The Evolutionary Genetics of the Genes Underlying Phenotypic Associations for Loblolly Pine (*Pinus taeda*, Pinaceae)

Andrew J. Eckert,*,[1] Jill L. Wegrzyn,[†,1] John D. Liechty,[†] Jennifer M. Lee,[‡] W. Patrick Cumbie,[§] John M. Davis, Barry Goldfarb,[††] Carol A. Loopstra,[‡‡] Sreenath R. Palle,[‡‡] Tania Quesada,** Charles H. Langley,[§§] and David B. Neale[†,2]**

*Department of Biology, Virginia Commonwealth University, Richmond, Virginia 23284, [†]Department of Plant Sciences and [§§]Department of Evolution and Ecology, University of California, Davis, California 95616, [‡]Computercraft, McLean, Virginia 22101, [§]ArborGen, Ridgeville, South Carolina 29472, **School of Forest Resources and Conservation, University of Florida, Gainesville, Florida 32611, [††]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, North Carolina 27695, and [‡‡]Department of Ecosystem Science and Management, Texas A&M University, College Station, Texas 77843

ABSTRACT A primary goal of evolutionary genetics is to discover and explain the genetic basis of fitness-related traits and how this genetic basis evolves within natural populations. Unprecedented technological advances have fueled the discovery of genetic variants associated with ecologically relevant phenotypes in many different life forms, as well as the ability to scan genomes for deviations from selectively neutral models of evolution. Theoretically, the degree of overlap between lists of genomic regions identified using each approach is related to the genetic architecture of fitness-related traits and the strength and type of natural selection molding variation at these traits within natural populations. Here we address for the first time in a plant the degree of overlap between these lists, using patterns of nucleotide diversity and divergence for >7000 unique amplicons described from the extensive expressed sequence tag libraries generated for loblolly pine (*Pinus taeda* L.) in combination with the >1000 published genetic associations. We show that loci associated with phenotypic traits are distinct with regard to neutral expectations. Phenotypes measured at the whole plant level (*e.g.*, disease resistance) exhibit an approximately twofold increase in the proportion of adaptive nonsynonymous substitutions over the genome-wide average. As expected for polygenic traits, these signals were apparent only when loci were considered at the level of functional sets. The ramifications of this result are discussed in light of the continued efforts to dissect the genetic basis of quantitative traits.

A primary goal of population and quantitative genetics is to understand the genetic architecture of ecologically relevant traits (Stinchcombe and Hoekstra 2008; Barrett and Hoekstra 2011; Neale and Kremer 2011). A primary step on the path to this goal is to link genetic with phenotypic variation, either through linkage mapping of quantitative trait loci using pedigrees or through linkage disequilibrium mapping in natural populations (Lander and Schork 1994), with the latter currently being the most utilized. A multitude of studies ranging across a diverse set of taxa have discovered myriad genotype–phenotype correlations (Hindorff *et al.* 2009; Ingvarsson and Street 2011; Neale and Kremer 2011). Each discovered variant, however, often explains only a small fraction of the heritable phenotypic variance, thus being consistent with a polygenic model for the genetic architecture of complex traits (Lynch and Walsh 1998). Concomitant with the discovery of these associations are population genomic scans documenting deviations from expectations derived using the neutral theory (Nielsen 2005). Such scans have also discovered large numbers of loci putatively underlying phenotypic traits in many different taxa (*e.g.*, Pollinger *et al.* 2005; Pritchard *et al.* 2010; Hufford *et al.* 2012), but in this case the link between genotype and phenotype is not explicit. A natural question thus arises about how much overlap exists between the lists generated using each approach.

doi: 10.1534/genetics.113.157198
Manuscript received September 5, 2013; accepted for publication September 26, 2013; published Early Online October 11, 2013.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.157198/-/DC1.
DNA sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession numbers listed in Supporting Information File S2.
[1]These authors contributed equally to this work.
[2]Corresponding author: Department of Plant Sciences, University of California, One Shields Ave., Davis, CA 95616. E-mail: dbneale@ucdavis.edu

Nonneutral signals of evolution at loci underlying polygenic, quantitative traits are expected to differ from those of selective sweeps due to directional selection (Maynard Smith and Haigh 1974; Bürger and Gimelfarb 1999; Pritchard *et al.* 2010; Pavlidis *et al.* 2012), with classic selective sweeps at these loci expected to be very rare (Chevin and Hospital 2008). Loci with small phenotypic effects are thus expected to rarely have advantageous alleles reach fixation, but rather are expected to have a polymorphic equilibrium reached (Pavlidis *et al.* 2012). As shown repeatedly, these loci are largely indistinguishable with regard to their patterns of nucleotide diversity and linkage disequilibrium from neutral loci (*e.g.*, Chevin and Hospital 2008, Pritchard *et al.* 2010), yet depending on a number of parameters (*e.g.*, the function relating phenotypic values to fitness) recently reached polymorphic equilibria may resemble incomplete sweeps (see Pavlidis *et al.* 2012). The ability to detect this, however, is expected to be rare, as the timing of sampling must coincide with the attainment of the equilibrium. A simple expectation is thus that the overlap between lists of genes correlated with phenotypic traits and lists of genes identified as outliers in genomic scans should be minimal.

In contrast to this expectation, genetic associations often involve SNPs located in distinctive portions of genomes (*e.g.*, Orozco *et al.* 2009; Chan *et al.* 2010a,b), with large effect genes often clearly displaying signals of selection (*e.g.*, Pollinger *et al.* 2005). For example, Casto and Feldman (2011) examined 1300 single-nucleotide polymorphisms (SNPs) identified as associated with a variety of human phenotypes and showed that those associated with certain phenotypes were located in genomic regions that would have been outliers with regard to measures such as F_{ST}. In an examination of the genetic associations with 292 primary metabolites for loblolly pine (*Pinus taeda* L.), however, a similar pattern was not found, as loci with elevated values of F_{ST} were not enriched in genetic associations (Eckert *et al.* 2012). In this case, genetic associations were enriched for rare alleles, both at the level of SNPs and at the level of the genes containing these SNPs, which could be indicative of synthetic associations (see Goldstein 2011) or some form of linked selection. A generalization of the expected pattern that should emerge from these types of studies is hampered by the uncertainty about the predominant type of natural selection that affects the phenotypic traits commonly investigated in combination with the underlying genetic architecture of those traits (Barton and Keightley 2002; Mitchell-Olds *et al.* 2008).

Both of the aforementioned uncertainties are intertwined, as, for example, the type of natural selection, as well as the patterning of the intensity of selection pressures across the range of a species, acting on a phenotypic trait will affect its underlying genetic architecture (Orr 1998; Barton and Keightley 2002; Mitchell-Olds *et al.* 2008; Eyre-Walker 2010). Knowledge of the type of selection acting on a phenotypic trait, however, requires information about the factors affecting fitness within natural populations (*i.e.*, the ecological context, see Endler 1986). Species of forest trees have a long history of ecological genetic investigation that is coupled with an even more developed history of quantitative genetic experimentation (White *et al.* 2007), so that these species of plants are optimal targets for examining the genetic architecture of ecologically relevant traits (Neale and Kremer 2011). This extensive research has identified the importance of locally adapted phenotypes to standing levels of genetic diversity across geographical ranges of many different tree species. As such, it is expected that conditional on appropriate sampling, many of the signals of selection and links between genotypes and phenotypes should reflect local adaptation.

Linkage disequilibrium mapping has been successively carried out for numerous forest tree species with a variety of phenotypes (reviewed by Neale and Ingvarsson 2008; González-Martínez *et al.* 2011; Ingvarsson and Street 2011). This success is due largely to the life history characteristics of many tree species (Neale and Savolainen 2004). For example, >1000 genetic associations for loblolly pine, each of small effect on the phenotype, have been discovered for a range of phenotypic traits that span the spectrum of trait complexity: expression profiles of genes related to lignin biosynthesis (Palle *et al.* 2011, 2013), primary metabolite concentrations (Eckert *et al.* 2012), drought tolerance and growth (González-Martínez *et al.* 2008; Cumbie *et al.* 2011), disease resistance (Quesada *et al.* 2010), and wood quality (González-Martínez *et al.* 2006a). In addition, much is known about the ecological context of the natural populations from which the trees used for the aforementioned studies were sampled, so that a set of complementary environmental associations has also been described (Eckert *et al.* 2010a,b). The set of associations in loblolly pine, therefore, is comprehensive with regard to phenotypic trait diversity. This includes both explicitly and implicitly measured phenotypic traits, as environmental associations are expected to reflect unmeasured phenotypic traits since phenotypic traits are selected upon in natural populations (see Eckert *et al.* 2009a for an example).

Loblolly pine is an ecologically and economically important tree species of southeastern North America. Its range extends from eastern Texas throughout the southeastern United States and north into Delaware and thus is distributed across diverse environmental gradients (Figure 1). As described previously, loblolly pine is one of the most intensively studied tree species with regard to the genetic architecture of phenotypic traits. The aforementioned phenotypic traits were measured from the same set of clonally replicated materials and correlated with the same set of SNPs (see Cumbie *et al.* 2011 for materials and Eckert *et al.* 2010a for SNPs). This study system thus offers an unparalleled chance to examine the degree of overlap between signals documented using linkage disequilibrium mapping and those discovered using population genomic scans. Here we

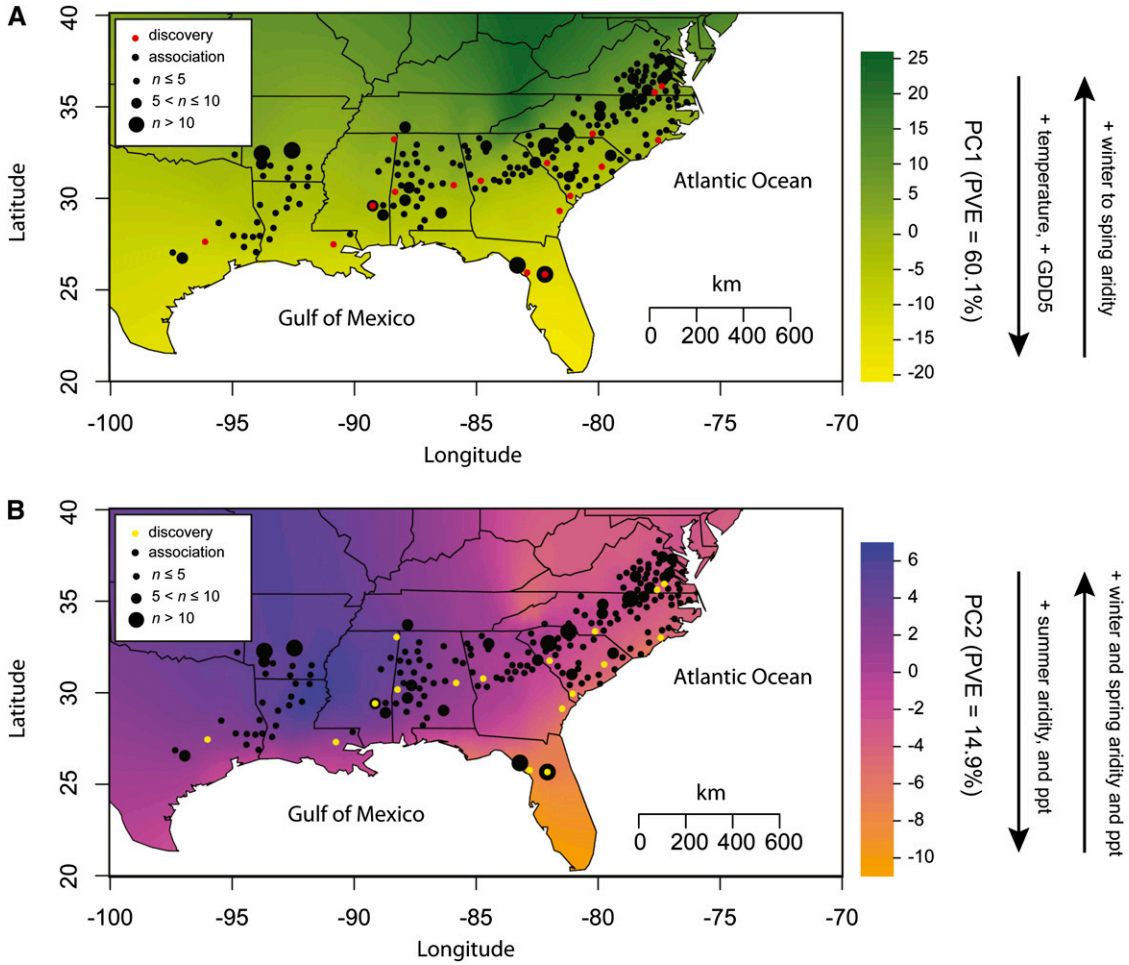

Figure 1 Sampling localities for loblolly pine for the samples used to discover single-nucleotide polymorphisms (SNPs) and to associate SNP genotypes with phenotypes. (A) Sample localities with respect to the first climatic principal component (PC), which largely represents temperature-related variables and seasonal aridity. (B) Sample localities with respect to the second climatic PC, which largely represents precipitation-related variables. PVE, percentage of variance explained.

address this question, using >7000 unique amplicons described from the extensive expressed sequence tag (EST) libraries generated for this species. Prior to addressing this question, we describe the current knowledge of the structure and diversity of the gene space of loblolly pine based upon these 7000 amplicons. We use these descriptions subsequently to examine patterns of neutrality, or the lack thereof, for individual genes, as well as sets of genes based upon their associations with phenotypes, and show that loci associated with phenotypic traits and environmental variables are indeed unique with regard to neutral expectations. The ramifications of this result are discussed in light of the continued efforts to dissect the genetic basis of quantitative traits in tree species as well as the evolutionary genetic ramifications of these studies (*sensu* Barton and Keightley 2002).

Materials and Methods

The results presented here are derived from data generated previously and thus represent a mixture of previously published results and novel population genetic analyses. In general, the following results were published previously: SNP genotypes based on a high-density SNP genotyping array (Eckert *et al.* 2010a,b, 2012; Quesada *et al.* 2010; Cumbie *et al.* 2011; Palle *et al.* 2013), the linkage map (Eckert *et al.* 2010a,b), and the association results (Quesada *et al.* 2010; Cumbie *et al.* 2011; Eckert *et al.* 2012; Palle *et al.* 2013). Results novel to this study are those related to nucleotide diversity, nucleotide divergence, linkage disequilibrium, and inferences of deviations from neutral models. We note in each subsection of *Materials and Methods* whether data or results are novel to this study.

Population sampling

SNP discovery panel: A total of 18 trees covering the natural range of loblolly pine were selected for resequencing (Figure 1). Multiple seeds were obtained from each sampled tree for which haploid megagametophyte tissue was manually excised. Megagametophyte tissue proliferates directly from one of the meiotic daughter cells (*i.e.*, the megaspore),

which also produces the female gamete, and thus the haploid genotype of the megagametophyte is that of the maternal contribution to the seed (see Pichot and El Maataoui 1997 for exceptions). Megagametophytes were grouped into 96-well plates, one per well, for DNA extraction. Total genomic DNA was extracted subsequently, using 96-well DNeasy Plant Mini Kits (QIAGEN, Valencia, CA), following the manufacturer's protocol. All tissue excision and DNA extractions were performed by the staff at the National Forest Genetics Electrophoresis Laboratory (U.S. Department of Agriculture Forest Service, Placerville, CA). A set of 7216 SNPs discovered using these samples was published previously and formed the basis for all inferences of genotype–phenotype associations (Quesada *et al.* 2010; Cumbie *et al.* 2011; Eckert *et al.* 2012; Palle *et al.* 2013).

Linkage-mapping panel: Two three-generation, outbred pedigrees have been used extensively for linkage mapping in loblolly pine (Devey *et al.* 1999). Parents for each pedigree were derived from crosses of outbred grandparents and were each crossed twice to produce two sets of full-sib progeny—the discovery set (n = 95–172 per pedigree) and the validation set (n = 500 per pedigree). We randomly selected 93 offspring from the discovery set from each pedigree to genotype with the newly designed SNP genotyping arrays. Data derived from this population were published previously (Eckert *et al.* 2010a,b).

Association mapping panel: A set of 498 largely unrelated genotypes has been used extensively for association mapping in loblolly pine (Figure 1). These genotypes were established from seeds collected from natural stands and form parts of the North Carolina State University Cooperative Tree Improvement and Western Gulf Forest Tree Improvement Programs. Germinated seedlings were grown for 1 year and then hedged for stem cuttings following standard protocols (Lebude *et al.* 2004), after which rooted cuttings were planted in replicates of two to four per genotype at three common garden localities: Texas A&M University (TAMU), North Carolina State University (NCSU), and the University of Florida (UF). Data and results from this population were published previously (Quesada *et al.* 2010; Cumbie *et al.* 2011; Eckert *et al.* 2012; Palle *et al.* 2013).

Development and application of genetic markers

Construction of EST clusters: A set of 40,000 EST sequences from loblolly pine was extracted from NCBI's dbEST repository and clustered using custom internal scripts (Beckman Coulter Genomics, Danvers, MA) into 20,500 unique EST contigs. Approximately 7900 high-quality amplicons were passed subsequently to the production pipeline from these 20,500 unique EST contigs (Supporting Information, File S1). The resulting 7900 amplicons were annotated functionally, using a comparative sequence analysis pipeline. These amplicons are synonymous with respect to the term loci (*i.e.*, we annotated 7900 loci). NCBI BLAST searches were performed against the curated, nonredundant RefSeq plant protein database (Release 43). Sequences without initial hits were compared against the NR database of GenBank (NCBI Release 192). All BLAST searches were subject to an *e*-value cutoff of 0.5 and 1*e*-05, respectively. Given putative functional annotations, coding regions were subsequently annotated using output from the SIBSim4 package (Florea *et al.* 1998). ESTs were trimmed based on matching to the PCR-generated amplicons, using matches to the amplification primer sequences. These ESTs were then aligned against the most common haplotype from the resequenced genomic alignments. Introns were subsequently annotated from the amplicon sequences based on the SIBSim4 predictions.

Resequencing and SNP discovery: A customized PCR and sequencing pipeline was established at Agencourt Biosciences (now Beckman Coulter Genomics) to resequence 7900 unique EST clusters for 18 loblolly pine trees via bidirectional Sanger sequencing (see File S1). A customized pipeline, PineSAP (Wegrzyn *et al.* 2009), which employs PHRED/PHRAP (Ewing *et al.* 1998; Ewing and Green 1998), CONSED (Gordon *et al.* 1998), POLYBAYES (Marth *et al.* 1999), POLYPHRED (Nickerson *et al.* 1997), and machine-learning tools, was used to generate sequence alignments and identify polymorphisms for these data. All reported polymorphisms were those detected after masking bases at a PHRED score of 30.

SNP genotyping: A total of 7216 SNPs from the set of those discovered as described previously were chosen for further analysis based on quality scores derived from the original sequence data and not on functional or site annotations. This ensured thorough coverage of the available EST cluster sequences for loblolly pine. Genotyping of SNPs utilizing the Infinium platform (Illumina, San Diego) was carried out at the University of California, Davis Genome Center DNA Technologies Core Facility (http://dnatech.genomecenter.ucdavis.edu/) for each of the genotypes that were established at the NCSU common garden and the 93 offspring from of each of the discovery linkage-mapping populations. These data were published previously (Eckert *et al.* 2010a,b, 2012; Quesada *et al.* 2010; Cumbie *et al.* 2011; Palle *et al.* 2013). Further information is given in File S1.

Phenotypic trait analysis

Phenotypic traits in four general categories, expression levels for lignin and cellulose-related genes (expression), primary metabolite concentrations (metabolite), drought tolerance and growth (drought), and disease resistance (disease), were evaluated in the same set of clonally replicated materials established at NCSU, TAMU, and UF, and the results from phenotypic analyses were published previously in a series of articles linking SNP genotypes to phenotypic traits, using association mapping (Quesada *et al.* 2010; Cumbie *et al.* 2011; Palle *et al.* 2011; Eckert *et al.*

2012; Palle *et al.* 2013), or to specific environments (Eckert *et al.* 2010a,b). More details about each phenotypic trait are given in File S1, as well as in the original publications.

Statistical analysis

Generation of linkage maps: We utilized a maximum-likelihood method that allows for genotyping errors (Cartwright *et al.* 2007) to create linkage maps composed of SNPs and restriction fragment length polymorphism (RFLP) framework markers (see Eckert *et al.* 2009b and references therein), following a double pseudotestcross strategy (Grattapaglia and Sederoff 1994). Additional information is provided in the original publications (Eckert *et al.* 2010a,b) and in File S1.

Identification of phenotypic associations: Methodologies used to discover genotype–phenotype and genotype–environment associations are described fully in the original publications (Eckert *et al.* 2010a,b, 2012; Quesada *et al.* 2010; Cumbie *et al.* 2011; Palle *et al.* 2013) and are summarized in File S1.

Estimation of nucleotide diversity and divergence: Standard estimates of nucleotide diversity [θ_W (Watterson 1975) and θ_π (Tajima 1983)], haplotype diversity [number of haplotypes, k, and haplotype diversity, H_d (see Nei 1987)], and nucleotide divergence [D_{xy} (Nei 1987)] were made from each alignment. Nucleotide divergence was estimated relative to *P. radiata* [~15–20 MY divergent or less (Gernandt *et al.* 2008)] and *P. lambertiana* [~72–87 MY divergent (Gernandt *et al.* 2008)]. Estimates were made for different categories of sites (*i.e.*, all, nonsynonymous, synonymous, and noncoding sites) for all statistics by calculating all statistics for each class of sites for each amplicon. When multiple mutations were segregating within a single codon and the pathways could not be unambiguously assigned as silent or replacement, we assumed a pathway requiring the fewest changes. We further assumed an infinite-sites model for all polymorphism calculations, and, thus, sites violating this model or with missing data were dropped prior to estimation. Alignment gaps consistent with insertion–deletion events (indels) were considered separately and were visually validated. Variation in coverage (*i.e.*, base counts at a PHRED score ≥30 at an aligned site across all sampled individuals and targeted base pair positions) was primarily concentrated among alignments, so that sites within alignments for a single amplicon with gaps were ignored and a weighting scheme was used to account for variation among amplicons. The resulting estimates from each alignment were combined into average estimates for groups of amplicons, such as genome-wide, categorical (*i.e.*, sets of genes), and genomic windows along the consensus linkage map, using weighted averages following the logic of Begun *et al.* (2007). The weights for each alignment were the maximal coverage class attained for the majority of the aligned sites. Genomic windows along the linkage map were defined in 5-cM bins with a step size of 2 cM. All calculations were conducted using DnaSAM (Eckert *et al.* 2010c) 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, with the average across pairwise comparisons between ingroup and outgroup sequences being presented for each statistic when multiple sequences were available from the outgroup. For different categories of sites, genome-wide estimates were averaged across amplicons, using the aforementioned weighting scheme. All results derived from these methodologies are novel to this study.

Linkage disequilibrium: Linkage disequilibrium within genes was investigated using the Z_{nS} statistic (Kelly 1997), which is the average r^2 among all possible pairwise comparisons of SNPs within an amplicon. Nonlinear regression was used to fit the expected decay of r^2 with physical distance (base pairs) following Remington *et al.* (2001), using R (R Development Core Team 2010). Separate analyses were conducted for each coverage class. All results derived from these methodologies are novel to this study.

Site-frequency spectrum and tests of neutrality: Folded and unfolded site-frequency spectra were estimated for alignments placed in each coverage class (*i.e.*, the weights used in the weighted averages described previously). Unfolded site-frequency spectra were determined separately for each outgroup comparison. Genome-wide and categorical estimates were made using the method implemented in SoFoS (http://www.scit.us/sofos/) (see Hufford *et al.* 2012, especially the supplemental information). Standard statistics based on the site-frequency spectrum [D (Tajima 1989) and H (Fay and Wu 2000)] were also estimated for each coverage class, using DnaSAM. Estimates of H were made separately for each outgroup, although we concentrate on those using radiata pine as the outgroup due to the much larger number of amplicons.

Fit of the data to the standard neutral model (SNM) was assessed using two types of tests that are sensitive to deviations from the SNM at different temporal scales. First, outlier tests for D and H were conducted per locus, using coalescent simulations (n = 100,000) conditional on the observed value of θ_π and assuming no recombination. A liberal threshold of $P = 0.05$ was used to assess significance of deviations from the SNM. We also incorporated an aspect of historical demography for these coalescent simulations by using the three-epoch model (TEM) given by Ersöz *et al.* (2010), which was parameterized based upon a set of 42 disease-resistance candidate genes resequenced for loblolly pine. In brief, this model describes a single lineage experiencing a bottleneck in the recent past, so that the three epochs refer to prebottleneck, bottleneck, and postbottleneck (see Ersöz *et al.* 2010). Simulations were performed and outliers were assessed and summarized by coverage classes for each model separately. All results derived from these methodologies are novel to this study.

Second, polymorphism and divergence at nonsynonymous and synonymous sites were contrasted using an extension of the McDonald–Kreitman approach (Welch 2006; Obbard *et al.* 2009). Forty-eight different models, based on whether parameters were fixed or allowed to vary across sets of amplicons (Table S10), were explored in this framework and the best model was chosen using the Akaike information criterion corrected for finite sample sizes (AICc) (see Burnham and Anderson 2004). Parameters of each model were estimated from four observations that were generated for each amplicon: the number of nonsynonymous polymorphisms (P_N), the number of synonymous polymorphisms (P_S), divergence at nonsynonymous sites (D_N), and divergence at synonymous sites (D_S). These values were also used to estimate the direction of selection (DoS) statistic for each amplicon (Stoletzki and Eyre-Walker 2011). The main parameter of interest of all models was the proportion of adaptive substitutions (α). Positive values of α represent the proportion of nonsynonymous substitutions driven to fixation by positive selection, while negative values result from the segregation of slightly deleterious variation and sampling error. The effect of the segregation of slightly deleterious variation on estimates of α was assessed by eliminating low-frequency variants from the data set at multiple thresholds (5% and 10%). All analysis was conducted using the MKtest-2.0 software (available at http://sitka.gen.cam.ac.uk/research/welch/GroupPage/Software.html), and all results derived from these methodologies are novel to this study.

Patterns among categories of alignments: Genome-wide levels of nucleotide diversity were assessed for constancy across amplicons, using a multilocus likelihood-ratio test (Hudson 1990, as used by Brown *et al.* 2004) (see https://github.com/RILAB/ThetaCurve for the Perl script; Ross-Ibarra, J., personal communication). Average levels of diversity were also assessed for differences across categories of amplicons defined by their putative functions, ability to be placed on the linkage map, linkage groups, whether or not SNPs were genotyped, presence or absence of indels, or presence or absence of genotype–phenotype associations. Gene Ontology (GO) categories (Ashburner *et al.* 2000) were assigned primarily through *molecular function* and were standardized to the same level in the GO hierarchy, to allow for appropriate comparisons. Following standardization, some categories were combined or split where appropriate to represent relevant high-level gene families (*e.g.*, heat shock, transcription factors, ATPases). Categories for genotype–phenotype associations were defined in numerous ways, ranging from a binary categorization (associated to unassociated) to a multistate categorization (unassociated, metabolite associations, expression associations, disease associations, and drought associations). For completeness, the environmental associations from Eckert *et al.* (2010a,b) were also included. Permutation tests (n = 10,000 permutations) were used to assess differences among averages or to validate inferences from statistical tests (*e.g.*, Kruskal–Wallis rank sum tests). Permutation tests were carried out by randomizing amplicons between (*e.g.*, for *t*-tests) or among (*e.g.*, for Kruskal–Wallis rank sum tests) categories and calculating the relevant test statistic 10,000 times to form a null distribution. Bootstrapping of amplicons (n = 10,000 replicates with replacement) within categories was also used to construct confidence intervals for categorical averages (Efron 1985). All results derived from these methodologies are novel to this study.

Results

Resequencing data summary

The number of amplicons passing design thresholds decreased from ~7900 to 7413 after requiring both forward (F) and reverse (R) reads to be present for each sample, which was followed by a further decrease to 6669 amplicons after screening for amplification primers in both reads. A total of 5773 amplicons passed our final quality thresholds, which also included screens for organellar contamination. A total of 22,621 SNPs were detected across the 5773 amplicons. Of these 22,621 SNPs, 10,591 could be annotated as nonsynonymous (n = 2915), synonymous (n = 3233), and noncoding (n = 4443). Detailed information is located in File S1 (see also Table 1, Table S1, Table S2, Table S3, Figure S1, Figure S2, Figure S3, Figure S4, and Figure S5). All sequence data used in the downstream analyses have been deposited at GenBank (accession numbers: File S2).

Nucleotide diversity and insertion–deletion events

Genome-wide patterns: Nucleotide diversity (per site), as measured by θ_π and θ_W, varied significantly among sample coverage classes (Kruskal–Wallis rank sum tests: $P < 1.0e\text{-}9$, $P_{perm} < 1.0e\text{-}4$). Thus, weighted estimates for averages and variances were used to construct genome-wide estimates. The overall weighted average ($\pm$1 SD) for each estimator was 0.0033 ($\pm$0.0048) and 0.0038 ($\pm$0.0049), respectively. These values are similar to those published previously for loblolly pine, using sets of candidate genes (Brown *et al.* 2004; González-Martínez *et al.* 2006a; Ersöz *et al.* 2010). Analysis by categories of sites also revealed similarity to previous estimates, with synonymous site diversity being the largest and nonsynonymous site diversity being the smallest (Figure S5 and Figure S6). Average estimates of diversity were also similar to those obtained using a multilocus maximum-likelihood method (Hudson 1990). Likelihood-ratio tests, however, rejected equality of nucleotide diversity across amplicons for all, nonsynonymous, synonymous, and noncoding sites (Table S4).

A total of 1572 indels occurred in 1080 of the 5773 amplicons, with sizes ranging from 1 to 352 bp. Indels accounted for 15,286 of the 2,139,768 aligned sites (0.71%). Most of the indels were $\leq$10 bp in size (79.45%), with a weighted average size of 10 bp, where

Table 1 Summary of sequence data generated for loblolly pine (*Pinus taeda* L.)

Catagories	All	Not genotyped	Mapped	Unassociated	Associated
No. amplicons[a]					
Total	5,773	3,154	1,306	1,930	689
Annotated	2,626	1,453	562	864	309
Not annotated	3,147	1,701	744	1,066	380
Sample size (n)[a]					
Total	12 ± 6	11 ± 6	12 ± 5	13 ± 5	13 ± 5
Annotated	12 ± 6	11 ± 6	13 ± 5	13 ± 5	13 ± 5
Not annotated	12 ± 6	11 ± 6	12 ± 5	13 ± 5	12 ± 5
Length (bp)[a]					
All	2,135,607 (370 ± 126)	1,107,387 (351 ± 120)	520,300 (398 ± 129	745,025 (390 ± 130)	283,195 (400 ± 131)
NS	583,159 (235 ± 83)	308,837 (230 ± 82)	132,704 (244 ± 84)	197,612 (238 ± 83)	76,710 (249 ± 84)
SY	160,814 (65 ± 24)	84,829 (63 ± 23)	36,439 (67 ± 24)	54,677 (66 ± 24)	21,308 (69 ± 24)
NC	417,915 (132 ± 153)	197,736 (114 ± 142)	110,076 (158 ± 165)	163,867 (156 ± 162)	56,312 (149 ± 161)
No. SNPs[a]					
All	22,621 (4 ± 5)	11,064 (4 ± 6)	6,079 (5 ± 5)	8,485 (4 ± 4)	3,072 (4 ± 4)
NS	2,915 (1 ± 2)	1,568 (1 ± 2)	709 (1 ± 2)	898 (1 ± 2)	449 (1 ± 2)
SY	3,233 (1 ± 2)	1,490 (1 ± 2)	897 (2 ± 2)	1,265 (1 ± 2)	478 (2 ± 2)
NC	4,443 (1 ± 3)	1,660 (1 ± 2)	1,462 (2 ± 3)	2,159 (2 ± 4)	624 (2 ± 3)

NC, noncoding; NS, nonsynonymous; SY, synonymous.
[a] Numbers are either total counts or arithmetic averages ± one standard deviation. Decimal remainders for averages and standard deviations were rounded to whole numbers according to >0.0–0.5 round down and >0.5–1.0 round up.

the weights were the sample frequencies for each unique indel length. The majority of amplicons with indels contained only a single indel (Figure S7), yet 16 amplicons contained ≥5 indels, with indel lengths ranging in size from 1 to 316 bp. A total of 547 of the 1572 indels were located in 393 amplicons with annotations. Indels were primarily located in the noncoding regions of these amplicons (86.76%), with 190 and 305 located in introns and UTRs, respectively. The 52 remaining indels were within coding regions and, as required during validation of the annotations, did not result in frameshift mutations. Indels occurred across coverage classes (n = 2–18), in approximately the same proportion as expected given the distribution of amplicons across coverage classes (Wilcoxon signed rank test: P = 0.059, P_{perm} = 0.087).

The number of indels within amplicons also affected diversity, divergence, and the shape of the site-frequency spectrum (Figure S8, Table S5; see also *Results*, *Departures from neutrality*). Conditional on an indel being present, the number of indels was positively correlated with average number of SNPs (Spearman's ρ = 0.83), average nucleotide diversity (Spearman's ρ = 0.77), average Tajima's D (Spearman's ρ = 0.82), the number of haplotypes (Spearman's ρ = 0.63), and haplotype diversity (Spearman's ρ = 0.83). These patterns could result from mutation rate variation introducing an autocorrelation. If this is true, then the trend with nucleotide diversity should disappear if diversity is scaled by divergence. This does not happen when using radiata pine (t = −4.577, d.f. = 1077.045, P = 5.257e-06, P_{perm} = 0.0001) or sugar pine (t = −4.238, d.f. = 145.239, P = 3.994e-05, P_{perm} = 0.0009) as outgroups, although the relative difference for the averages is reduced (*i.e.*, the average diversity is 2.6-fold greater in amplicons with indels, while the average diversity scaled by divergence is only 1.2- to 1.5-fold greater, depending on the outgroup comparison; see Figure S8). No differences in these patterns were apparent across sets of amplicons based on phenotypic categories or based on any other categorization (Wilcoxon signed rank or Kruskal–Wallis rank sum tests: $P < 1.45e$-05, $P_{perm} < 0.0001$).

Comparisons across categories of amplicons: Nucleotide diversity varied across sets of amplicons (Table 2; Figure 2). The consensus linkage map for loblolly pine included a total of 1268 amplicons with more than one sampled allele. These amplicons had average levels of nucleotide diversity that were larger than those of the entire set of 5773 amplicons, with θ_π and θ_W being 1.2- and 1.1-fold larger, respectively. Nucleotide diversity varied across linkage groups for loblolly pine, ranging from 0.00323 (θ_π) and 0.00320 (θ_W) for linkage group 1 to 0.00458 (θ_π) and 0.00473 (θ_W) for linkage group 10 (Figure S9 and Figure S10; Table S6). These differences, however, were not significant for diversity at all, nonsynonymous, synonymous, or noncoding sites (Kruskal–Wallis rank sum tests: $P > 0.20$, $P_{perm} > 0.10$), yet were significantly correlated with linkage group length for synonymous sites (Spearman's ρ = 0.420, P_{perm} = 0.0052).

Table 2 Nucleotide diversity across the loblolly pine (*Pinus taeda* L.) genome

Amplicons (*L* = all/annotated)	Nucleotide diversity (θ_π)				Nucleotide diversity (θ_W)			
	All: 95% C.I.[a]	NS: 95% C.I.[a]	SY: 95% C.I.[a]	NC: 95% C.I.[a]	All: 95% C.I.[a]	NS: 95% C.I.[a]	SY: 95% C.I.[a]	NC: 95% C.I.[a]
All	0.0033	0.0013	0.0059	0.0021	0.0038	0.0015	0.0067	0.0024
(*L* = 5773/2626)	0.0032–0.0034	0.0012–0.0014	0.0056–0.0062	0.0019–0.0023	0.0037–0.0039	0.0014–0.0016	0.0064–0.0071	0.0023–0.0025
Not genotyped	0.0030	0.0012	0.0050	0.0016	0.0035	0.0015	0.0059	0.0019
(*L* = 3154/1453)	0.0027–0.0033	0.0011–0.0013	0.0045–0.0055	0.0014–0.0018	0.0033–0.0037	0.0014–0.0016	0.0054–0.0063	0.0017–0.0021
Mapped[b]	0.0040	0.0016	0.0081	0.0031	0.0042	0.0017	0.0083	0.0032
(*L* = 1306/562)	0.0038–0.0042	0.0014–0.0018	0.0074–0.0088	0.0027–0.0034	0.0040–0.0044	0.0015–0.0019	0.0076–0.0090	0.0029–0.0035
Unassociated	0.0037	0.0013	0.0070	0.0028	0.0041	0.0015	0.0078	0.0031
(*L* = 1930/864)	0.0035–0.0039	0.0012–0.0014	0.0064–0.0076	0.0026–0.0030	0.0039–0.0042	0.0013–0.0016	0.0072–0.0084	0.0029–0.0034
Associated	0.0034	0.0014	0.0065	0.0022	0.0039	0.0017	0.0073	0.0024
(*L* = 689/309)	0.0031–0.0037	0.0012–0.0016	0.0057–0.0075	0.0018–0.0025	0.0036–0.0041	0.0015–0.0019	0.0065–0.0081	0.0021–0.0028
Expression	0.0033	0.0017	0.0077	0.0016	0.0042	0.0023	0.0084	0.0022
(*L* = 76/39)	0.0028–0.0038	0.0013–0.0021	0.0049–0.0102	0.0009–0.0022	0.0037–0.0047	0.0018–0.0028	0.0057–0.0110	0.0014–0.0029
Metabolite	0.0033	0.0014	0.0067	0.0021	0.0038	0.0017	0.0075	0.0024
(*L* = 576/257)	0.0031–0.0035	0.0012–0.0016	0.0060–0.0074	0.0018–0.0024	0.0036–0.0040	0.0015–0.0019	0.0068–0.0082	0.0021–0.0027
Drought	0.0017	0.0006	0.0084	0.0007	0.0020	0.0009	0.0090	0.0012
(*L* = 12/6)	0.0012–0.0022	0.0001–0.0012	0.0017–0.0130	0.0001–0.0012	0.0014-0.0025	0.0001–0.0017	0.0032–0.0132	0.0001–0.0024
Disease	0.0037	0.0006	0.0029	0.0029	0.0044	0.0009	0.0033	0.0037
(*L* = 9/5)	0.0020–0.0051	0.0001–0.0014	0.0013–0.0046	0.0008–0.0057	0.0026-0.0060	0.0001–0.0021	0.0011–0.0057	0.0015–0.0074

C.I., confidence interval; L, number of amplicons; NC, noncoding; NS, nonsynonymous; SY, synonymous.

[a] Nonparametric bootstrapping (*n* = 1000 replicates) across amplicons was used to estimate 95% C.I.'s.

[b] Estimates per linkage group are given in Table S6.

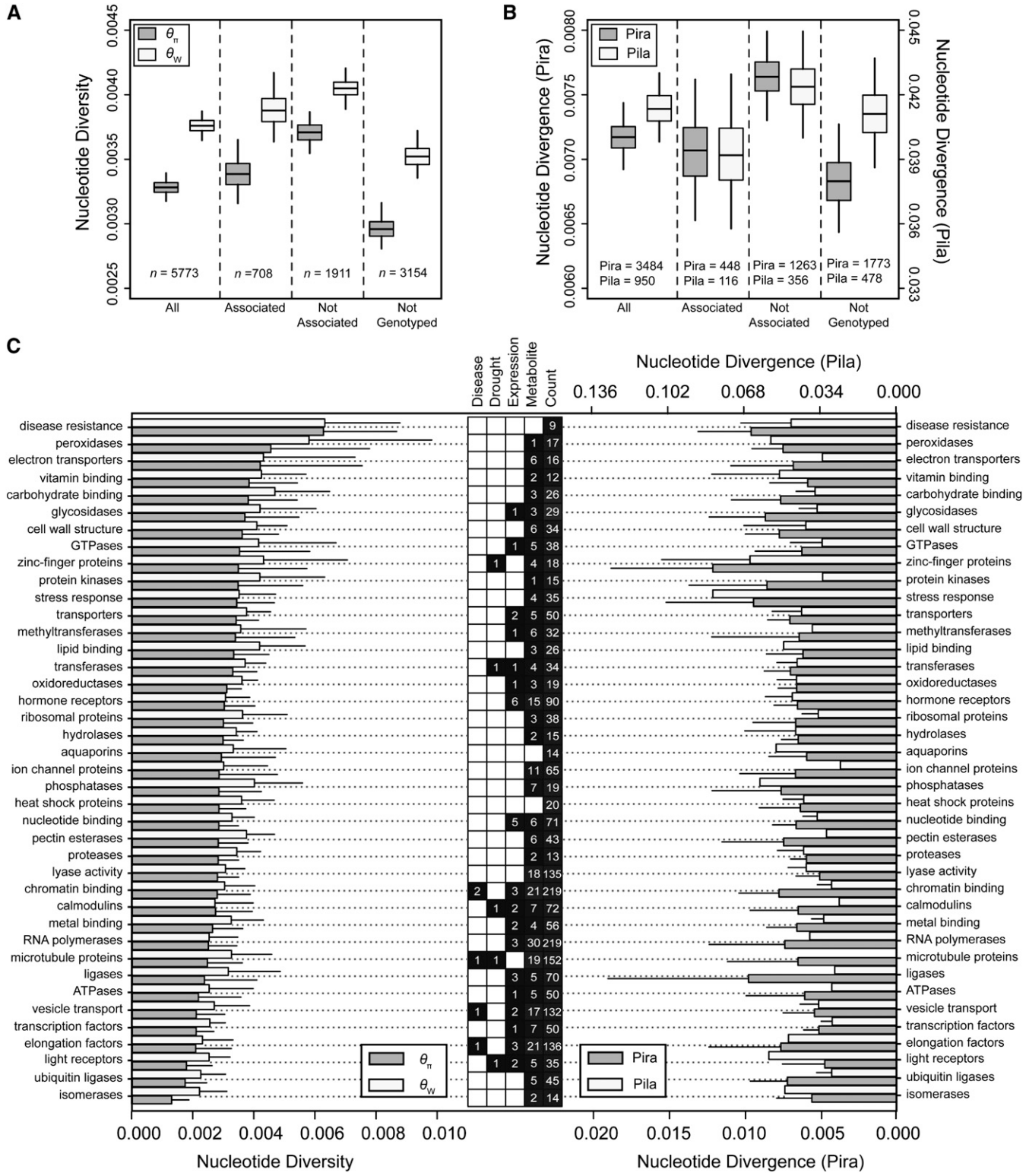

Figure 2 Nucleotide diversity and divergence across categories of amplicons for loblolly pine. (A) Average levels of total nucleotide diversity by category of amplicons. Whiskers represent 95% confidence intervals generated via bootstrapping across amplicons. (B) Average levels of total nucleotide divergence by category of amplicons and outgroup (Pira, *P. radiata*; Pila, *P. lambertiana*). Whiskers represent 95% confidence intervals generated via bootstrapping across amplicons. (C) Average levels of total nucleotide diversity (left) and divergence (right) by functional categories in relation to categories of phenotypic associations (center). Numbers in the center box represent numbers of amplicons for each functional category (Count) and the number of phenotypic associations (solid, associated; open, no association). Error bars represent 95% confidence intervals generated via bootstrapping across amplicons.

Nucleotide diversity also varied among categories of amplicons based on those that were annotated *vs.* those that were not annotated, those that were associated to at least one type of phenotypic trait, and those associated with each of the four phenotypic trait categories (Table 2). Most differences, however, were not significant (Kruskal–Wallis rank sum tests: $P > 0.30$, $P_{perm} > 0.20$). The only significant difference observed was for nonsynonymous and noncoding diversities (Kruskal–Wallis rank sum tests: $P < 0.001$, $P_{perm} < 0.005$), with amplicons containing SNPs related to expression and metabolite phenotypes having higher nonsynonymous diversities and lower noncoding diversities. Total nucleotide diversity, however, did differ significantly among functional categories of amplicons (Kruskal–Wallis rank sum tests: $P < 0.001$, $P_{perm} < 0.005$; Figure 2), with the highest levels of diversity being in amplicons whose gene products were associated with disease resistance and the lowest levels of diversity being in amplicons whose gene products were classified as isomerases. Similar patterns were observed for divergence at different categories of sites.

Nucleotide divergence

Genome-wide patterns: Nucleotide divergence as measured by D_{xy} relative to radiata and sugar pines varied significantly among sample coverage classes (Kruskal–Wallis rank sum tests: $P < 1.00e\text{-}9$, $P_{perm} < 1.00e\text{-}4$). Nucleotide divergence was ~2–12 times the level of nucleotide diversity, depending upon choice of outgroup (Table 3), and, as with nucleotide diversity, was affected by the presence or absence of an indel (Table S5). On average (±1 SD), sequences from loblolly pine differed with respect to radiata pine at 0.72% (±0.79%) of sites, while for sugar pine they differed at 4.1% (±2.3%) of sites. As with nucleotide diversity, nucleotide divergence was largest at synonymous sites (radiata pine, 1.2 ± 1.6%; sugar pine, 8.7 ± 6.1%) and lowest at nonsynonymous sites (radiata pine, 0.28 ± 0.58%; sugar pine, 1.9 ± 1.8%; Table 3, Figure S10). The average K_a/K_s ratio (±1 SD) thus varied from 0.59 (±1.68) to 0.27 (±0.30) for radiata and sugar pines, respectively.

Comparisons across categories of amplicons: Nucleotide divergence across all sites relative to radiata pine varied across sets of amplicons (Table 3; Figure 2, Figure S11, and Figure S12) as defined by their presence or absence on the consensus linkage map (Wilcoxon signed rank test: $P = 7.56e\text{-}08$, $P_{perm} < 0.001$), their being annotated or not (Wilcoxon signed rank test: $P < 2.20e\text{-}16$, $P_{perm} < 0.001$), their being associated or unassociated with at least one phenotypic trait (Wilcoxon signed rank test: $P = 0.048$, $P_{perm} = 0.032$), their being associated to different phenotypic categories (Kruskal–Wallis rank sum test: $P = 6.75e\text{-}09$, $P_{perm} < 0.0001$), and their classification into functional categories (Kruskal–Wallis rank sum test: $P = 5.53e\text{-}07$, $P_{perm} < 0.0001$). On average, nucleotide divergence at all sites relative to radiata pine was lower for amplicons that were mapped, those that were annotated, those associated to at least one phenotypic trait, and those that encoded light receptors or ion channels (Figure 2). For those amplicons that were mapped, differences among linkage groups (see Table S7) were significant (Kruskal–Wallis rank sum test: $P = 0.0197$, $P_{perm} = 0.009$). Among phenotypic trait categories, amplicons associated with disease resistance had the largest nucleotide divergence, while those associated with drought had the lowest (Table 3). Nucleotide divergence was largely similar for different categories of sites across different categories of amplicons. For example, nucleotide divergence at nonsynonymous, synonymous, and noncoding sites was not significantly different for amplicons associated with at least one phenotypic trait *vs.* those that were unassociated (Wilcoxon signed rank tests: $P > 0.40$, $P_{perm} > 0.50$). This contradiction with respect to nucleotide divergence for all sites can be explained by the fact that annotated amplicons had a lower level of nucleotide divergence relative to amplicons that were not annotated, so that the previously reported difference for overall nucleotide divergence differing between sets of amplicons associated *vs.* unassociated to at least one phenotypic trait was driven by amplicons that were not annotated. With regard to amplicons associated with specific phenotypic categories, the average (±1 SD) K_a/K_s was largest for amplicons associated with disease phenotypes (0.67 ± 0.89), while it was smallest for amplicons related to drought phenotypes (0.15 ± 0.22). Similar numerical patterns were noted across categories of amplicons when using sugar pine as the outgroup. These differences, however, were not statistically significant (Wilcoxon signed rank tests and Kruskal–Wallis rank sum tests: $P > 0.15$, $P_{perm} > 0.20$).

Linkage disequilibrium

Genome-wide patterns: Levels of intragenic linkage disequilibrium varied across amplicons, but were within the range of those previously published for conifers (*e.g.*, Neale and Savolainen 2004; Namroud *et al.* 2010; Pyhäjärvi *et al.* 2011). For example, the weighted average (±1 SD) value of Kelly's Z_{nS} statistic across amplicons with two or more polymorphic sites was 0.311 (±0.288). The expected rapid decay with physical distance, however, was apparent only when including singletons in the analysis (Table S8, Figure S13, and Figure S14). When singletons were included, the expected value of r^2 decayed to half its initial value in 102–496 bp, depending upon coverage class, indicating that recombination events were ~1.4- to 9.1-fold more likely than mutation events. When excluding singletons, the decay of linkage disequilibrium with physical distance was much reduced to nonexistent, with only three coverage classes experiencing any decay (Table 4). For these analyses, recombination events were ~1.2- to 4.8-fold less likely than mutation events, so that r^2 would decay to half its initial value in 1279–4633 bp. These patterns are related to the relatively short physical distances for most amplicons (*i.e.*, <500 bp; see *Results, Resequencing data summary*). Further

Table 3 Nucleotide divergence across the loblolly pine (*Pinus taeda* L.) genome

Amplicons (*L* = all/annotated)	Nucleotide divergence (D_{xy}: Pira)				Nucleotide divergence (D_{xy}: Pila)			
	All: 95% C.I.[a]	NS: 95% C.I.[a]	SY: 95% C.I.[a]	NC: 95% C.I.[a]	All: 95% C.I.[a]	NS: 95% C.I.[a]	SY: 95% C.I.[a]	NC: 95% C.I.[a]
All	0.0072	0.0028	0.0117	0.0080	0.0410	0.0186	0.0871	0.0492
(Pira: 3484/1659)	0.0069–0.0075	0.0025–0.0031	0.0109–0.0126	0.0074–0.0085	0.0396–0.0426	0.0171–0.0204	0.0825–0.0928	0.0461–0.0526
(Pila: 950/497)								
Not genotyped	0.0068	0.0028	0.0115	0.0074	0.0408	0.0191	0.0907	0.0464
(Pira: 1773/857)	0.0064–0.0072	0.0024–0.0033	0.0102–0.0127	0.0065–0.0083	0.0382–0.0432	0.0169–0.0215	0.0826–0.0987	0.0408–0.0518
(Pila: 478/255)								
Mapped[b]	0.0078	0.0032	0.0129	0.0081	0.0429	0.0198	0.0864	0.0503
(Pira: 836/365)	0.0073–0.0082	0.0027–0.0037	0.0112–0.0147	0.0071–0.0091	0.0396–0.0464	0.0156–0.0248	0.0749–0.0986	0.0446–0.0561
(Pila: 212/111)								
Unassociated	0.0076	0.0028	0.0120	0.0086	0.0420	0.0174	0.0849	0.0515
(Pira: 1263/588)	0.0073–0.0080	0.0025–0.0032	0.0109–0.0131	0.0079–0.0095	0.0396–0.0444	0.0152–0.0199	0.0768–0.0938	0.0473–0.0561
(Pila: 356/208)								
Associated	0.0071	0.0029	0.0117	0.0082	0.0391	0.0206	0.0787	0.0521
(Pira: 448/214)	0.0065–0.0076	0.0023–0.0035	0.0101–0.0133	0.0068–0.0098	0.0360–0.0422	0.0163–0.0255	0.0662–0.0918	0.0425–0.0611
(Pila: 116/62)								
Expression	0.0071	0.0027	0.0155	0.0072	0.0370	0.0205	0.0696	0.0510
(Pira: 44/23)	0.0061–0.0085	0.0016–0.0038	0.0110–0.0203	0.0044–0.0101	0.0319–0.0431	0.0145–0.0282	0.0536–0.0918	0.0356–0.0578
(Pila: 17/9)								
Metabolite	0.0070	0.0028	0.0115	0.0085	0.0390	0.0203	0.0814	0.0517
(Pira: 373/180)	0.0066–0.0075	0.0023–0.0032	0.0010–0.0130	0.0071–0.0098	0.0364–0.0415	0.0164–0.0240	0.0704–0.0919	0.0435–0.0598
(Pila: 92/49)								
Drought	0.0036	0.0022	0.0143	0.0031	0.0878	0.0499	0.1496	0.1025
(Pira: 7/4)	0.0031–0.0070	—	—	—	—	—	—	—
(Pila: 2/1)								
Disease	0.0105	0.0074	0.0111	0.0065	0.0463	0.0581	0.0734	—
(Pira: 6/3)	0.0062–0.0147	—	—	—	—	—	—	—
(Pila: 2/1)								

C.I., confidence interval; *L*, number of amplicons; NC, noncoding; NS, nonsynonymous; Pila, *Pinus lambertiana*; Pira, *P. radiata*; SY, synonymous.

[a] Nonparametric bootstrapping (n = 1000 replicates) across amplicons was used to estimate 95% C.I.'s.

[b] Estimates per linkage group are given in Table S7.

Table 4 Estimates for the intralocus, per site crossing-over rate ($C = 4N_e r$) for coverage classes where a decay of r^2 was observed over physical distance (bp)

Coverage class	Amplicons	C	LD-half (bp)	Z_{nS}	C/θ_π
18	898	0.0021 (0.0006–0.0044)	1,279 (629–4,477)	0.266 (0.243–0.288)	0.868 (0.253–1.831)
17	755	0.0016 (0.0004–0.0031)	1,706 (881–6,824)	0.270 (0.247–0.292)	0.555 (0.151–1.057)
15	374	0.0006 (0.0001–0.0025)	4,633 (1,112–27,800)	0.286 (0.253–0.320)	0.208 (0.034–0.902)

Decay of r^2 with physical distance was not observed for amplicons in all other coverage classes. In these cases, the regression-based estimate of C was approximately zero or negative. Table S8 gives estimates when singletons were included for coverage classes ranging from 11 to 18. Values in parentheses are 95% nonparametric bootstrap (n = 10,000 replicates) confidence intervals based on resampling of amplicons with replacement. LD-half represents the distance in base pairs required for r^2 to decrease to half its initial value. Values of θ_π represent arithmetic averages across amplicons within each coverage class.

descriptions of linkage disequilibrium, including those for categories of amplicons, can be found in File S1.

Departures from neutrality

Genome-wide patterns: The folded site-frequency spectrum for all sites deviated strongly from the SNM (Figure S15). This was also apparent for different categories of sites (data not shown), with deviations most apparent for categories corresponding to low-frequency variants. Similar patterns were observed for the unfolded site-frequency spectrum, but deviations were observed for low- and high-frequency derived variants. This pattern was consistent across unfolded site-frequency spectra based on each outgroup (Figure S16 and Figure S17). Summarizing site-frequency spectra by locus using Tajima's D gave genome-wide weighted averages (± 1 SD) of -0.474 (± 0.951), -0.465 (± 0.925), -0.354 (± 0.984), and -0.417 (± 0.961) for all, nonsynonymous, synonymous, and noncoding sites, respectively. Genome-wide estimates of Fay and Wu's H were similarly negative, with weighted averages (± 1 SD) of -0.223 (± 0.987), -0.268 (± 1.052), -0.198 (± 0.945), and -0.247 (± 0.967) for all, nonsynonymous, synonymous, and noncoding sites, respectively.

Comparison of amplicon-specific values by coverage class of Tajima's D to the distribution predicted under a strict neutral model revealed that 3.9–8.0% of amplicons were outliers ($n = 260$ amplicons total) at a significance level of $P = 0.05$ (Figure S18A). None of these survived corrections for multiple tests, using a false-discovery rate (FDR) correction (Storey and Tibshirani 2003). Incorporation of a simple TEM of demography from Ersöz *et al.* (2010), however, removed the vast majority of these outliers (Figure S18B), so that only 52 of the previously described 260 outliers remained as outliers at a significance level of $P = 0.05$ after incorporation of the TEM. These amplicons tended to be in either tail (*i.e.*, the extreme positive or extreme negative tail) of the genome-wide distribution for Tajima's D conditional on coverage class. This was expected as the TEM fitted the data better than the SNM (Table S9), even though the formulation of the TEM by Ersöz *et al.* (2010) was based on samples collected only from the eastern portion of the range of loblolly pine and disease-related candidate genes (but see Figure S19). In addition, none of the significant amplicons survived corrections for multiple tests, using a FDR approach.

Analysis using Fay and Wu's H produced a similar trend for the SNM, but with the number of significant outliers increasing from 178 to 692 when the TEM was incorporated into the null model. This resulted in 21.3–38.9% of the observed values of Fay and Wu's H being outliers across coverage classes at a significance threshold of $P = 0.05$, which exceeds the expected number of false positives by approximately fivefold. As with Tajima's D, these amplicons were in the tails of the genome-wide distributions for Fay and Wu's H conditional on coverage class (Figure S20). Using a FDR correction for the TEM resulted in 57 amplicons remaining significant, using a threshold of $Q = 0.05$. Part of this difference relative to Tajima's D is due to the poor predictive ability of the TEM for the variance across amplicons for Fay and Wu's H [*i.e.*, the observed value for the weighted variance of H across amplicons is in the upper 0.01% tail of the simulated distribution ($n = 1.0 \times 10^5$ simulations) under the TEM].

Joint consideration of Tajima's D and Fay and Wu's H under the TEM identified 31 amplicons with D and H both significant at $P = 0.05$ (Table S10). All 31 amplicons had negative values for D and H that were ~3.8- (D) to 22.3-fold (H) more negative than the average across amplicons. These 31 amplicons spanned the spectrum of putative gene functions ranging from heat-shock proteins (*e.g.*, 0_10631_01) and universal stress protein family proteins (*e.g.*, 0_12117_01) to protein kinases (*e.g.*, CL2463Contig1_03), with 15 of these 31 amplicons, however, not being annotated with respect to putative gene function or categories of sites. Although not a formal multidimensional test (see Zeng *et al.* 2007), this list gives the most extreme outliers with respect to D and H when considered jointly. When compared with the list of amplicons associated with phenotypic traits, there was no enrichment for associated amplicons to be in the lower tail of the distributions of D and H (permutation tests: $P > 0.20$).

Comparisons of polymorphism and divergence at multiple classes of sites using an extension of the McDonald–Kreitman test revealed that the loblolly pine genome contains significant signatures of past positive selection (Figure 3). Using levels of nucleotide diversity and divergence at nonsynonymous and synonymous sites suggested that on average 29% [95% confidence interval (C.I.): 22–37%] of new nonsynonymous substitutions have been fixed by positive, directional selection within the loblolly pine genome. This

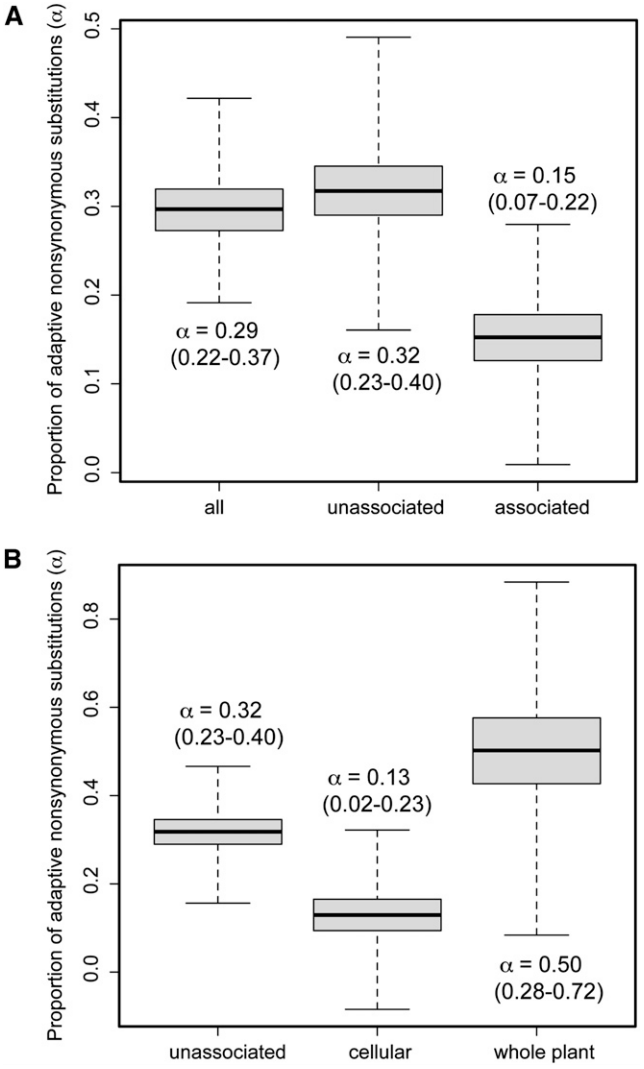

Figure 3 The proportion of adaptive nonsynonymous substitutions (α) is significantly different from zero for the loblolly pine genome and varies across classes of amplicons. Numbers in parentheses give 95% bootstrap confidence intervals (n = 10,000 replicates) for α in both panels. (A) Whereas α is significantly larger than zero for all classes of amplicons, it is highest for the set of amplicons without associations to any of the phenotypes. (B) The patterns observed in A, however, are driven by amplicons associated to one of the cellular phenotypes (expression and metabolites), where the point estimate of α is 3.8 times lower than the point estimate for whole plant phenotypes (drought, disease, environmental associations).

estimate is derived from a model of polymorphism and divergence where each amplicon has a unique value of $\theta = 4N_e u$ (see Table S11) and a fixed level of divergence (ut = 0.0083, 95% C.I.: 0.0078–0.0089), selective constraint (f = 0.22, 95% C.I.: 0.15–0.29), and α. Of the 48 models tested, where these parameters were set to zero, fixed, or allowed to vary across amplicons (see Table S11), this model was one of the best and, when fixing α across amplicons, it was the best (AICc: 18,936.13 *vs.* next best model AICc: 19,383.60, so ΔAICc = 447.47). Modifying the data to exclude segregating variation below a frequency threshold of 10% did not change this estimate dramatically (α = 25%, 95% C.I.: 19–35%), which suggests that the segregation of slightly deleterious mutations is not driving this result.

Comparisons across categories of amplicons: As noted previously, nucleotide diversity differed significantly between amplicons associated with at least one phenotypic trait relative to those that were not (see *Results*, *Nucleotide diversity and insertion–deletion events*, Figure 2, and Figure S21). This translated into significant differences for Tajima's D (Figure S21) and Fay and Wu's H (Wilcox rank sum tests: all, $P = 0.254$, $P_{perm} = 0.196$; nonsynonymous, $P < 2.2e\text{-}16$, $P_{perm} < 0.001$; synonymous, $P = 0.137$, $P_{perm} = 0.098$; noncoding, $P = 0.051$, $P_{perm} = 0.048$) between these classes. The strongest differences occurred for nonsynonymous sites, which exhibited significantly more negative values for D and H for amplicons associated with phenotypic traits relative to amplicons not associated with a phenotypic trait for both statistics.

With respect to polymorphism and divergence, the best model of those examined allowed α and θ to vary and for the strength of selective constraint and divergence to be constant across amplicons (AICc = 18,884.98). The estimate for the average value of α across amplicons in this model (average α = 0.32, 95% C.I.: 0.19–0.45) was similar to that for fixed α. This result, however, established that α varied across amplicons. To investigate this variation further, we utilized classes of amplicons based on whether or not they were associated with a phenotypic trait and fitted a model where α was allowed to vary between classes of amplicons. This model fitted the data almost as well as the best-fit model with variable α drawn from a β-distribution (AICc = 18,953.98, so ΔAICc = 69.00), which suggests that much of the variability of α across amplicons is driven by differences between these classes of amplicons. Estimates of α for each class in this model illustrate a twofold higher value for amplicons unassociated with a phenotype (α = 0.32) compared to amplicons associated with a phenotype (α = 0.15; Figure 3A). If we again divide amplicons from the class entitled "associated with a phenotypic trait" into two categories, whole plant phenotypic traits and cellular phenotypic traits, and fit a model with three classes of amplicons, the fit of the resulting model is again almost as good as that of the best-fit model (AICc = 18,960.11, so ΔAICc = 75.13). In this case, the estimate of α for amplicons unassociated with a phenotypic trait does not change, but the estimates for α for the two classes of amplicons associated with a phenotypic trait were significantly different (Figure 3B). The whole plant phenotypic traits were associated with amplicons with abundant signals of positive, directional selection (α = 0.50, 95% C.I.: 0.28–0.72), while the cellular phenotypic traits were associated with amplicons displaying marginal signals of positive, directional selection (α = 0.13, 95% C.I.: 0.02–0.23). In accordance with α varying between these classes, so did the level of selective constraint, with cellular phenotypic traits being associated to amplicons with much higher

levels of selective constraint (f = 0.305, 95% C.I.: 0.256–0.354) than those associated with whole plant phenotypic traits (f = 0.106, 95% C.I.: 0.053–0.148).

Similar trends were apparent when amplicons containing a SNP associated to at least one phenotypic trait were clustered into categories based on putative gene functions (Table S12, Figure 4). In general, the higher the fraction of amplicons associated to a phenotypic trait, the more extreme were deviations from neutrality based on the site-frequency spectrum (Figure 4, left) and polymorphism and divergence (Figure 4, right). The site-frequency spectrum was most skewed, as measured using the weighted average of Tajima's D, for synonymous sites, with many of the functional categories, especially those at extreme values of the DoS statistic, being significantly more skewed than expected by chance. For polymorphism and divergence, 11 functional categories had 95% bootstrap confidence intervals for the DoS statistic excluding zero (7 with significantly positive values of DoS and 4 with significantly negative values of DoS), and those tended to be categories with more amplicons associated to at least one phenotypic trait and those with more skewed site-frequency spectra (see Table S12). These categories contained genes encoding proteins such as microtubules (+DoS), ferredoxins (+DoS), metal-binding proteins (+DoS), ligases (+DoS), phosphatases (+DoS), protein kinases (+DoS), glyoxyl oxidases (+DoS), zinc-finger proteins (−DoS), isomerases (−DoS), ATPases (−DoS), and vitamin-binding proteins (−DoS).

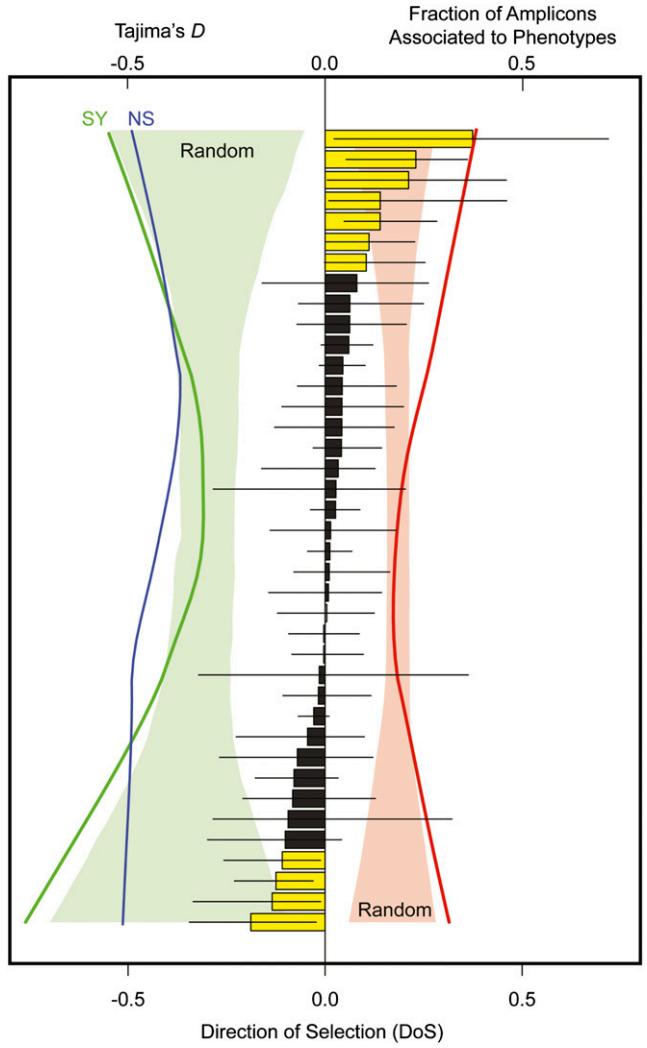

Figure 4 Nonneutral evolution was apparent by functional categories of amplicons, so that as the fraction of amplicons associated to at least one phenotype increased within a functional category, so did patterns of divergence and polymorphism at nonsynonymous (NS) and synonymous (SY) sites as well as the skew in the site-frequency spectrum at synonymous sites. The bar plot gives the weighted average of the direction of selection (DoS) statistic for amplicons within the 40 functional categories. All amplicons within one category (methyltransferases) did not have outgroup data available, so that DoS was undefined. Error bars are 95% bootstrap confidence intervals (n = 10,000 replicates over amplicons). Yellow bars have 95% confidence intervals excluding zero, while black bars do not (see Table S11). The bottom x-axis is for the DoS statistic, while the top x-axis is split between Tajima's D (left) and the fraction of amplicons associated with at least one phenotype (right). The red colored area to the right gives the 95% confidence interval for the null distribution, based on 10,000 permutations of amplicons among categories, for the fraction of amplicons associated to at least one phenotype across functional categories. Similarly, the green area on the left does the same for the weighted average value of Tajima's D at SY sites. Lines give observed values (green, Tajima's D at SY sites; blue, Tajima's D at NS sites; red, fraction of amplicons associated to at least one phenotype). All lines were loess smoothed, including those comprising the null distributions. The null distribution for the weighted average of Tajima's D at NS sites was similar to that for synonymous sites, so was omitted for clarity.

Discussion

Understanding the genetic basis of ecologically relevant traits is a primary focus of evolutionary genetics, with ramifications for the basic understanding of genetic diversity, including the origin and maintenance of this diversity (reviewed by Barton and Turelli 1989; Barton and Keightley 2002), as well as the conservation and management of this diversity in natural populations (González-Martínez *et al.* 2006b; Neale 2007; Neale and Kremer 2011). Here we have shown that the genetic variation correlated with a wide range of different phenotypic traits for loblolly pine appears to be nonneutral, but also that the magnitude of the skew from neutral expectations by this genetic variation depends upon the type of phenotypic trait under investigation.

Nucleotide diversity, nucleotide divergence, and linkage disequilibrium

Genome-wide patterns of nucleotide diversity were similar to those reported previously for loblolly pine (Brown *et al.* 2004; González-Martínez *et al.* 2006a; Ersöz *et al.* 2010), as well as for conifers in general (Savolainen and Pyhäjärvi 2007). Novel to this study, however, is the observation that indel-associated mutations or a combination of indel-associated mutations and variation in selective constraint likely drive associations between indels and nucleotide diversity (see

Table S5; Tian *et al.* 2008; Hollister *et al.* 2010). Inspection of a large number of amplicons (n = 5773), moreover, allowed for the dissection of this pattern by types of amplicon. As noted by Ersöz *et al.* (2010) and as illustrated here (see Figure 2 and Figure S21), nucleotide diversity varied by amplicon category, with this being true for functional categories and for categories based on whether amplicons were associated with a phenotypic trait. With respect to the latter, this was observed for both noncoding and nonsynonymous sites (see Figure S21). For amplicons associated with a phenotypic trait, nucleotide diversity, as measured with θ_W, was significantly higher at nonsynonymous sites, while for noncoding sites it was significantly lower. This is indicative of an increase in the level of rare segregating variation at nonsynonymous sites for amplicons associated with a phenotypic trait. This pattern could reflect either an increase in segregating deleterious variation for these amplicons or the action of linked, positive selection where the targeted SNP was not typed (*e.g.*, another nonsynonymous variant not seen in the sample).

Teasing apart segregating deleterious variation from linked, positive selection, however, depends, at least partly, on the patterns of linkage disequilibrium within conifer genomes. As shown repeatedly, linkage disequilibrium decays over physical distance quickly within coding regions (Neale and Savolainen 2004; but see Namroud *et al.* 2010; Pyhäjärvi *et al.* 2011), yet conifers appear to have extremely low recombination rates (Jaramillo-Correa *et al.* 2010) and effective population sizes as estimated from estimates of Θ are moderate (Savolainen and Pyhäjärvi 2007). The same pattern was observed here, as linkage disequilibrium decayed quite often within 500 bp to half its initial value. This pattern, however, was apparent only when including singletons in the analysis and removal of these types of polymorphisms resulted in fairly strong linkage disequilibrium being detected. Inspection of values less sensitive to sample size (*e.g.*, D', data not shown), moreover, showed little decay with physical distance, so that linkage disequilibrium may be more prominent in some regions than thought previously within conifer genomes (*e.g.*, Moritsuka *et al.* 2012). The observation that average levels of linkage disequilibrium varied across sets of amplicons defined as being associated with certain phenotypic traits is consistent with this conclusion (File S1), as is the observation that autocorrelation in values of nucleotide diversity and divergence occurred at the centimorgan scale (Figure S9; autocorrelation >0 extended upward of 3 cM for each measure, results not shown). Further studies on a genome-wide scale with larger samples, however, are needed to resolve the issue fully with regard to patterns of linkage disequilibrium and recombination (see Mackay *et al.* 2012).

Genome-wide levels of nucleotide divergence were also as expected given recent estimates of the divergence times among pine species (Gernandt *et al.* 2008; Parks *et al.* 2009). For example, the per site divergence of loblolly relative to radiata pine was 0.72%, and although severalfold greater than nucleotide diversity, was relatively low. This was also reflected in the low number of fixed differences observed between these two species, which is likely a function of their recent divergence time ($\sim \leq$20 MY). When broken down by categories of sites, the largest divergence was seen at synonymous sites and the lowest was observed at nonsynonymous sites, which supports the ubiquitous nature of purifying selection (average $K_a/K_s << 1$, Stephan 2010; Charlesworth 2013). With regard to categories of amplicons, however, several categories exhibited average K_a/K_s ratios that were severalfold larger than the genome-wide average (*e.g.*, disease-associated amplicons). Such ratios have been used as evidence of selection previously in species of *Pinus* (*e.g.*, Palmé *et al.* 2008) and are also consistent with previous studies that documented strong signals of positive selection in genes related to disease resistance for loblolly pine (Ersöz *et al.* 2010). When sugar pine was used as an outgroup, however, these quantitative patterns largely disappeared. This is likely a function of the decreased number of amplicons available for analysis (3484 *vs.* 950), as well as the bias imposed by use of PCR primers from diverged taxa (Kern 2009; Eckert *et al.* 2013).

Evolutionary genetics of gene regions underlying phenotypic traits

Genome-wide patterns of diversity and divergence were used to infer that ~29% of new nonsynonymous substitutions on average were fixed due to the action of positive directional selection. This is the first report of this quantity on a genome-wide scale for a conifer (but see Eckert *et al.* 2013), which will add to the emerging literature of adaptive amino acid evolution estimates for plant species (Gossmann *et al.* 2010; Slotte *et al.* 2010; Strasburg *et al.* 2011). In addition, a suite of 31 outliers with respect to summaries of the site-frequency spectrum was also identified after accounting for a simple demographic null model (see Table S10). As such, this represents one of the first comprehensive scans of a conifer genome for deviations from neutrality (see also Pavy *et al.* 2013), with the conclusion that signals of positive directional selection are apparent and moderately abundant.

The central question addressed here, however, is how amplicons identified using population genomic scans relate to those identified using linkage disequilibrium mapping. The genetic architecture of phenotypic traits amenable to linkage disequilibrium mapping is likely composed of mostly additive genetic variance (Hill *et al.* 2008; but see Hansen 2013). As such, marginal signals of selection for the amplicons comprising the underlying genetic architecture of these phenotypic traits are expected to be minimal (Hermisson and Pennings 2005; Chevin and Hospital 2008; Pavlidis *et al.* 2012). This was supported by the results presented here, as the individual amplicons themselves did not emerge from scans of summary statistics such as Tajima's D or Fay and Wu's H. The joint effects of multiple amplicons grouped together because of their associations with phenotypic traits,

however, do emerge as different from those of amplicons not associated with phenotypic traits. This implies that it is the multilocus attributes that are important when considering the links between phenotypic traits, genetic associations, and scans for positive selection (Bürger and Gimelfarb 1999; Bürger 2000; Pritchard *et al.* 2010). This pattern emerged on several levels of analysis: summaries of nucleotide diversity and divergence, average summary statistics based on the site-frequency spectrum, and results from a MacDonald–Kreitman-type analysis. Moreover, as the fraction of amplicons associated with a phenotypic trait increased within functional categories, so did signals deviating from neutrality (Figure 4). In general, the differences were consistent with deviations from neutrality supporting the presence of positive, directional selection, but also in many instances that of purifying selection.

While this lends credence to the methods employed previously for linkage disequilibrium mapping, as signals in these studies appear to be driven by biological indicators, it does not explain fully the type, magnitude, and dynamics of natural selection that could be driving the results. Specifically, we have not addressed the role of population structure, both neutral patterns due to phylogeographic history and adaptive divergence due to local selection pressures, as influencing the observed patterns. As mentioned previously, forest trees, and specifically loblolly pine, display clear patterns of local adaptation (Neale and Kremer 2011). These patterns likely drive the environmental associations detected previously (*e.g.*, Eckert *et al.* 2010a,b), as well as many of the genotype–phenotype associations (Quesada *et al.* 2010; Cumbie *et al.* 2011; Palle *et al.* 2011, 2013; Eckert *et al.* 2012). This may thus explain the disconnect between summaries of the site-frequency spectrum for individual amplicons and their frequency of being associated with a phenotype. In this case, our sample design does not likely have the power to detect the effects of local selection pressures (Städler *et al.* 2009), as this power depends on sampling intensity and pattern with respect to the largely unknown important selective gradients (see discussion in Sork *et al.* 2013).

Population structure, however, is unlikely to explain the observed patterns for α. Depending on the method used to estimate α, population structure would need to be pronounced, the sampling relatively even across diverged populations, and magnitudes of population structure would need to differ between categories of amplicons due to demographic history (see Eckert *et al.* 2013). Further work based on population-level sampling, however, would be needed to quantify any effects of population structure on our inferences.

It is clear from the data and analysis presented here that long-term signals of positive selection were apparent for amplicons associated with phenotypic traits, while those for recent selective sweeps (in terms of $4N_e$ generations) were only marginally different from genome-wide estimates. This implies that the amplicons underlying these phenotypic traits have experienced fixation of adaptive alleles over the divergence time separating loblolly and radiata pines. As such, segregating variants within these amplicons are subject to this historical context, so that an interesting question arises with regard to whether these variants, which are correlated with phenotypic traits, represent the segregation of slightly deleterious variants (*i.e.*, new deleterious mutations deviating from the adaptive fixed allele) or indeed reflect positively selected alleles (*i.e.*, some form of recurrent selection). For example, the putatively synthetic associations documented by Eckert *et al.* (2012) for the loblolly pine primary metabolome could reflect linked selection upon rare deleterious mutations, but also could reflect complex patterns of recent positive selection. Both processes would affect standing patterns of linkage disequilibrium, so that answering this question without a truly genomic resource is difficult, especially since the landscape of linkage disequilibrium, despite the data presented here, remains largely unknown for this species (see Goldstein and Weale 2001; but see also Moritsuka *et al.* 2012). One tempting glimpse into the answer of this question, however, is found by comparing estimates of long-term selection, as measured by α, across different phenotypic trait categories.

As the genome-wide estimate of α was significantly greater than zero, it is not surprising that amplicons associated with phenotypic traits were estimated to have statistically significant signals of long-term positive selection. This similarity also carried over to the distribution of fitness effects for newly arising nonsynonymous mutations. Using method 2 of Eyre-Walker and Keightley (2009), as implemented in the DoFE software ver. 3 (see http://www.lifesci.sussex.ac.uk/home/Adam_Eyre-Walker/Website/Software.html and Eckert *et al.* 2013 for a description of how DoFE was implemented), this distribution was estimated to be similar regardless of whether the data were from amplicons associated with a phenotype or not, with most new nonsynonymous variants being highly ($N_e s > 100$, ~65% for both associated and unassociated amplicons) or slightly ($N_e s < 1$, ~20% for both associated and unassociated amplicons) deleterious. One of the most striking results, however, is that phenotypic trait class was differentiated with respect to estimates of α relative to the genome-wide average. For example, whole plant phenotypes (including the environmental associations) were almost twofold greater than the genome-wide average, while cellular phenotypes were almost twofold less than the genome-wide average. Differences of this magnitude, moreover, were unable to be replicated by randomly permuting the full data set among category labels ($P_{perm} < 0.001$) so that it is unlikely these patterns are statistical artifacts. This implies that the strength, type, and dynamics of natural selection for amplicons within these classes differ (see Casto and Feldman 2011 for a similar case for humans), as was shown, for example, by the larger estimate of selective constraint for amplicons associated with cellular phenotypes.

The implication would then be that genetic associations discovered for cellular phenotypes represent the segregation of slightly deleterious mutations, whereas this would not likely be the case for whole plant phenotypes. This is consistent with the lower estimates of α, the significant enrichment for rare alleles in genetic associations, and the increased estimates of selective constraint for amplicons associated with cellular phenotypes (Park *et al.* 2010). In addition, an analysis using method 2 of Eyre-Walker and Keightley (2009) shows that the distribution of newly arising nonsynonymous mutations is skewed toward an increase of slightly deleterious mutations (*i.e.*, $N_e s$ in the range 0–1) for amplicons associated with cellular phenotypes (*i.e.*, this proportion was 0.32 with a 95% confidence interval of 0.27–0.37) compared to those associated with whole plant phenotypes (this proportion was 0.16 with a 95% confidence interval of 0.09–0.23). With respect to α, however, the lower estimate for amplicons may instead just reflect an increased abundance of segregating slightly deleterious variants that is independent from the discovered genetic associations (*i.e.*, these are not driving the discovery of the genetic associations), so that estimates for α are downwardly biased. The segregation of slightly deleterious mutations will produce this effect (Charlesworth and Eyre-Walker 2008). Successive removal of low-frequency variants (*i.e.*, singletons) for just the amplicons associated with cellular phenotypes, however, did not increase the estimate for α beyond 0.15 (95% C.I.: 0.08–0.25). In addition, estimates of α using method 2 of Eyre-Walker and Keightley (2009), which models the segregation of deleterious variants explicitly, were similar to those reported here (*i.e.*, changes of $<$20% of the point estimates shown in Figure 3 with all estimates of α still significantly >0), so that it is unlikely that segregating deleterious variation is solely driving the observed patterns. Cellular phenotypes thus appear to be different with regard to their genetic architecture and patterns of selection compared with whole plant phenotypes.

Limitations and conclusions

The analysis presented here is limited with respect to sample design and power. The disconnect between results from population genomic scans and those from linkage disequilibrium mapping with regard to amplicons identified by each approach may thus be an artifact. For example, only 18 trees were sampled for the population genomic scan utilized here. While much of the underlying information about a sample genealogy can be obtained from a relatively small sample size, sample size does have a direct effect on estimates of the summary statistics used to test neutrality and on inferences of linkage disequilibrium (Nielsen 2005). As such, estimates were stratified by sample size and weighted averages were used (Begun *et al.* 2007). Caution is thus needed when interpreting the results presented here, especially since sample coverage affected most statistics (see Table S3). This effect, moreover, is likely related to standing levels of population structure, our conservative sequence analysis pipeline, and inclusion of small coverage classes (*i.e.*, n = 2–5). We have also used a recently diverged outgroup, and only a single sequence from this outgroup for analysis, that may in fact not be reciprocally monophyletic (Syring *et al.* 2007). As such, estimates of divergence may be biased and thus by extension so would estimates of long-term positive selection. With that said, however, the genome-wide estimate of α using sugar pine as the outgroup remained significantly positive (α = 0.10, 95% C.I.: 0.01–0.20), with the difference likely attributed to the increase of conserved amplicons in this set (Kern 2009; Eckert *et al.* 2013). A more thorough analysis, however, was not performed, as only 950 amplicons were available, and many of the amplicons associated with phenotypes were not in this set. Finally, the models considered during the estimation of α did not explicitly account for the segregation of slightly deleterious mutations. The same quantitative patterns, however, emerge as those presented in Figure 3 as rare mutations are excluded from the data set (data not shown for singletons and doubletons removed) and when using a method that specifically models the segregation of deleterious mutations. In these cases, estimates of α change by $<$15–20%.

The continued application of linkage disequilibrium mapping will help to uncover the identity of genes putatively composing the genetic architecture of complex traits for populations of forest trees. As shown here, the amplicons identified by these studies appear to be nonneutral with respect to patterns of nucleotide diversity and divergence, yet further work will be needed to make general statements across groups of taxa. This nonneutrality, moreover, appears to be detectable most clearly over long periods of time (*e.g.*, since the divergence with the outgroup). As noted by Hahn (2008), the most interesting steps forward, once truly genome-wide data are available, will not be to reject neutrality, but to determine why this null model was rejected.

Acknowledgments

The authors thank the staff at Agencourt Biosciences (now Beckman Coulter Genomics) for help in sequencing and the initial bioinformatics. Comments from the associate editor and two anonymous reviewers greatly improved this manuscript. Funding for this project was made available from the National Science Foundation (NSF) Integrative Organismal Systems (IOS) Plant Genome Research Program (PGRP) (NSF-IOS-PGRP: 0501763).

Literature Cited

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25: 25–29.

Barrett, R. D. H., and H. E. Hoekstra, 2011 Molecular spandrels: tests of adaptation at the genetic level. Nat. Rev. Genet. 12: 767–780.

Barton, N. H., and P. D. Keightley, 2002 Understanding quantitative genetic variation. Nat. Rev. Genet. 3: 11–21.

Barton, N. H., and M. Turelli, 1989 Evolutionary quantitative genetics: How little do we know? Annu. Rev. Genet. 23: 337–370.
Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y.-P. Poh *et al.*, 2007 Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 5: e310.
Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley, and D. B. Neale, 2004 Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc. Natl. Acad. Sci. USA 101: 15255–15260.
Bürger, R., 2000 *The Mathematical Theory of Selection, Recombination, and Mutation*. John Wiley & Sons, West Sussex, UK.
Bürger, R., and A. Gimelfarb, 1999 Genetic variation maintained in multilocus models of additive quantitative traits under stabilizing selection. Genetics 152: 807–820.
Burnham, K. P., and D. R. Anderson, 2004 *Model Selection and Multimodal Inference: A Practical Information-Theoretic Approach*, Ed. 2. Springer-Verlag, New York.
Cartwright, D. A., M. Troggio, R. Velasco, and A. Gutin, 2007 Genetic mapping in the presence of genotyping errors. Genetics 176: 2521–2527.
Casto, A. M., and M. W. Feldman, 2011 Genome-wide association study SNPs in the human genome diversity project populations: Does selection affect unlinked SNPs with shared trait associations? PLoS Genet. 7: e1001266.
Chan, E. K. F., H. C. Rowe, and D. J. Kliebenstein, 2010a Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. Genetics 185: 991–1007.
Chan, E. K. F., H. C. Rowe, B. G. Hansen, and D. J. Kliebenstein, 2010b The complex genetic architecture of the metabolome. PLoS Genet. 6: e1001198.
Charlesworth, B., 2013 Background selection 20 years on – The Wilhelmine E. Key 2012 Invitational Lecture. J. Hered. 104: 161–171.
Charlesworth, J., and A. Eyre-Walker, 2008 The McDonald-Kreitman test and slightly deleterious mutations. Mol. Biol. Evol. 25: 1007–1015.
Chevin, L.-M., and F. Hospital, 2008 Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics 180: 1645–1660.
Comeron, J., 1995 A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. 41: 1152–1159.
Cumbie, W. P., A. J. Eckert, J. L. Wegrzyn, R. Whetten, D. B. Neale *et al.*, 2011 Association genetics of carbon isotope discrimination, height, and foliar nitrogen in a natural population of *Pinus taeda* L. Heredity 107: 105–114.
Devey, M. E., M. M. Sewell, T. L. Uren, and D. B. Neale, 1999 Comparative mapping in loblolly and radiata pine using RFLP and microsatellite markers. Theor. Appl. Genet. 99: 656–662.
Eckert, A. J., A. D. Bower, B. Pande, K. D. Jermstad, K. V. Krutovsky *et al.*, 2009a Association genetics of coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. Genetics 182: 1289–1302.
Eckert, A. J., B. Pande, E. S. Ersöz, M. H. Wright, V. K. Rashbrook *et al.*, 2009b High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). Tree Genet. Genomes 5: 225–234.
Eckert, A. J., J. van Heerwaarden, J. L. Wegrzyn, C. D. Nelson, J. Ross-Ibarra *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, A. J., A. D. Bower, S. C. González-Martínez, J. L. Wegrzyn, G. Coop *et al.*, 2010b Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). Mol. Ecol. 19: 3789–3805.
Eckert, A. J., J. D. Liechty, B. R. Tearse, B. Pande, and D. B. Neale, 2010c DnaSAM: software to perform neutrality testing for large datasets with complex null models. Mol. Ecol. Res. 10: 542–545.
Eckert, A. J., J. L. Wegrzyn, W. P. Cumbie, B. Goldfarb, D. A. Huber *et al.*, 2012 Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. New Phytol. 193: 890–902.
Eckert, A. J., A. D. Bower, K. D. Jermstad, J. L. Wegrzyn, B. J. Knauss *et al.*, 2013 Multilocus analyses reveal little evidence for lineage wide adaptive evolution within major clades of soft pines (*Pinus* subgenus *Strobus*). Mol. Ecol. DOI: 10.1111/mec.12514.
Efron, B., 1985 Bootstrap confidence intervals for a class of parametric problems. Biometrika 72: 45–58.
Endler, J. A., 1986 *Natural Selection In the Wild*. Princeton University Press, Princeton, NJ.
Ersöz, E. S., M. H. Wright, S. C. González-Martínez, C. H. Langley, and D. B. Neale, 2010 Evolution of disease response genes in loblolly pine: insights from candidate genes. PLoS ONE 5: e14234.
Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using PHRED. II. Error probabilities. Genome Res. 8: 186–194.
Ewing, B., L. Hillier, M. C. Wendl, and P. Green, 1998 Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. Genome Res. 8: 175–185.
Eyre-Walker, A., 2010 Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc. Natl. Acad. Sci. USA 107: 1752–1756.
Eyre-Walker, A., and P. D. Keightley, 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26: 2097–2108.
Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.
Florea, L., G. Hartzell, Z. Zhang, G. M. Rubin, and W. Miller, 1998 A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. 8: 967–974.
Gernandt, D. S., S. Magallon, G. G. Lopez, O. Z. Flores, A. Willyard *et al.*, 2008 Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. Int. J. Plant Sci. 169: 1086–1099.
Goldstein, D. B., 2011 The importance of synthetic associations will only be resolved empirically. PLoS Biol. 9: e1001008.
Goldstein, D. B., and M. E. Weale, 2001 Population genomics: linkage disequilibrium holds the key. Curr. Biol. 11: R576–R579.
González-Martínez, S. C., E. Ersöz, G. R. Brown, N. C. Wheeler, and D. B. Neale, 2006a DNA sequence variation and selection of tag SNPs at candidate genes for drought-stress response in *Pinus taeda*. Genetics 172: 1915–1926.
González-Martínez, S. C., K. V. Krutovsky, and D. B. Neale, 2006b Forest tree population genomics and adaptive evolution. New Phytol. 170: 227–238.
González-Martínez, S. C., D. Huber, E. Ersoz, J. M. Davis, and D. B. Neale, 2008 Association genetics in *Pinus taeda* L. II. Water Use Efficiency. Heredity 101: 19–26.
González-Martínez, S. C., S. Dillon, P. Garnier-Géré, K. Krutovsky, R. Alía *et al.*, 2011 Patterns of nucleotide diversity and association mapping, pp. 239–275 in *Genetics, Genomics, and Breeding of Conifers* (Series on Genomics of Industrial Crops), edited by C. Plomion and J. Bousquet. Science Publishers, Enfield, New Hampshire.
Gordon, D., C. Abajian, and P. Green, 1998 Consed: a graphical tool for sequence finishing. Genome Res. 8: 195–202.
Gossmann, T. I., B.-H. Song, A. J. Windsor, T. Mitchell-Olds, C. J. Dixon *et al.*, 2010 Genome wide analyses reveal little evidence

for adaptive evolution in many plant species. Mol. Biol. Evol. 27: 1822–1832.
Grattapaglia, D., and R. Sederoff, 1994 Genetic linkage maps of *Eucalyptus grandis* and *E. urophylla* using a pseudo-testcross mapping strategy and RAPD markers. Genetics 137: 1121–1137.
Kelly, J., 1997 A test of neutrality based on interlocus associations. Genetics 146: 1197–1206.
Kern, A. D., 2009 Correcting the site frequency spectrum for divergence-based ascertainment. PLoS ONE 4: e5152.
Hahn, M. W., 2008 Toward a selection theory of molecular evolution. Evolution 62: 255–265.
Hansen, T. F., 2013 Why epistasis is important for selection and adaptation. Evolution DOI: 10.1111/evo.12214.
Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169: 2335–2352.
Hill, W. G., M. E. Goddard, and P. M. Visscher, 2008 Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 4: e1000008.
Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta *et al.*, 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA 106: 9362–9367.
Hollister, J. D., J. Ross-Ibarra, and B. S. Gaut, 2010 Indel-associated mutation rate varies with mating system in flowering plants. Mol. Biol. Evol. 27: 409–416.
Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7: 1–44.
Hufford, M. B., X. Xun, J. van Heerwaarden, T. Pyhäjärvi, J.-M. Chia *et al.*, 2012 Comparative population genomics of maize domestication and improvement. Nat. Genet. 44: 808–811.
Ingvarsson, P. K., and N. R. Street, 2011 Association genetics of complex traits in plants. New Phytol. 189: 909–922.
Jaramillo-Correa, J. P., M. Verdu, and S. C. Gonzalez-Martinez, 2010 The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. BMC Evol. Biol. 10: 22.
Lander, E. S., and N. J. Schork, 1994 Genetic dissection of complex traits. Science 265: 2037–2048.
Lebude, A. V., B. Goldfarb, F. A. Blazich, J. Frampton, and F. C. Wise, 2004 Mist, substrate water potential, and cutting water potential influence rooting of stem cuttings of loblolly pine. Tree Physiol. 24: 823–831.
Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.
Mackay, J., J. F. Dean, C. Plomion, D. G. Peterson, F. M. Canovas *et al.*, 2012 Towards decoding the conifer giga-genome. Plant Mol. Biol. 80: 555–569.
Marth, G. T., I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu *et al.*, 1999 A general approach to single nucleotide polymorphism discovery. Nat. Genet. 23: 452–456.
Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. 23: 23–35.
Mitchell-Olds, T., M. Feder, and G. Wray, 2008 Evolutionary and ecological functional genomics. Heredity 100: 101–102.
Moritsuka, E., Y. Hisataka, M. Tamura, K. Uchiyama, A. Watanabe *et al.*, 2012 Extended linkage disequilibrium in noncoding regions in a conifer, *Cryptomeria japonica*. Genetics 190: 1145–1148.
Namroud, M.-C., C. Guillet-Claude, J. Mackay, N. Isabel, and J. Bousquet, 2010 Molecular evolution of regulatory genes in spruces from different species and continents: heterogeneous patterns of linkage disequilibrium and selected but correlated recent demographic changes. J. Mol. Evol. 70: 371–386.
Neale, D. B., 2007 Genomics to tree breeding and forest health. Curr. Opin. Genet. Dev. 17: 539–544.
Neale, D. B., and P. K. Ingvarsson, 2008 Population, quantitative and comparative genomics of adaptation in forest trees. Curr. Opin. Plant Biol. 11: 149–155.
Neale, D. B., and A. Kremer, 2011 Forest tree genomics: growing resources and applications. Nat. Rev. Genet. 12: 111–122.
Neale, D. B., and O. Savolainen, 2004 Association genetics of complex traits in conifers. Trends Plant Sci. 9: 325–330.
Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.
Nickerson, D. A., V. O. Tobe, and S. L. Taylor, 1997 PolyPHRED: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based re-sequencing. Nucleic Acids Res. 25: 2745–2751.
Nielsen, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. 39: 197–218.
Obbard, D. J., J. Welch, K. W. Kim, and F. M. Jiggins, 2009 Quantifying adaptive evolution in the *Drosophila* immune system. PLoS Genet. 5: e1000698.
Orozco, L. D., S. J. Cokus, A. Ghazalpour, L. Ingram-Drake, S. Wang *et al.*, 2009 Copy number variation influences gene expression and metabolic traits in mice. Hum. Mol. Genet. 18: 4118–4129.
Orr, H. A., 1998 The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution 52: 935–949.
Palle, S. R., C. M. Seeve, A. J. Eckert, W. P. Cumbie, B. Goldfarb *et al.*, 2011 Natural variation in expression of genes involved in xylem development in loblolly pine (*Pinus taeda* L.). Tree Genet. Genomes 7: 193–206.
Palle, S. R., C. M. Seeve, A. J. Eckert, J. L. Wegrzyn, D. B. Neale *et al.*, 2013 Association of loblolly pine xylem development gene expression with single nucleotide polymorphisms. Tree Physiol. 33: 763–774.
Palmé, A. E., M. Wright, and O. Savolainen, 2008 Patterns of divergence among conifer ESTs and polymorphism in *Pinus sylvestris* identify putative selective sweeps. Mol. Biol. Evol. 25: 2567–2577.
Park, J.-H., M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung *et al.*, 2010 Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc. Natl. Acad. Sci. USA 108: 18026–18031.
Parks, M., R. Cronn, and A. Liston, 2009 Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 7: 84.
Pavlidis, P., D. Metzler, and W. Stephan, 2012 Selective sweeps in multilocus models of quantitative traits. Genetics 192: 225–239.
Pavy, N., A. Deschenes, S. Blais, P. Lavigne, J. Beaulieu *et al.*, 2013 The landscape of nucleotide polymorphism among 13,500 genes of the conifer *Picea glauca*, relationships with functions and comparison with *Medicago trunculata*. Genome Biol. Evol. 5: 1910–1925.
Pichot, C., and M. El Maataoui, 1997 Flow cytometric evidence for multiple ploidy levels in the endosperm of some gymnosperm species. Theor. Appl. Genet. 94: 865–870.
Pollinger, J. P., C. D. Bustamante, A. Fledel-Alon, S. Schmutz, M. M. Gray *et al.*, 2005 Selective sweep mapping of genes with large phenotypic effects. Genome Res. 15: 1809–1819.
Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. 20: R208–R215.
Pyhäjärvi, T., S. T. Kujala, and O. Savolainen, 2011 Revisiting protein heterozygosity in plants – nucleotide diversity in allozyme coding genes of conifer *Pinus sylvestris*. Tree Genet. Genomes 7: 385–397.
Quesada, T., V. Gopal, W. P. Cumbie, A. J. Eckert, J. L. Wegrzyn *et al.*, 2010 Association mapping of quantitative disease resistance in a natural population of loblolly pine (*Pinus taeda* L.). Genetics 186: 677–686.

R Development Core Team, 2010 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available at: http://www.R-project.org.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98: 11479–11484.

Savolainen, O., and T. Pyhäjärvi, 2007 Genomic diversity in forest trees. Curr. Opin. Plant Biol. 10: 162–167.

Slotte, T., J. P. Foxe, K. M. Hazzouri, and S. I. Wright, 2010 Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol. Biol. Evol. 27: 1813–1821.

Sork, V. L., S. N. Aitken, R. J. Dyer, A. J. Eckert, P. Legendre *et al.*, 2013 Putting the landscape into the genomics of forest trees: approaches for understanding local adaptation and population responses to a changing climate. Tree Genet. Genomes 9: 901–911.

Städler, T., B. Haubold, C. Merino, W. Stephen, and P. Pfaffelhuber, 2009 The impact of sampling schemes on the site-frequency spectrum in nonequilibrium subdivided populations. Genetics 182: 205–216.

Stephan, W., 2010 Genetic hitchhiking *vs.* background selection: the controversy and its implications. Philos. Trans. R. Soc. B 365: 1245–1253.

Stinchcombe, J. R., and H. E. Hoekstra, 2008 Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100: 158–170.

Stoletzki, N., and A. Eyre-Walker, 2011 Estimation of the neutrality index. Mol. Biol. Evol. 28: 63–70.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genome-wide studies. Proc. Natl. Acad. Sci. USA 100: 9440–9445.

Strasburg, J. L., N. C. Kane, A. R. Raduski, A. Bonin, R. Michelmore *et al.*, 2011 Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol. Biol. Evol. 28: 1569–1580.

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

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.

Tajima, F., 1989 Statistical method for testing the neutral 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.

Tian, D., Q. Wang, P. Zhang, H. Araki, S. Yang *et al.*, 2008 Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455: 105–108.

Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. 7: 256–276.

Wegrzyn, J. L., J. M. Lee, J. D. Liechty, and D. B. Neale, 2009 PineSAP - Pine alignment and SNP Identification Pipeline. Bioinformatics 25: 2609–2610.

Welch, J. J., 2006 Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. Genetics 173: 821–837.

White, T. L., W. T. Adams, and D. B. Neale, 2007 *Forest Genetics*. CABI Publishing, Cambridge, MA.

Zeng, K., S. Shi, and C.-I. Wu, 2007 Compound tests for the detection of hitchhiking under positive selection. Mol. Biol. Evol. 24: 1898–1908.

Communicating editor: S. I. Wright

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.157198/-/DC1

The Evolutionary Genetics of the Genes Underlying Phenotypic Associations for Loblolly Pine (*Pinus taeda*, Pinaceae)

Andrew J. Eckert, Jill L. Wegrzyn, John D. Liechty, Jennifer M. Lee, W. Patrick Cumbie, John M. Davis, Barry Goldfarb, Carol A. Loopstra, Sreenath R. Palle, Tania Quesada, Charles H. Langley, and David B. Neale

DOI: 10.1534/genetics.113.157198

File S1
Materials and Methods

The following text represents supplemental information with respect to the **Materials and Methods**. Citations are found at the end of the Supplemental Text in this document. Supplemental figures and tables are found after the Supplemental Text in this document.

Development and application of genetic markers

Construction of EST clusters: An internally developed primer design package (*wt_primer*) was used to design polymerase chain reaction (PCR) primers from the 20,500 unique cDNA cluster consensus sequences. In total, 14,000 primer pairs were successfully designed with the following default conditions: primer length of 16-28 bp, a maximum melting temperature of 66°C, and a maximum difference of 5°C between melting temperatures of forward (F) and reverse (R) primers to maximize the likelihood that amplicons would amplify using standard PCR pipelines. Amplicon size was set to 450 bp, and for longer contigs, multiple overlapping amplicons were designed and the primer pair with the best score was selected for validation. If a predicted contig was less than 450 bp, the largest possible amplicon was chosen for further investigation. The best-scoring oligo pairs were tagged with M13F (GTAAAACGACGGCCAGT) and M13R (CAGGAAACAGCTATGACC) primers for high-throughput sequencing. Primers were validated using loblolly pine DNA from a single tree at a concentration of 2.5 ng/µl. The resulting PCR product was sequenced with M13F and M13R. Sequence quality was assessed using overall PHRED quality score of 20 and signal strength. Passing sequences were compared to the loblolly pine cluster consensus sequences using BLAST to ensure specificity.

Re-sequencing and SNP discovery: Genomic DNA was amplified in 384-well format PCR setup. Each PCR reaction contained 10 ng DNA, 1x HotStar buffer, 0.8 mM dNTPs, 1 mM $MgCl_2$, 0.2U HotStar enzyme (Qiagen) and 0.2 uM F and R primers in a 10 ul total reaction volume. PCR cycling parameters were: one cycle of 95°C for 15 min, 35 cycles of 95°C for 20 s, 60°C for 30 s and 72°C for one min, followed by one cycle of 72°C for three min. The resultant PCR products were purified using solid phase reversible immobilization chemistry followed by dye-terminator fluorescent sequencing with universal M13 primers. Sequencing reactions proceeded as follows: 95°C for 15 min to start followed by 40 cycles of 95°C for 10 s, 50°C for five seconds, 60°C for 2.5 min. Reactions were cleaned using solid phase reversible immobilization (Beckman Coulter Genomics) and the resulting sequencing fragments were detected via capillary electrophoresis using ABI Prism 3730xl DNA analyzers (Applied Biosystems, Foster City CA).

A customized pipeline, PineSAP (Wegrzyn *et al.* 2009), which employs PHRED/PHRAP (Ewing *et al.* 1998, Ewing and Green 1998), CONSED (Gordon *et al.* 1998), POLYBAYES (Marth *et al.* 1999), POLYPHRED (Nickerson *et al.* 1997), and machine

learning tools was used to generate sequence alignments and identify polymorphisms for these data. Custom scripts were added to the PHRED/PHRAP pipeline to identify the downstream amplification primer regions and to quantify the presence of secondary signal in the chromatograms. This allowed for improved screening of sequencing primer sequence and sequence due to mispriming in the PCR product and allowed us to reject chromatograms with high secondary signal that could indicate the presence of signal due to unintentional amplification of a paralogous locus.

Specifically, an integrated automatic and human input pipeline was designed to identify amplicons as putatively paralogous using the ratio of primary to secondary signal at each peak in a chromatogram. One sign that a pair of amplification primers is amplifying more than one site, given that we are working with haploid material, is the presence of secondary peaks in the chromatograms. We generated our assemblies using a single forward and reverse read, whereas programs that are designed to look for the presence of secondary peaks and call heterozygous SNPs often require more than two reads for input. Samples with many low secondary peaks can also indicate the presence of a paralog that may not be amplifying as strongly as the primary sequence. To address both of these issues, we ran PHRED and instructed it to generate a .poly file for each chromatogram (output intended for the POLYPHRED program) that contains information about the signal strength of each of the four fluorophores at the time each base was called. We then use the information in the .poly file to create custom tags that appear in CONSED that indicate the relative intensity of the most prominent secondary signal present. These tags are then visible when reviewing the trace files, and can also be detected automatically. If we look at the ratio of secondary to primary signal, if it was greater than 0.6, it was marked as 'high', if not but greater than 0.47, it was marked as 'medium', and if not but greater than 0.35, it was marked as 'low'. We then rejected any sample that had one or more 'high' sites, more than 2 'medium' sites or more than 6 'low' sites.

SNP genotyping: Total genomic DNA for each sample for genotyping was obtained from either pooled megagametophytes or needle tissue using Qiagen 96-well DNeasy Plant Mini Kits. Arrays were imaged on a Bead Array reader (Illumina) and genotype calling was performed using BeadStudio v. 3.1.3.0 (Illumina). Sample independent controls were assessed on each array to ensure assay integrity. We used a threshold of 55% for the call rate (CR) and 0.15 for the GenCall50 (GC50) scores for inclusion of SNP amplicons in the final dataset (see Eckert *et al.* 2009). More information about these data is available in Eckert *et al.* (2010a).

Generation of linkage maps

We utilized a maximum likelihood method that allows for genotyping errors (Cartwright *et al.* 2007) to create linkage maps comprised of SNPs and restriction fragment length polymorphism (RFLP) framework markers (see Eckert *et al.* 2009b and references therein) following a double pseudo-testcross strategy (Grattapaglia and Sederoff 1994). Each marker was tested for

segregation distortion using a goodness-of-fit test (significance threshold: $P = 0.01$). Once distorted markers were removed from each pedigree ($n = 35$ QTL, $n = 32$ base), phasing was performed for each pedigree separately, markers were grouped using a minimum LOD of 5.0 and a maximum distance of 20 cM, and marker ordering was improved within groups following the methods described by Cartwright *et al.* (2007). The Kosambi mapping function was used subsequently to transform recombination frequencies into map distances (Kosambi 1944). Resulting linkage maps were deposited in the TreeGenes CMAP database (http://dendrome.ucdavis.edu/cmap/) and first appeared in Eckert *et al.* (2010a,b).

Phenotypic trait analysis

Expression levels of lignin and cellulose related genes (expression): Transcript levels of 112 genes putatively involved with lignin and cellulose production were determined for each of the replicated genotypes using quantitative real-time PCR (qRT-PCR). Transcript levels were quantified with duplicate reactions (i.e. technical replicates) carried out on a GeneAmp 7900HT Sequence Detection System (Applied Biosystems, Carlsbad, CA, USA) using SYBER-Green PCR Master Mix (Applied Biosystems). Two ramets per genotype were used as biological replicates, and the final estimate of transcript level per gene for each genotype was the arithmetic average of the biological and technical replicates. All expression levels were standardized to 18S rRNA and β-actin controls. This process quantified transcript levels for 112 genes putatively involved with lignin and cellulose production for 400 of the 498 targeted genotypes (Palle *et al.* 2011).

Primary metabolite concentrations (metabolite): Gas chromatography coupled with time-of-flight mass spectrometry (GC-TOF-MS) was used to determine primary metabolite concentrations from pulverized xylem tissue collected from each replicated genotype established in the NCSU common garden. All analysis was conducted at the University of California Davis Genome Center Metabolomics Core Facility (http://www.metacore.ucdavis.edu/techno1). Two technical and biological replicates were used in these analyses. Resulting GC-TOF-MS data were processed following the methods outlined by Fiehn *et al.* (2008). Mixed linear models were used subsequently to adjust clonal least-square means for evaluation dates and experimental design. This process resulted in concentration estimates for 292 primary metabolites, of which 82 were known, assayed in 297 of the 498 targeted genotypes (Eckert *et al.* 2012).

Drought-tolerance and growth (drought): Estimates of carbon isotope ratios, height and foliar nitrogen content were assayed in each of the replicated genotypes established at the NCSU common garden. Isotope (^{13}C and ^{12}C) and nitrogen content (%N) analyses were based on 3 mg of needle tissue and were carried out at the COIL (http://www.cobsil.com/) stable isotope facility located at Cornell University. Total tree height (cm) was determined at the end of the second growing season. Phenotypic values for each genotype were estimated using mixed linear models that accounted for experimental design and

spatial heterogeneity in the common garden. This process resulted in estimates of $\delta^{13}C$, foliar nitrogen content and height after the second year for 425 of the 498 targeted genotypes (Cumbie *et al.* 2011).

Disease resistance (disease): Lengths of lesions (mm) produced in response to inoculation with pitch canker (*Fusarium circinatum* Nirenberg & O'Donnell) were estimated at four, eight and 12 weeks post inoculation for each of the replicated genotypes established in the NCSU common garden. These estimates were taken as a measure of disease resistance, and phenotypic values for each genotype were estimated as best linear unbiased predictors (BLUPs) using mixed linear models incorporating effects due to experimental design. This process resulted in lesion length estimates at multiple time points for 404 of the targeted 498 genotypes of which the estimates at 12 weeks post inoculation were used in association analyses (Quesada *et al.* 2010).

Identification of phenotypic associations: Phenotypic associations were identified using a two-stage approach where clonal values were predicted and then associated with SNPs using linear models. Most often these were general or mixed linear models (Yu *et al.* 2006) with fixed effects for SNPs as implemented in the program TASSEL (Bradbury *et al.* 2007), although Eckert *et al.* (2012) used the method of Price *et al.* (2006). All analyses included effects due to population structure, as described in Eckert *et al.* (2010a), and were largely based on single locus tests of fixed effects for each SNP. When kinship was included, the kinship matrix was estimated using EMMA (Kang *et al.* 2008) and included as a matrix of random effects. Exceptions to this were the multilocus models used by Quesada *et al.* (2010) and Eckert *et al.* (2012). Multiple testing was accounted for during single locus testing using the positive false discovery rate (FDR) with a threshold of $Q = 0.10$ (Storey and Tibishirani 2003). Additional information about the statistical methodologies used for each phenotypic trait is given in the original publications (disease resistance: Quesada *et al.* 2010; drought-tolerance and growth: Cumbie *et al.* 2011; primary metabolites: Eckert *et al.* 2012) or in the supplemental online materials (gene expression: Palle *et al.* 2013).

Results

The following text represents supplemental information with respect to the **Results**. Citations are found at the end of the Supplemental Text in this document. References to figures and tables are for those in the main text unless noted otherwise. Supplemental figures and tables are found after the Supplemental Text in this document.

Re-sequencing data summary

The number of amplicons passing design thresholds decreased from approximately 7,900 to 7,413 after requiring both F and R reads to be present for each sample followed by a further decrease to 6,669 amplicons after screening for amplification primers in both reads. A total of 5,773 amplicons passed our final quality thresholds, which also included screens for organellar contamination. The average (± 1 standard deviation [sd]) sample size per amplicon was 12 (± 6), with the frequency distribution of sample sizes being skewed towards larger sample sizes (Figure S1). The average sample size also

changed little across different categories of amplicons. Of these 5,773 amplicons, 1,306 could be positioned on the linkage map (22.6%), 2,626 could be annotated to level of coding and noncoding regions (45.5%), 3,484 had a putative ortholog for radiata pine (60.3%) and 950 had a putative ortholog for sugar pine (16.4%). Only a moderate fraction of the total number of amplicons (45.4%) and the number of amplicons that were annotated (44.6%) were represented by at least one SNP on the genotyping chip. Of the total number of amplicons represented on the genotyping chip (*n* = 2,619), 689 (26.3%) unique amplicons had at least one SNP associated to at least one phenotype. The fraction of annotated and unannotated amplicons on the genotyping chip (*n* = 1,173 and 1,446, respectively) with at least one SNP associated to at least one phenotype were similar, with 309 (26.3%) and 380 (26.3%) being associated to at least one phenotype. For the 689 amplicons containing at least one SNP associated to at least one phenotype, 76 were associated to expression, 576 to metabolite, 12 to drought, and nine to disease related phenotypes. The remaining 16 amplicons were associated to environmental variables (see Eckert *et al.* 2010a, 2010b). One hundred and ninety-five (28.3%) out of the 689 amplicons were also associated with more than one phenotype and/or environmental variable (range: 1 to 6).

At the level of individual sites, a total of 2,135,607 aligned sites were analyzed across the 5,773 amplicons. The average (± 1 sd) length of amplicons was 370 (± 126) bp, with longer amplicons being more likely to have at least one SNP genotyped (Table 1) and more likely to be annotated. Of these sites, 1,161,888 could be annotated (54.4%), with 583,159 (50.2%), 160,814 (13.8%), and 417,915 (36.0%) sites being nonsynonymous, synonymous, and noncoding, respectively. The higher percentage of annotated sites relative to annotated amplicons is accounted for by the observation that annotated amplicons were longer than the genome-wide average (432 bp versus 370 bp). This difference, however, was not statistically significant (P_{perm} = 0.348). The same patterns were observed for classes of amplicons (Table 1).

A total of 22,621 SNPs were detected across the 5,773 amplicons. There was little to no effect of sequence quality on the number of SNPs per amplicon, as the correlation between the number of SNPs called for each amplicon using a PHRED threshold of 30 versus a PHRED threshold of 40 was large (Pearson's $r > 0.85$ for all, nonsynonymous, synonymous and noncoding SNPs) and the slope was almost equal to one (Figure S2). Coverage, however, affected several alignment metrics related to nucleotide diversity and divergence (Figures S3-S5; Tables S1-S3). This level of polymorphism was similar to that reported previously for loblolly pine (Brown *et al.* 2004, González-Martínez *et al.* 2006), with one SNP per 94 bp on average. Of these 22,621 SNPs, 10,591 could be annotated as nonsynonymous (*n* = 2,915), synonymous (*n* = 3,233) and noncoding (*n* = 4,443). On a per site basis, SNPs were more common at synonymous sites (one SNP per 50 bp on average) and noncoding (one SNP per 94 bp on average) relative to nonsynonymous (one SNP per 200 bp on average) sites. Patterns were similar across different categories of amplicons (Table 1). At the level of SNPs selected genotyping (*n* = 7,216), there was no enrichment of

certain types of SNPs in the set of associated amplicons (n = 873 SNPs associated to at least one phenotype), so that the numbers of nonsynonymous (n = 127), synonymous (n = 160) and noncoding SNPs (n = 201) associated to at least one phenotype were no different than those expected by randomly subsampling the annotated SNPs on the Illumina genotyping array ($P_{perm} > 0.15$).

Linkage Disequilibrium

Genome-wide patterns: Intragenic linkage disequilibrium, as assessed with Kelly's Z_{nS} (Kelly 1997), was positively correlated with nucleotide diversity (Spearman's $\rho > 0.30$, $P_{perm} < 0.005$), while it was negatively correlated with the number of haplotypes (Spearman's $\rho = -0.427$, $P_{perm} = 0.008$). Breaking the range of observed values for Z_{nS} into high and low categories, the correlation becomes significantly negative between nucleotide diversity and linkage disequilibrium when $Z_{nS} > 0.50$ (Spearman's $\rho < -0.20$, $P_{perm} < 0.05$). Correlations with nucleotide diversity at different categories of sites were approximately 2.5-fold smaller, yet still positive, and non-significant (Spearman's $\rho < 0.05$, $P_{perm} > 0.20$). These correlations, however, changed when considering only amplicons with $Z_{nS} > 0.50$, so that nucleotide diversity at nonsynonymous and noncoding sites was significantly, negatively correlated with linkage disequilibrium when $Z_{nS} > 0.50$ (Spearman's $\rho < -0.25$, $P_{perm} < 0.05$). Correlations of linkage disequilibrium with nucleotide divergence were close to zero and non-significant ($-0.05 <$ Spearman's $\rho < 0.05$, $P_{perm} > 0.40$), even when breaking Z_{nS} into low and high categories ($-0.10 <$ Spearman's $\rho < 0.10$, $P_{perm} > 0.35$).

Comparisons across categories of amplicons: Linkage disequilibrium varied across sets of amplicons defined by whether or not they were located on a linkage map (Mann-Whitney U-test: $P = 0.0371$, $P_{perm} = 0.011$), whether or not they were annotated (Mann-Whitney U-test: $P = 1.486e\text{-}05$, $P_{perm} < 0.001$), and whether or not they were associated with at least one phenotype (Mann-Whitney U-test: $P = 0.0493$, $P_{perm} = 0.026$). On average (± 1 sd), linkage disequilibrium was higher for amplicons that were mapped (Z_{nS}: 0.327 ± 0.276 vs. 0.304 ± 0.292), while it was lower for those that were annotated (Z_{nS}: 0.302 ± 0.287 vs. 0.336 ± 0.288) and for those associated with at least one phenotype (Z_{nS}: 0.282 ± 0.263 vs. 0.313 ± 0.277). Significant differences in the level of linkage disequilibrium were also noted among amplicons grouped into categories based on the types of phenotypes to which they were associated (Kruskal-Wallis rank sum test: $P = 0.032$, $P_{perm} = 0.012$), with amplicons associated with disease phenotypes having the lowest ($Z_{nS} = 0.144$) and amplicons associated with expression phenotypes having the highest ($Z_{nS} = 0.308$) average levels of linkage disequilibrium. In general, correlations between levels of linkage disequilibrium and diversity and divergence estimates within different categories of amplicons were similar to genome-wide patterns.

Literature Cited

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–2635.

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley, and D. B. Neale, 2004 Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc. Natl. Acad. Sci. USA 101: 15255–15260.

Cartwright, D. A., M. Troggio, R. Velasco, and A. Gutin, 2007 Genetic mapping in the presence of genotyping errors. Genetics 176: 2521–2527.

Cumbie, W. P., A. J. Eckert, J. L. Wegrzyn, R. Whetten, D. B. Neale, *et al.*, 2011 Association genetics of carbon isotope discrimination, height, and foliar nitrogen in a natural population of *Pinus taeda* L. Heredity 107: 105–114.

Eckert, A. J., B. Pande, E. S. Ersöz, M. H. Wright, V. K. Rashbrook, *et al.*, 2009b High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). Tree Genet. Genomes 5: 225–234.

Eckert, A. J., J. van Heerwaarden, J. L. Wegrzyn, C. D. Nelson, J. Ross-Ibarra, *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, A. J., A. D. Bower, S. C. González-Martínez, J. L. Wegrzyn, G. Coop, *et al.*, 2010b Back to nature: Ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). Mol. Ecol. 19: 3789–3805.

Eckert, A. J., J. L. Wegrzyn, W. P. Cumbie, B. Goldfarb, D. A. Huber, *et al.*, 2012 Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. New Phytol. 193: 890–902.

Ersöz, E. S., M. H. Wright, S. C. González-Martínez, C. H. Langley, and D. B. Neale, 2010 Evolution of disease response genes in loblolly pine: Insights from candidate genes. PLoS ONE 5: e14234.

Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using PHRED. II. Error probabilities. Genome Res.8: 186–194.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green, 1998 Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. Genome Res.8: 175–185.

Fiehn, O., G. Wohlgemuth, M. Scholz, T. Kind, D. Y. Lee, *et al.*, 2008 Quality control for plant metabolomics: reporting MSI-compliant studies. Plant J. 53: 691–704.

González-Martínez, S. C., E. Ersöz, G. R. Brown, N. C. Wheeler, and D. B. Neale, 2006a DNA sequence variation and selection of tag SNPs at candidate genes for drought-stress response in *Pinus taeda*. Genetics 172: 1915–1926.

Gordon, D., C. Abajian, and P. Green, 1998 Consed: A graphical tool for sequence finishing. Genome Res.8: 195–202.

Grattapaglia, D., and R. Sederoff, 1994 Genetic linkage maps of *Eucalyptus grandis* and *E. urophylla* using a pseudo-testcross mapping strategy and RAPD markers. Genetics 137: 1121–1137.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. 7: 1–44.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, *et al.*, 2008 Efficient control of population structure in model organism association mapping. Genetics 178: 1709–1723.

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

Kosambi, D. D., 1944 The estimation of map values from recombination values. Ann. Eugenic. 12: 172–175.

Marth, G. T., I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, *et al.*, 1999 A general approach to single nucleotide polymorphism discovery. Nat. Genet. 23: 452–456.

Nickerson, D. A., V. O. Tobe, and S. L. Taylor, 1997 PolyPHRED: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based re-sequencing. Nucleic Acids Res. 25: 2745–2751.

Palle, S. R., C. M. Seeve, A. J. Eckert, W. P. Cumbie, B. Goldfarb, *et al.*, 2011 Natural variation in expression of genes involved in xylem development in loblolly pine (*Pinus taeda* L.). Tree Genet. Genomes 7: 193–206.

Palle, S. R., C. M. Seeve, A. J. Eckert, J. L. Wegrzyn, D. B. Neale, and C. A. Loopstra, 2013 Association of loblolly pine xylem development gene expression with single nucleotide polymorphisms. Tree Physiol. 33: 763–774.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909.

Quesada, T., V. Gopal, W. P. Cumbie, A. J. Eckert, J. L. Wegrzyn, *et al.*, 2010 Association mapping of quantitative disease resistance in a natural population of loblolly pine (*Pinus taeda* L.). Genetics 186: 677–686.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt, *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98: 11479–11484.

Stoletzki, N., and A. Eyre-Walker, 2011 Estimation of the neutrality index. Mol. Biol. Evol. 28: 63–70.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genome-wide studies. Proc. Natl. Acad. Sci. USA 100: 9440–9445.

Wegrzyn, J. L., J. M. Lee, J. D. Liechty, and D. B. Neale, 2009 PineSAP - Pine alignment and SNP Identification Pipeline. Bioinformatics 25: 2609–2610.

Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208.

Table S1 Summary statistics across sample coverage classes.

Coverage class[a]	Count	L_{total} (bp)	L_{mean} (bp)	Masked Bases[b]	Masked SNPs[b]	S_{total}	S_{mean}	r	L_{total}/S_{total}	Indels
18	898	295869	329	324	1	3032	3.38	0.359	97.58	137
17	755	258801	343	613	4	2989	3.96	0.395	86.58	147
16	559	195733	350	526	3	2162	3.87	0.297	90.53	105
15	374	135247	362	462	3	1611	4.31	0.401	83.95	77
14	339	125825	371	670	1	1602	4.73	0.283	78.54	84
13	278	105873	381	579	4	1313	4.72	0.339	80.63	69
12	276	105623	383	729	3	1158	4.20	0.304	91.21	58
11	236	88298	374	744	2	1181	5.00	0.203	74.76	47
10	215	83498	388	791	4	967	4.50	0.273	86.34	47
9	186	73161	393	522	4	988	5.31	0.168	74.05	36
8	199	76123	383	470	4	969	4.87	0.268	78.55	45
7	170	65349	384	405	3	943	5.55	0.351	69.30	50
6	169	68561	406	779	6	1048	6.20	0.264	65.42	40
5	177	69670	394	695	9	762	4.31	0.122	91.43	37
4	178	71503	402	575	0	715	4.02	0.178	100.00	35
3	197	82293	418	686	2	704	3.57	0.216	116.89	36
2	250	105399	422	899	0	835	3.34	0.113	126.23	30
1	316	132942	421	753	NA	NA	NA	NA	NA	NA

Abbreviations: bp, base pairs; Indels, insertion-deletion events; *L*, length; *r*, Pearson's correlation coefficient between the number of segregating sites and the length of the amplicon (bp); *S*, segregating sites; SNPs, single nucleotide polymorphisms.

[a]Sample size in the alignment (i.e. the number of sequences).

[b]Masked bases are the number of aligned sites with at least one base masked due to its quality score < 30.

Table S2 Summary by coding versus noncoding regions for each coverage class.

Coverage Class[a]	Count	*L* total	*L* coding	*L* noncoding	Masked *L* coding[b]	Masked *L* noncoding[b]	*S* coding	*S* noncoding	Masked *S* coding[b]	Masked *S* noncoding[b]
18	541	183938	125940	57998	138	76	1045	641	0	0
17	422	153028	100226	52802	262	131	891	674	2	1
16	302	111593	71956	39637	220	74	521	568	0	0
15	218	84201	56940	27261	232	62	474	344	2	0
14	192	74973	45635	29338	293	154	360	498	0	1
13	140	54967	38671	16296	205	30	361	211	1	0
12	160	63364	42051	21313	364	142	353	229	3	0
11	120	48291	30583	17708	334	115	247	222	1	0
10	115	44879	30935	13944	491	54	227	137	1	0
9	88	38335	24047	14288	163	71	254	139	1	0
8	103	41122	23613	17509	132	161	165	158	1	0
7	79	31723	21421	10302	129	101	257	134	1	0
6	80	34431	21666	12765	262	136	283	140	1	1
5	103	42626	27736	14890	254	198	222	169	3	4
4	93	37578	24879	12699	193	89	198	94	0	0
3	95	41177	27362	13815	223	159	208	86	1	0
2	136	56143	36340	19803	277	93	206	95	0	0
1	174	75124	46686	28438	197	139	NA	NA	NA	NA

Abbreviations: *L*, length; *S*, segregating sites.

[a]Sample size in the alignment (i.e. the number of sequences).

[b]Masked bases are the number of aligned sites with at least one base masked due to its quality score < 30.

Table S3 A summary of statistical tests used to assess the effects of coverage variation on basic alignment summaries.

Measure	Statistical Test	Statistical Test Results	Interpretation
Nucleotide diversity (θ_π)	Kruskal-Wallis	χ^2 = 97.63, df=16, *P* = 9.62e-14	Average ranks of diversity vary significantly across coverage classes
Nucleotide divergence (Pira)	Kruskal-Wallis	χ^2 = 95.61, df=16, *P* = 2.29e-14	Average ranks of divergence vary significantly across coverage classes
Nucleotide divergence (Pila)	Kruskal-Wallis	χ^2 = 29.41, df=16, *P* = 0.02134	Average ranks of divergence vary significantly across coverage classes
The number of SNPs	Kruskal-Wallis	χ^2 = 174.31, df=16, *P* < 2.2e-16	Average ranks of SNPs vary significantly across coverage classes
Alignment length (bp)	Kruskal-Wallis	χ^2 = 236.24, df=16, *P* < 2.2e-16	Average ranks of alignment lengths vary significantly across coverage classes
Noncoding sites (bp)	Kruskal-Wallis	χ^2 = 24.24, df=16, *P* = 0.08437	Average ranks of noncoding sites do not vary significantly across coverage classes
Coding sites (bp)	Kruskal-Wallis	χ^2 = 42.08, df=16, *P* = 0.00038	Average ranks of coding sites vary significantly across coverage classes
The proportion of masked bases	Goodness-of-fit	χ^2 = 3559.68, df=17, *P* < 2.2e-16	Too few masked bases with high coverage, too many masked bases with low coverage
Proportion of annotated genes	Goodness-of-fit	χ^2 = 16.64, df=17, *P* = 0.47900	Annotated genes within each coverage class occurred in proportion to overall fraction of genes that were annotated
Indels	Goodness-of-fit	χ^2 = 36.32, df=16, *P* = 0.00415	Too many indels at intermediate coverage classes
OG (Pira)	Goodness-of-fit	χ^2 = 418.84, df=17, *P* < 2.2e-16	Too many genes with Pira outgroup when coverage was high and too few when coverage was low.
OG (Pila)	Goodness-of-fit	χ^2 = 171.03, df=17, *P* < 2.2e-16	Too many genes with Pila outgroup when coverage was high and too few when coverage was low.

Abbreviations: bp, base pairs; Indels, insertion-deletion events; OG, outgroup present (i.e. either a single sequence of *Pinus lambertiana* or *P. radiata* or both is available for the amplicon); Pila, *Pinus lambertiana*; Pira, *Pinus radiata*; SNPs, single nucleotide polymorphisms.

Table S4 Likelihood scores for assessing models of genome-wide nucleotide diversity ($\theta = 4N_e u$) that are constant or variable across loci using the method outlined by Hudson (1990). Estimates of nucleotide diversity are per locus.

Model	log*L*	-2log*L*	*P*
All sites			
Constant θ	-15627.58		
	θ = 1.33		
Variable θ	-8301.572	14652.02	*P*<2.2e-16
(*df* = 5,455)			
NS sites			
Constant θ	-4175.475		
	θ = 0.39		
Variable θ	-1745.141	4860.67	*P*<2.2e-16
(*df* = 2,480)			
SY sites			
Constant θ	-4030.923		
	θ = 0.44		
Variable θ	-2086.158	3889.53	*P*<2.2e-16
(*df* = 2,480)			
NC sites			
Constant θ	-6102.177	8125.56	
	θ = 0.49		
Variable θ	-2039.397		*P*<2.2e-16
(*df* = 2,480)			

Abbreviations: *df*, degrees of freedom; NC, noncoding; NS, nonsynonymous; SY, synonymous.

Table S5 Indels affected levels of nucleotide diversity and divergence. Illustrated are results from Student *t*-tests (*t*) with Welch corrections for unequal variances. *P*-values were determined parametrically (*P*) and non-parametrically (P_{perm})using permutations. The permutation-based tests randomized the data with respect to presence or absence of indels and then used the distribution for the *t*-statistic based on 10,000 randomizations as the null distribution with which to compare to the observed *t*-statistic. Note that parametric *t*-tests were used here because we were interested in comparing means (which the Wilcoxon-rank sum test does not). Use of nonparametric tests gave the same results (data not shown).

Statistic	mean (- indels)	mean (+ indels)	*t*	*df*	*P*	P_{perm}
S	3.03	7.79	-19.34	1200.157	<2.2e-16	<1.0e-04
h_1	1.64	4.09	-13.33	1179.734	<2.2e-16	<1.0e-04
θ_π	0.0028	0.0074	-13.33	1178.909	<2.2e-16	<1.0e-04
D_{xy} (Pira)	0.0073	0.0106	-7.24	875.017	9.537e-13	<1.0e-04
D_{xy} (Pila)	0.0415	0.0515	-4.26	157.005	3.441e-05	0.0008
k	2.82	4.17	-19.85	1447.415	<2.2e-16	<1.0e-04
H_d	0.39	0.62	-24.17	1854.041	<2.2e-16	<1.0e-04
Tajima's *D*	-0.47	-0.36	-3.15	1538.623	0.001660	0.009
n	11.99	12.10	-1.56	1851.381	0.116912	0.225
Noncoding (bp)	203.46	226.93	-2.86	578.456	0.004265	0.047

Abbreviations: bp, base pairs; *df,* degrees of freedom; D_{xy}, nucleotide divergence; H_d, haplotypic diversity; indels, insertion-deletion events; *k*, the number of haplotypes; *n*, sample size; h_1, singletons or the first class of the folded site-frequency spectrum; Pila, *Pinus lambertiana*; Pira, *Pinus radiata*; *S,* segregating sites; θ_π, nucleotide diversity based on the average number of pairwise differences (per site).

Table S6 Nucleotide diversity (per site) across linkage groups of loblolly pine (*Pinus taeda* L.). Values are weighted averages (see Materials and Methods).

	l	*l*	*l*	*n*	*n*	*S*	*S*	*S*	*S*	*S*	θ_π	θ_π	θ_π	θ_π	θ_W	θ_W	θ_W	θ_W
LG	All	*n*>1	Ann	n>1	Ann	All	Ann	NS	SY	NC	All	NS	SY	NC	All	NS	SY	NC
1	72	72	36	13.0	13.6	3.6	3.5	1.4	1.1	1.4	0.00323	0.00167	0.00550	0.00194	0.00320	0.00169	0.00478	0.00192
2	115	111	55	12.5	13.2	5.6	4.4	1.4	1.6	1.8	0.00467	0.00167	0.00792	0.00269	0.00486	0.00192	0.00708	0.00282
3	116	111	55	12.2	12.0	4.4	4.0	0.9	1.7	1.7	0.00383	0.00116	0.00892	0.00271	0.00381	0.00137	0.00855	0.00303
4	94	94	58	12.8	12.7	5.1	4.8	1.1	1.7	2.2	0.00421	0.00180	0.00839	0.00357	0.00418	0.00164	0.00784	0.00358
5	117	113	64	12.7	12.8	5.4	5.2	1.4	1.8	2.4	0.00390	0.00150	0.00843	0.00427	0.00430	0.00171	0.00895	0.00447
6	118	114	59	13.2	13.5	4.8	4.6	1.5	1.4	2.5	0.00394	0.00165	0.00710	0.00246	0.00431	0.00201	0.00771	0.00299
7	113	108	57	12.5	12.9	4.7	4.5	1.3	1.6	2.2	0.00369	0.00127	0.00793	0.00312	0.00390	0.00161	0.00749	0.00345
8	136	131	70	12.7	13.4	4.5	4.5	1.1	1.8	2.4	0.00389	0.00154	0.00865	0.00270	0.00394	0.00148	0.00929	0.00272
9	99	97	54	12.7	12.6	5.3	4.9	1.0	2.0	2.4	0.00399	0.00126	0.00802	0.00259	0.00422	0.00129	0.01039	0.00281
10	114	112	59	13.0	12.6	5.1	5.1	1.5	1.8	2.5	0.00458	0.00205	0.00859	0.00411	0.00473	0.00215	0.00790	0.00400
11	101	99	49	12.6	13.4	5.3	6.1	2.2	2.3	2.9	0.00416	0.00241	0.00910	0.00356	0.00434	0.00259	0.00998	0.00365
12	111	106	57	12.6	13.5	4.7	4.3	1.3	1.8	1.9	0.00414	0.00143	0.00838	0.00283	0.00422	0.00166	0.00847	0.00259

Abbreviations: All, all aligned sites; Ann, annotated; *l*, number of loci; LG, linkage group; *n*, sample size; NC, noncoding; NS, nonsynonymous; *S*, segregating sites; SY, synonymous, θ_π, nucleotide diversity based on the average number of pairwise differences; θ_W, nucleotide diversity based on the number of segregating sites following Watterson (1975).

Table S7 Nucleotide divergence (per site) across linkage groups of loblolly pine (*Pinus taeda* L.). Values are weighted averages (see Materials and Methods).

	l	*l*	*l*	*l*	*l*	*l*	*l*	*n*	*n*	*n*	*n*	*n*	*n*	D_{xy}_pira	D_{xy}_pira	D_{xy}_pira	D_{xy}_pira	D_{xy}_pila	D_{xy}_pila	D_{xy}_pila	D_{xy}_pila
LG	All	OG	Ann	Pira	Ann	Pila	Ann	OG	Ann	Pira	Ann	Pila	Ann	All	NS	SY	NC	All	NS	SY	NC
1	72	56	28	55	27	13	11	14.2	14.3	14.2	14.3	13.0	13.0	0.00907	0.00466	0.01236	0.01106	0.04078	0.02505	0.07176	0.04304
2	115	75	42	71	40	22	17	13.7	14.0	13.8	14.2	13.6	13.8	0.00912	0.00285	0.01564	0.00893	0.05205	0.03017	0.11096	0.05828
3	116	76	39	74	38	18	12	13.6	13.5	13.6	13.4	14.5	14.0	0.00746	0.00289	0.01396	0.00524	0.03902	0.01590	0.09227	0.06016
4	94	64	37	62	36	22	14	13.9	14.2	14.0	14.5	15.1	14.9	0.00760	0.00336	0.01064	0.00802	0.03617	0.01130	0.06300	0.04315
5	117	76	43	72	39	17	13	13.8	13.7	14.1	14.3	13.2	12.0	0.00748	0.00305	0.01612	0.01107	0.04278	0.01286	0.07615	0.06620
6	118	75	40	70	36	15	10	14.6	14.6	14.5	14.4	16.2	16.5	0.00682	0.00305	0.01164	0.00474	0.03951	0.03065	0.11051	0.04546
7	113	62	34	61	33	17	12	14.0	14.2	14.0	14.2	15.4	14.7	0.00888	0.00386	0.01302	0.00846	0.05446	0.02924	0.10919	0.05925
8	136	96	51	89	47	31	21	13.7	14.6	13.8	14.5	13.7	15.3	0.00802	0.00318	0.01093	0.00881	0.04171	0.01661	0.07830	0.04353
9	99	72	41	70	40	14	9	13.8	13.6	13.8	13.7	11.7	10.6	0.00657	0.00187	0.01001	0.00624	0.04235	0.01622	0.05880	0.03843
10	114	81	47	78	45	17	13	13.6	13.4	13.8	13.8	12.2	10.8	0.00849	0.00490	0.02144	0.00941	0.04195	0.01747	0.10651	0.05827
11	101	65	35	63	34	11	7	14.0	14.5	14.1	14.4	14.5	15.6	0.00727	0.00268	0.00901	0.00713	0.03943	0.02152	0.06913	0.03439
12	111	72	37	71	36	15	10	13.5	14.3	13.5	14.2	12.5	14.5	0.00643	0.00229	0.00846	0.00724	0.04396	0.01373	0.07085	0.05323

Abbreviations: All, all aligned sites; Ann, annotated; D_{xy}, nucleotide divergence; *l*, number of loci; LG, linkage group; *n*, sample size; NC, noncoding; NS, nonsynonymous; OG, outgroup present (i.e. either a single sequence of *Pinus lambertiana* or *P. radiata* or both is available for the amplicon); Pila, *Pinus lambertiana*; Pira, *Pinus radiata*; SY, synonymous.

Table S8 Estimates for per site crossing over rate ($C = 4N_e r$) and additional summary statistics related to linkage disequilibrium for each sample coverage class where $n > 10$. Values in parentheses are 95% confidence intervals based on bootstrapping across loci (n = 10,000 replicates). Singletons were included.

Coverage	Loci	C	LD-half (bp)[a]	Z_{nS}	C/θ_π
18	898	0.026	102	0.266	9.107
		(0.017 – 0.042)	(63 – 154)	(0.246 – 0.289)	(7.161 – 17.585)
17	755	0.023	117	0.270	8.129
		(0.017 – 0.032)	(85 – 159)	(0.248 – 0.295)	(5.957 – 11.376)
16	559	0.007	386	0.271	2.317
		(0.001 – 0.025)	(110 – 1390)	(0.243 – 0.299)	(0.489 – 8.911)
15	374	0.021	137	0.286	5.740
		(0.011 – 0.041)	(69 – 255)	(0.253 – 0.319)	(3.577 – 13.717)
14	339	0.010	288	0.314	2.871
		(0.004 – 0.020)	(144 – 701)	(0.282 – 0.349)	(1.159 – 5.771)
13	278	0.008	360	0.308	2.564
		(0.001 – 0.031)	(95 – 1392)	(0.271 – 0.347)	(0.473 – 9.252)
12	276	0.006	496	0.327	1.426
		(0.001 – 0.015)	(193 - 1405)	(0.285 – 0.371)	(0.517 – 5.033)
11	236	0.012	274	0.361	3.039
		(0.004 – 0.028)	(111 – 807)	(0.319 – 0.403)	(0.910 – 6.856)

Abbreviations: bp, base pairs; LD, linkage disequilibrium; θ_π, nucleotide diversity from the average number of pairwise differences; Z_{nS}, Kelly's statistic representing the average pairwise LD among SNPs within an amplicon.
[a]The distance in bp where the expected value of allelic correlations (r^2) dropped to half its initial value.

Table S9 Fit of the SNM and Ersöz *et al.* (2010) model to all and the trimmed data. Note that loci with less than four alleles and less than two SNPs were excluded from both analyses. Means and variances are weighted by the sample coverage class in each case.

		All (l = 3,360)			Trimmed[a] (l = 3,133)	
Statistic	Obs	*P* (SNM)	*P* (TEM)	Obs	*P* (SNM)	*P* (TEM)
Mean						
θ_π	0.0045	0.995	0.359	0.0047	0.997	0.415
D	-0.487	< 0.001	0.077	-0.467	< 0.001	0.092
Z_{nS}	0.305	> 0.999	0.098	0.298	> 0.999	0.115
Variance						
θ_π	2.28e-05	0.887	0.087	2.26e-05	0.874	0.068
D	0.918	0.997	0.003	0.912	0.001	0.001
Z_{nS}	0.080	> 0.999	0.104	0.079	0.089	0.089

Abbreviations: *D*, Tajima's *D*; *l*, number of loci or amplicons; Obs, observed value; SNM, standard neutral model; θ_π, nucleotide diversity from the average number of pairwise differences; TEM, three epoch model from Ersöz *et al.* (2010); Z_{nS}, Kelly's statistic representing the average pairwise linkage disequilibrium (LD) among SNPs within an amplicon.

[a]Trimmed data refer to data where samples west of the Mississippi River were excluded. This caused some amplicons to be dropped. The reason for excluding these samples is that the model of Ersöz *et al.* (2010) was fit to data derived from samples exclusively collected from the eastern portion of the range of loblolly pine.

Table S10 List of joint outliers with respect to Tajima's *D* and Fay and Wu's *H* that were annotated with respect to putative gene products.

Amplicon	Putative gene product	Tajima's *D*	Fay and Wu's *H*
0_10631_01	HSP7NAT-2 (HEAT-SHOCK PROTEIN 7NAT-2); ATP binding	-1.84	-5.18
0_10631_02	HSP7NAT-2 (HEAT-SHOCK PROTEIN 7NAT-2); ATP binding	-1.71	-3.42
0_12117_01	universal stress protein (USP) family protein	-1.70	-5.25
0_3461_01	DIN1NA (DARK INDUCIBLE 1NA); hydrolase, hydrolyzing O-glycosyl compounds	-2.14	-8.18
0_8408_01	glyoxal oxidase-related	-1.82	-3.22
0_8694_01	sodium:solute symporter family protein	-2.08	-8.60
0_9825_01	DIR1 (DEFECTIVE IN INDUCED RESISTANCE 1); lipid binding / lipid transporter	-1.85	-5.23
2_4925_01	zinc finger (C3HC4-type RING finger) family protein	-1.72	-6.87
2_6183_01	CRK1NA (CYSTEINE-RICH RLK1NA); ATP binding / kinase/ protein kinase/ protein serine/threonine kinase/ protein tyrosine kinase	-2.08	-4.85
2_9466_01	membrane-associated zinc metalloprotease, putative	-1.85	-3.34
CL1344Contig1_03	PFK2 (PHOSPHOFRUCTOKINASE 2); 6-phosphofructokinase	-1.71	-3.45
CL162Contig1_01	pectinesterase family protein	-1.82	-5.08
CL2463Contig1_03	TMKL1 (transmembrane kinase-like 1); ATP binding / kinase/ protein serine/threonine kinase	-1.69	-3.34
CL4663Contig1_02	FTSZ1-1; protein binding / structural molecule	-1.85	-3.34
UMN_3361_01	DNA-binding protein, putative	-1.82	-5.08
UMN_5367_02	chaperonin, putative	-1.69	-5.20

Table S11 Summary of model fitting in a McDonald-Kreitman framework for all amplicons that were annotated and had *Pinus radiata* as an outgroup (*l* = 1,623).

Model	*k*	ln*L*	AICc
θ = constant, *ut* = constant, *f* = 0, α = 0	2	-12064.81	24133.62
θ = constant, *ut* = constant, *f* = constant, α = 0	3	-10067.49	20140.99
θ = constant, *ut* = constant, *f* = unique, α = 0	1625	-7548.09	19432.19
θ = unique, *ut* = constant, *f* = 0, α = 0	1624	-9320.81	22974.07
θ = unique, *ut* = constant, *f* = constant, α = 0	1625	-7330.06	18996.13
θ = unique, *ut* = constant, *f* = unique, α = 0	3247	-5725.23	24446.46
θ = constant, *ut* = constant, *f* = 0, α = constant	3	-11203.01	22412.02
θ = constant, *ut* = constant, *f* = constant, α = constant	4	-10061.02	20130.04
θ = constant, *ut* = constant, *f* = unique, α = constant	1626	-7522.02	19383.60
θ = unique, *ut* = constant, *f* = 0, α = constant	1625	-8668.88	21673.77
θ = unique, *ut* = constant, *f* = constant, α = constant	1626	-7298.35	18936.13
θ = unique, *ut* = constant, *f* = unique, α = constant	3248	-5716.25	24436.52
θ = constant, *ut* = constant, *f* = 0, α = beta	4	-11949.69	23907.38
θ = constant, *ut* = constant, *f* = constant, α = beta	5	-9834.78	19679.57
θ = constant, *ut* = constant, *f* = unique, α = beta	1627	-7510.3	19363.73
θ = unique, *ut* = constant, *f* = 0, α = beta	1626	-9309.12	22957.81
θ = unique, *ut* = constant, *f* = constant, α = beta	1627	-7270.93	18884.98
θ = unique, *ut* = constant, *f* = unique, α = beta	3249	-5714.08	24440.19
θ = constant, *ut* = constant, *f* = 0, α = two-spike	5	-11006.98	22023.97
θ = constant, *ut* = constant, *f* = constant, α = two-spike	6	-9869.41	19750.84
θ = constant, *ut* = constant, *f* = unique, α = two-spike	1628	-7494.66	19336.01
θ = unique, *ut* = constant, *f* = 0, α = two-spike	1627	-8591.69	21526.51
θ = unique, *ut* = constant, *f* = constant, α = two-spike	1628	-7274.38	18895.45
θ = unique, *ut* = constant, *f* = unique, α = two-spike	3250	-5705.77	24431.60
θ = constant, *ut* = unique, *f* = 0, α = 0	1624	-10830.79	25994.03
θ = constant, *ut* = unique, *f* = constant, α = 0	1625	-8836.12	22008.24
θ = constant, *ut* = unique, *f* = unique, α = 0	3247	-6828.13	26652.26
θ = unique, *ut* = unique, *f* = 0, α = 0	3246	-8541.03	30070.05
θ = unique, *ut* = unique, *f* = constant, α = 0	3247	-6551.34	26098.68
θ = unique, *ut* = unique, *f* = unique, α = 0	4869	-5098.53	49173.07
θ = constant, *ut* = unique, *f* = 0, α = constant	1625	-9865.29	24066.59
θ = constant, *ut* = unique, *f* = constant, α = constant	1626	-8836.08	22011.72
θ = constant, *ut* = unique, *f* = unique, α = constant	3248	-6820.54	26645.10
θ = unique, *ut* = unique, *f* = 0, α = constant	3247	-7662.37	28320.75
θ = unique, *ut* = unique, *f* = constant, α = constant	3248	-6548.89	26101.79
θ = constant, *ut* = unique, *f* = unique, α = constant	4870	-5093.23	49194.52

θ = constant, *ut* = unique, *f* = 0, α = beta	1626	-10830.79	26001.15
θ = constant, *ut* = unique, *f* = constant, α = beta	1627	-8835.43	22013.99
θ = constant, *ut* = unique, *f* = unique, α = beta	3249	-6820.56	26653.16
θ = unique, *ut* = unique, *f* = 0, α = beta	3248	-8541.03	30086.07
θ = unique, *ut* = unique, *f* = constant, α = beta	3249	-6550.46	26112.96
θ = unique, *ut* = unique, *f* = unique, α = beta	4871	-5093.25	49226.66
θ = constant, *ut* = unique, *f* = 0, α = two-spiked	1627	-9863.55	24070.22
θ = constant, *ut* = unique, *f* = constant, α = two-spike	1628	-8833.74	22014.17
θ = constant, *ut* = unique, *f* = unique, α = two-spike	3250	-6820.54	26661.14
θ = unique, *ut* = unique, *f* = 0, α = two-spike	3249	-7660.96	28333.96
θ = unique, *ut* = unique, *f* = constant, α = two-spike	3250	-6547.15	26114.35
θ = unique, *ut* = unique, *f* = unique, α = two-spike	4872	-5093.23	49258.75

Abbreviations: AICc, corrected Akaike Information Criterion; α, fraction new mutations driven to fixation by positive selection; beta, the Beta distribution; constant, a constant value of a parameter across all amplicons; *f*, the fraction of amplicons not under strong purifying selection; *k*, number of model parameters; *l*, number of loci or amplicons; ln*L*, log-likelihood; θ, expected neutral diversity; two-spike, two-spiked multimodal distribution; unique, unique value of a parameter for each amplicon; *ut*, expected neutral divergence.

Table S12 Functional categories of amplicons and signatures of selection. These are the raw data used in Figure 4 to which loess smoothing was applied. Values for the Direction of Selection statistic (DoS) and Tajima's *D* are weighted averages where the weights are the sample size.

Functional category	Assoc	Unassoc	Total	DoS	*D* (NS)	*D* (SY)
zinc.finger.proteins	4	10	29	-0.189	-0.524	-0.838
isomerase.activity.topoisomerase.epimerase.isomerases.	3	6	19	-0.134	-0.931	-0.587
ATPases	4	6	18	-0.125	-0.495	0.009
vitamin.binding.Vitamin.B6.anthocyanidin.flavin.dependent.beta.carotene.	5	13	34	-0.109	-0.366	-0.509
calmodulin.binding.proteins.Calmodulin.	6	8	34	-0.101	-0.063	0.037
ion.channel.glutamate.gated.ion.channel.KAB.potassium.channels.CLC.	2	8	15	-0.094	-0.417	-0.981
vesicle.mediated.transport.VAMP.VPS.exosomes.coatomer.NECAP.exocysts.	2	12	26	-0.082	-0.708	-1.032
Glycosidase.chitinase.glycosylase.glycosidase.glucanase.	7	16	50	-0.079	-0.634	-0.394
signal.transducers.tranducins.protein.kinases.	4	17	38	-0.071	-0.697	-0.362
pigment.binding.light.receptors.chlorophyll.	0	9	14	-0.045	-0.508	-1.183
structural.constituent.of.ribosome.ribosomal.subunits.	5	23	43	-0.029	-0.363	-0.499
structural.constituent.of.cell.wall.actin.tubulin.extensin.expansin. arabinogalactans.	5	24	56	-0.017	-0.384	-0.191
stress.response.USP.	3	11	26	-0.015	-0.339	-0.495
hormone.binding.auxin.receptors.ethylene.receptors.brassinosteroid.receptors.	11	24	72	-0.004	-0.341	-0.228
Protease.Peptidase.serine.threonine.kinases.endopeptidases.aspartyl.metalloproteases.TMK.cysteine.proteases.	18	65	152	-0.003	-0.602	-0.701
peroxidases.cationic.peroxidase.Haem.peroxidase.	10	14	38	0.004	-0.456	-0.626
ubiquitin.ligase.PUB.f.box.ubiquitin.protein.ligase.	5	25	65	0.008	-0.647	-0.647
hydrolases.HAD.hydrolases.	19	28	90	0.010	-0.450	-0.551
transporters.ABC.transporter.OPT.POT.nodulin.amino.acid.transporter.MATE.MDR.hexose.transporter.permease.	22	51	136	0.012	-0.269	-0.455
pectin.esterases.pectinesterase.	1	10	17	0.014	-0.690	-0.817
oxidoreductases.cytochrome.P450.cytochrome.c.catalases.dehydrogenases. reductases.	27	70	219	0.027	-0.509	-0.291
disease.resistance.NBS.	3	19	45	0.027	-0.214	-0.464
GTPase.GTP.binding.RAB.GTPase.RAS.GTPase.RAN.GTPase.	3	11	35	0.033	-0.315	-0.236

transcription.factors.Myb.Myc.GRAS.WRKY.bZIP.ARR.	19	52	132	0.041	-0.536	-0.324
chromatin.binding.RCC1.chromosome.condensation.complex.chromatin.	5	17	35	0.043	-0.317	0.519
remodeling.histone.proteins.						
RNA.polymerase.RDR.RNA.polymerase.	2	9	14	0.043	0.027	0.037
lyase.activity.dehydratase.pectate.lyase.carbon.sulfer.lyase.	8	25	70	0.043	-0.388	-0.583
transferases.PFK.glucuronosyltransferases.SEC.	22	80	219	0.045	-0.410	-0.300
nucleic.acid.nucleotide.binding.Anth.retinoblastoma.argonaute.BLHL.	16	48	135	0.060	-0.378	-0.084
VARICOSE.SWAP.DNA.Polymerase.						
lipid.binding.lipase.phospholipase.EXL.clathrin.associated.complex.	6	17	50	0.062	-0.616	-0.370
translation.Initiation.elongation.Factors.EIFG.elongation.factors.	3	6	19	0.063	-0.512	-0.416
water.channel.aquaporins.MIP.TIP.HOS.	0	7	9	0.081	-0.412	0.125
electron.transporter.photosystems.cytochrome.b6.photo.assimilate.	6	6	16	0.104	-0.350	-1.164
heat.shock.HSP.DnaJ.	7	33	71	0.112	-0.574	-0.357
carbohydrate.binding.sucrase.glyoxyl.oxidase.VTC.INT.lectin.protein.kinase.	5	16	32	0.140	-0.665	-0.831
carbohydrate.protein.kinase.						
phosphatase.regulator.activity.phosphatase.2.pho1.phosphatases.NIF.	2	9	15	0.140	-0.986	-1.019
ligase.activity.synthetases.ligases.	0	5	20	0.212	-0.566	0.084
metal.cluster.binding.embryo.defective.proteins.germins.Rieske.ALS. ferredoxins.	5	25	50	0.230	-0.623	-0.258
microtubule.motor.proteins.kinesin.microtubule.	3	6	12	0.374	-0.790	-0.713

Abbreviations: Assoc, associated to at least one phenotype; *D*, Tajima's *D*; DoS, Direction of Selection statistic (Stoletzki and Eyre-Walker 2011); NS, nonsynonymous; SY, synonymous; Unassoc, unassociated to at least one phenotype.

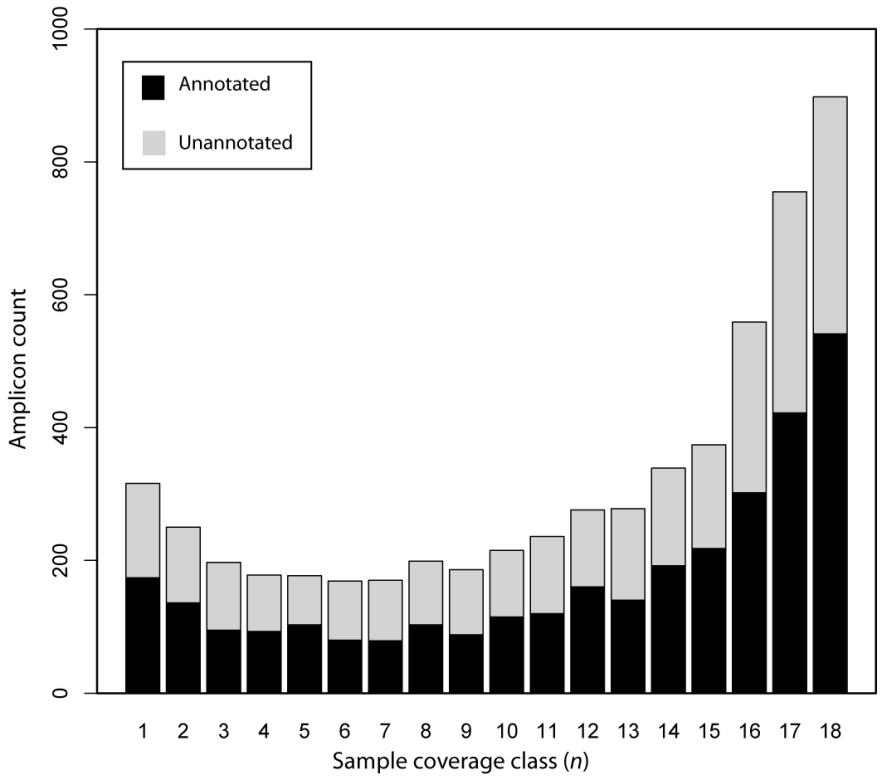

Figure S1 The frequency distribution of sample sizes across all 5,773 amplicons reveals that the majority of amplicons have 10 or more samples. Colors distinguish amplicons for which coding and noncoding regions could be annotated (black) from those that could not (gray).

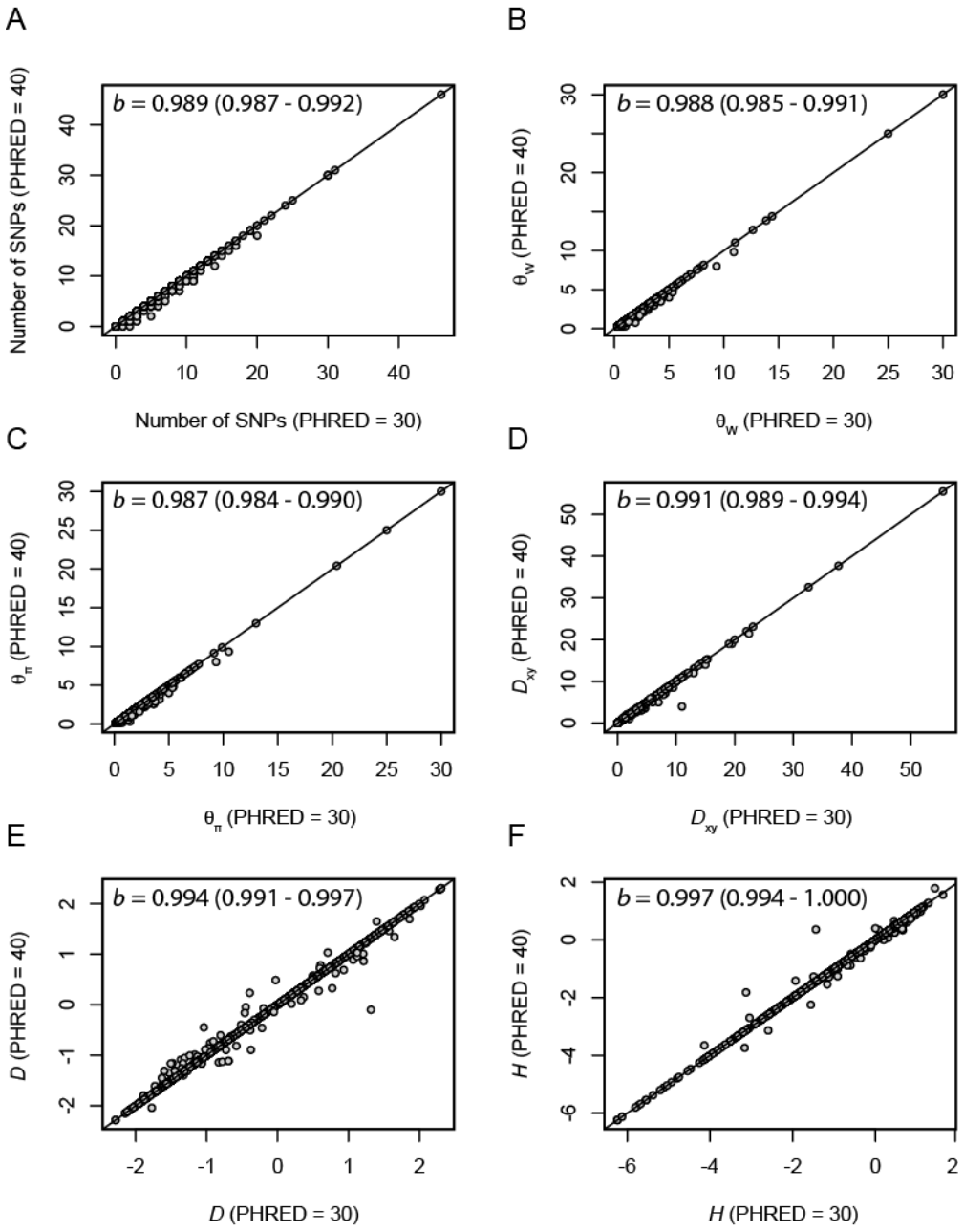

Figure S2 The correlation between various estimates of nucleotide diversity and divergence for two different cutoffs of base calling quality (PHRED 30 and PHRED 40). Summary statistics (*b* = slope (95% confidence interval)) of linear models are given in the upper left of each plot.

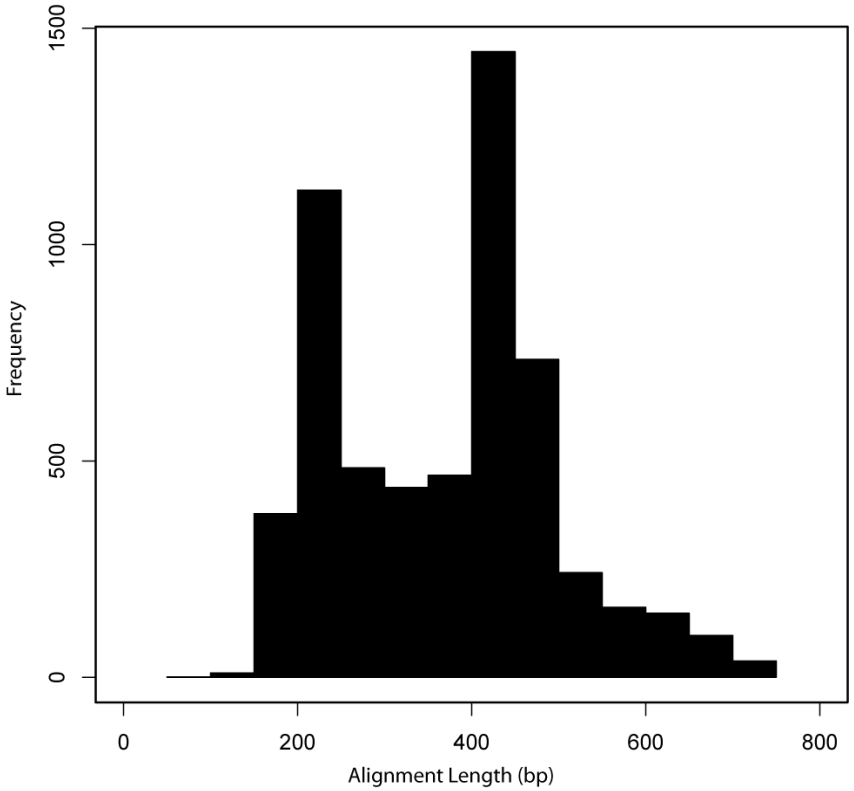

Figure S3 The distribution of the number of aligned sites across amplicons is strongly bimodal.

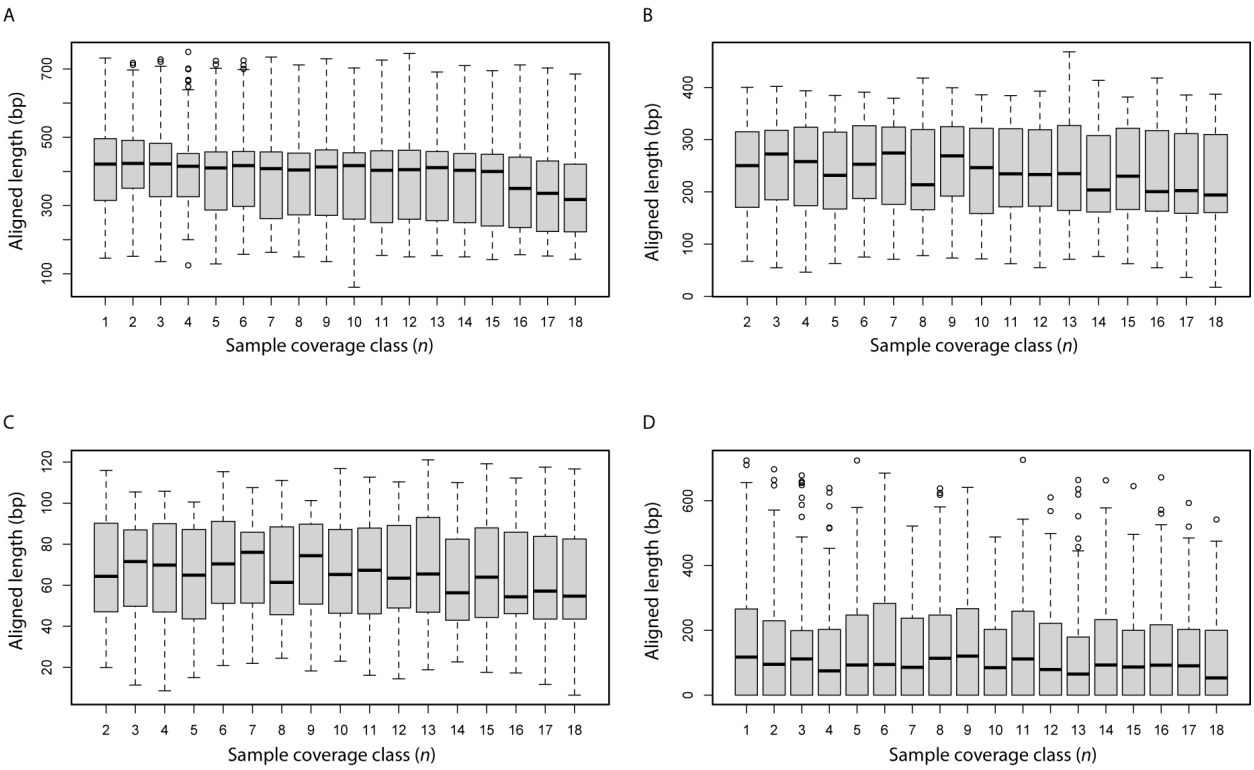

Figure S4 Alignment length (bp) varied significantly among sample coverage classes for all (A; Kruskal-Wallis test: $\chi^2 = 274.87$, *df* = 17, *P* < 2.2e-16), nonsynonymous (B; Kruskal-Wallis test: $\chi^2 = 41.71$, *df* = 16, *P* = 0.00044), and synonymous (C; Kruskal-Wallis test: $\chi^2 = 40.09$, *df* = 16, *P* = 0.00076) sites. Alignment length for noncoding sites (D; Kruskal-Wallis test: $\chi^2 = 27.42$, *df* = 17, *P* = 0.05221), however, did not differ significantly among coverage classes. Whiskers extend to 1.5 times the interquartile range. Note that coverage class one does not contain amplicons with coding regions that were annotated.

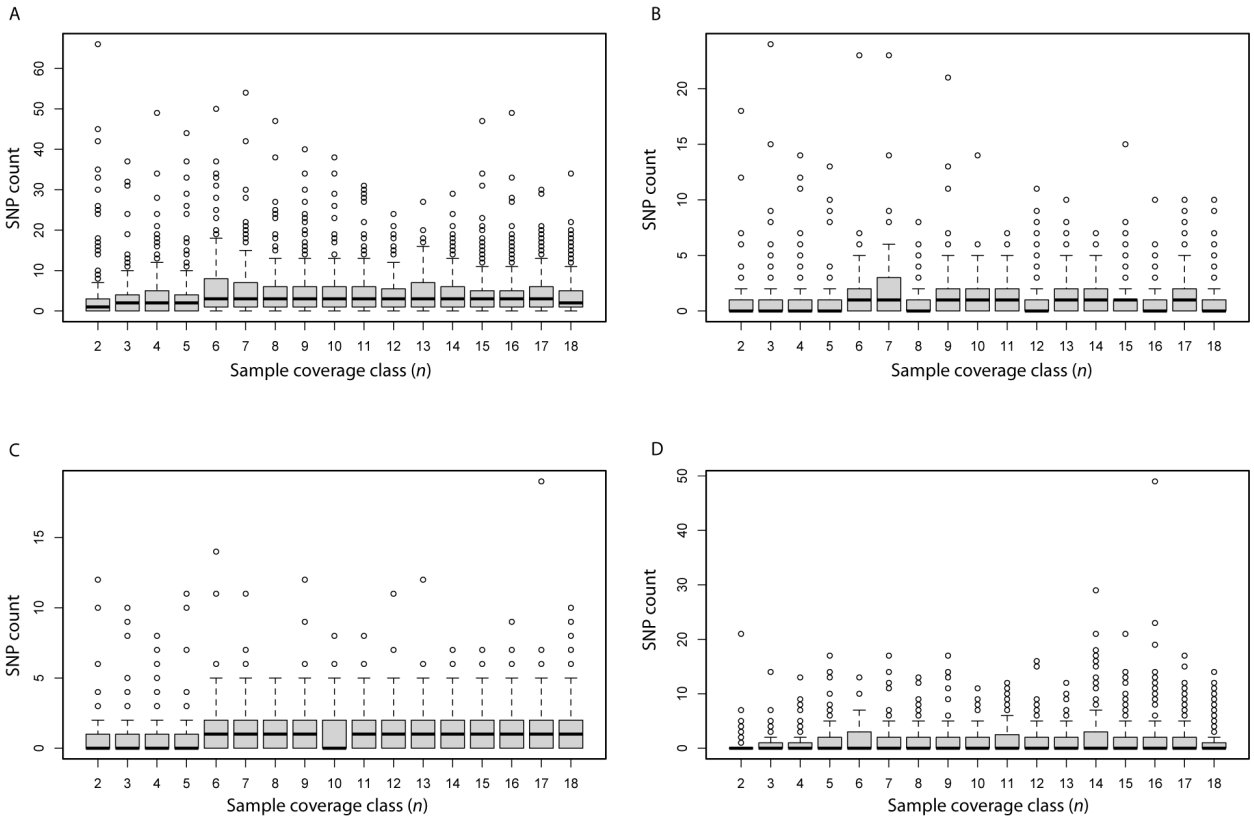

Figure S5 The number of SNPs varied significantly among sample coverage classes for all (Kruskal-Wallis test: χ^2 = 174.31, *df* = 16, *P* < 2.2e-16), nonsynonymous (Kruskal-Wallis test: χ^2 = 174.31, *df* = 16, *P* = 0.00035), synonymous (Kruskal-Wallis test: χ^2 = 72.05, *df* = 16, *P* = 4.3e-09) and noncoding (Kruskal-Wallis test: χ^2 = 48.71, *df* = 16, *P* = 0.00004) sites. Counts of SNPs included those that were tri- and tetra-allelic, as well as those associated with masked bases or indels. Note that the sample coverage class with one allele has been omitted. Retaining only biallelic SNPs did not change these results (data not shown). Whiskers extend to 1.5 times the interquartile range.

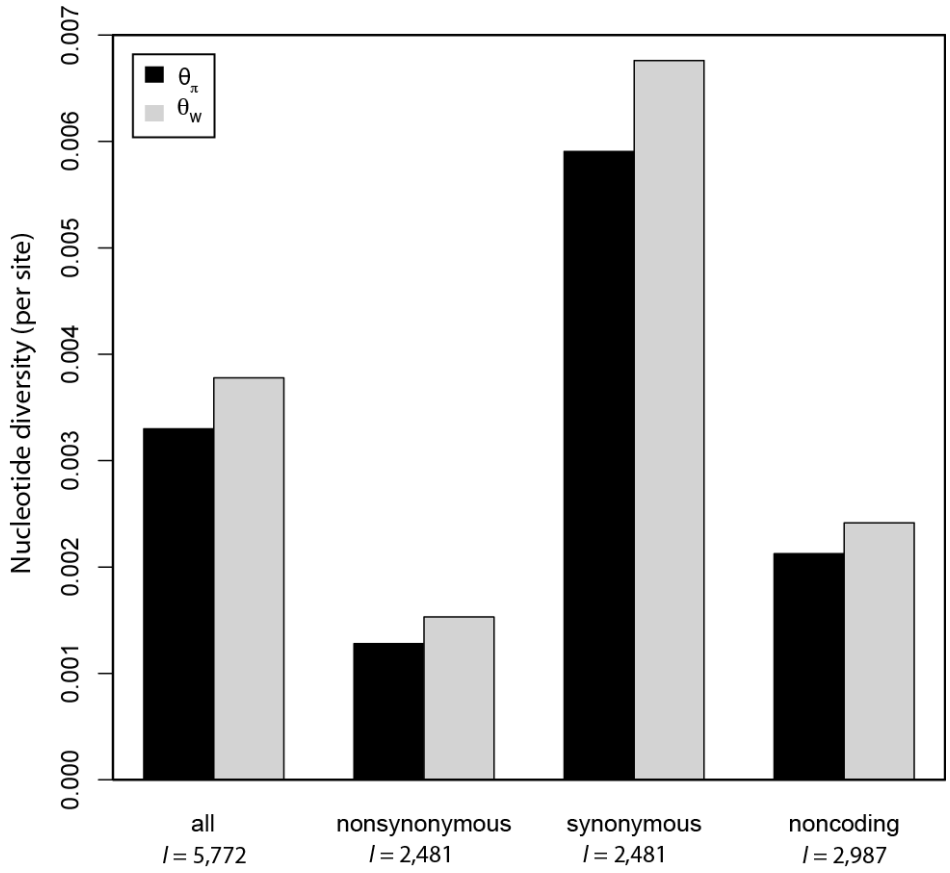

Figure S6 Average nucleotide diversity for all and annotated amplicons (*l* = number of loci or amplicons). Averages are weighted averages using coverage classes as the weights.

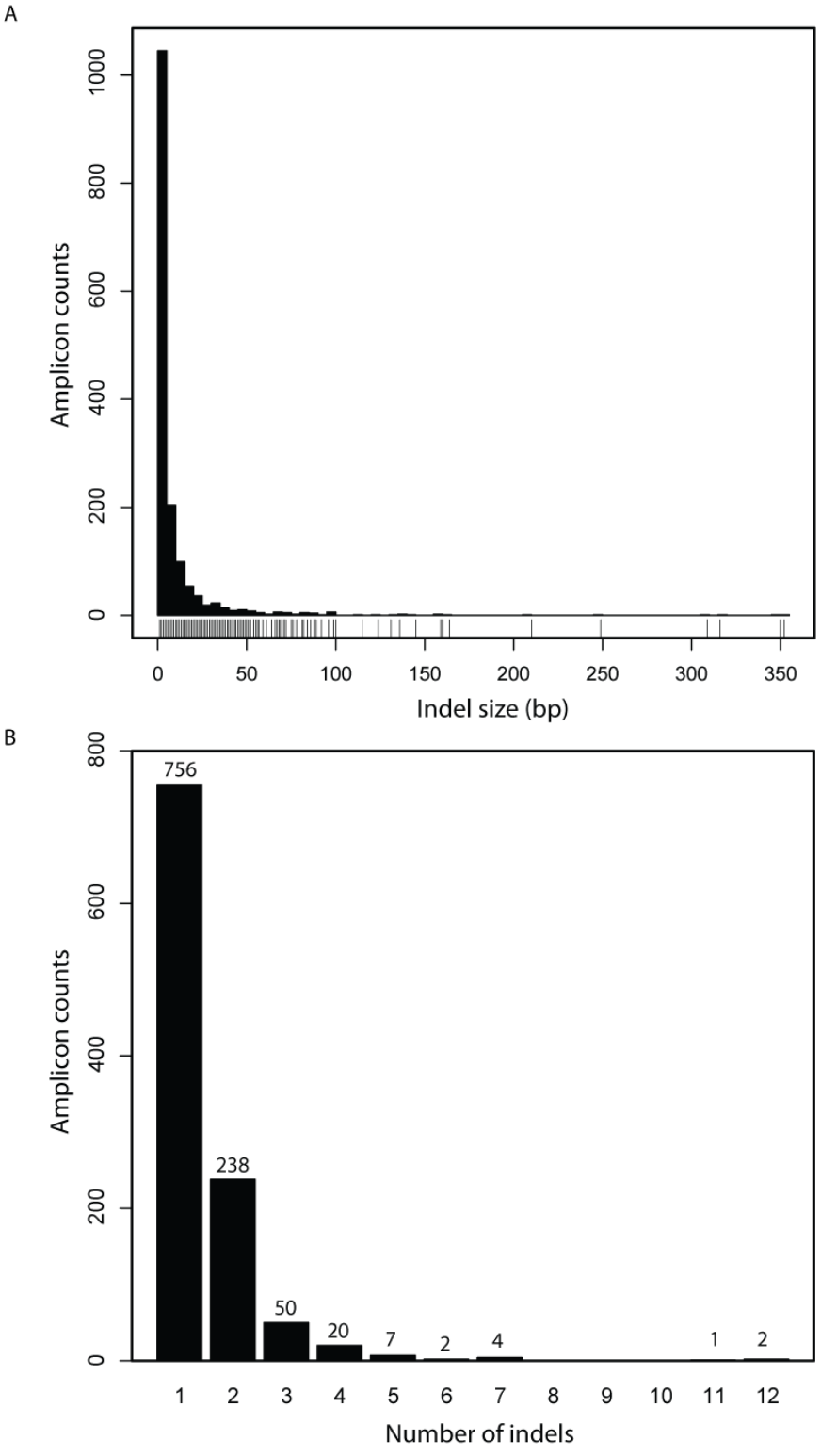

Figure S7 Frequency distribution for indel size (A) and the number of indels per amplicon (B). The rug plot in panel A identifies bins in the histogram that are difficult to differentiate.

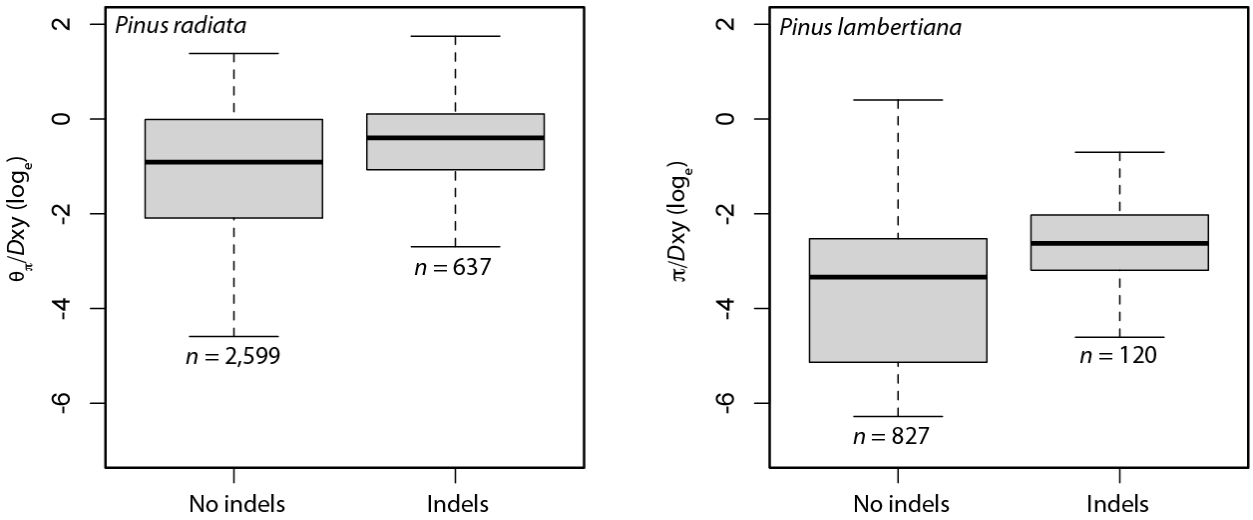

Figure S8 Nucleotide diversity scaled by divergence differs between classes of amplicons defined based on the presence of at least one indel. The patterns are the same for divergence relative to *Pinus radiata* or *P. lambertiana*. Note that values are on a log-scale (base *e*). Sample sizes for the number of amplicons in each category are below the lower whisker for each box. Whiskers extend to the data extremes.

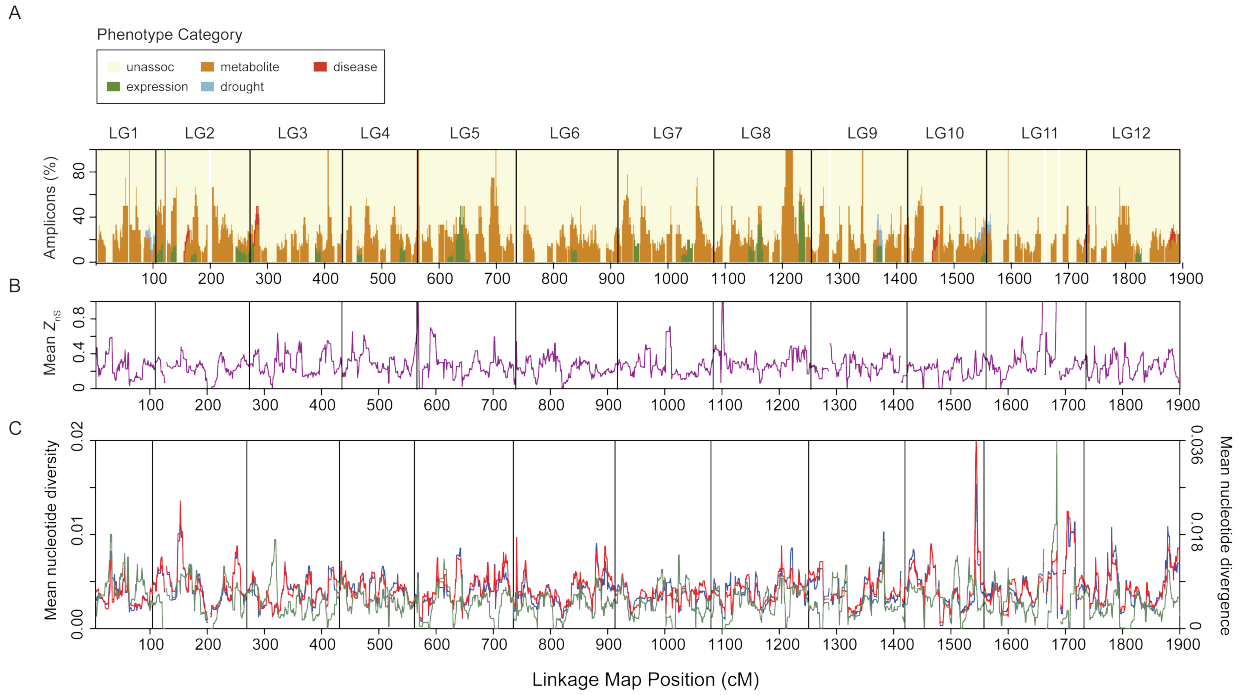

Figure S9 Distributions of gene categories, linkage disequilibrium, nucleotide diversity and nucleotide divergence across the consensus linkage map of loblolly pine based on sliding windows (5 cM size in steps of 2 cM). (A) Stacked bar plot of gene categories across the consensus linkage map. (B) Intragenic linkage disequilibrium, as assessed using Kelly's Z_{nS} statistic, across the consensus linkage map. (C) Nucleotide diversity (θ_π = red, θ_W = blue) and nucleotide divergence with respect to *P. radiata* (green) across the consensus linkage map. The number of amplicons where nucleotide divergence relative to *P. lambertiana* was defined was too small to plot across linkage groups.

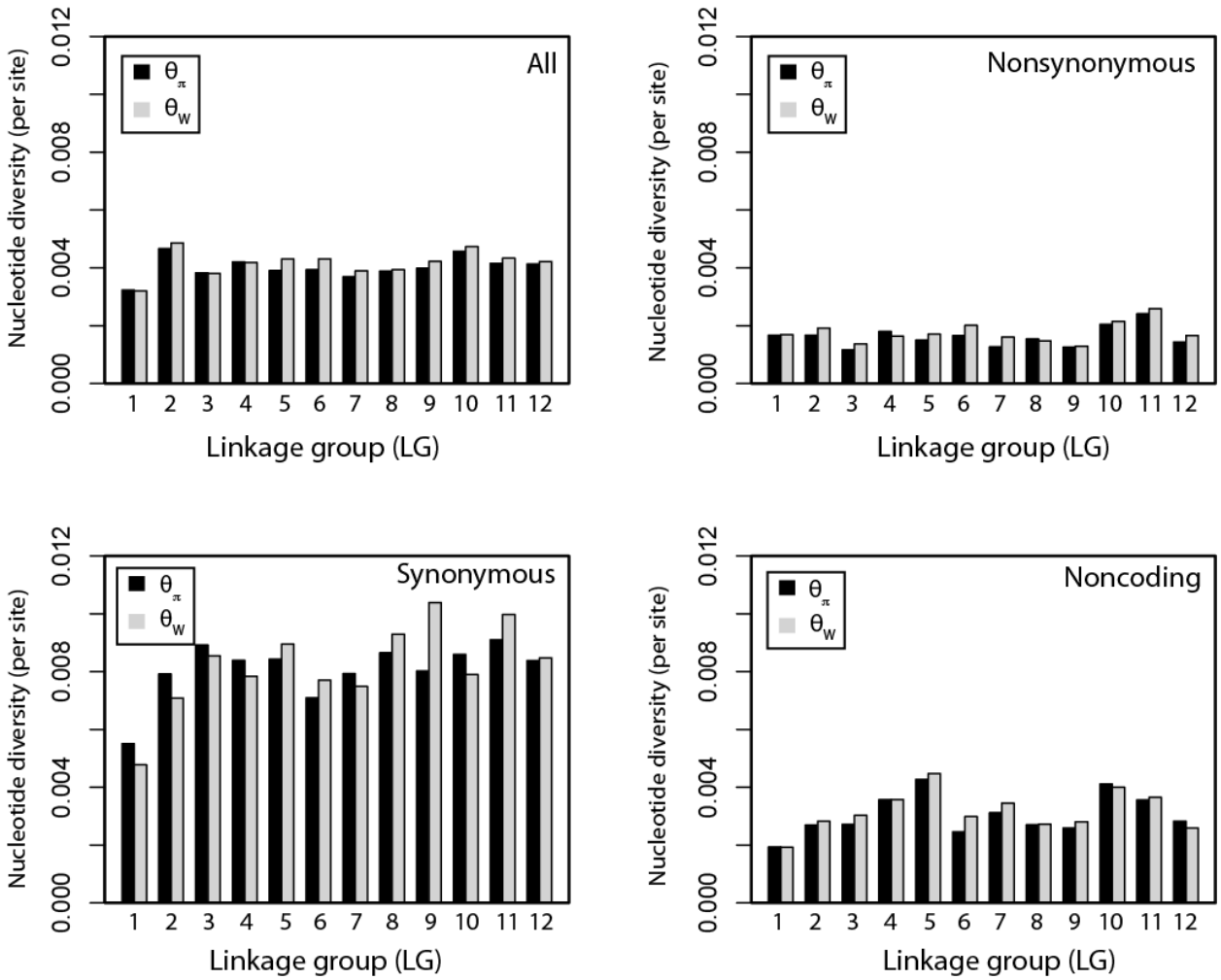

Figure S10 Average nucleotide diversity across linkage groups. Averages are weighted averages using coverage classes as the weights.

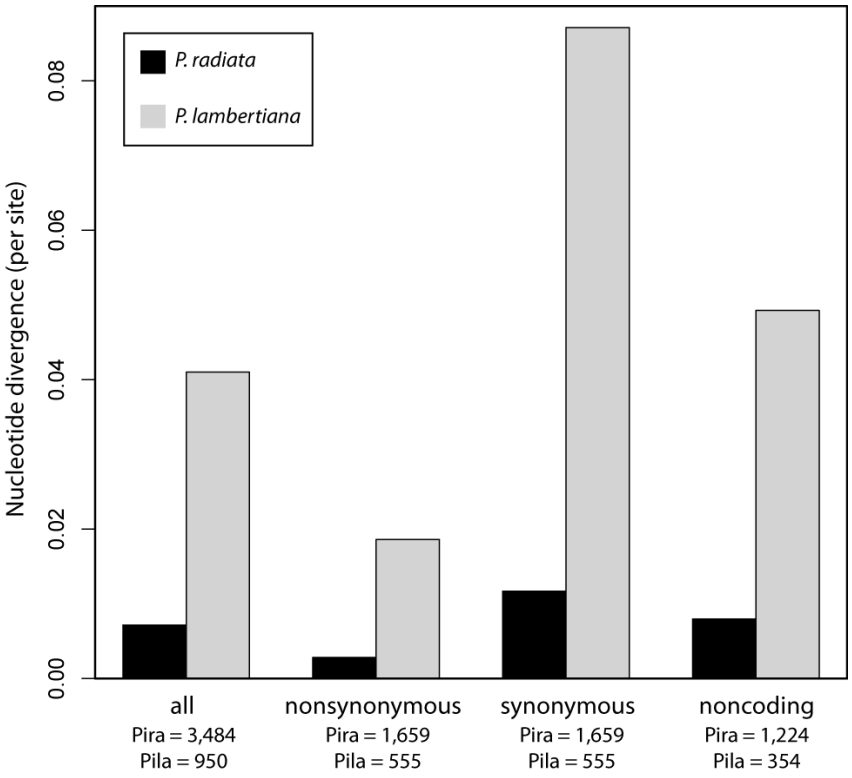

Figure S11 Average nucleotide divergence (D_{xy}) for all and annotated amplicons for each outgroup (l = number of loci or amplicons). Averages are weighted averages using coverage classes as the weights. Pila, *Pinus lambertiana*; Pira, *Pinus radiata*.

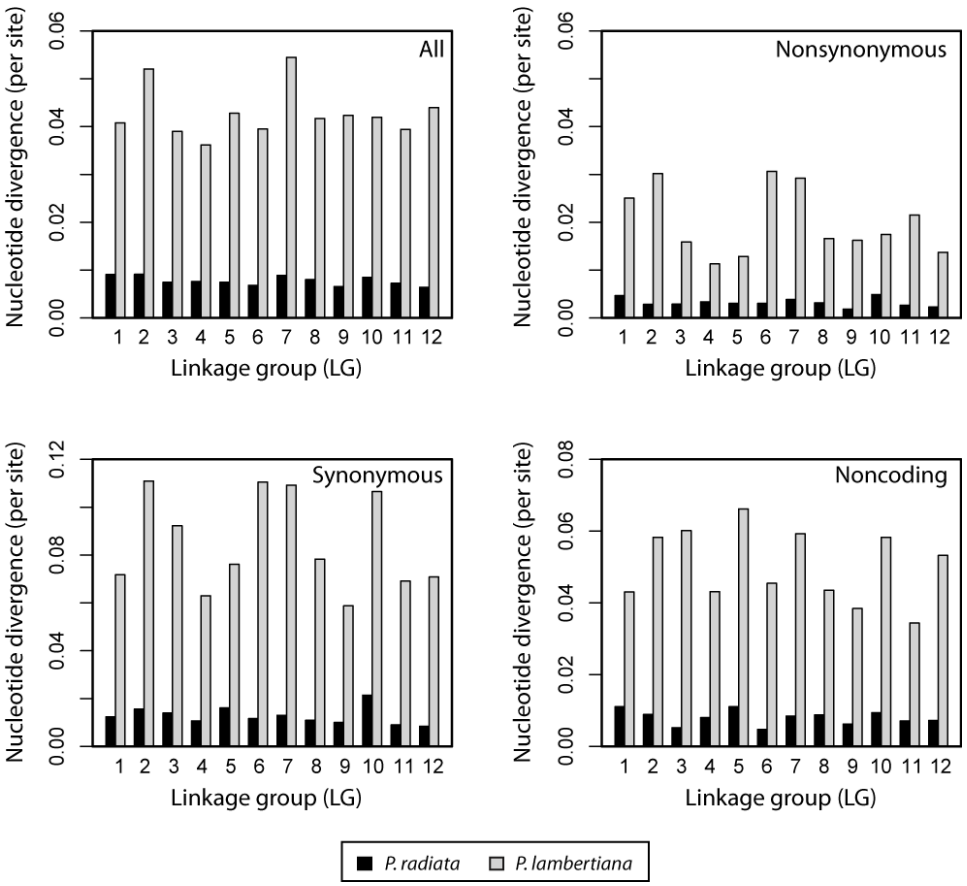

Figure S12 Average nucleotide divergence (D_{xy}) for all and annotated amplicons across linkage groups. Averages are weighted averages using coverage classes as the weights.

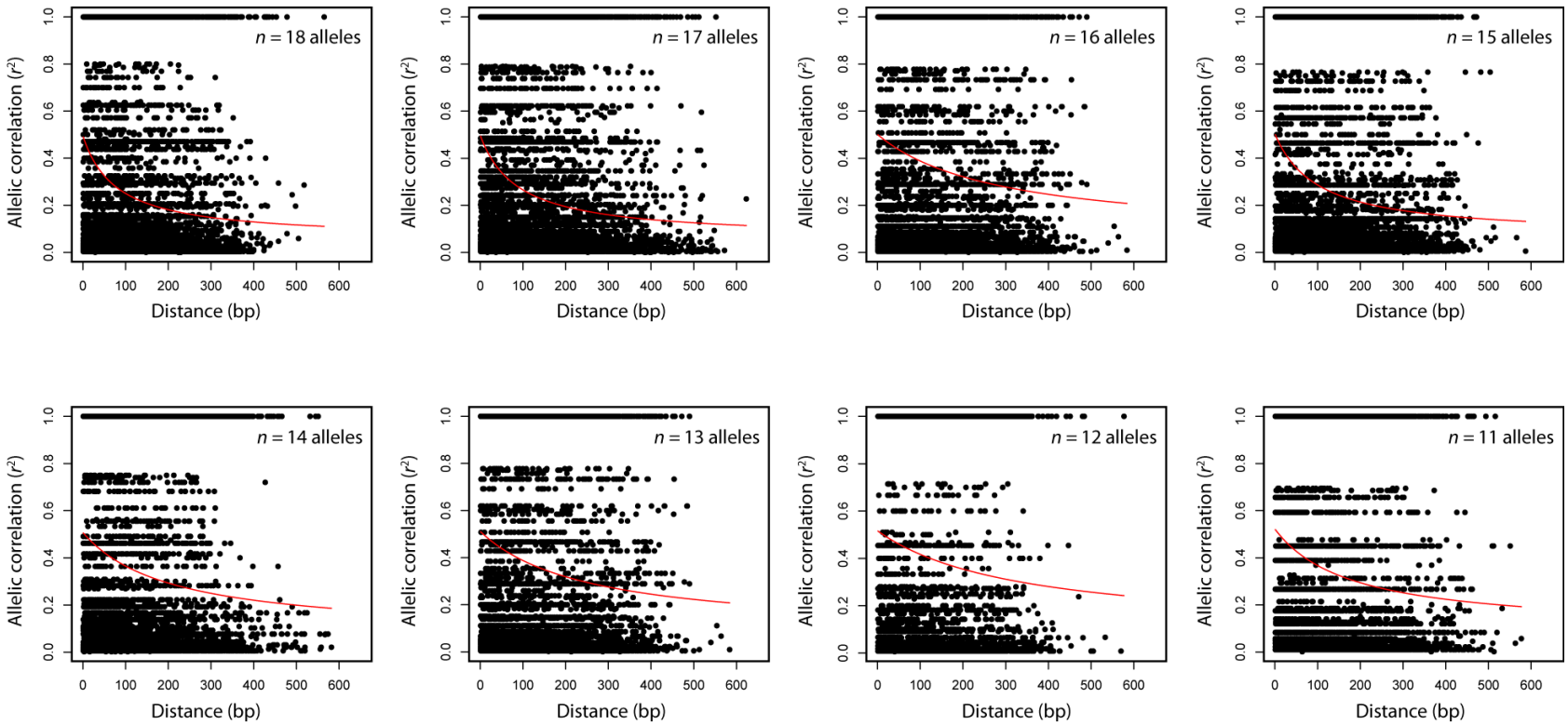

Figure S13 decay of linkage disequilibrium, as measured using pairwise allelic correlations (r^2), with physical distance (bp) across coverage classes as estimated with data including singletons. Red lines give the expected value of r^2 following Remington *et al.* (2001). Intragenic pairs of SNPs were pooled across amplicons.

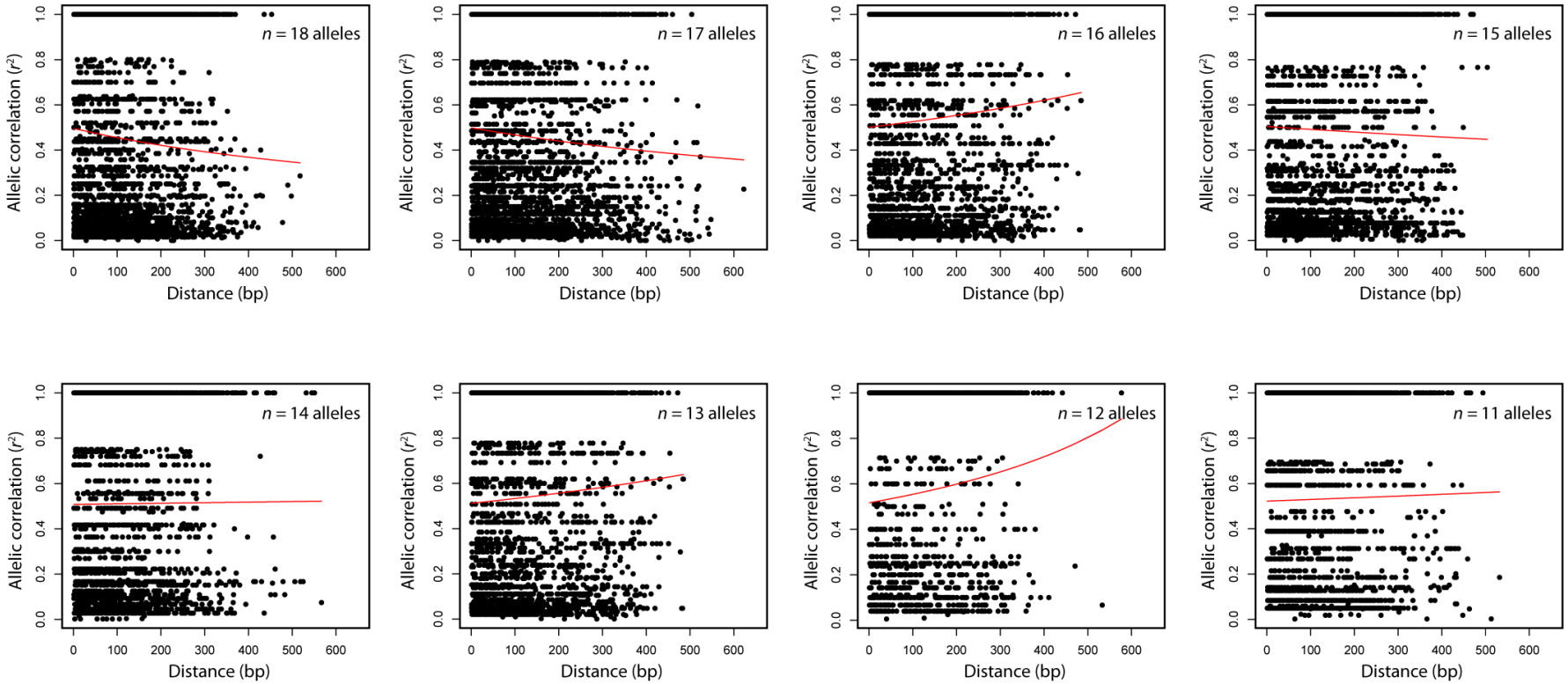

Figure S14 decay of linkage disequilibrium, as measured using pairwise allelic correlations (r^2), with physical distance (bp) across coverage classes as estimated with data excluding singletons. Red lines give the expected value of r^2 following Remington *et al.* (2001). Intragenic pairs of SNPs were pooled across amplicons.

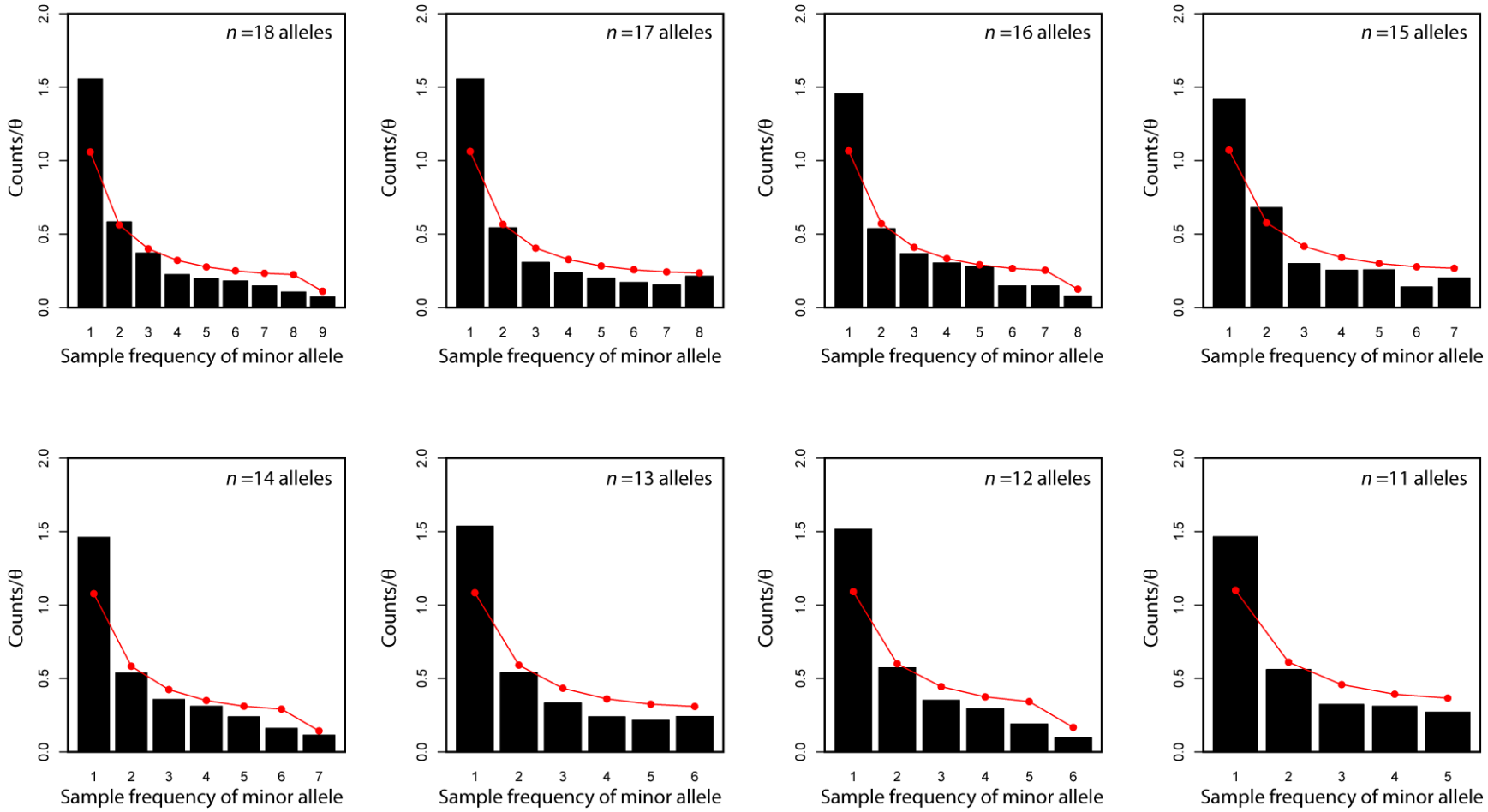

Figure S15 Observed (black bars) and expected (red points and lines) folded site-frequency spectra for sample coverage classes from 11 to 18 sampled alleles. Counts were standardized by θ for ease of comparison across sample coverage classes. Goodness-of-fit tests to the expected distribution reveal that all spectra deviate from neutral expectations (class 18: χ^2 = 365.09, *P* < 2.20e-16; class 17: χ^2 = 323.45, *P* < 2.20e-16; class 16: χ^2 = 172.31, *P* < 2.20e-16; class 15: χ^2 = 136.98, *P* < 2.20e-16; class 14: χ^2 = 117.99, *P* < 2.20e-16; class 13: χ^2 = 130.32, *P* < 2.20e-16; class 12: χ^2 = 114.23, *P* < 2.20e16; class 11: χ^2 = 82.15, *P* < 2.20e-16). The degrees of freedom for each test are the number of bins on the *x*-axis minus one.

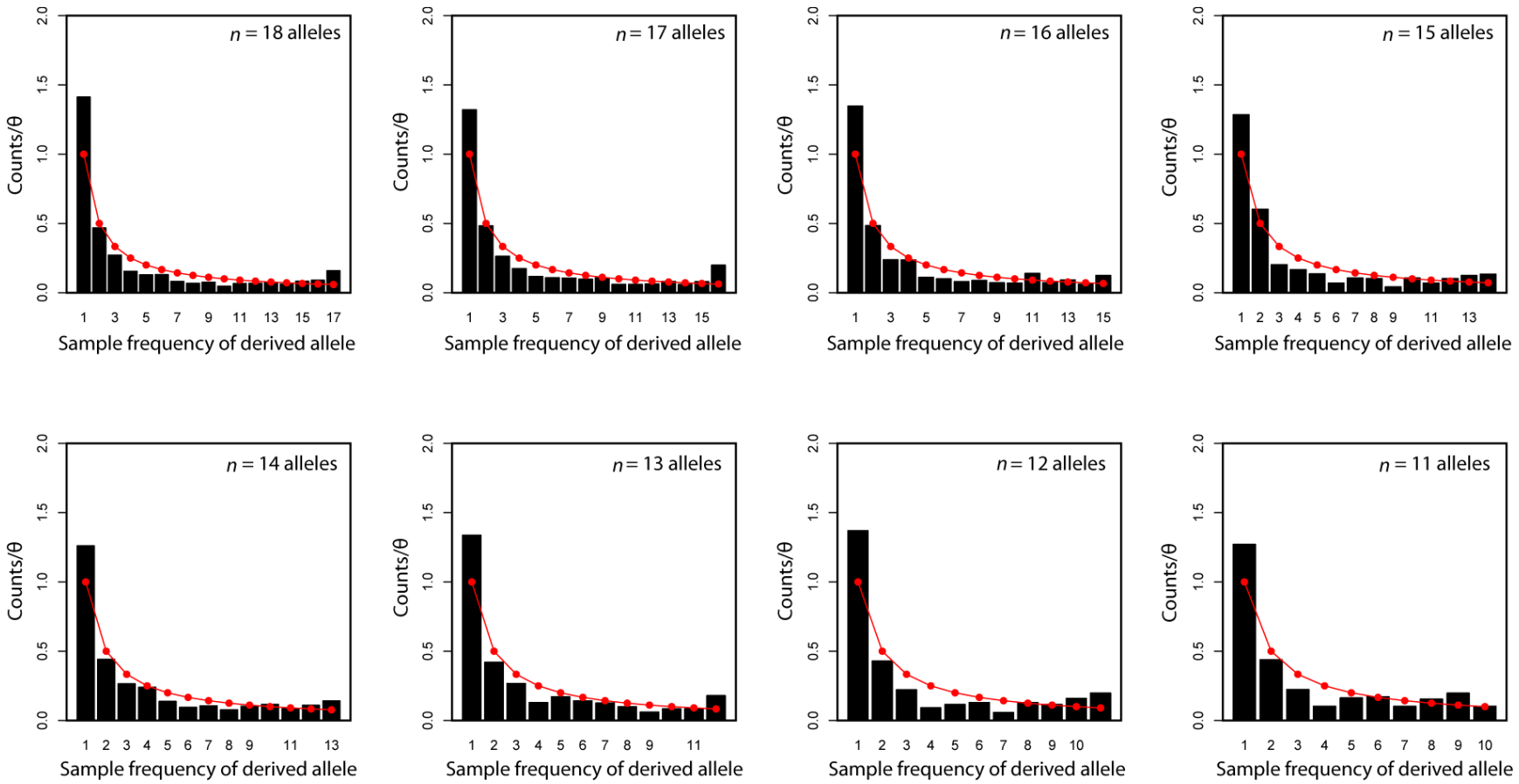

Figure S16 Unfolded site-frequency spectra across sample coverage classes using *Pinus radiata* as the outgroup. Counts were standardized by θ for ease of comparison across sample coverage classes. Goodness-of-fit tests to the expected distribution reveal that all spectra deviate from neutral expectations (class 18: χ^2 = 392.01, *P* < 2.20e-16; class 17: χ^2 = 381.01, *P* < 2.20e-16; class 16: χ^2 = 175.17, *P* < 2.20e-16; class 15: χ^2 = 134.08, *P* < 2.20e-16; class 14: χ^2 = 65.87, *P* = 1.88e-09; class 13: χ^2 = 91.33, *P* = 9.13e-15; class 12: χ^2 = 111.84, *P* < 2.20e-16; class 11: χ^2 = 34.49, *P* = 7.32e-05). The degrees of freedom for each test are the number of bins on the *x*-axis minus one.

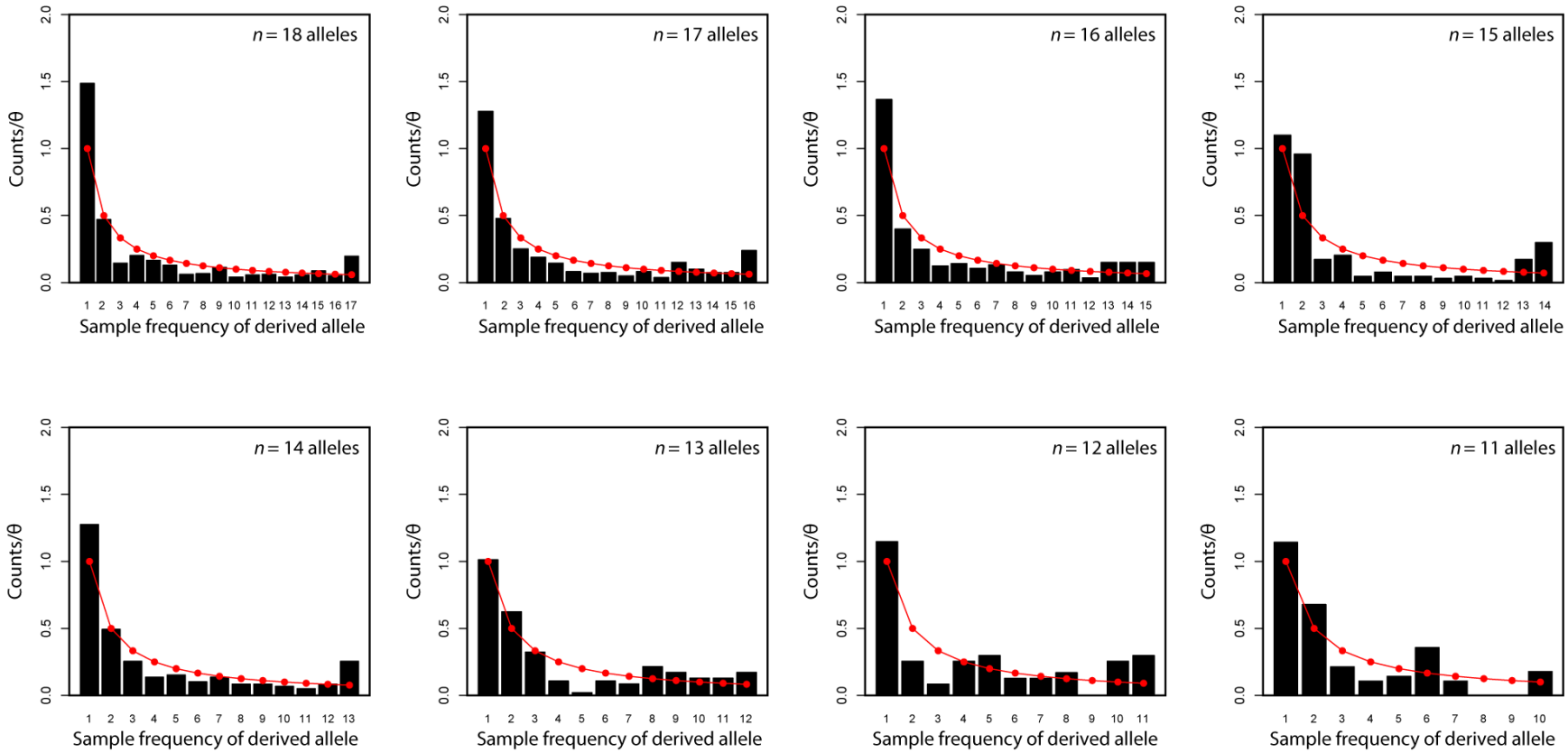

Figure S17 Unfolded site-frequency spectra across sample coverage classes using *Pinus lambertiana* as the outgroup. Counts were standardized by θ for ease of comparison across sample coverage classes. Goodness-of-fit tests to the expected distribution reveal that all spectra deviate from neutral expectations (class 18: χ^2 = 163.05, *P* < 2.20e-16; class 17: χ^2 = 136.63, *P* < 2.20e-16; class 16: χ^2 = 70.18, *P* = 1.79e-09; class 15: χ^2 = 115.33, *P* < 2.20e-16; class 14: 37.75, *P* = 1.68e-4; class 13: χ^2 = 24.80, *P* = 9.75e-3; class 12: χ^2 = 28.78, *P* =1.35e-03; class 11: χ^2 = 20.98, *P* = 0.01). The degrees of freedom for each test are the number of bins on the *x*-axis minus one.

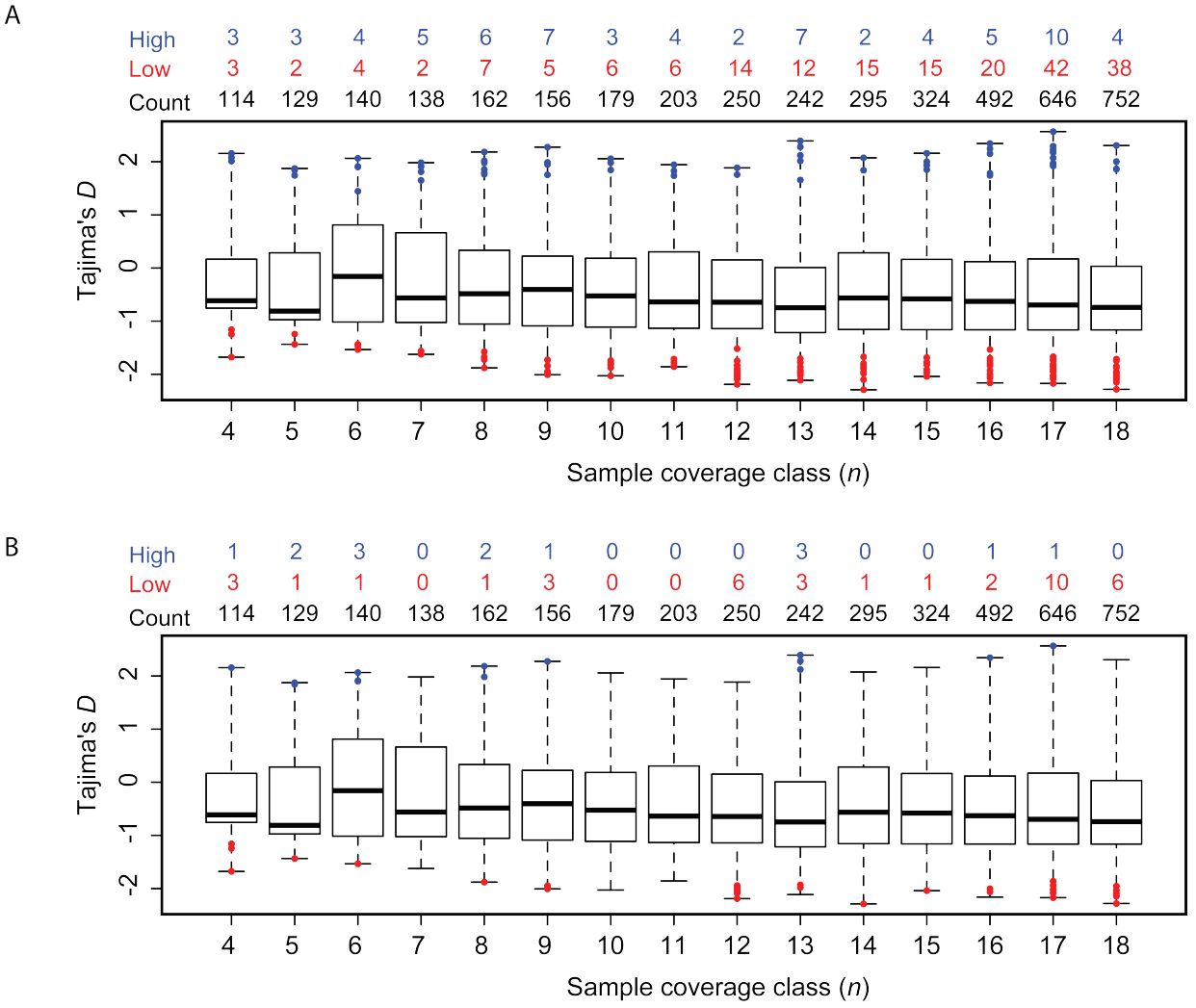

Figure S18 Summary of the folded site-frequency spectrum using Tajima's *D*. Boxplots give the observed distributions across amplicons, while colored points denote outliers (red = lower tail, blue = upper tail) at the *P* = 0.05 level. Whiskers extend to the extremes of the observed data. (A) The standard neutral model (SNM). Counts are given above this panel. (B) The three-epoch model (TEM) from Ersöz *et al.* (2010). Counts are given above this panel.

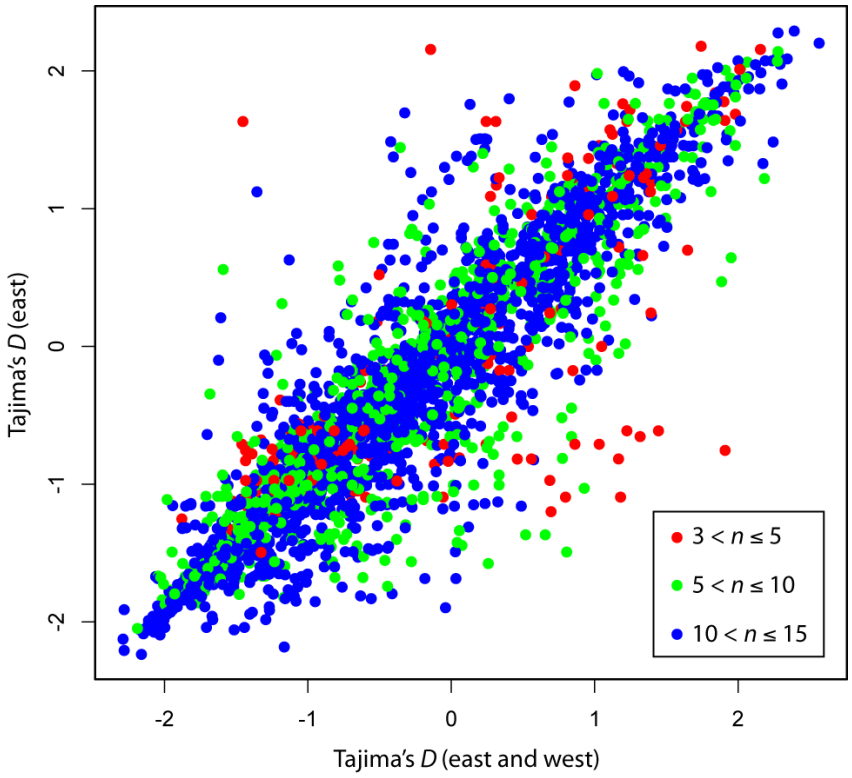

Figure S19 Summary statistics of the site-frequency spectrum are correlated between the full sample set and the sample set trimmed to just samples obtained from east of the Mississippi River. Colors denote bins of sample coverage classes. The overall correlation structure does not differ among the three classes of sample coverage (ANCOVA: $F = 1.27$, $df_1 = 2$, $df_2 = 3{,}129$, $P = 0.28$, $P_{perm} = 0.32$). This suggests that fitting the three-epoch model (TEM) of Ersöz *et al.* (2010) to the full data set, which includes samples from west of the Mississippi River, is likely appropriate.

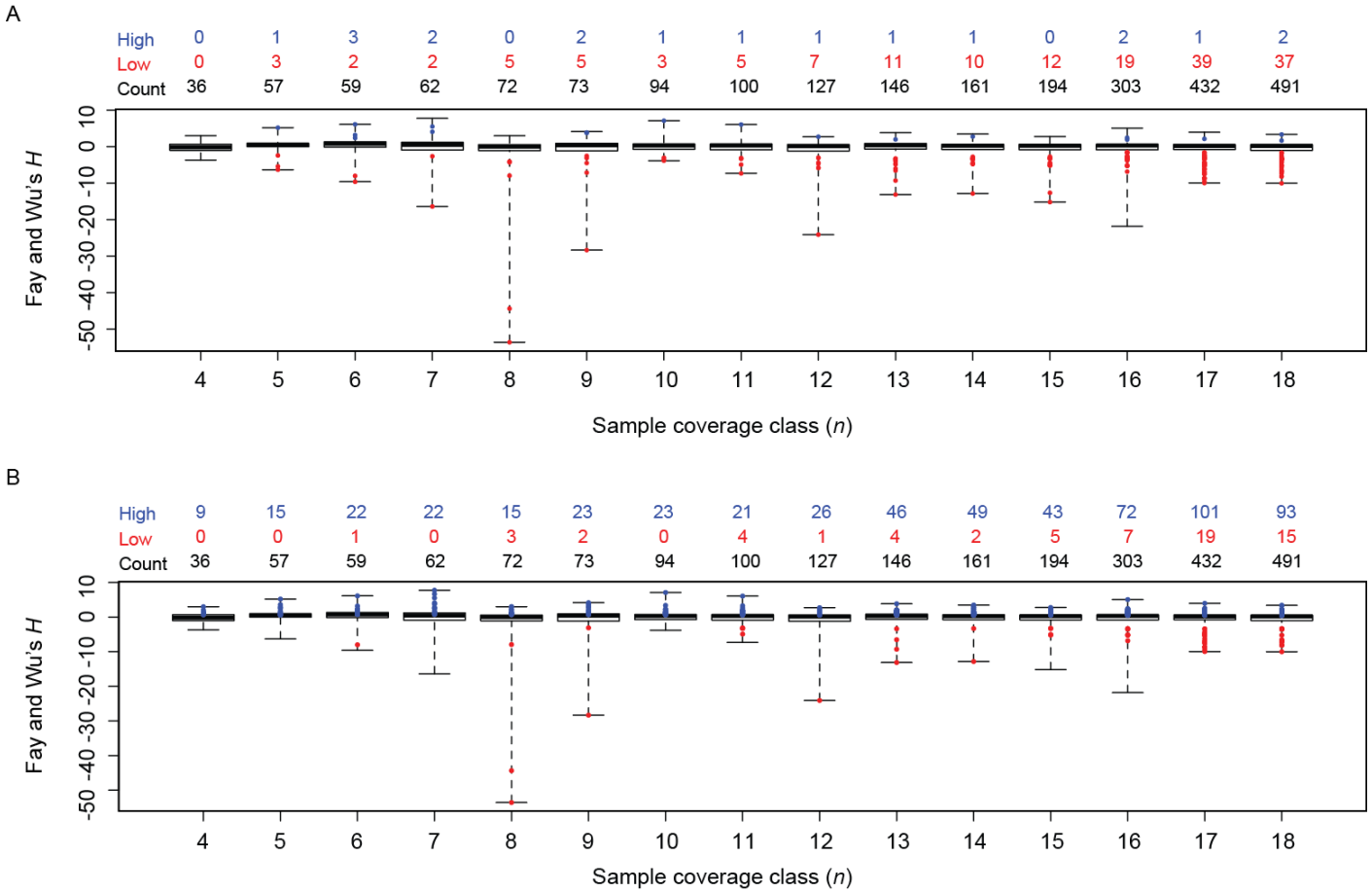

Figure S20 Summary of Fay and Wu's *H* by coverage class for two neutral models – the standard neutral model (SNM) and the three-epoch model from Ersöz *et al.* (2010). Boxplots give the observed distributions across amplicons, while colored points denote outliers (red = lower tail, blue = upper tail) at the *P* = 0.05 level. Whiskers extend to the extremes of the observed data. (A) The standard neutral model (SNM). Counts are given above this panel. (B) The three-epoch model (TEM) from Ersöz *et al.* (2010). Counts are given above this panel.

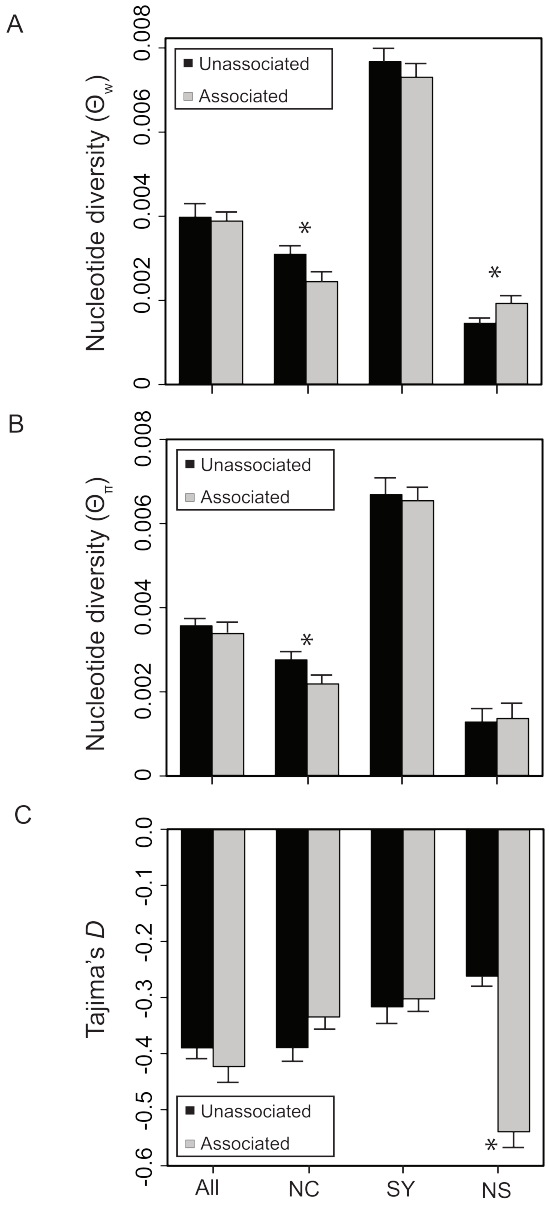

Figure S21 Summary of differences in the site-frequency spectra for amplicons associated to at least one phenotype and those unassociated to a phenotype. Amplicons associated to at least one phenotype have too many rare variants at nonsynonymous sites, whereas they have too few rare variants at noncoding sites. This pattern causes Tajima's *D* to be more negative for nonsynonymous sites and less negative for noncoding sites.

File S2
DNA Sequence Data

Available for download as an Excel file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.157198/-/DC1.