P. taeda* was in a separate clade, albeit with low bootstrap support (BS). Consequently, the monophyly of the four species was placed under question.

Evidence from cpDNA used in recent studies additionally questioned monophyly of the subsection *Australes* as defined by Little and Critchfield (1969), and suggested that the 11 *Australes* species may be scattered throughout a larger clade of over twice as many species (Gernandt *et al.* 2005; Eckert and Hall 2006; Hernández-León *et al.* 2013). Chloroplast genomes, however, may follow different evolutionary trajectories than nuclear genomes, with potentially confounding effects in phylogenetic studies (Rieseberg and Soltis 1991). In pines, chloroplast genomes are strictly paternally inherited, thus having smaller effective population sizes than nuclear genomes. Additionally, cpDNA experiences lower substitution rates (Willyard *et al.* 2007). Consequently, these factors may lead to discordant patterns of polymorphism between the two genomes (Soltis *et al.* 1992; Hong *et al.* 1993*b*). Moreover, cpDNA differentiation among populations of a species can be high (Hong *et al.* 1993*a*), and the cpDNA sequence variation in pines may be unevenly distributed throughout the genome requiring a representative locus sampling (Whittall *et al.* 2010). Finally, foreign chloroplast genomes introduced through hybridization may undergo a rapid fixation—chloroplast capture (Rieseberg and Soltis 1991; Matos and Schaal 2000; Liston *et al.* 2007).

Apart from the specificity of cpDNA evolution, additional problems with taxonomic classification of *Australes* may stem from incomplete lineage sorting, a phenomenon reported in pines (Syring *et al.* 2007; Willyard *et al.* 2009), especially considering the relatively recent evolutionary history of the subsection. Radiation within *Australes* is thought to have begun only 7–15 million years ago (MYA) (Willyard *et al.* 2007; Hernández-León *et al.* 2013). Large overlap in present-day areas of the four species studied here has fostered occasional or historic hybridization between some of them that may have additionally contributed to insufficient resolution in phylogenetic studies. Natural hybridization has been observed between *P. echinata* and *P. taeda* (Zobel 1953; Mergen *et al.* 1965; Smouse and Saylor 1973; Edwards-Burke *et al.* 1997), between *P. palustris* and *P. taeda* (Chapman 1922) and between *P. elliottii* and *P. palustris* (Mergen 1958). Almost all possible hybrids for these species can be artificially generated in controlled crosses, with the exception of *P. echinata* and *P. palustris* that are not considered interfertile despite a reported hybrid case (Snyder and Squillace 1966; Campbell *et al.* 1969). Additionally, similar environmental constraints could lead to a parallel support for beneficial alleles and/or removal of detrimental ones. Nevertheless, sympatry could have been repeatedly disturbed in the past; for example, during the Pleistocene, *P. taeda* was likely constricted to two refugia, *P. elliottii* and *P. palustris* could have been separated, each one occupying one of the two refugia, while *P. echinata*'s range was probably continuous due to its cold-hardiness (Wells *et al.* 1991; Schmidtling and Hipkins 1998; Schmidtling 2003). Each of these factors, or their combination, could confound phylogenetic inference.

The challenge with inferring relationships among *Australes* indicates that more work is needed. Nuclear markers are recommended to use when introgression, cytoplasmic in particular, or lineage sorting phenomena are present (Soltis *et al.* 1992). Consequently, we used 11 nuclear protein coding genes. Our objective was to refine and potentially clarify the phylogenetic relationships among *P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda*. *Pinus pinaster* was used as an outgroup. Despite the larger dataset and application of advanced methods, untangling their phylogeny was not straightforward. We discuss potential factors that likely contributed to this problem and could explain the difficulties in inferring phylogenies using multiple genes observed in previous studies.

Methods

Source of data

Data from 32 genes recently sequenced and annotated in four pines from subsection *Australes*, namely *P. taeda* (Brown *et al.* 2004; González-Martínez *et al.* 2006), *P. echinata*, *P. elliottii* and *P. palustris* (Koralewski *et al.* 2014), were used to query the NCBI GenBank database (Benson *et al.* 2013) for orthologs in related pine (genus *Pinus*) taxa that could be used as appropriate outgroup species. Eleven putative orthologous genes were identified in *P. pinaster* (Table 1), which was used as an outgroup. Given that the ingroup species belong to a single subsection, it is possible that a species from a more closely related subsection could be a more optimal outgroup for phylogenetic analyses; however, *P. pinaster* is the best studied species with respect to the genes investigated in the southern pines, allowing us to utilize more sequence data in the analyses. We assumed orthology of the genes based on exon–intron structure and very high sequence similarity. For each putative ortholog, the *E*-value was 3E-46 or less and identity score was 95 % or more in at least one comparison of an ingroup species

Table 1. Genes used in the study and their NCBI accession numbers. Genomic DNA sequences for the ingroup (*P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda*) were newly generated and presented in Koralewski *et al.* (2014), and both genomic DNA and mRNA sequences for the outgroup (*P. pinaster*) were already available in NCBI.

Gene name	Gene abbreviation	*Pinus echinata*	*Pinus elliottii*	*Pinus palustris*	*Pinus taeda*	*Pinus pinaster*
4-coumarate:CoA ligase	*4cl*	KF158811	KF158813	KF158814	KF158816	HM482497
arabinogalactan 4	*agp-4*	KF158819	KF158821	KF158822	KF158824	AM501931
trans-cinnamate 4-hydroxylase 2	*c4h-2*	KF158875	KF158877	KF158878	KF158880	JN013973
cinnamyl alcohol dehydrogenase	*cad*	KF158882	KF158883	KF158884	KF158885	FN824799
cellulose synthase	*cesA3*	KF158898	KF158900	KF158901	KF158903	FN257074
caffeate O-methyltransferase	*comt-2*	KF158906	KF158908	KF158909	KF158911	HE574557
dehydrin 2	*dhn-2*	KF158924	KF158926	KF158927	KF158929	HE796687
early response to drought 3	*erd3*	KF158931	KF158932	KF158933	KF158934	EU020011
glycine hydroxymethyltransferase	*glyhmt*	KF158935	KF158936	KF158937	KF158938	HE574564
ABII protein phosphatase 2C-like	*pp2c*	KF158952	KF158953	KF158954	KF158955	EU020014
chloroplast Cu/Zn superoxide dismutase	*sod-chl*	KF158978	KF158979	KF158980	KF158981	AF434186

with *P. pinaster*; additionally, in comparisons where the *E*-value was higher than E-100, the identity score was 99 %. However, well-annotated, mapped and well-assembled whole-genome data are ultimately needed to validate orthology. Such data are still unavailable, although the first incomplete draft genome assembly was published recently for loblolly pine (Neale *et al.* 2014; Zimin *et al.* 2014).

Multiple alignment

The FASTA sequences for individual genes were aligned using BioEdit (ver. 7.0.9.0) (Hall 1999) and merged into one dataset in SeaView (ver. 4.0) (Galtier *et al.* 1996). Conversion from FASTA to NEXUS was done in SeaView, and from FASTA to PHYLIP manually in the text editor Notepad++ (ver. 5.9.3) (Ho 2011). Coding sites were assigned based on annotation data in NCBI GenBank for each individual gene separately using DnaSP (ver. 5.10.01) (Librado and Rozas 2009). The merged NEXUS file was further manually annotated, and one of five categories was assigned to each site: a codon position (1, 2 or 3), intron or 3′UTR. All sites with missing data were removed from the analysis **[see Supporting Information—Dataset S1]**.

Partitioning

Partitioning schemes and models of molecular evolution were evaluated using PartitionFinder (ver. 1.1.1) (Lanfear *et al.* 2012). Model selection was restricted to the set of models implemented in the software in which we intended to run the phylogenetic analysis (parameter 'model =' in PartitionFinder). At most 42 models were evaluated simultaneously based on Akaike information criterion (AIC) and Bayesian information criterion (BIC). The parameters describing gamma distribution of rates among sites (Γ or G) and a proportion of invariable sites (I) were considered among the models exclusively, i.e. none of the evaluated models accounted for both Γ and I jointly, because of reported problems with independent optimization of these two parameters (see discussions and relevant references in Stamatakis 2008: pp. 20–21 and in Stamatakis 2014: p. 49). Additionally, in the group of models evaluated for GARLI (Zwickl 2006) (see below), one or more of the K81 models (K81, K81 + I and K81 + G, depending on the subset of data; a few data subsets were affected) caused convergence problems and were not considered. Depending on the intended phylogenetic analysis, alternative partitioning schemes were then implemented **[see Supporting Information—Table S1]**: (i) by-gene-site: best partitioning schemes for 39 sets corresponding to different genes and site categories within genes, or (ii) by-gene: best models identified for each of the 11 genes in the set (no partitioning within a gene). Additionally, in the cases where each gene was analysed separately, models were identified for every gene independently with sites assigned to one of the five site categories (by-site).

Phylogenetic analysis

The combined dataset of 11 genes was subjected to two gene tree methods with partitioning by-gene-site: maximum likelihood (ML; GARLI, ver. 2.01) (Zwickl 2006) and

Bayesian inference (BI; MrBayes, ver. 3.2.2) (Ronquist *et al.* 2012). Species trees were reconstructed using the Bayesian method BEST (ver. 2.3) (Liu 2008) with partitioning by-gene. To account for potential polytomies, and thus potential for the star tree paradox (Suzuki *et al.* 2002; Lewis *et al.* 2005), we ran Phycas (ver. 1.2.0) (Lewis *et al.* 2010), also a Bayesian method, with partitioning by-gene-site. We then analysed each gene separately using GARLI and MrBayes (partitioning by-site). The MrBayes outputs were further investigated using Bayesian concordance analysis (Ané *et al.* 2007), as implemented in BUCKy (ver. 1.4.0, mbsum ver. 1.4.2) (Larget *et al.* 2010), for seven values of α: 0.01, 0.5, 1, 2, 5, 10 and 10 000. The value of α corresponds to the probability that loci share the same tree (the lower α the higher the probability, and vice versa), thus affecting the clade support (the higher the probability the higher support). Formally, BUCKy is not considered a species-tree method; however, the resulting primary concordance tree can be generally considered comparable with species trees (for an in-depth discussion about the species methods, see Mateos *et al.* 2012).

All analyses were run in Windows 7 **[see Supporting Information—Table S1]** except for BUCKy, which was compiled and run under Cygwin. The tree figures were visualized in FigTree (ver. 1.4) (Rambaut 2013) and further edited manually to increase their readability and compactness.

Clade support and convergence

In order to determine the clade support, BS was calculated in GARLI using 1000 replicates **[see Supporting Information—Table S1]**. SumTrees (ver. 3.3.1) (Sukumaran and Holder 2010) was used to generate majority consensus trees. Posterior probability (PP) estimates were used for the BI methods. To verify that the Markov chain Monte Carlo analysis converged on a stationary distribution and that the sampling of the distribution was adequate, the following criteria were applied for MrBayes and BEST: (i) stable PP values, (ii) small and stable average standard deviation of the split frequencies of independent runs, (iii) potential scale reduction factor close to 1 and (iv) an effective sample size of at least 200 for the posterior probabilities. The conditions (i) and (iv) were evaluated also for Phycas, and additionally the split PP plot and split sojourn plot were examined. Samples prior to reaching stationarity were discarded as 'burnin'. The conditions (i) and (iv) were evaluated in Tracer (ver. 1.5.0) (Rambaut and Drummond 2009). Average standard deviation of mean sample-wide concordance factor (CF) was examined for BUCKy, and the CF was used to determine clade support.

Results

Combined evidence from all genes

Three methods, GARLI, MrBayes and Phycas, recovered the clade *echinata*–*elliottii* for the concatenated matrix of 11 genes. Support from the ML method was much lower (highest BS = 65 %, AIC partitioning) than that from the BI methods (lowest PP = 0.98 in MrBayes, BIC partitioning). The BI methods additionally supported the clade *echinata*–*elliottii*–*palustris*, but support varied considerably (from PP < 0.50 in Phycas, AIC, to PP = 0.94 in MrBayes, BIC). Significant differences between the ML and BI methods in support for the clade *echinata*–*elliottii*, and in the case of the clade *echinata*–*elliottii*–*palustris*, also between Phycas and MrBayes (Fig. 2) are characteristic signatures of true or approximate star phylogenies (Suzuki *et al.* 2002; Lewis *et al.* 2005).

BEST supported the clade *echinata*–*taeda*, absent in the gene-tree methods, with PP = 0.53 (AIC and BIC; Fig. 3). This clade was present also in the primary concordance tree in BUCKy (CF ranging from 0.35 to 0.44, depending on α and model selection criterion). Another clade present in the BUCKy's primary concordance tree was *elliottii*–*palustris* (CF ranging from 0.24 to 0.44, depending on α and model selection criterion).

Individual gene approach

Given the results from the analysis of the combined dataset resulting in mostly unresolved relationships among the studied species, and the conflicts found between

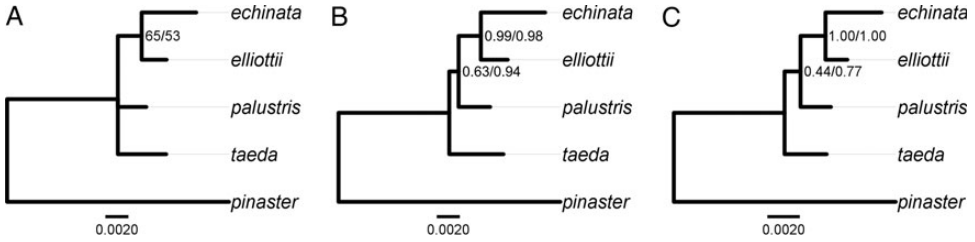

Figure 2. Joint analysis of 11 genes in *P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda* using GARLI (A), MrBayes (B) and Phycas (C). Numbers at nodes correspond to BS for GARLI (%; AIC/BIC), and PP for MrBayes and Phycas (AIC/BIC). Branch lengths are shown for the BIC partitioning schemes.

the gene-tree and species-tree methods, we looked at each gene separately using ML (GARLI) and BI (MrBayes). The two approaches produced consistent results for each gene, although the difference in clade support varied among the genes **[see Supporting Information—Fig. S1]**. We, therefore, clustered genes into three groups following similarity in topologies. The genes *4cl*, *c4h-2* and *cesA3* (Group A) supported the clade *echinata*–*taeda*, the genes *agp-4* and *sod-chl* (Group B) supported the clade *echinata*–*elliottii* and the genes *dhn-2* and *erd3* (Group C) supported the clade *echinata*–*palustris*. Group C genes disagreed in the placement of the species *P. elliottii* and *P. taeda*; however, we decided to focus on the similarity. Additional analysis in the R environment (ver. 2.15.2) (R Core Team 2012) supported these groupings—principal coordinate analysis (function pcoa, package ape ver. 3.0-11) (Paradis *et al.* 2004) run on Robinson–Foulds distance matrix for the individual gene trees (function RF.dist, package phangorn ver. 1.99-7) (Schliep 2011) placed the three groups of genes in distinct clusters **[see Supporting Information—Fig. S2]**. Each set was partitioned by-gene-site, and analysed using GARLI and MrBayes. In general, the BS for the clades jointly supported by genes within each group increased, the PP reached (or stayed at) 1.00 and the level of support for these clades became consistent between GARLI and MrBayes (Fig. 4) **[see Supporting Information—Fig. S3]**. In the case of Group B, the clade *palustris*–*taeda*, previously present only in the *agp-4* gene tree, was also recovered, although with low support. In Group C, *P. elliottii* and *P. taeda* were placed as sister taxa in a clade with low BS and high PP, which reflects the conflict between the two individual gene trees.

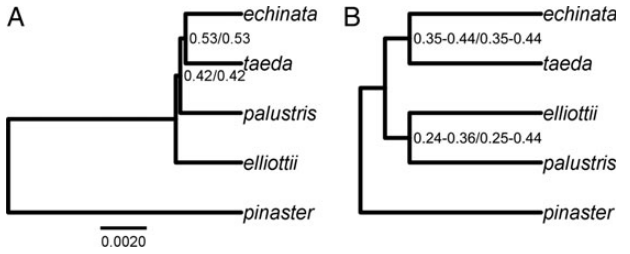

Figure 3. Joint analysis of 11 genes in *P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda* using BEST (A) and BUCKy (cladogram; B). Numbers at nodes correspond to PP for BEST (AIC/BIC) and CFs for BUCKy (AIC/BIC; range of CFs for α values from 10 000 to 0.01).

Discussion

The discordance among various approaches applied in this study mirrors the conflicting results from previously published work (see Introduction). Comparison of the topologies supported by Group A vs. Group B vs. Group C genes shows highly consistent results within each group across different methods, yet the conclusions contradict each other among the groups. The clade *echinata*–*taeda*, supported by Group A genes, was previously recovered by Dvorak *et al.* (2000) using RAPD markers and the neighbour-joining method with BS = 0.90. The species in the clade *palustris*–*taeda*, supported by the gene *agp-4* (Group B), were placed in one clade by Adams and Jackson (1997) based on parsimony analysis of morphological characters, and by Grotkopp *et al.* (2004) through a supertree approach (57 % of individual trees agreed with the supertree at the node). Both clades *echinata*–*palustris* and *elliottii*–*taeda* supported by the Group C genes corresponded to grouping in Eckert and Hall (2006), who used chloroplast data.

Given life history of the subsection (see Introduction), long-distance gene flow through pollen, hybridization events that still occur today and large (but variable in the past) population sizes of the investigated species, incomplete lineage sorting could be the simplest and most straightforward explanation of the pattern observed in our study. It is a known phenomenon in pines (Syring *et al.* 2007; Willyard *et al.* 2009), and it is likely to confound phylogenetic inference in a clade as recent as *Australes*. Interestingly, however, the genes forming the three groups

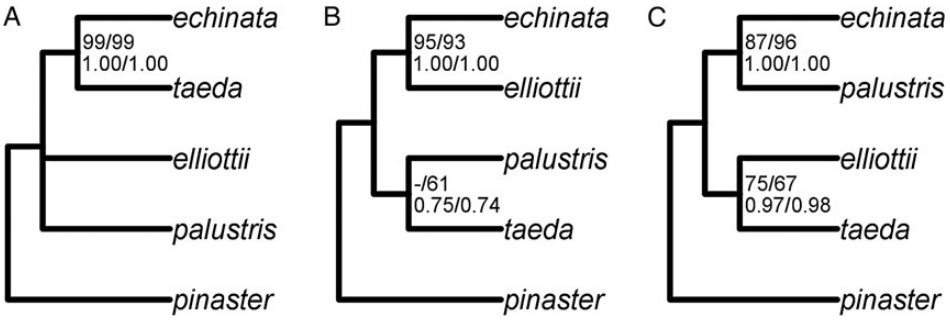

Figure 4. Analysis of three groups of genes: Group A (*4cl*, *c4h-2* and *cesA3*; A), Group B (*agp-4* and *sod-chl*; B) and Group C (*dhn-2* and *erd3*; C) in *P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda* using GARLI and MrBayes. Cladograms are shown. Numbers at nodes correspond to clade support: GARLI (AIC/BIC; top row) and MrBayes (AIC/BIC; bottom row). The BS for the clade *palustris*–*taeda* in Group B was <0.50.

Table 2. Groups of genes based on phylogenetic analysis of *P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda*, their selected putative functions and numbers of nonsynonymous (N) and silent (S) nucleotide substitutions in each group.

Group	Genes	Putative function	N	S
A	*4cl*, *c4h-2* and *cesA3*	Wood properties	2	10
B	*agp-4* and *sod-chl*	Plant defence, water management	4	10
C	*dhn-2* and *erd3*	Water stress recognition and response	2	11

are also connected at the putative functional level (Table 2). The Group A genes all directly affect wood properties: *4cl* and *c4h-2* are both part of the lignin biosynthesis pathway (Whetten and Sederoff 1995; Boerjan *et al.* 2003), and *cesA3* is involved in cellulose synthesis (Pot *et al.* 2005). Another gene in our dataset, *comt-2*, is also involved in lignin biosynthesis (Whetten and Sederoff 1995; Boerjan *et al.* 2003) but did not recover the *echinata–taeda* clade. Unlike the other three genes in Group A, it supported the clade *elliottii–palustris*. An additional joint analysis of all four genes, using the same methodology, showed that the support for the previously recovered clade *echinata–taeda* almost did not change, while the clade *elliottii–palustris* received a low-to-moderate support **[see Supporting Information—Fig. S4]**. The latter two species were previously placed in one clade (BS $\leq$ 0.50) by Hernández-León *et al.* (2013) based on an ML method applied to plastid data. Conversely, the Group B and C genes are not involved in the lignin biosynthesis pathway directly. The Group B genes (*agp-4* and *sod-chl*) appear to contribute to plant defence, water management and other functions. Proteins from the arabinogalactan family play signalling and protective roles, and participate in xylem development, cell growth and expansion, programmed cell death and other processes (Loopstra and Sederoff 1995; Loopstra *et al.* 2000; Zhang *et al.* 2000, 2003; Showalter 2001; Seifert and Roberts 2007). *Sod-chl* has antioxidative properties (Huttunen and Heiska 1988; Bowler *et al.* 1992). It was also drought responsive (Costa *et al.* 1998) and hence was considered a drought-tolerance candidate gene in loblolly pine (González-Martínez *et al.* 2006). *Dhn-2* and *erd3* (Group C) are also associated with water stress recognition and response. *Dhn-2*, a *dehydrin*, is responsive to dehydration stress and also plays other protective functions (González-Martínez *et al.* 2006; Eveno *et al.* 2008). The role of *erd3* is less known, but it has been found active in an early stage of dehydration stress (González-Martínez *et al.* 2006; Eveno *et al.* 2008).

Regardless of the extent of the potential effects of shared ancestral polymorphisms, two other hypotheses of parallel evolution within the functional gene groups could be pursued in follow-up studies. The functionally bound clades could have resulted from a transfer of adaptive alleles via introgressive hybridization from an adapted pine donor, followed by positive selection and subsequent purifying or balancing selection acting in both donor and acceptor populations. Alternatively, speciation could have begun while the locus was already under purifying or balancing selection, which continued working simultaneously in both populations in parallel, given common habitat locations and environmental pressures. Sympatry, shifting locale of the optimum habitat, population size changes and the recent evolutionary origin of the southern pines could have facilitated hybridization, while crucial roles of the genes belonging to the three groups could have resulted in preservation of adaptive variants, especially under similar environmental selection pressures. A tight physical linkage among the loci within each group can be excluded as an alternative possible explanation for the observed phenomenon. Most of the genes studied here are mapped to different linkage groups in *P. taeda*, or far from each other, although no linkage data were found for *dhn-2* and *erd3* **[see Supporting Information—Table S2]**. These two hypotheses require more assumptions to explain the observed pattern when compared with the hypothesis of incomplete lineage sorting, and therefore, the latter should be preferred, although some form of interplay of all three is certainly possible.

In order to reject the hypothesis of incomplete lineage sorting, and to pursue an alternative, limitations of our study need to be overcome. Multiple individuals sampled for each species would allow for thorough identification of shared ancestral polymorphisms and variation fixed at the species level. Including more genes per functional group would allow to examine whether the observed pattern holds also for other members of a given pathway or process, or if it is purely stochastic. Samples from all members of the clade *Australes* would likely help improve overall robustness of the phylogenetic inference. Additionally, to test the monophyly of the clade *Australes* in the traditional sense (Little and Critchfield 1969), which has been questioned by cpDNA-based studies (Gernandt *et al.* 2005; Eckert and Hall 2006; Hernández-León *et al.* 2013), species traditionally classified as *Attenuatae*, *Leiophyllae* and *Oocarpae* should be included in the

nuclear marker-based study. Alternative outgroup species could also be considered, especially those from subsections *Contortae* and *Ponderosae* that are more closely related to the ingroup than *P. pinaster*. Solving these caveats, however, requires additional resources and cannot be done purely analytically. Given the limitations of the data, our primary intention was to apply a spectrum of methods to a gene sample that would maximize genome coverage.

Recent estimates of radiation within *Australes* suggested that it could have begun as recently as 7 MYA (*P. taeda*–*P. radiata* split), although the split between the ancestors of *Ponderosae* and *Australes* might have happened as early as 15 MYA (Willyard *et al.* 2007). This timeframe overlaps with the mid- to late-Miocene (about 14–15 MYA), starting at or directly following the middle Miocene climate transition (MMCT), a period of cooling and ice-sheet expansion that took place about 14 MYA (Shevenell *et al.* 2004). Pines experienced habitat locale shifts both before and after the MMCT, for example, during the Eocene (56–34 MYA), interpreted as the major stimulus for pine divergence at the time, and during the Pleistocene (2.6–0.01 MYA) (Millar 1993). The MMCT likely affected population sizes, species range and distribution of the allelic variation, and probably had the momentum to trigger radiation within *Australes*. The potential for hybridization events resulting in introgression was likely greater back in time, especially when the ancestral species were far less diverged and stressed by recurrent changes in environmental pressures and by range shifts. Range expansions could have then brought multiple genetic effects (Excoffier *et al.* 2009) including increase in frequency of rare (and also newly introduced) variants. This process would have happened much faster if the newly acquired alleles were advantageous.

The adaptations shared among the southern pines and shaped by the vibrant historic climate are particularly interesting in the light of the ongoing and forthcoming climate changes, amidst the discussion on assisted migration (Vitt *et al.* 2010; Krutovsky *et al.* 2012; Koralewski *et al.* 2015). The historical events might have led to increased standing genetic variation in these species, directly influencing their level of adaptability and making them somewhat 'climate-change ready'. Additional inquiries directed towards the loci studied here could improve breeding strategies in the face of climate change (Krutovsky *et al.* 2013).

Conclusions

Incomplete lineage sorting, introgression and parallel evolution can explain inconsistencies observed in the phylogenetic analysis of the four southern pines. However, more data are needed to discriminate among these hypotheses. The conflicting signals were vigorously tested, but evidence in the current data was not robust enough to support potent claims, and thus, the simplest hypothesis of incomplete lineage sorting may be preferred, while the alternatives may be pursued in future studies. To overcome limitations of our study, additional sampling should include multiple individuals per species, additional species that form one clade with the four pines investigated here, less distant outgroup species and additional functionally related genes. Our work provided new insights into the *Australes* phylogeny, but their evolutionary history remains elusive.

Sources of Funding

The project was supported by the United States Department of Agriculture (USDA) Cooperative State Research, Education, and Extension Service (CSREES) and Texas Agricultural Experiment Station (TAES) McIntire-Stennis Project (TEX09122-0210381). The Pine Integrated Network: Education, Mitigation, and Adaptation Project (PINEMAP), a Coordinated Agricultural Project funded by the USDA National Institute of Food and Agriculture, Award #2011-68002-30185, provided support during preparation of this manuscript.

Contributions by the Authors

T.E.K., M.M. and K.V.K. conceived and designed the study. T.E.K. arranged the data and ran the software. T.E.K., M.M. and K.V.K. interpreted results. T.E.K. drafted the manuscript. T.E.K., M.M. and K.V.K. edited, read and approved the final version of the manuscript.

Conflict of Interest Statement

None declared.

Acknowledgements

We are grateful to Dr Alan E. Pepper, Dr Clare A. Gill (Texas A&M University) and Dr Ruzong Fan (National Institute of Child Health and Human Development, National Institutes of Health) for valuable discussions, suggestions and comments. We thank Dr Earl M. 'Fred' Raley (Texas A&M Forest Service) for comments on the final version of the manuscript. T.E.K. would like to thank Dr Thomas D. Byram (Texas A&M Forest Service, Texas A&M University) for his support during this study. We thank anonymous reviewers, the Associate Editor, Dr Chelsea D. Specht (University of California – Berkeley), and the Chief Editor, Dr J. Hall Cushman (Sonoma State University), for suggestions that helped us improve the manuscript.

Supporting Information

The following additional information is available in the online version of this article —

Dataset S1. Multiple alignment in NEXUS format.

Figure S1. Separate analysis of each gene for the four *Australes* species: *P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda*. Cladograms are shown. From top to bottom: *4cl*, *c4h-2* and *cesA3* (A); *agp-4* and *sod-chl* (B); *dhn-2* and *erd3* (C); *cad*, *glyhmt*, *pp2c* and *comt-2* (D). Numbers at nodes correspond to clade support: GARLI (AIC/BIC; top row) and MrBayes (AIC/BIC; bottom row).

Figure S2. Principal coordinate analysis run on Robinson–Foulds distance matrix for the individual gene trees. Group A gene names (*4cl*, *c4h-2* and *cesA3*) are in green, Group B gene names (*agp-4* and *sod-chl*) are in orange and Group C gene names (*dhn-2* and *erd3*) are in blue.

Figure S3. Joint analysis of the genes from Group A (*4cl*, *c4h-2* and *cesA3*; A), Group B (*agp-4* and *sod-chl*; B) and Group C (*dhn-2* and *erd3*; C) in *P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda* using GARLI and MrBayes. From top to bottom: GARLI (AIC), GARLI (BIC), MrBayes (AIC) and MrBayes (BIC). Numbers at nodes correspond to clade support.

Figure S4. Joint analysis of the genes from Group A (*4cl*, *c4h-2* and *cesA3*) and *comt-2* in *P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda* using GARLI and MrBayes. From top to bottom: GARLI (AIC), GARLI (BIC), MrBayes (AIC) and MrBayes (BIC). Numbers at nodes correspond to clade support.

Table S1. Partitioning schemes and software settings.

Table S2. Linkage information for the studied genes.

Literature Cited

Adams DC, Jackson JF. 1997. A phylogenetic analysis of the southern pines (*Pinus* subsect. *Australes* Loudon): biogeographical and ecological implications. *Proceedings of the Biological Society of Washington* **110**:681–692.

Alexander SJ, Pilz D, Weber NS, Brown E, Rockwell VA. 2002. Mushrooms, trees, and money: value estimates of commercial mushrooms and timber in the Pacific Northwest. *Environmental Management* **30**:129–141.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* **24**:412–426.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Research* **41**: D36–D42.

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Saint Yuen MM, Keeling CI, Brand D, Vandervalk BP, Kirk H, Pandoh P, Moore RA, Zhao YJ, Mungall AJ, Jaquish B, Yanchuk A, Ritland C, Boyle B, Bousquet J, Ritland K, MacKay J, Bohlmann J, Jones SJM. 2013. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics* **29**:1492–1497.

Boerjan W, Ralph J, Baucher M. 2003. Lignin biosynthesis. *Annual Review of Plant Biology* **54**:519–546.

Borders BE, Bailey RL. 2001. Loblolly pine—pushing the limits of growth. *Southern Journal of Applied Forestry* **25**:69–74.

Bowler C, Van Montagu M, Inzé D. 1992. Superoxide dismutase and stress tolerance. *Annual Review of Plant Physiology and Plant Molecular Biology* **43**:83–116.

Brockerhoff EG, Jactel H, Parrotta JA, Quine CP, Sayer J. 2008. Plantation forests and biodiversity: oxymoron or opportunity? *Biodiversity and Conservation* **17**:925–951.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences of the United States of America* **101**:15255–15260.

Bürgi M, Gimmi U, Stuber M. 2013. Assessing traditional knowledge on forest uses to understand forest ecosystem dynamics. *Forest Ecology and Management* **289**:115–122.

Campbell TE, Hamaker JM, Schmitt DM. 1969. Longleaf pine × shortleaf pine—a new hybrid. *Bulletin of the Torrey Botanical Club* **96**: 519–524.

Chapman HH. 1922. A new hybrid pine (*Pinus palustris* × *Pinus taeda*). *Journal of Forestry* **20**:729–734.

Costa P, Bahrman N, Frigerio J-M, Kremer A, Plomion C. 1998. Water-deficit-responsive proteins in maritime pine. *Plant Molecular Biology* **38**:587–596.

Critchfield WB, Little EL. 1966. *Geographic distribution of the pines of the world*. Miscellaneous Publication 991. Washington, DC: USDA Forest Service, U.S. Government Printing Office.

Croitoru L. 2007. How much are Mediterranean forests worth? *Forest Policy and Economics* **9**:536–545.

Dvorak WS, Jordon AP, Hodge GP, Romero JL. 2000. Assessing evolutionary relationships of pines in the *Oocarpae* and *Australes* subsections using RAPD markers. *New Forests* **20**:163–192.

Eckert AJ, Hall B. 2006. Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. *Molecular Phylogenetics and Evolution* **40**:166–182.

Edwards-Burke MA, Hamrick JL, Price RA. 1997. Frequency and direction of hybridization in sympatric populations of *Pinus taeda* and *P. echinata* (Pinaceae). *American Journal of Botany* **84**:879–886.

Eveno E, Collada C, Guevara MA, Léger V, Soto A, Díaz L, Léger P, González-Martínez SC, Cervera MT, Plomion C, Garnier-Géré PH. 2008. Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Molecular Biology and Evolution* **25**:417–437.

Excoffier L, Foll M, Petit RJ. 2009. Genetic consequences of range expansions. *Annual Review of Ecology Evolution and Systematics* **40**:481–501.

Fox TR, Jokela EJ, Allen HL. 2007. The development of pine plantation silviculture in the southern United States. *Journal of Forestry* **105**: 337–347.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* **60**:433–453.

Galtier N, Gouy M, Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Computer Applications in the Biosciences* **12**:543–548.

Gernandt DS, López GG, García SO, Liston A. 2005. Phylogeny and classification of *Pinus*. *Taxon* **54**:29–42.

González-Martínez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB. 2006. DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. *Genetics* **172**:1915–1926.

Goodale CL, Apps MJ, Birdsey RA, Field CB, Heath LS, Houghton RA, Jenkins JC, Kohlmaier GH, Kurz W, Liu SR, Nabuurs G-J, Nilsson S, Shvidenko AZ. 2002. Forest carbon sinks in the Northern Hemisphere. *Ecological Applications* **12**:891–899.

Grotkopp E, Rejmánek M, Sanderson MJ, Rost TL. 2004. Evolution of genome size in pines (*Pinus*) and its life-history correlates: Supertree analyses. *Evolution* **58**:1705–1729.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**:95–98.

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**:e70501.

Ho D. 2011. *Notepad++ v5.9.3*. http://notepad-plus-plus.org (11 October 2011).

Hong Y-P, Hipkins VD, Strauss SH. 1993*a*. Chloroplast DNA diversity among trees, populations and species in the California closed-cone pines (*Pinus radiata*, *Pinus muricata* and *Pinus attenuata*). *Genetics* **135**:1187–1196.

Hong Y-P, Krupkin AB, Strauss SH. 1993*b*. Chloroplast DNA transgresses species boundaries and evolves at variable rates in the California closed-cone pines (*Pinus radiata*, *P. muricata*, and *P. attenuata*). *Molecular Phylogenetics and Evolution* **2**:322–329.

Huttunen S, Heiska E. 1988. Superoxide dismutase (SOD) activity in Scots pine (*Pinus sylvestris* L.) and Norway spruce (*Picea abies* L. Karst.) needles in northern Finland. *European Journal of Forest Pathology* **18**:343–350.

Knoth J, Frampton J, Moody R. 2002. Genetic improvement of Virginia pine planting stock for Christmas tree production in South Carolina. *HortTechnology* **12**:675–678.

Koralewski TE, Brooks JE, Krutovsky KV. 2014. Molecular evolution of drought tolerance and wood strength related candidate genes in loblolly pine (*Pinus taeda* L.). *Silvae Genetica* **63**:59–66.

Koralewski TE, Wang H-H, Grant WE, Byram TD. 2015. Plants on the move: assisted migration of forest trees in the face of climate change. *Forest Ecology and Management* **344**:30–37.

Krutovsky KV, Burczyk J, Chybicki I, Finkeldey R, Pyhäjärvi T, Robledo-Arnuncio JJ. 2012. Gene flow, spatial structure, local adaptation, and assisted migration in trees. In: Schnell RJ, Priyadarshan PM, eds. *Genomics of tree crops*. New York: Springer Science, Inc., 71–116.

Krutovsky K, Byram T, Whetten R, Wheeler N, Neale D, Lu M, Koralewski T, Loopstra C. 2013. PINEMAP + PineRefSeq = future forests. In: Sommer E, ed. *PINEMAP (Pine Integrated Network: Education, Mitigation, and Adaptation Project) Year 2 Annual Report | March 2012-February 2013 "Mapping the future of southern pine management in a changing world"*, 26–27.

Krutovsky KV, Oreshkova NV, Putintseva Yu. A, Ibe AA, Deych KO, Shilkina EA. 2014. Preliminary results of *de novo* whole genome sequencing of the Siberian larch (*Larix sibirica* Ledeb.) and the Siberian stone pine (*Pinus sibirica* Du Tour). *Siberian Journal of Forest Science* **1**:79–83.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* **29**:1695–1701.

Larget BR, Kotha SK, Dewey CN, Ané C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* **26**:2910–2911.

Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and Bayesian phylogenetic inference. *Systematic Biology* **54**:241–253.

Lewis PO, Holder MT, Swofford DL. 2010. *Phycas ver. 1.2.0*. http://www.phycas.org (17 September 2013).

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. *Science Advances* **1**:e1501084.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**:1451–1452.

Liston A, Parker-Defeniks M, Syring JV, Willyard A, Cronn R. 2007. Interspecific phylogenetic analysis enhances intraspecific phylogeographical inference: a case study in *Pinus lambertiana*. *Molecular Ecology* **16**:3926–3937.

Little EL, Critchfield WB. 1969. *Subdivisions of the genus Pinus (pines)*. Miscellaneous Publication No. 1144. Washington, DC: USDA Forest Service, U.S. Government Printing Office.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* **24**:2542–2543.

Loopstra CA, Sederoff RR. 1995. Xylem-specific gene expression in loblolly pine. *Plant Molecular Biology* **27**:277–291.

Loopstra CA, Puryear JD, No E-G. 2000. Purification and cloning of an arabinogalactan-protein from xylem of loblolly pine. *Planta* **210**:686–689.

Lowe WJ, Byram TD, Bridgwater FE. 1999. Selecting loblolly pine parents for seed orchards to minimize the cost of producing pulp. *Forest Science* **45**:213–216.

Martinsen GD, Whitham TG, Turek RJ, Keim P. 2001. Hybrid populations selectively filter gene introgression between species. *Evolution* **55**:1325–1335.

Mateos M, Hurtado LA, Santamaria CA, Leignel V, Guinot D. 2012. Molecular systematics of the deep-sea hydrothermal vent endemic brachyuran family Bythograeidae: a comparison of three Bayesian species tree methods. *PLoS ONE* **7**:e32066.

Matos JA, Schaal BA. 2000. Chloroplast evolution in the *Pinus montezumae* complex: a coalescent approach to hybridization. *Evolution* **54**:1218–1233.

McKeand S, Mullin T, Byram T, White T. 2003. Deployment of genetically improved loblolly and slash pines in the South. *Journal of Forestry* **101**:32–37.

McKinley DC, Ryan MG, Birdsey RA, Giardina CP, Harmon ME, Heath LS, Houghton RA, Jackson RB, Morrison JF, Murray BC, Pataki DE, Skog KE. 2011. A synthesis of current knowledge on forests and carbon storage in the United States. *Ecological Applications* **21**:1902–1924.

Mergen F. 1958. Genetic variation in needle characteristics of slash pine and in some of its hybrids. *Silvae Genetica* **7**:1–9.

Mergen F, Stairs GR, Snyder EB. 1965. Natural and controlled loblolly × shortleaf pine hybrids in Mississippi. *Forest Science* **11**:306–314.

Millar CI. 1993. Impact of the Eocene on the evolution of *Pinus* L. *Annals of the Missouri Botanical Garden* **80**:471–498.

Nagasaka K. 2013. Comparative economic value estimation of matsutake mushroom and timber production in Swedish Scots pine forest. Master Thesis no. 218, Alnarp, Southern Swedish Forest Research Centre, Swedish University of Agricultural Sciences, Sweden.

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martínez-García PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu L-S, Gilbert D, Marçais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JFD, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, Main D, Loopstra CA, Mockaitis K, deJong PJ, Yorke JA, Salzberg SL, Langley CH. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* **15**:R59.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hällman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Käller M, Luthman J, Lysholm F, Niittylä T, Olson Å, Rilakovic N, Ritland C, Rosselló JA, Sena J, Svensson T, Talavera-López C, Theißen G, Tuominen H, Vanneste K, Wu Z-Q, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Gil RG, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**:579–584.

Outcalt KW. 2000. The longleaf pine ecosystem of the South. *Native Plants Journal* **1**:42–53.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**:289–290.

Pot D, McMillan L, Echt C, Le Provost G, Garnier-Géré P, Cato S, Plomion C. 2005. Nucleotide variation in genes involved in wood formation in two pine species. *New Phytologist* **167**: 101–112.

Rambaut A. 2013. *Figtree v. 1.4*. http://tree.bio.ed.ac.uk/software/figtree (16 January 2014).

Rambaut A, Drummond AJ. 2009. *Tracer v. 1.5*. http://beast.bio.ed.ac.uk/Tracer (26 April 2013).

R Core Team. 2012. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org (31 October 2012).

Rieseberg LH, Soltis DE. 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants* **5**: 65–84.

Ronquist F, Teslenko M, Van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**:539–542.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**:592–593.

Schmidtling RC. 2003. The southern pines during the Pleistocene. *ISHS Acta Horticulturae* **615**:203–209.

Schmidtling RC, Hipkins V. 1998. Genetic diversity in longleaf pine (*Pinus palustris*): influence of historical and prehistorical events. *Canadian Journal of Forest Research* **28**:1135–1145.

Schultz RP. 1997. *Loblolly pine. The ecology and culture of loblolly pine (Pinus taeda L.)*. Agricultural Handbook 713. Washington, DC: USDA Forest Service, U.S. Government Printing Office.

Seifert GJ, Roberts K. 2007. The biology of arabinogalactan proteins. *Annual Review of Plant Biology* **58**:137–161.

Sharitz RR, Boring LR, Van Lear DH, Pinder JE III. 1992. Integrating ecological concepts with natural resource management of southern forests. *Ecological Applications* **2**:226–237.

Shevenell AE, Kennett JP, Lea DW. 2004. Middle Miocene Southern Ocean cooling and Antarctic cryosphere expansion. *Science* **305**:1766–1770.

Showalter AM. 2001. Arabinogalactan-proteins: structure, expression and function. *Cellular and Molecular Life Sciences* **58**:1399–1417.

Smouse PE, Saylor LC. 1973. Studies of the *Pinus rigida-serotina* complex II. Natural hybridization among the *Pinus rigida-serotina* complex, *P. taeda* and *P. echinata*. *Annals of the Missouri Botanical Garden* **60**:192–203.

Snyder EB, Squillace AE. 1966. Cone and seed yields from controlled breeding of southern pines. New Orelans, LA: Southern Forest Experiment Station, USDA Forest Service (USDA For. Serv. Res. Pap. SO-22).

Soltis DE, Soltis PS, Milligan BG. 1992. Intraspecific chloroplast DNA variation: systematic and phylogenetic implications. In: Soltis PS, Soltis DE, Doyle JJ, eds. *Molecular systematics of plants*. New York, NY: Chapman & Hall, 117–150.

Stamatakis A. 2008. *The RAxML 7.0.4 manual*. http://sco.h-its.org/exelixis/resource/download/oldPage/RAxML-Manual.7.0.4.pdf (16 January 2014).

Stamatakis A. 2014. *The RAxML v8.0.X manual*. http://sco.h-its.org/exelixis/resource/download/NewManual.pdf (16 January 2014).

Sternitzke HS, Nelson TC. 1970. The southern pines of the United States. *Economic Botany* **24**:142–150.

Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**:1569–1571.

Suzuki Y, Glazko GV, Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of the National Academy of Sciences of the United States of America* **99**:16138–16143.

Syring J, Farrell K, Businsky R, Cronn R, Liston A. 2007. Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. *Systematic Biology* **56**:163–181.

Tyrväinen L, Väänänen H. 1998. The economic value of urban forest amenities: an application of the contingent valuation method. *Landscape and Urban Planning* **43**:105–118.

Vitt P, Havens K, Kramer AT, Sollenberger D, Yates E. 2010. Assisted migration of plants: changes in latitudes, changes in attitudes. *Biological Conservation* **143**:18–27.

Wear DN, Greis JG. 2002. Southern forest resource assessment: summary of findings. *Journal of Forestry* **100**:6–14.

Wells OO, Switzer GL, Schmidtling RC. 1991. Geographic variation in Mississippi loblolly pine and sweetgum. *Silvae Genetica* **40**:105–119.

Whetten R, Sederoff R. 1995. Lignin biosynthesis. *Plant Cell* **7**: 1001–1013.

Whittall JB, Syring J, Parks M, Buenrostro J, Dick C, Liston A, Cronn R. 2010. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* **19**:100–114.

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Molecular Biology and Evolution* **24**:90–101.

Willyard A, Cronn R, Liston A. 2009. Reticulate evolution and incomplete lineage sorting among the ponderosa pines. *Molecular Phylogenetics and Evolution* **52**:498–511.

Wood TE, Burke JM, Rieseberg LH. 2005. Parallel genotypic adaptation: when evolution repeats itself. *Genetica* **123**: 157–170.

Zhang Y, Sederoff RR, Allona I. 2000. Differential expression of genes encoding cell wall proteins in vascular tissues from vertical and bent loblolly pine trees. *Tree Physiology* **20**:457–466.

Zhang Y, Brown G, Whetten R, Loopstra CA, Neale D, Kieliszewski MJ, Sederoff RR. 2003. An arabinogalactan protein associated with secondary cell wall formation in differentiating xylem of loblolly pine. *Plant Molecular Biology* **52**:91–102.

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ, Neale DB, Salzberg SL, Yorke JA, Langley CH. 2014. Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics* **196**:875–890.

Zobel BJ. 1953. Are there natural loblolly-shortleaf pine hybrids? *Journal of Forestry* **51**:494–495.

Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD Dissertation, The University of Texas at Austin, USA.

