MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 5635–5650 doi: 10.1111/mec.12514

Multilocus analyses reveal little evidence for lineage-wide adaptive evolution within major clades of soft pines (*Pinus* subgenus *Strobus*)

ANDREW J. ECKERT,* ANDREW D. BOWER,† KATHLEEN D. JERMSTAD,‡ JILL L. WEGRZYN,§ BRIAN J. KNAUS,¶†† JOHN V. SYRING** and DAVID B. NEALE§

*Department of Biology, Virginia Commonwealth University, Richmond, VA 23284, USA, †USDA Forest Service, Olympic National Forest, Olympia, WA 98512, USA, ‡USDA Forest Service, Pacific Southwest Research Station, Institute of Forest Genetics, Placerville, CA 95667, USA, §Department of Plant Sciences, University of California at Davis, Davis, CA 95616, USA, ¶USDA Forest Service, Pacific Northwest Research Station, Corvallis, OR 97331, USA, **Department of Biology, Linfield College, McMinnville, OR 97128, USA

Abstract

Estimates from molecular data for the fraction of new nonsynonymous mutations that are adaptive vary strongly across plant species. Much of this variation is due to differences in life history strategies as they influence the effective population size (N_e). Ample variation for these estimates, however, remains even when comparisons are made across species with similar values of N_e. An open question thus remains as to why the large disparity for estimates of adaptive evolution exists among plant species. Here, we have estimated the distribution of deleterious fitness effects (DFE) and the fraction of adaptive nonsynonymous substitutions (α) for 11 species of soft pines (subgenus *Strobus*) using DNA sequence data from 167 orthologous nuclear gene fragments. Most newly arising nonsynonymous mutations were inferred to be so strongly deleterious that they would rarely become fixed. Little evidence for long-term adaptive evolution was detected, as all 11 estimates for α were not significantly different from zero. Nucleotide diversity at synonymous sites, moreover, was strongly correlated with attributes of the DFE across species, thus illustrating a strong consistency with the expectations from the Nearly Neutral Theory of molecular evolution. Application of these patterns to genome-wide expectations for these species, however, was difficult as the loci chosen for the analysis were a biased set of conserved loci, which greatly influenced the estimates of the DFE and α. This implies that genome-wide parameter estimates will need truly genome-wide data, so that many of the existing patterns documented previously for forest trees (e.g. little evidence for signature of selection) may need revision.

Keywords: adaptive evolution, distribution of deleterious fitness effects, nearly neutral theory of molecular evolution, *Pinus*, purifying selection, subgenus *Strobus*

Received 9 April 2012; revision received 27 August 2013; accepted 29 August 2013

Introduction

Understanding the relative roles of evolutionary forces within natural populations is a major goal of evolutionary genetics. While the neutral and nearly neutral theories of molecular evolution (Kimura 1968, 1983; Ohta 1973) have dominated research in population genetics for decades, a resurgence of alternative models incorporating a greater role for Darwinian natural selection has been advocated (Gillespie 2000, 2001; Hahn 2008; Sella *et al.* 2009). Fuelling this resurgence are empirical results from methods contrasting polymorphism and divergence for categories of sites that are evolving differently

Correspondence: David B. Neale, Fax: (530) 754 8431; E-mail: dbneale@ucdavis.edu

††Present address: USDA Agricultural Research Service, Horticultural Crops Research Unit, Corvallis, OR 97331, USA

(McDonald & Kreitman 1991). Analyses such as these have been applied to a diverse set of animal, plant and microbial species, with estimates for the fraction of new nonsynonymous mutations that are adaptive (α) varying greatly among species (e.g. Boyko *et al.* 2008; Axelsson & Ellegren 2009; Eyre-Walker & Keightley 2009; Gossmann *et al.* 2010; Halligan *et al.* 2010; Slotte *et al.* 2010), yet with many estimates illustrating a surprisingly large fraction of amino acid substitutions being fixed due to positive selection (e.g. 75% in sunflowers, Strasburg *et al.* 2009). An open question remains, however, as to why the large disparity for estimates of α exists among species (Sella *et al.* 2009).

One explanation is that variation in the effective population size (N_e) among species drives variation in the efficacy of selection (Ohta 1973, 1992; Kimura 1983). This is because the strength of selection is the product of N_e and the selection coefficient, s, so that for a given s, increasing N_e will increase the strength of selection. In this case, adaptive mutations will fix more rapidly and deleterious mutations will be eliminated more efficiently. In addition, the waiting time for a beneficial mutation will decrease as N_e increases (Gillespie 1999; Charlesworth 2009; but see Gillespie 2001). The effect of N_e on the efficacy of selection has been shown routinely in the theoretical and empirical literature, especially for N_e and observed levels of protein sequence conservation (e.g. Ohta & Gillespie 1996; Ohta 2002). A relationship between N_e and efficacy of positive selection has also been observed, so that species with large N_e tend to have higher estimates α as compared to species with smaller N_e (e.g. Strasburg *et al.* 2011). For example, estimates for α range from ≈50% for *Drosophila* (reviewed by Sella *et al.* 2009) to ≈30% for *Populus tremula* (Ingvarsson 2010), both of which have relatively large N_e, to <10% in humans, which have a comparatively lower N_e (Boyko *et al.* 2008; Eyre-Walker & Keightley 2009). Some caution, however, is needed, because the number of effectively neutral mutations is also correlated with N_e (Piganeau & Eyre-Walker 2009), and the theoretical relationship between α and N_e is nonlinear when considering further complexities such as linkage and fluctuating selection coefficients (Gillespie 2000).

Striking exceptions to the observed relationship between N_e and α, however, are the estimates for many plant species. The results concerning the prevalence of adaptive amino acid evolution for plants are mixed (Bustamante *et al.* 2002; Hamblin *et al.* 2006; Foxe *et al.* 2008; Ross-Ibarra *et al.* 2009; Strasburg *et al.* 2009, 2011; Gossmann *et al.* 2010, 2012; Ingvarsson 2010; Slotte *et al.* 2010), with point estimates for α (or a closely related quantity, ω_a) varying more than 10-fold across species. A recent comparative analysis of nine plant species, which varied 30-fold in estimates of N_e, highlighted a surprising dearth of adaptive amino acid evolution (Gossmann *et al.* 2010). Only a single species, *Helianthus petiolaris*, had an estimated value of α significantly different from zero (Fig. 2a in Gossmann *et al.* 2010). A relationship between N_e and α was also lacking, yet there was a strong positive relationship between N_e and the strength of purifying selection. This anomaly was attributed to a number of reasons—effects of N_e, long-term population size changes, population structure, problematic designation of orthology or too little divergence between ingroup and outgroup. In contrast, a multispecies analysis within the genus *Helianthus* identified a strong linear trend between N_e and α (Strasburg *et al.* 2011) as did an analysis across 13 pairs of eukaryotes species for ω_a (Gossmann *et al.* 2012), although the relationships in both of these studies were driven by negative values (cf. Fig. 1 from Gossmann *et al.* 2012). While differences among methodological approaches (e.g. Bierne & Eyre-Walker 2004; Eyre-Walker & Keightley 2009) may be driving variation for estimates of α for plants, it is clear that a wider taxonomic sampling is needed to address fully why plants deviate so much from estimates for animals, fungi and bacteria.

Life history characteristics of plants affect standing levels of genetic diversity (Hamrick & Godt 1996; Ledig 1998; Gaut *et al.* 2007). It is reasonable, therefore, to expect that estimates of α will also vary among plants with differing life histories. Most studies to date have focused solely on domesticated species (e.g. *Zea mays*), inbreeding annuals (e.g. *Arabidopsis thaliana*), or outcrossing, short-lived perennials (e.g. *Arabidopsis lyrata*). An exception is the study of Ingvarsson (2010) for *P. tremula*, which is a long-lived, undomesticated and outcrossing forest tree species with large N_e. Forest trees have a suite of life history characteristics that facilitate large N_e, low population structure and high levels of genetic diversity (Neale & Kremer 2011). Such characteristics enable a more thorough investigation of the reasons highlighted by Gossmann *et al.* (2010) and contrasted by Strasburg *et al.* (2011) for the uniqueness or nonuniqueness of plants with respect to estimated rates of adaptive evolution.

To address these concerns, we resequenced 167 orthologous nuclear loci in 11 species of closely related pines (genus *Pinus*). Within *Pinus,* most genomic research has focused on economically important species such as loblolly pine (*Pinus taeda* L.). Largely ignored are members of subgenus *Strobus*, which arguably contains species that occupy the widest set of ecological niches within the genus (Mirov 1967). We use these data to: (i) characterize the levels of nucleotide diversity and divergence for each of the 11 species considered, (ii) infer the fraction of amino acid evolution that is adaptive conditional on the distribution of deleterious

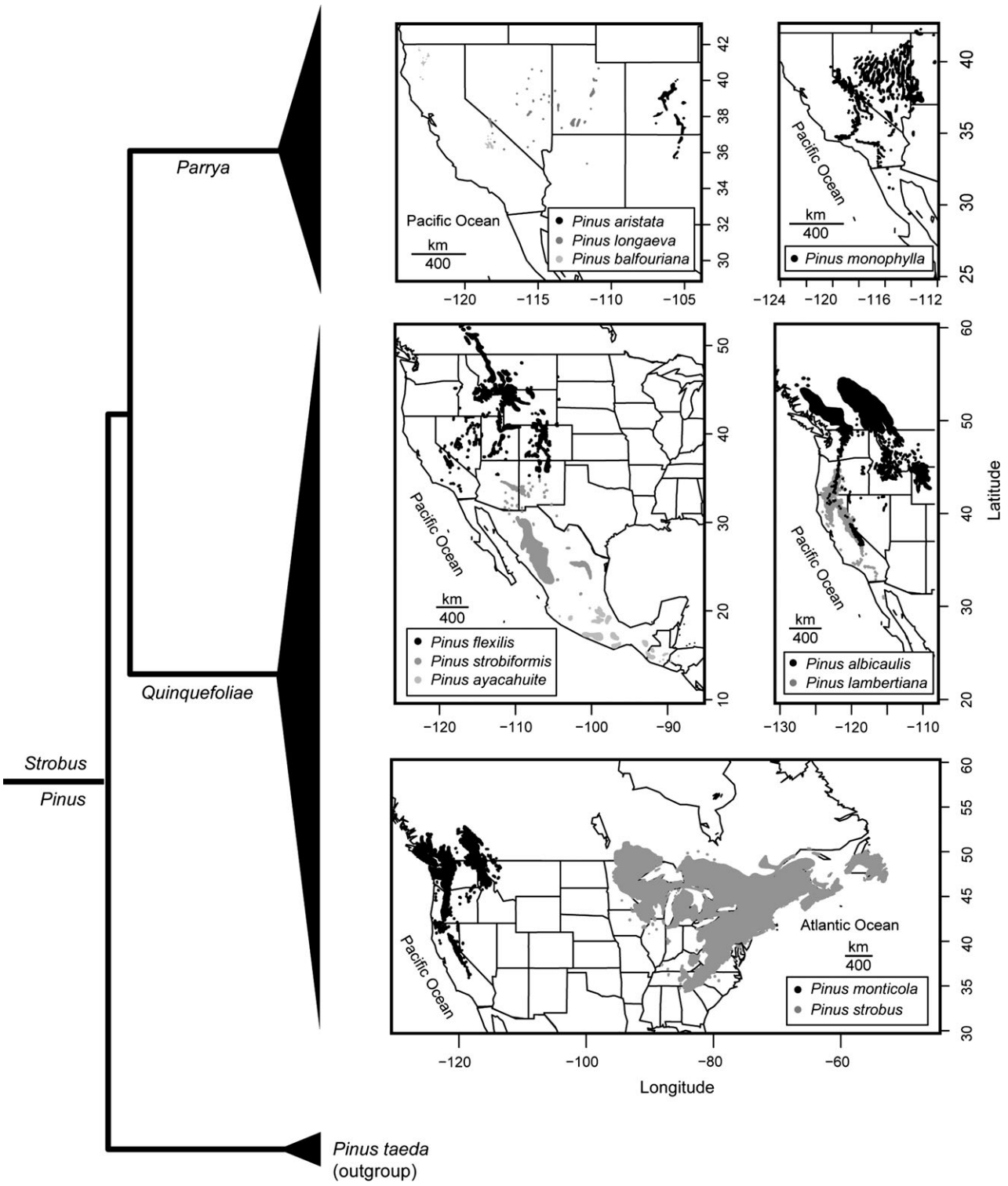

Fig. 1 Distribution and phylogeny for 11 species of soft pines (subgenus *Strobus*) used for multilocus inferences of adaptive evolution. Taxonomy follows that given by Gernandt *et al.* (2005).

fitness effects (DFE), and (iii) show that these inferences are confounded by population structure and various forms of ascertainment and methodological biases. In doing so, we produce a large set of orthologous nuclear gene markers for an ecologically important, yet understudied, clade of pine species.

Materials and methods

Sampling

Samples of trees were assembled into SNP discovery panels of eight to 19 trees for each species (see Appendix S1, Supporting information). Samples were chosen to be range-wide when possible, with single to only a few trees being sampled per geographical region per species. Exceptions to this included equal sample sizes for the regional populations of *Pinus balfouriana* (Eckert *et al.* 2008) and geographically restricted sampling for Rocky Mountain bristlecone pine. Sampled trees are described in further detail in Table S1 (Supporting information).

Geographical Information Systems (GIS) shape files encapsulating range maps for each species were obtained from the United States Geological Survey (http://esp.cr.usgs.gov/data/atlas/little/). These are digitized versions of the range maps developed by Critchfield & Little (1966). These shape files were used for the visualization of species' ranges and validation of range coverage for selected samples, as well as for the estimation of range size (km^2). The latter was carried out using the MAPTOOLS, RGDAL and PBSMAPPING geospatial libraries in R (R Core Development Team 2010).

For each sampled tree, separate DNA extractions from five to 10 haploid megagametophytes were performed using 96-well DNeasy Plant Mini Kits (Qiagen, Inc., Valencia, CA) following the manufacturer's protocol. Multiple megagametophytes per tree were excised so as to obtain enough DNA for resequencing. All DNA extractions were carried out at the National Forest genetics Electrophoresis Laboratory (Institute of Forest Genetics, USDA National Forest Service, Placerville, CA).

Sequence analysis

Single-pass bidirectional Sanger sequencing was carried out on haploid DNA templates using a subset of the 7535 PCR primer sets developed for loblolly pine and the protocols described by Eckert *et al.* (2010a). Loci chosen for sequencing were those that PCR amplified and sequenced cleanly from a target set of 800 primer pairs for single samples of several soft pines (*Pinus albicaulis*, *P. aristata*, *P. balfouriana*, *P. cembroides*, *P. lambertiana* and *P. longaeva*). This initial screen resulted in 245 primer sets, of which 238 were from unique loci. These 245 primer sets were used to resequence the SNP discovery panels described previously using ABI 3730XL Genome Analyzers (Life Technologies Corporation, Carlsbad, CA) by Beckman Coulter Genomics (Beverly, MA, USA). We chose loblolly pine as the outgroup for all 11 species due to the greater likelihood of reciprocal monophyly (Syring *et al.* 2007). Data for loblolly pine were obtained from the ADEPT2 ftp website (http://loblolly.ucdavis.edu/bipod/ftp/).

Sequence data were processed and organized using PineSAP (Wegrzyn *et al.* 2009), followed by manual assembly and alignment using CODONCODE ver. 2.0.4 (CodonCode Corporation, Dedham, MA). Bases were called using only nucleotides that had a probability of <0.01 of being miscalled (PHRED $Q \geq 20$) (Ewing & Green 1998; Ewing *et al.* 1998) SNPs were called using a combination of POLYPHRED (Nickerson *et al.* 1997), POLYBAYES (Marth *et al.* 1999) and visual validation. Of the 245 loci, 37 were dropped from further consideration due to overall low quality of sequence reads, 15 were dropped due to identification as chloroplast or mitochondrial contamination using BLAST analysis against pine organellar sequences deposited in GenBank (Parks *et al.* 2009), and one locus was dropped due to its sequence similarity to a retroelement-like protein. This process resulted in a core set of 192 high-quality nuclear gene alignments for the 11 target species. All alignments for each species and gene were further aligned to resequencing data from loblolly pine using the profile-profile option in MUSCLE ver. 3.8.31 (Edgar 2004a,b). All data are deposited in GenBank (Appendix S2, Supporting information).

Coding regions could confidently be deduced for 167 of the 192 loci by further analysis using BLASTX against gene models for *Arabidopsis*, *Oryza*, *Populus* and *Picea*. For these 167 loci, the gene model for the highest scoring BLAST hit, in combination with the expressed sequence tag from loblolly pine, was used to deduce coding regions. Site annotations for each alignment were validated with BLASTp analysis of the amino acid sequences derived from the deduced coding intervals against the gene model that was used to derive the site annotations. Evolutionary genetic analysis was restricted to polymorphism and divergence data for the 167 fully annotated loci. More information about these loci is given in Appendix S1 (Tables S2–S4, Supporting information).

Polymorphism and divergence

Levels of polymorphism were quantified using two estimators of $\Theta = 4N_e u$: θ_W (Watterson 1975) and θ_π (Tajima 1983). Divergence was calculated using pairwise nucleotide divergence (D_{xy}; Nei 1987). Estimates were made for different categories of sites for all statistics. When multiple mutations were segregating within a single codon and the pathways could not be unambiguously assigned as silent or replacement, we assumed the most conservative pathway. We further assumed an infinite site model for all polymorphism calculations, and thus, sites violating this model were dropped prior to estimation. Intragenic linkage disequilibrium (LD) for each of

the 11 species was described using the r^2 and Z_{nS} statistics (Hill & Robertson 1968; Kelly 1997; Appendix S1, Supporting information). Unfolded site-frequency spectra were determined separately for each gene and species using loblolly pine as an outgroup and were further summarized using two summary statistics: Tajima's D (Tajima 1989) and Fay and Wu's H (Fay & Wu 2000). All calculations were conducted using DNASAM (Eckert *et al.* 2010b) and the libsequence, analysis, and gestimator C++ libraries (Thornton 2003). The latter C++ library uses the method of Comeron (1995) to calculate nonsynonymous (K_a) and synonymous (K_s) divergences. Site-frequency spectra were combined across loci for each species by adjusting sample sizes (n) to the mean size across all loci using the resampling methods of Nielsen *et al.* (2005) and J. Ross-Ibarra and R. A. Cartwright (unpublished, http://www.scit.us/sofos/).

The distribution of deleterious fitness effects and the fraction of adaptive substitutions

We estimated the distribution of DFE, α and the fraction of adaptive substitutions relative to synonymous substitutions (ω_a) for each of the 11 species using the method of Eyre-Walker & Keightley (2009). This method accounts for the segregation and fixation of deleterious mutations by first estimating the DFE from polymorphism data. The DFE was binned into four categories: $0 < N_e s < 1$, $1 < N_e s < 10$, $10 < N_e s < 100$ and $N_e s > 100$, where N_e is the effective population size and s is the selection coefficient. Historical demography is accounted for heuristically by comparing the observed site-frequency spectrum at synonymous sites to that expected for a neutral coalescent process and assuming subsequently that demography affects equally the site-frequency spectrum at nonsynonymous sites (Eyre-Walker *et al.* 2006). Specifically, we used the Bayesian implementation of method 2 from Eyre-Walker & Keightley (2009), which is available as part of the DOFE ver. 3.0 software (http://www.lifesci.sussex.ac.uk/home/Adam_Eyre-Walker/Website/Software.html). We explicitly base our conclusions on this method, as it has fewer assumptions (e.g. it does not specify a demographic model), and make comparisons between this method and those that specify demographic models (see Investigation of bias).

Posterior distributions for α and the DFE were generated via Markov chain Monte Carlo (MCMC). A two-step procedure was used to verify convergence to similar posterior distributions for model parameters. First, three independent MCMC analyses were run from different starting points for 10^6 steps after a burn-in of 10^5 steps. Samples were retained every 10^3 steps. Differences among the resulting posterior distributions for α and the DFE were assessed with Kolmogorov–Smirnov tests. Second, if large-scale differences were not discovered among the posterior distributions resulting from the three initial runs, a single run of the MCMC sampler was initiated using the previous point estimates for model parameters as starting values for 10^7 steps after a burn-in of 10^5 steps. Samples were retained every 10^4 steps. Posterior distributions were summarized using the median, standard deviation and the 95% credible interval as calculated using the CODA library in R.

Investigation of bias

Three sources of bias for our estimates of α were investigated. First, we addressed a form of ascertainment bias inherent to the primer design. As described previously, loci were selected based upon primers developed for loblolly pine (i.e. the outgroup). This results in biased estimates of genome-wide values for diversity and divergence (Kern 2009) and could reflect an increased level of functional constraint for the analysed loci (Gossmann *et al.* 2010; Haddrill *et al.* 2010). To investigate these biases, we created 10 000 data sets of 167 loci by randomly sampling the resequencing data set for loblolly pine (n = 5772 loci of which 3162 were annotated) and estimated levels of polymorphism and divergence as described for the observed data. We compared observed means for diversity and divergence to these distributions to assess the degree of bias.

Second, we compared estimates of model parameters using method 1 and method 2 of Eyre-Walker & Keightley (2009). Method 1 assumes an explicit demographic model, while method 2 performs an approximate correction for demography. Both methods have recently been employed for plant species and have produced contradictory results (e.g. Gossmann *et al.* 2010; Slotte *et al.* 2010), although the set of genes used for the analysis and sampling schemes differed between these two studies. The same data sets as used with method 2 were analysed using method 1 as implemented through the DFE-alpha web server application (http://homepages.ed.ac.uk/eang33/). This method assumes a population size change model with a single change from N_2 to N_1 occurring t generations in the past (Keightley & Eyre-Walker 2007). This method is approximately unbiased with respect to α when assuming the single-step demographic model described above, even when this model is wrong, although the simulations did not examine the effects of population structure (Messer & Petrov 2013). Maximum likelihood is then used to jointly estimate the demographic model parameters and the DFE from which an estimate of α is made as for method 2. Nonparametric bootstrapping over loci was used to assess uncertainty in parameter estimates (n = 100 replicates).

Third, we assessed the effects of underlying population structure on inferences of the DFE, α and ω_a using both methods of Eyre-Walker & Keightley (2009). We searched for population structure in each species prior to this assessment using STRUCTURE ver. 2.3.2.1 (Falush *et al.* 2003) with the independent allele frequencies and no admixture models (see Appendix S1, Supporting information). Previous work has assessed the effect of population structure on estimates of the DFE using samples from single populations or balanced samples across multiple populations (Eyre-Walker *et al.* 2006; Gossmann *et al.* 2010). The way samples are gathered across populations, however, affects inferences based on the site-frequency spectrum (Städler *et al.* 2009). We assessed sampling intensities across clusters inferred using the results from STRUCTURE by assigning trees to clusters using the maximum admixture coefficient and estimating the fraction of the sampled trees assigned to each of the two clusters.

We validated the effect of population structure and sampling intensity on estimation of the DFE, α and ω_a using simulations of two neutral populations of size N_e = 10 000 sharing genes at a rate in which to produce an observed F_{ST} of 0.15. Two sampling schemes were employed (n = 8 from each population vs. n = 3 from one population and n = 13 from the other), with each sampling scheme randomly choosing 200 of the simulated 5000 genes for the analysis. Each gene in each population was assumed to have a selection coefficient drawn from a discretized distribution (n = 4 bins) resembling the observed values of the DFE estimated for *P. balfouriana* (see Results). For comparison, a single neutral population of size N_e = 10 000 was also simulated. For each data set, α was calculated using method 1 and method 2 as described previously. A total of 100 simulations for each scenario were conducted using SFS_CODE (Hernandez 2008) followed by processing of SFS_CODE output using convertSFS_CODE with the –MK and –SFS options. Output from convertSFS_CODE was thus the input for inference of selection parameters. All the results of the simulations, therefore, are based on the known, and hence true, values for the site-frequency spectrum and levels of divergence. This differs from the observed data, where parsimony was used to reconstruct the ancestral state based on an outgroup (i.e. *Pinus taeda*).

Results

Polymorphism and divergence

Mean nucleotide diversities (θ_π) across loci were similar to those previously reported for conifers (Savolainen & Pyhäjärvi 2007), ranging from 0.14% to 0.37% (Table 1). Diversities at synonymous sites were approximately three- to fivefold larger than at nonsynonymous sites. The mean Tajima's D and Fay and Wu's H at synonymous sites were negative, with the only exception being the slightly positive value of Tajima's D for *Pinus aristata*. This is indicative of site-frequency spectra that are U-shaped due to a skew towards low- and high-frequency variants (Fig. S1, Supporting information). There also appeared to be no influence of rescaling on the shape of site-frequency spectra across species, as site-frequency spectra across sample coverage classes largely resembled those of the adjusted spectra (Figs. S1 and S2, Supporting information). Mean estimates of Tajima's D at nonsynonymous sites were more negative on average than at synonymous sites, while mean values of Fay and Wu's H varied between slightly positive and negative depending on the species (Table 1).

Mean nucleotide divergence (D_{xy}) across loci ranged from 3.7% to 3.9% across species (Table 1). Divergence at synonymous sites (K_s) was 4.5-fold greater than at nonsynonymous sites (K_a) on average, with values of K_a and K_s ranging from 1.3% (*Pinus monticola*) to 1.9% (*Pinus monophylla*) and from 6.1% (*P. monticola*) to 8.5% (*Pinus ayacahuite*), respectively. Nucleotide divergence at all synonymous and nonsynonymous sites was also highly correlated with loci across species (pairwise Spearman's $\rho > 0.95$, $P < 0.001$ [P-values estimated using 10 000 permutations]).

Evidence for selective sweeps from polymorphism summaries was lacking for each species. Nucleotide diversity at synonymous sites was uncorrelated with divergence at nonsynonymous sites (Spearman's ρ: 0.01–0.18, $P > 0.05$). LD for most species decayed rapidly and, after removing singletons, decayed to ~0.25 after 500 bp for most species (Fig. S3, Supporting information). The exceptions to this pattern were those species with the strongest levels of inferred population structure (see below). Evidence for selective sweeps from polymorphism and divergence summaries was also lacking for each species. Only 4 (*Pinus aristata*) to 11 (*Pinus strobus*) loci per species had $K_a/K_s > 1$, with all species having an average $K_a/K_s < 0.40$. The segregation of slightly deleterious mutations and the influence of the effective population size (N_e) on this pattern, however, were apparent for all species. Nucleotide diversity for all synonymous and nonsynonymous sites was strongly correlated with the $\log_{10}$ of range size (Spearman's $\rho > 0.50$, $P < 0.005$), even after correcting for differences in sample size [i.e. Spearman's $\rho > 0.40$ ($P < 0.01$) using residuals from regressing out influence of sample size and diversity with $\log_{10}$ range size], while the ratio of nonsynonymous to synonymous polymorphisms (P_n/P_s) was negatively correlated with θ_W (Spearman's $\rho = -0.64$, $P = 0.003$) using the method of

Table 1 Polymorphism and divergence across 167 genes for 11 species sampled from subgenus *Strobus**

	Quinquefoliae							*Parrya*			
Statistic	Pial	Piay	Pifl	Pila	Pimn	Pisb†	Pist	Piar	Pibf	Pilo	Pimo
n	12	15	10	12	16	15	16	8	14	10	13
Size‡	5.60	4.77	5.37	5.03	5.53	5.26	6.28	4.30	3.41	3.93	4.88
All											
bp	397 (66 347)	397 (66 418)	398 (66 449)	398 (66 455)	398 (66 533)	398 (66 008)	398 (66 537)	399 (66 608)	398 (66 439)	398 (66 452)	397 (66 377)
S	4 (663)	4 (652)	5 (756)	3 (576)	5 (901)	5 (884)	5 (778)	1 (232)	3 (456)	3 (564)	5 (799)
θ_π	0.0033	0.0031	0.0036	0.0028	0.0037	0.0037	0.0033	0.0014	0.0021	0.0028	0.0035
θ_W	0.0035	0.0032	0.0041	0.0030	0.0042	0.0043	0.0038	0.0015	0.0023	0.0031	0.0038
D_{xy}	0.0389	0.0385	0.0383	0.0378	0.0376	0.0381	0.0380	0.0367	0.0372	0.0361	0.0388
D	−0.346	−0.309	−0.512	−0.398	−0.548	−0.569	−0.474	−0.158	−0.355	−0.417	−0.396
H	−0.270	−0.482	−0.320	−0.273	−0.290	−0.449	−0.611	−0.399	−0.037	−0.409	−0.322
Z_{nS}	0.428	0.362	0.371	0.357	0.280	0.340	0.389	0.556	0.343	0.431	0.269
Syn											
bp	65 (10 838)	65 (10 871)	64 (10 751)	65 (10 900)	65 (10 889)	65 (10 768)	65 (10 827)	65 (10 939)	65 (10 866)	65 (10 836)	65 (10 816)
S	1 (240)	2 (274)	2 (291)	1 (245)	2 (335)	2 (353)	2 (311)	<1 (73)	1 (145)	1 (187)	2 (311)
θ_π	0.0071	0.0079	0.0083	0.0071	0.0093	0.0089	0.0082	0.0025	0.0041	0.0050	0.0088
θ_W	0.0075	0.0081	0.0095	0.0076	0.0094	0.0102	0.0089	0.0004	0.0043	0.0058	0.0092
K_s	0.0831	0.0854	0.0841	0.0833	0.0611	0.0843	0.0829	0.0783	0.0800	0.0779	0.0808
D	−0.199	−0.048	−0.370	−0.266	−0.247	−0.316	−0.195	0.004	−0.121	−0.442	−0.185
H	−0.256	−0.338	−0.242	−0.262	−0.249	−0.380	−0.314	−0.359	−0.031	−0.287	−0.236
Z_{nS}	0.398	0.385	0.338	0.334	0.269	0.248	0.359	0.383	0.385	0.322	0.181
Nonsyn											
bp	236 (39 391)	237 (39 505)	234 (39 063)	237 (39 619)	237 (39 579)	236 (39 096)	235 (39 327)	238 (39 753)	237 (39 502)	236 (39 420)	236 (39 400)
S	1 (235)	1 (247)	2 (278)	1 (180)	2 (319)	2 (338)	2 (282)	<1 (83)	1 (174)	1 (196)	2 (311)
θ_π	0.0020	0.0017	0.0020	0.0013	0.0020	0.0021	0.0018	0.0007	0.0011	0.0014	0.0022
θ_W	0.0018	0.0019	0.0021	0.0015	0.0024	0.0027	0.0022	0.0008	0.0014	0.0017	0.0025
K_a	0.0185	0.0185	0.0187	0.0183	0.0134	0.0186	0.0188	0.0178	0.0179	0.0175	0.0197
D	−0.514	−0.449	−0.583	−0.400	−0.689	−0.654	−0.576	−0.152	−0.777	−0.583	−0.558
H	0.103	−0.077	0.040	0.046	0.039	0.008	−0.295	−0.271	−0.063	−0.131	−0.048
Z_{nS}	0.342	0.333	0.351	0.314	0.208	0.233	0.333	0.371	0.409	0.318	0.267

bp, alignment length in base pairs; *D*, Tajima's *D*; D_{xy}, nucleotide divergence; *H*, Fay and Wu's *H*; K_a, nucleotide divergence at nonsynonymous sites; K_s, nucleotide divergence at synonymous sites; *n*, sample size; *S*, number of segregating sites; θ_π, nucleotide diversity; θ_W, Watterson's estimate of nucleotide diversity; Z_{nS}, average pairwise r^2. Pial, *Pinus albicaulis*; Piar, *P. aristata*; Piay, *P. ayacahuite*; Pibf, *P. balfouriana*; Pifl, *P. flexilis*; Pimn, *P. monticola*; Pimo, *P. monophylla*; Pisb, *P. strobiformis*; Pist, *P. strobus*.

*Numbers are averages across genes (n = 167), with totals given in parentheses. Averages were rounded to the closest whole number for values that should be discrete (e.g. sequence length).

†Pisb only has 166 loci due to lack of PCR amplification for locus *CL3007Contig1_02*.

‡Estimate of range size in units of km^2. Note these are on a $\log_{10}$ scale.

Piganeau & Eyre-Walker (2009). If the mutation rate is constant across lineages, these correlations establish that N_e varies as a function of the $\log_{10}$ of range size and that P_n/P_s is influenced by N_e. Thus, there should be a negative correlation between P_n/P_s and range size, which is what is observed (Spearman's $\rho = -0.56$, $P = 0.012$).

The distribution of deleterious fitness effects and the fraction of adaptive substitutions

The distribution of DFE for new nonsynonymous mutations was estimated using a Bayesian method that accounts heuristically for historical demography (Eyre-Walker & Keightley 2009). Following previous studies, we binned the DFE into four categories based on the magnitude of N_es. The estimates for the DFE for all species were characterized by most (>50%) new nonsynonymous mutations being so strongly deleterious ($N_es > 100$) that they would very rarely become fixed (Table 2, Fig. 2). Most other new nonsynonymous mutations (20–30%) were inferred to be weakly deleterious ($N_es < 1$). Few (<20%) new nonsynonymous mutations were moderately deleterious ($1 < N_es < 100$). There were differences between clades of soft pines for estimates of the DFE, with species in section *Parrya* having a smaller fraction of new nonsynonymous mutations strongly selected against on average (57% vs. 61% for section *Quinquefoliae*) and a larger fraction of new nonsynonymous mutations weakly selected against on average (27% vs. 22% for section *Quinquefoliae*). Estimates were largely consistent, however, across species and independent runs of the MCMC sampler (Fig. S4, Supporting information).

The fraction of adaptive, nonsynonymous substitutions (α) was estimated conditional on the DFE for each species. Point estimates of α ranged from -0.233 (*P. aristata*) to 0.118 (*P. strobus*), with an average across species of -0.015 (Table 2). When standardized against the fraction of synonymous substitutions (ω_a), estimates ranged from -0.0477 (*P. aristata*) to 0.0325 (*P. strobus*), with estimates of α and ω_a being strongly correlated (Spearman's $\rho = 0.755$, $P < 0.001$, see Table S6, Supporting information). Estimates of α were also correlated with nucleotide diversity at synonymous sites (Spearman's $\rho = 0.536$, $P = 0.019$). This correlation was driven by the negative values estimated for α, so that setting negative values for α to zero decreased the correlation (Spearman's $\rho = 0.372$, $P = 0.048$), and removal of negative values caused the correlation to become negative (Spearman's $\rho = -0.771$, $P = 0.029$). Estimates of α were most dependent upon bins in the DFE corresponding to $N_es < 1$ (Spearman's $\rho = -0.936$, $P < 0.001$) and $N_es > 100$ (Spearman's $\rho = 0.825$, $P < 0.001$), which were negatively correlated with one another (Spearman's $\rho = -0.866$, $P = 0.007$). The differences in DFE estimates mentioned previously, therefore, manifested themselves as differences in the mean estimates of α between clades (*Parrya*: -0.123, *Quinquefoliae*: 0.047). Similar correlations were also observed between these quantities and ω_a (e.g. θ_π and ω_a: Spearman's $\rho = 0.547$, $P = 0.002$). The 95% credible intervals for α were large and included zero, however, for all species (Table 2, Fig. 3). The same was true for estimates of ω_a.

Table 2 Estimates of the distribution of deleterious fitness effects (DFE) and the fraction of adaptive substitutions (α)*

	DFE				
Species	$0 < N_es < 1$	$1 < N_es < 10$	$10 < N_es < 100$	$N_es > 100$	α
Quinquefoliae					
Pial	0.243 (0.196, 0.297)	0.064 (0.058, 0.108)	0.081 (0.073, 0.160)	0.604 (0.501, 0.656)	−0.028 (−0.294, 0.200)
Piay	0.217 (0.174, 0.265)	0.065 (0.057, 0.123)	0.084 (0.073, 0.202)	0.625 (0.478, 0.676)	0.067 (−0.174, 0.281)
Pifl	0.238 (0.197, 0.286)	0.062 (0.058, 0.080)	0.078 (0.073, 0.108)	0.616 (0.559, 0.662)	−0.006 (−0.249, 0.191)
Pila	0.213 (0.138, 0.273)	0.067 (0.057, 0.268)	0.087 (0.073, 0.474)	0.612 (0.401, 0.678)	0.101 (−0.190, 0.478)
Pimn	0.224 (0.168, 0.269)	0.065 (0.058, 0.176)	0.084 (0.073, 0.328)	0.616 (0.322, 0.664)	0.047 (−0.192, 0.350)
Pisb	0.228 (0.187, 0.270)	0.071 (0.058, 0.115)	0.094 (0.073, 0.179)	0.606 (0.495, 0.662)	0.031 (−0.192, 0.224)
Pist	0.215 (0.159, 0.261)	0.065 (0.057, 0.183)	0.084 (0.073, 0.349)	0.624 (0.304, 0.673)	0.118 (−0.103, 0.403)
Parrya					
Piar	0.295 (0.211, 0.414)	0.063 (0.058, 0.081)	0.077 (0.069, 0.105)	0.558 (0.442, 0.645)	−0.233 (−0.776, 0.143)
Pibf	0.282 (0.175, 0.370)	0.067 (0.059, 0.131)	0.083 (0.071, 0.255)	0.542 (0.351, 0.620)	−0.189 (−0.601, 0.352)
Pilo	0.264 (0.211, 0.330)	0.063 (0.059, 0.079)	0.078 (0.072, 0.104)	0.588 (0.519, 0.646)	−0.113 (−0.429, 0.138)
Pimo	0.247 (0.206, 0.296)	0.064 (0.059, 0.094)	0.081 (0.073, 0.131)	0.601 (0.527, 0.647)	0.044 (−0.180, 0.230)

Pial, *Pinus albicaulis*; Piar, *P. aristata*; Piay, *P. ayacahuite*; Pibf, *P. balfouriana*; Pifl, *P. flexilis*; Pimn, *P. monticola*; Pimo, *P. monophylla*; Pisb, *P. strobiformis*; Pist, *P. strobus*.

*Numbers are point estimates with 95% credible intervals in parentheses.

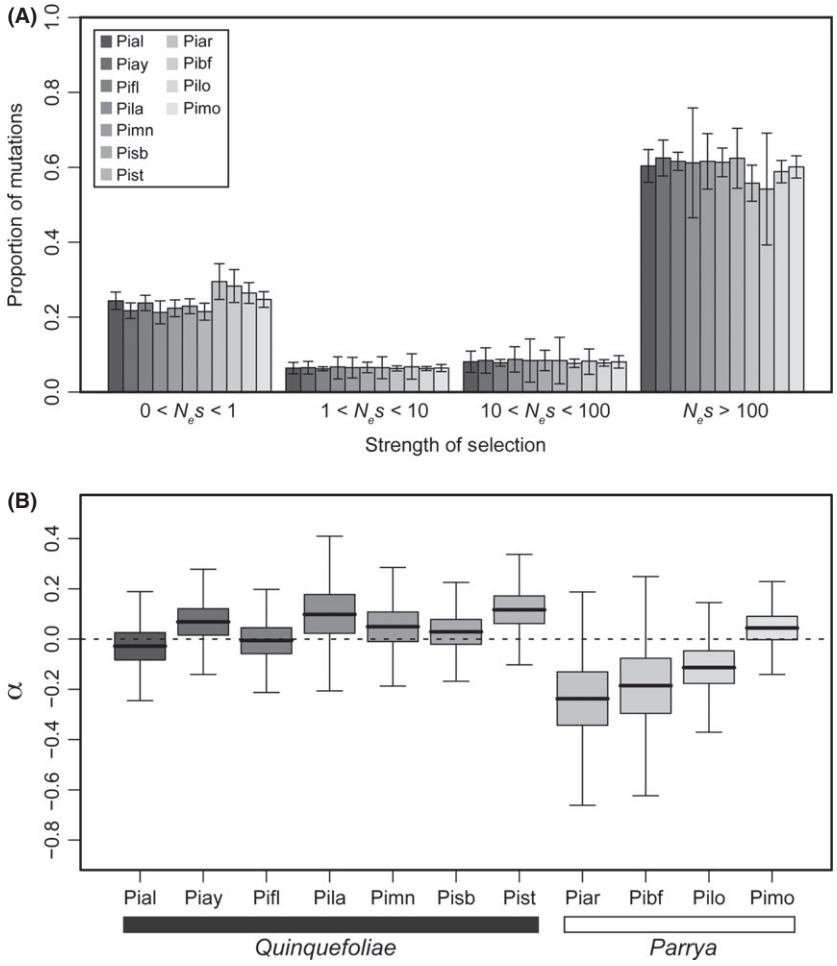

Fig. 2 Distribution of deleterious fitness effects (DFE) and the fraction of adaptive substitutions (α) for 11 species of soft pines (subgenus *Strobus*) as inferred using method 2 of Eyre-Walker & Keightley (2009). (A) The inferred DFE for each species, with point estimates given by the mean (± one standard deviation) of the posterior distribution. (B) The distribution of α for each species. Whiskers on the boxplots extend to the limits of the posterior distribution in each case, with point estimates given as the median of the posterior distribution. Pial, *Pinus albicaulis*; Piar, *P. aristata*; Piay, *P. ayacahuite*; Pibf, *P. balfouriana*; Pifl, *P. flexilis*; Pimn, *P. monticola*; Pimo, *P. monophylla*; Pisb, *P. strobiformis*; Pist, *P. strobus*.

Investigation of bias

Nucleotide diversity in loblolly pine was downwardly biased for the loci chosen for resequencing (Fig. 3A). This bias was accounted for by diversity at nonsynonymous sites being twofold less diverse on average than sets of 167 alignments drawn at random from all annotated alignments for loblolly pine (n = 3162). There was no effect on diversity at synonymous sites. This translated into the ratio of nonsynonymous to synonymous diversity also being significantly lower on average for loci chosen for this analysis (P = 0.033). Nucleotide divergence between loblolly and sugar pines was also downwardly biased for the loci selected for resequencing (Fig. 3B). This downward bias was approximately proportional for nonsynonymous and synonymous sites, so that their average ratio remained similar to that for sets of 167 loci drawn at random from all annotated alignments including both loblolly and sugar pine (n = 647).

Estimates of α using the two methods of Eyre-Walker & Keightley (2009) were substantially different. Parameter estimates for the demographic model assumed by method 1 are located in Table S7 (Supporting information). On average, point estimates of α were greater when using method 1 vs. method 2 (Fig. 4A), although the ranks of estimates for α across species were correlated (Spearman's ρ = 0.473, P = 0.028). In one extreme case, the point estimate of α for *Pinus balfouriana* changed from −0.189 (method 2) to 0.965 (method 1). Choice of method also affected inference of the statistical significance for α and ω_a, with estimates for *P. balfouriana* and *P. lambertiana* from method 1 having bootstrap confidence intervals excluding zero (Fig. S5, Supporting information). The differences in α estimates between methods were due to differences in the inferred DFE (Fig. 4B), as the differences in α were highly correlated with the summed differences across bins in the DFE inferred by each method (Spearman's ρ = 0.954, P < 0.001). Higher estimates of α were due to shifts in the DFE towards the increased prevalence of weakly and moderately selected sites (Fig. 4B).

One possible cause for different estimates of the DFE is underlying population structure. Population

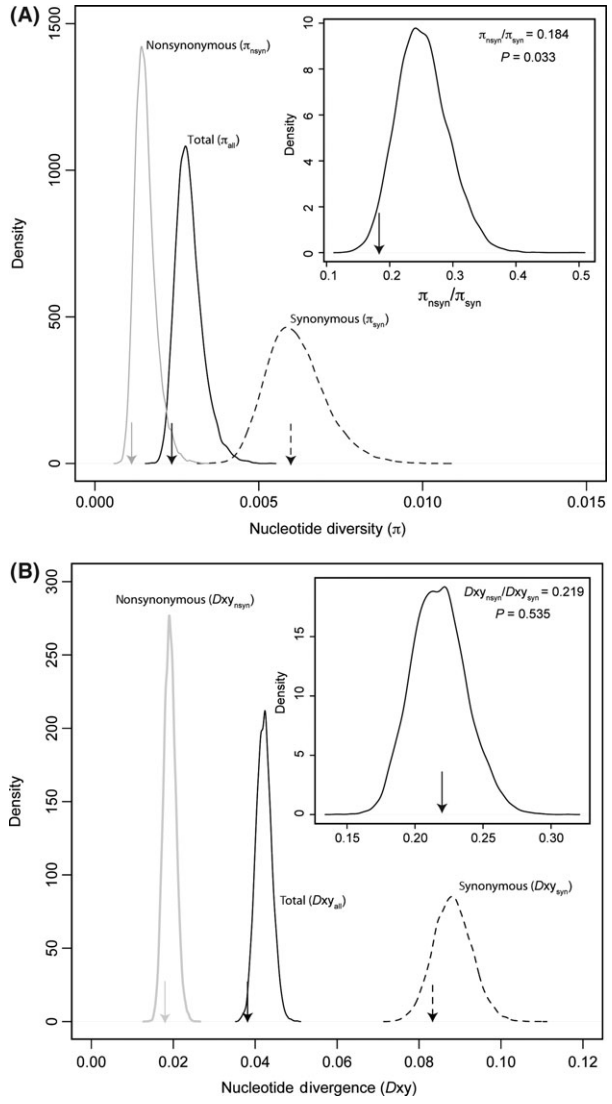

Fig. 3 Genes chosen for analysis are biased with respect to diversity and divergence due to an ascertainment of PCR primers developed in the outgroup, which were applied to the ingroup. (A) Bootstrap distributions of mean nucleotide diversity for different categories of sites for sets of 167 genes subsampled with replacement from a larger set of annotated genes resequenced for loblolly pine (n = 3162). Arrows denote the observed means for the orthologous loci used in this analysis. Genes selected for analysis were those with lower nonsynonymous diversity and thus a lower ratio of nonsynonymous to synonymous diversity (inlaid graph). (B) Bootstrap distributions for mean nucleotide divergence for different categories of sites for sets of 167 genes subsampled with replacement from a larger set of annotated genes successfully resequenced for loblolly and sugar pines (n = 647). Arrows denote the observed means for the orthologous loci used in this analysis. Genes selected for analysis were those with overall lower diversity that was proportionally less for both nonsynonymous and synonymous sites, so that their ratio was unchanged relative to random subsamples of genes (inlaid graph).

structure, as inferred using STRUCTURE, was detected for all species assuming that the number of genetic clusters (K) was two (see Table S8, Supporting information). This value of K was chosen so as to match the simulations reported elsewhere (Eyre-Walker *et al.* 2006; Keightley & Eyre-Walker 2007; Gossmann *et al.* 2010). For each species, the sampling intensity across the two underlying populations varied from equitable (e.g. *P. balfouriana*) to skewed (e.g. *P. ayacahuite*). The degree of skew was correlated with the differences in estimates of α (Fig. S6A, Supporting information), with those species characterized by more equitable sampling schemes having the largest differences for estimates of α between the methods (Fig. S6B; Table S8, Supporting information).

To test this effect, we divided the data for *P. balfouriana* into two populations based on the results from STRUCTURE (n = 8 for each population) for which we calculated the DFE, α and ω_a using each method as described previously. The differences between methods in estimates of α disappeared, with both methods for each population giving slightly negative estimates of α (method 1: −0.029 and −0.019, method 2: −0.074 and −0.058) and similar estimates of ω_a (method 1: 0.0278 and 0.0367, method 2: 0.0145 and 0.0231). Simulations confirmed this effect on the estimation of α, with population structure causing an increase in the difference between methods relative to a single population (Fig. S7, Supporting information). This effect was also observed for ω_a, as the inferred value using method 1 was 0.1835, while it was −0.0183 with method 2 (Table S6, Supporting information). This effect was not related to decreased sample size, because subsampling the full data for *P. balfouriana* (n = 100 replicates), so as to have eight sampled trees with four trees per population, retained the large differences between methods for estimates of α (mean for method 1: 0.554; mean for method 2: −0.084). A similar trend was observed for ω_a (data not shown).

Discussion

Soft pines are important components of coniferous forests distributed throughout the Northern Hemisphere. To date, however, genomic resources for these species are largely nonexistent. This is despite the ecological importance of these species to many ecosystems threatened by climate change (e.g. Larson 2011) and nonnative pathogens (e.g. blister rust). Here, we have leveraged existing genomic resources for loblolly pine in an effort to address this disparity and to better understand how these resources can be used to make inferences about the processes affecting extant patterns of genetic diversity for these species.

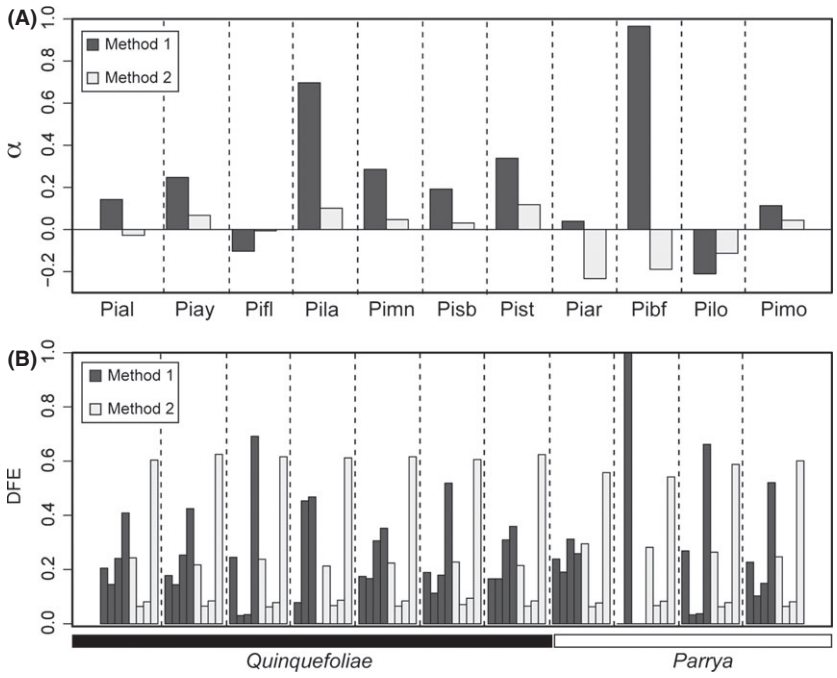

Fig. 4 Estimation of the distribution of deleterious fitness effects (DFE) and the fraction of adaptive substitutions (α) for 11 species of soft pines (subgenus *Strobus*) is affected by choice of method. (A) Point estimates for α vary between the two methods of Eyre-Walker & Keightley (2009). (B) The patterns in panel A were related to differences for the point estimates of the DFE between the two methods of Eyre-Walker & Keightley (2009). As described in the main text, the DFE was broken into four bins for each method. Pial, *Pinus albicaulis*; Piar, *P. aristata*; Piay, *P. ayacahuite*; Pibf, *P. balfouriana*; Pifl, *P. flexilis*; Pimn, *P. monticola*; Pimo, *P. monophylla*; Pisb, *P. strobiformis*; Pist, *P. strobus*.

Polymorphism and divergence

Estimates of nucleotide diversity and divergence form the base of evolutionary genetic inferences, as well as provide guidance for applications in conservation and management. As compared to flowering plants, relatively little is known about genome-wide patterns of polymorphism and divergence across the large and complex genomes of conifer species (reviewed by Savolainen & Pyhäjärvi 2007; Neale & Kremer 2011). Most knowledge to date is based on the sets of candidate genes resequenced for economically important conifers. In a similar vein, we have resequenced 167 orthologous loci across 11 soft pine species, but without prior selection of targets based on functional information from model species. The overall patterns were similar to those described previously for conifers, with average synonymous site diversity on the order of 0.007 differences/site. Diversity was also strongly correlated with range size, establishing that standing levels of diversity were correlated with an indicator of N_e, as were further summaries of diversity such as the ratio of diversities at nonsynonymous and synonymous sites.

The loci selected for resequencing were biased with respect to patterns of diversity and divergence. These biases were consistent with the targeted loci being more conserved on average than the average gene across all available genomic resources for loblolly pine. This level of conservation is probably a function of purifying selection as well as variation in mutation rates, although teasing apart the relative contributions for each of these to the observed pattern would be difficult. The bias towards loci with lower polymorphism and divergence originated from an ascertainment step for the PCR primers (Kern 2009; Cai & Petrov 2010); therefore, much of the inference about patterns of diversity and divergence for these 11 species was estimated from a nonrandom set of loci. The prevalence of similar biases in estimates of nucleotide diversity for other conifers, however, is unclear, as the majority of studies focus on a single species, yet also use outgroups that are ascertained using similar methods to those detailed here. Future work, especially that using next generation technologies not reliant on target-specific PCR, would eliminate much of this bias, especially as de novo assemblers' not requiring use of a known genome (e.g. that from a closely related model species) improve.

The fraction of adaptive substitutions

Forest trees exhibit striking adaptations to the environments in which they live (Morgenstern 1996). Some of the most well-studied examples come from the conifers that dominate temperate and boreal forest ecosystems. Decades of common garden experimentation have established the genetic basis of adaptive phenotypes for these species. Efforts to identify the genetic variants underlying the observed heritability have only emerged recently and in tandem with developments and applications of high-throughput technologies to natural tree populations (reviewed by Ingvarsson & Street 2011; Neale & Kremer 2011), yet results from scans for

positive selection using patterns of diversity and divergence across sets of gene sequences and genotype–phenotype associations are mixed. For example, selective sweeps resulting from historical positive selection are rarely documented for conifers, although ample evidence of genotype–phenotype correlations has been identified across a multitude of fitness-related phenotypes (Neale & Kremer 2011), as have instances of local adaptation (Savolainen *et al.* 2007). With respect to the latter, range-wide samplings of taxa known to have extensive evidence supporting locally adapted populations, such as many conifers (Morgenstern 1996), may bias results.

The recent focus on hitchhiking resulting from recent positive selection has overwhelmed investigation into other forms of selection for forest trees. As aptly detailed by Palmé *et al.* (2009), and expected from basic population genetic theory (e.g. Ohta 1992), purifying selection is ubiquitous across plant genomes (reviewed by Wright & Andolfatto 2008). Here, we expand our knowledge of this ubiquitous form of selection by estimating the DFE for newly arising deleterious mutations for 11 species of soft pines and have shown patterns that are strongly consistent with those expected for the Nearly Neutral Theory of Molecular Evolution (Ohta 1973; Gillespie 1999). Most newly arising nonsynonymous mutations were inferred to be so strongly deleterious that they would very rarely become fixed (Fig. 2A). Conditional on the estimates for the DFE, α was estimated to be low and not significantly different from zero for all 11 lineages (Fig. 2B). As noted previously (e.g. Strasburg *et al.* 2011), the DFE and α were significantly correlated with diversity at synonymous sites, but the correlation for α was largely driven by negative estimates that are difficult to interpret biologically and removal of those estimates changed the correlation dramatically. This was also true for ω_a. Thus, our results are most consistent with those of Gossmann *et al.* (2010).

As highlighted earlier and confirmed for these 11 soft pine species, identification of polymorphism and divergence patterns consistent with the action of historical directional selection has largely been absent for conifers. However, estimates of α for other plant species with large N_e, such as *Capsella grandiflora* (Slotte *et al.* 2010), multiple sunflower and lettuce species (Strasburg *et al.* 2011) and *Populus tremula* (Ingvarsson 2010), are all positive and significantly greater than zero. Conifers are thought to have values for N_e that rivals those in these species (Savolainen & Pyhäjärvi 2007), yet the same signals have yet to be documented. One tempting explanation for this disconnect is that estimates of N_e for conifers are upwardly biased, as they are dependent upon estimates of absolute mutation rates and generation times both of which have enormous uncertainties. Some of this disconnect is also probably related to the type of selection for which we are searching for within conifer genomes (e.g. hard vs. soft sweeps, polygenic adaptation, local adaptation, see Pritchard *et al.* 2010). Our results, however, are also consistent with biases in gene choices (see Wright & Gaut 2005) that are largely unavoidable until fully sequenced genomes of several conifer species are available.

One possibility for the lack of adaptive evolution detected here, therefore, is that loci not meeting the experimental standards set forth in PCR and sequencing protocols across highly diverged taxa are those that are enriched for signals of adaptive evolution. In effect, we have estimated the DFE and α for a highly conserved set of loci that do not cover the enormous gene space of conifers nor represent a random sample from this space. In support of this argument are the relatively low levels of nucleotide diversity observed for some species (cf. Savolainen & Pyhäjärvi 2007). Nevertheless, these are the loci, at least in the near future, which will be used for downstream analysis attempting to dissect complex phenotypes into their genetic components, and knowledge of the DFE directly informs empirical expectations of the types of loci underlying complex traits. For example, the shape of the DFE influences the expected minor allele frequency of polymorphisms underlying complex traits (Eyre-Walker 2010), which may help explain the enrichment of low-frequency variants underlying metabolomic profiles for loblolly pine (Ohta 2002; Eckert *et al.* 2012).

The remaining question is then how reliable are the estimates of these parameters for the resequenced loci? Choice of method with which to estimate the DFE and α influenced the results greatly, with the demographic assumptions of each method being the main driver of the observed differences (but see Messer & Petrov 2013). This influence was related to the appropriateness of the underlying demographic model. For example, method 1 from Eyre-Walker & Keightley (2009) assumes an explicit model of population size change. Application of this method to samples gathered across structured populations can produce wildly different results relative to a method that does not assume an explicit model. This was observed for *Pinus balfouriana*, where estimates of α varied from approximately -0.20 to almost 1. In this case, samples of *P. balfouriana* came from two distinct populations that are quite diverged (Eckert *et al.* 2008). Further analysis illustrated that the difference in estimates between methods was related to sampling scheme across the unknown populations, with equitable sampling schemes correlated with the largest differences. This is intuitive as larger samples within each population are more likely to capture rare variants that mimic the signal of segregating slightly deleterious

mutations and thus affect inference of the DFE. More work, however, is needed to fully understand how population structure affects inference of the DFE and α. Thus, as with other tests of neutrality reliant on the site-frequency spectrum (Städler *et al.* 2009), estimation of the DFE and α appears to also be sensitive to the sampling scheme across populations. Extensive simulations, however, would be needed to quantify this effect (but see Fig. S7, Supporting information), as well as to rigorously test the performance of each method. For the limited simulations performed here, both methods overestimated α in the presence of population structure, but method 2 was less biased (e.g. with two populations, and even sampling the mean estimate for α using method 2 was ~0.05, while for method 1, it was 0.27).

Future work and conclusions

We have assumed for each species that a single DFE underlies all genes in the genome, which is unlikely (Cai & Petrov 2010), and that the ecological context driving selective pressures has been constant over the time frame relevant to the observed alleles. This time frame in conifers, however, can be extensive (e.g. tens of millions of years, see Syring *et al.* 2007). For example, models incorporating mixtures of distributions comprising various selection parameters have been developed (e.g. Welch 2006), but their application to nonmodel conifer species awaits higher dimensional data that should become available once the genomes of several conifers have been sequenced, or until full, or at least longer, gene sequences are obtained for many nuclear loci. It is also unlikely that different populations within the same species have a single underlying DFE. Work with yeast has established significant variation in the efficacy of purifying selection across populations as a function of local N_e (Elyashiv *et al.* 2010). Such a scenario is also likely for the species here and should be considered when studying the genetic architecture of complex traits across naturally structured populations. Lastly, phylogeny-based inferences would probably provide a more in-depth view of long-term patterns of selection across this clade (e.g. Palmé *et al.* 2009), but rampant incomplete lineage sorting for these species complicates use of a single phylogram for inference (Syring *et al.* 2007). For example, members of subsection *Balfourianae* share (i.e. segregate the same SNP alleles) ~8.2% (*P. balfouriana*—*P. aristata*) to 11.6% (*P. longaeva*—*P. aristata*) of polymorphic sites across species.

This represents a first step towards providing orthologous nuclear gene markers across an understudied clade of pine species, while jointly inferring the evolutionary forces responsible for the observed variation. We have shown that patterns of nucleotide diversity and divergence across these markers are consistent with expectations of the Nearly Neutral Theory. No evidence was apparent for adaptive evolution during the split between the hard (e.g. *Pinus taeda*) and soft pines for any of the 11 species. It is important to note that we cannot de-emphasize the role of adaptive evolution in the lineage leading to these species, as we have shown that the loci used for analysis were biased with respect to patterns of diversity and divergence. In addition, we have employed a suite of methods that look for adaptation at a certain evolutionary timescale, so that different methods, especially those that also model demography, may reveal a view with a prominent role for recent adaptive evolution. As Gillespie (1999) pointed out, and as highlighted recently by numerous authors (e.g. Hahn 2008), there is a diversity of DNA sequence patterns that deviate from purely neutral models, so that our view of adaptation and the processes underlying this phenomenon would benefit greatly from a more balanced consideration.

Acknowledgements

Samples were kindly provided by Paul Berrang, Gaetan Doust, Joan Dunlap, Tom Ledig, Aaron Liston, Sierra Curtis-McLean, Donna Palamarek, Ken Roeder, Anna Schoettle, Dale Simpson and Det Vogler. Erik Grimstad and Ismael Grachico contributed many hours of work on determining and confirming gene structure. This work was supported by the National Science Foundation (IOS-PGRP 0638502 to D. B. Neale and C. H. Langley).

References

Axelsson E, Ellegren H (2009) Quantification of adaptive evolution of genes expressed in avian brain and the population size effect on the efficacy of selection. *Molecular Biology and Evolution*, **26**, 1073–1079.

Bierne N, Eyre-Walker A (2004) The genomic rate of adaptive amino acid substitution in *Drosophila*. *Molecular Biology and Evolution*, **21**, 1350–1360.

Boyko AR, Williamson SH, Indap AR *et al.* (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics*, **4**, e1000083.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL (2002) The cost of inbreeding in *Arabidopsis*. *Nature*, **416**, 531–534.

Cai JJ, Petrov DA (2010) Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biology and Evolution*, **12**, 393–409.

Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**, 195–205.

Comeron J (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of Molecular Evolution*, **41**, 1152–1159.

Critchfield WB, Little EL Jr (1966) *Geographic Distribution of the Pines of the World*, pp. 1–97. U.S. Department of Agriculture Miscellaneous Publication 991, Washington DC, USA.

Eckert AJ, Tearse BR, Hall BD (2008) A phylogeographical analysis of the range disjunction for foxtail pine (*Pinus balfouriana*, Pinaceae): the role of Pleistocene glaciation. *Molecular Ecology*, **17**, 1983–1997.

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

Eckert AJ, Liechty JD, Tearse BR, Pande B, Neale DB (2010b) DnaSAM: software to perform neutrality testing for large datasets with complex null models. *Molecular Ecology Resources*, **10**, 542–545.

Eckert AJ, Wegrzyn JL, Cumbie WP *et al.* (2012) Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. *New Phytologist*, **193**, 890–902.

Edgar RC (2004a) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Edgar RC (2004b) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Elyashiv E, Ballaughey K, Sattath S, Rinott Y, Przeworski M, Sella G (2010) Shifts in the intensity of purifying selection: an analysis of genome-wide polymorphism data from two closely related yeast species. *Genome Research*, **20**, 1558–1573.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186–194.

Ewing B, Hiller L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175–185.

Eyre-Walker A (2010) Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy Sciences USA*, **107**, 1752–1756.

Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution*, **26**, 2097–2108.

Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, **173**, 891–900.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fay JC, Wu C-I (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.

Foxe JP, Dar VU, Zheng H, Nordborg M, Gaut BS, Wright SI (2008) Selection on amino acid substitutions in *Arabidopsis*. *Molecular Biology and Evolution*, **25**, 1375–1383.

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK (2007) Recombination: an underappreciated factor in the evolution of plant genomes. *Nature Reviews Genetics*, **8**, 77–84.

Gernandt DS, Lopez GG, Garcia SO, Liston A (2005) Phylogeny and classification of *Pinus*. *Taxon*, **54**, 29–42.

Gillespie JH (1999) The role of population size in molecular evolution. *Theoretical Population Biology*, **55**, 145–156.

Gillespie JH (2000) Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics*, **155**, 909–919.

Gillespie JH (2001) Is the population size of a species relevant to its evolution? *Evolution*, **55**, 2161–2169.

Gossmann TI, Song B-H, Windsor AJ *et al.* (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, **27**, 1822–1832.

Gossmann TI, Keightley PD, Eyre-Walker A (2012) The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biology and Evolution*, **4**, 658–667.

Haddrill PR, Loewe L, Charlesworth B (2010) Estimating the parameters of selection on nonsynonymous mutations in *Drosophila pseudoobscura* and *D. miranda*. *Genetics*, **185**, 1381–1396.

Hahn MW (2008) Toward a selection theory of molecular evolution. *Evolution*, **62**, 255–265.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD (2010) Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genetics*, **6**, e1000825.

Hamblin MT, Casa AM, Sun H *et al.* (2006) Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. *Genetics*, **173**, 953–964.

Hamrick JL, Godt MJW (1996) Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **351**, 1291–1298.

Hernandez R (2008) A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**, 2786–2787.

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, **38**, 226–231.

Ingvarsson P (2010) Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. *Molecular Biology and Evolution*, **27**, 650–660.

Ingvarsson P, Street NR (2011) Association genetics of complex traits in plants. *New Phytologist*, **189**, 909–922.

Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, **177**, 2251–2261.

Kelly J (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.

Kern AD (2009) Correcting the site-frequency spectrum for divergence-based ascertainment. *PLoS ONE*, **4**, e5152.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, New York City, New York.

Larson ER (2011) Influences of the biophysical environment on blister rust and mountain pine beetle, and their interactions, in whitebark pine forests. *Journal of Biogeography*, **38**, 453–470.

Ledig FT (1998) Genetic variation in *Pinus*. In: *Ecology and Biogeography of Pinus* (ed. Richardson DM), pp. 251–280. Cambridge University Press, New York City, New York.

Marth GT, Korf I, Yandell MD *et al.* (1999) A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, **23**, 452–456.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, **351**, 652–654.

Messer PW, Petrov DA (2013) Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences USA*, **110**, 8615–8620.

Mirov NT (1967) *The Genus Pinus*. The Ronald Press Company, New York City, New York.
Morgenstern EK (1996) *Geographic Variation in Forest Trees*. UBC Press, Vancouver, British of Columbia, Canada.
Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Reviews Genetics*, **12**, 111–122.
Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York City, New York.
Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research*, **25**, 2745–2751.
Nielsen R, Bustamante C, Clark AG *et al.* (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, **3**, e170.
Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature*, **246**, 96–98.
Ohta T (1992) The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, **23**, 263–286.
Ohta T (2002) Near-neutrality in evolution of genes and gene regulation. *Proceedings of the National Academy Sciences USA*, **99**, 16134–16137.
Ohta T, Gillespie JH (1996) Development of the neutral and nearly neutral theories. *Theoretical Population Biology*, **49**, 128–142.
Palmé AE, Pyhäjärvi T, Wachowiak W, Savolainen O (2009) Selection on nuclear genes in a *Pinus* phylogeny. *Molecular Biology and Evolution*, **26**, 898–905.
Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology*, **7**, 84.
Piganeau G, Eyre-Walker A (2009) Evidence for variation in the effective population size of animal mitochondrial DNA. *PLoS ONE*, **4**, e4396.
Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, R208–R215.
R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Available from http://www.R-project.org.
Ross-Ibarra J, Wright SI, Foxe JP *et al.* (2009) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE*, **3**, e2411.
Savolainen O, Pyhäjärvi T (2007) Genomic diversity in forest trees. *Current Opinions in Plant Biology*, **10**, 162–167.
Savolainen O, Pyhäjärvi T, Knürr T (2007) Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 595–619.
Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the *Drosophila* genome? *PLoS Genetics*, **5**, e1000495.
Slotte T, Foxe JP, Hazzouri KM, Wright SI (2010) Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Molecular Biology and Evolution*, **27**, 1813–1821.
Städler T, Haubold B, Merino C, Stephen W, Pfaffelhuber P (2009) The impact of sampling schemes on the site-frequency spectrum in nonequilibrium subdivided populations. *Genetics*, **182**, 205–216.
Strasburg JL, Scotti-Saintagne C, Scotti I, Lai Z, Rieseberg LH (2009) Genomic patterns of adaptive divergence between chromosomally differentiated sunflower species. *Molecular Biology and Evolution*, **26**, 1341–1355.
Strasburg JL, Kane NC, Raduski Bonin A, Michelmore R, Rieseberg LH (2011) Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Molecular Biology and Evolution*, **28**, 1569–1580.
Syring JV, Farrell K, Businsky R, Cronn R, Liston A (2007) Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. *Systematic Biology*, **56**, 163–181.
Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.
Tajima F (1989) Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
Thornton K (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, **19**, 2325–2327.
Watterson GA (1975) On the number of segregation sites. *Theoretical Population Biology*, **7**, 256–276.
Wegrzyn JL, Lee JM, Liechty J, Neale DB (2009) PineSAP – sequence alignment and SNP identification pipeline. *Bioinformatics*, **25**, 2609–2610.
Welch JJ (2006) Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics*, **173**, 821–837.
Wright SI, Andolfatto P (2008) The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 193–213.
Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution*, **22**, 506–519.

A.J.E., A.D.B., K.D.J., and D.B.N. conceived and designed the project. A.J.E., B.J.K., J.L.W., and J.V.S. performed the data analyses. All authors contributed to the writing of the manuscript. D.B.N. obtained funding to carry out the project.

Data accessibility

GenBank accession numbers are available in Appendix S2 (Supporting information). Sequence alignment files are available at Dryad: doi:10.5061/dryad.80ph2. Metadata about sequence alignments are available in Appendix S1 (Supporting information).

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Supplemental text, Tables, and Figures.

Table S1 A summary of diversity panels assembled for polymorphism discovery in 11 species of pines from subgenus *Strobus*.

Table S2 PCR primers for the 167 genes sequenced for 11 species of pines within subgenus *Strobus*.

Table S3 Annotation information for the 167 loci sequenced for 11 species of pine within subgenus *Strobus*.

Table S4 Site annotations for 167 genes sequenced for 11 species of pine within subgenus *Strobus*.

Table S5 Summary statistics for 167 genes sequenced for 11 species of pines within subgenus *Strobus*.

Table S6 Estimates of ω_a from each method proposed by Eyre-Walker and Keightley (2009).

Table S7 Maximum likelihood parameter estimates for the single step population size change model where $N1$ is set to 100 (Eyre-Walker and Keightley 2009).

Table S8 Effects of population structure ($K = 2$) on the estimation of α and ω_a.

Fig. S1 Unfolded site-frequency spectra for all, synonymous (syn), and nonsynonymous (nsyn) sites across 11 species of soft pines (subgenus *Strobus*) and the outgroup.

Fig. S2 Site-frequency spectra for the first six sample coverage classes ($n \in \{13,18\}$) for *P. strobiformis* reveals that the U-shaped spectra are not an artifact of re-scaling, as the U shape is apparent within coverage classes where the sample size is constant.

Fig. S3 The decay of intragenic linkage disequilibrium, as measured using squared allelic correlation coefficients (r^2), with physical distance (bp) for 167 genes sequenced for 11 species of pines in subgenus *Strobus*.

Fig. S4 Summary of pairwise Kolmogorov–Smirnov tests for convergence among three independent runs of the Markov chain Monte Carlo (MCMC) algorithm used to estimate the distribution of deleterious fitness effects (DFE) and the proportion of adaptive substitutions (α) for 11 species of pines in subgenus *Strobus*.

Fig. S5 The distribution of deleterious fitness effects (DFE) and the proportion of adaptive substitutions (α) for 11 species of soft pines (subgenus *Strobus*) as inferred using method 1 of Eyre-Walker and Keightley (2009).

Fig. S6 Population structure and sampling frequency across populations affects the estimation of the proportion of adaptive substitutions (α) for 11 species of soft pines (subgenus *Strobus*).

Fig. S7 The effect of population structure on estimation of the proportion of adaptive substitutions (α) as assessed using simulations.

Appendix S2 GenBank accession numbers for all data, with accession numbers organized in tabular format by locus and individual.