# The evolutionary genetics of the genes underlying phenotypic associations for loblolly pine (*Pinus taeda*, Pinaceae)

Andrew J. Eckert[1,§], Jill L. Wegrzyn[2,§], John D. Liechty[2], Jennifer M. Lee[3], W. Patrick Cumbie[4], John M. Davis[5], Barry Goldfarb[6], Carol A. Loopstra[7], Sreenath R. Palle[7], Tania Quesada[5], Charles H. Langley[8], and David B. Neale[2,9]

[1]Department of Biology, Virginia Commonwealth University, Richmond, VA 23284

[2]Department of Plant Sciences, University of California at Davis, Davis, CA 95616

[3]Computercraft, McLean, VA 22101

[4]ArborGen Inc., Ridgeville, SC 29472

[5]School of Forest Resources and Conservation, University of Florida, Gainesville, FL 32611

[6]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC 27695

[7]Department of Ecosystem Science and Management, Texas A&M University, College Station, TX 77843

[8]Department of Evolution and Ecology, University of California at Davis, Davis, CA 95616

[§]These authors contributed equally to this work.

DNA sequence data are available at GenBank: located in File S2.

**Running title:** Evolutionary genetics of phenotypic associations

**Key words:** association mapping, complex traits, evolutionary genetics, loblolly pine, *Pinus taeda*, natural selection, population genomics, quantitative genetics

[9]**Author for correspondence:**

David B. Neale
University of California at Davis
Department of Plant Sciences
One Shields Avenue
Davis, CA 95616

phone: (530) 754-8431
fax: (530) 754-9366
email: dbneale@ucdavis.edu

## 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 δ$^{13}$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 ($P_{\mathrm{perm}}$ = 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 ($P_{\text{perm}} > 0.15$).

**Linkage Disequilibrium**

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

**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: $P = 0.0371$, $P_{\text{perm}} = 0.011$), whether or not they were annotated (Mann-Whitney U-test: $P = 1.486\text{e-}05$, $P_{\text{perm}} < 0.001$), and whether or not they were associated with at least one phenotype (Mann-Whitney U-test: $P = 0.0493$, $P_{\text{perm}} = 0.026$). On average (± 1 sd), linkage disequilibrium was higher for amplicons that were mapped ($Z_{\text{nS}}$: 0.327 ± 0.276 vs. 0.304 ± 0.292), while it was lower for those that were annotated ($Z_{\text{nS}}$: 0.302 ± 0.287 vs. 0.336 ± 0.288) and for those associated with at least one phenotype ($Z_{\text{nS}}$: 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: $P = 0.032$, $P_{\text{perm}} = 0.012$), with amplicons associated with disease phenotypes having the lowest ($Z_{\text{nS}} = 0.144$) and amplicons associated with expression phenotypes having the highest ($Z_{\text{nS}} = 0.308$) 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 | $L_{\text{total}}$ (bp) | $L_{\text{mean}}$ (bp) | Masked Bases[b] | Masked SNPs[b] | $S_{\text{total}}$ | $S_{\text{mean}}$ | $r$ | $L_{\text{total}}/S_{\text{total}}$ | 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 | Count | *L* | *L* | *L* | Masked *L* | Masked *L* | *S* | *S* | Masked *S* | Masked *S* |
|---|---|---|---|---|---|---|---|---|---|---|
| Class[a] | | total | coding | noncoding | coding[b] | noncoding[b] | coding | noncoding | coding[b] | 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 ($\theta_\pi$) | Kruskal-Wallis | $\chi^2$ = 97.63, df=16, *P* = 9.62e-14 | Average ranks of diversity vary significantly across coverage classes |
| Nucleotide divergence (Pira) | Kruskal-Wallis | $\chi^2$ = 95.61, df=16, *P* = 2.29e-14 | Average ranks of divergence vary significantly across coverage classes |
| Nucleotide divergence (Pila) | Kruskal-Wallis | $\chi^2$ = 29.41, df=16, *P* = 0.02134 | Average ranks of divergence vary significantly across coverage classes |
| The number of SNPs | Kruskal-Wallis | $\chi^2$ = 174.31, df=16, *P* < 2.2e-16 | Average ranks of SNPs vary significantly across coverage classes |
| Alignment length (bp) | Kruskal-Wallis | $\chi^2$ = 236.24, df=16, *P* < 2.2e-16 | Average ranks of alignment lengths vary significantly across coverage classes |
| Noncoding sites (bp) | Kruskal-Wallis | $\chi^2$ = 24.24, df=16, *P* = 0.08437 | Average ranks of noncoding sites do not vary significantly across coverage classes |
| Coding sites (bp) | Kruskal-Wallis | $\chi^2$ = 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 | $\chi^2$ = 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 | $\chi^2$ = 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 | $\chi^2$ = 36.32, df=16, *P* = 0.00415 | Too many indels at intermediate coverage classes |
| OG (Pira) | Goodness-of-fit | $\chi^2$ = 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 | $\chi^2$ = 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 ($P_{\text{perm}}$)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<br>(- indels) | mean<br>(+ indels) | *t* | *df* | *P* | $P_{\text{perm}}$ |
|---|---|---|---|---|---|---|
| *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 |
| $\theta_\pi$ | 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; $\theta_\pi$, 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* | $\theta_\pi$ | $\theta_\pi$ | $\theta_\pi$ | $\theta_\pi$ | $\theta_W$ | $\theta_W$ | $\theta_W$ | $\theta_W$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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, $\theta_{\pi,}$ nucleotide diversity based on the average number of pairwise differences; $\theta_{W,}$ 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 ($C$ = 4$N_e r$) 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] | $Z_{nS}$ | $C/\theta_\pi$ |
|---|---|---|---|---|---|
| 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; $\theta_\pi$, 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<br>(*l* = 3,360) | | | Trimmed[a]<br>(*l* = 3,133) | | |
|---|---|---|---|---|---|---|
| Statistic | Obs | *P*<br>(SNM) | *P*<br>(TEM) | Obs | *P*<br>(SNM) | *P*<br>(TEM) |
| Mean | | | | | | |
| $\theta_\pi$ | 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 |
| $Z_{\mathrm{nS}}$ | 0.305 | > 0.999 | 0.098 | 0.298 | > 0.999 | 0.115 |
| Variance | | | | | | |
| $\theta_\pi$ | 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 |
| $Z_{\mathrm{nS}}$ | 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; $\theta_\pi$, nucleotide diversity from the average number of pairwise differences; TEM, three epoch model from Ersöz *et al.* (2010); $Z_{\mathrm{nS}}$, 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. remodeling.histone.proteins. | 5 | 17 | 35 | 0.043 | -0.317 | 0.519 |
| 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. VARICOSE.SWAP.DNA.Polymerase. | 16 | 48 | 135 | 0.060 | -0.378 | -0.084 |
| 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. carbohydrate.protein.kinase. | 5 | 16 | 32 | 0.140 | -0.665 | -0.831 |
| 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.

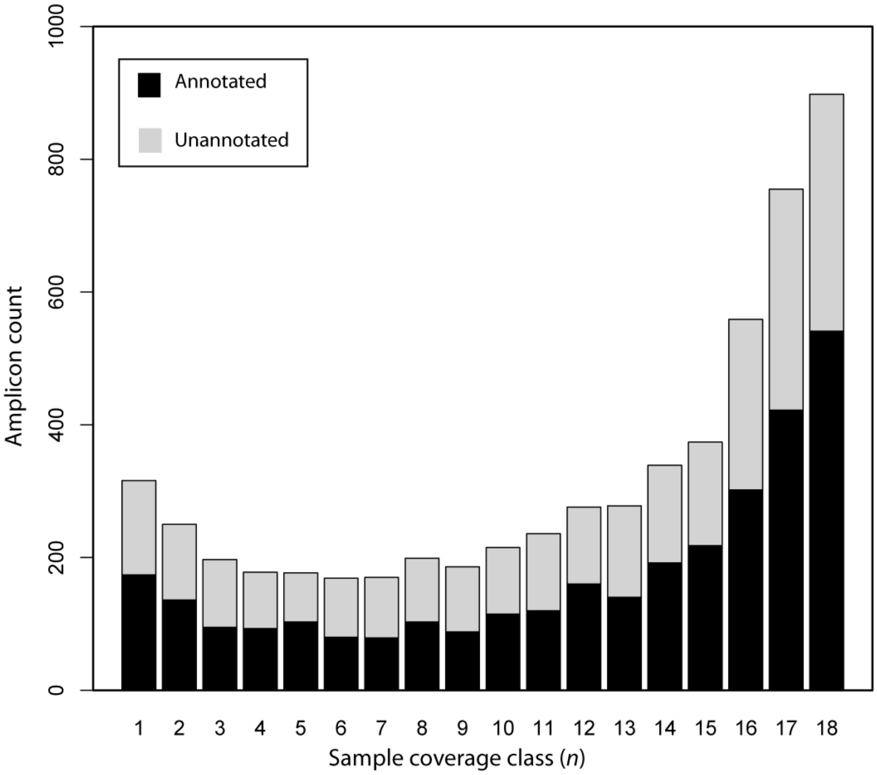

**Figure S1** The frequency distribution of sample sizes across all 5,773 amplicons reveals that the majority of amplicons have 10 or more samples. Colors distinguish amplicons for which coding and noncoding regions could be annotated (black) from those that could not (gray).

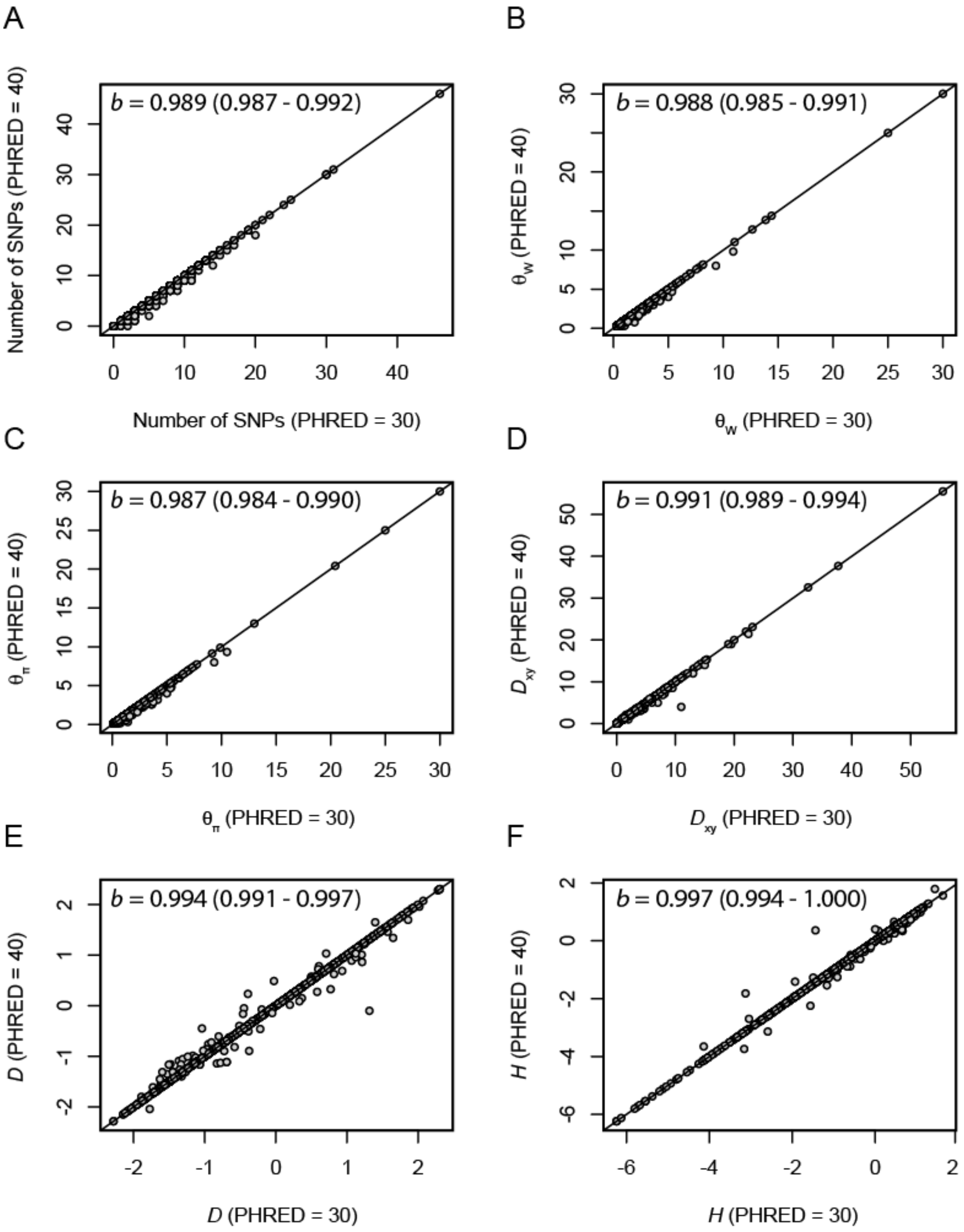

**Figure S2** The correlation between various estimates of nucleotide diversity and divergence for two different cutoffs of base calling quality (PHRED 30 and PHRED 40). Summary statistics (*b* = slope (95% confidence interval)) of linear models are given in the upper left of each plot.

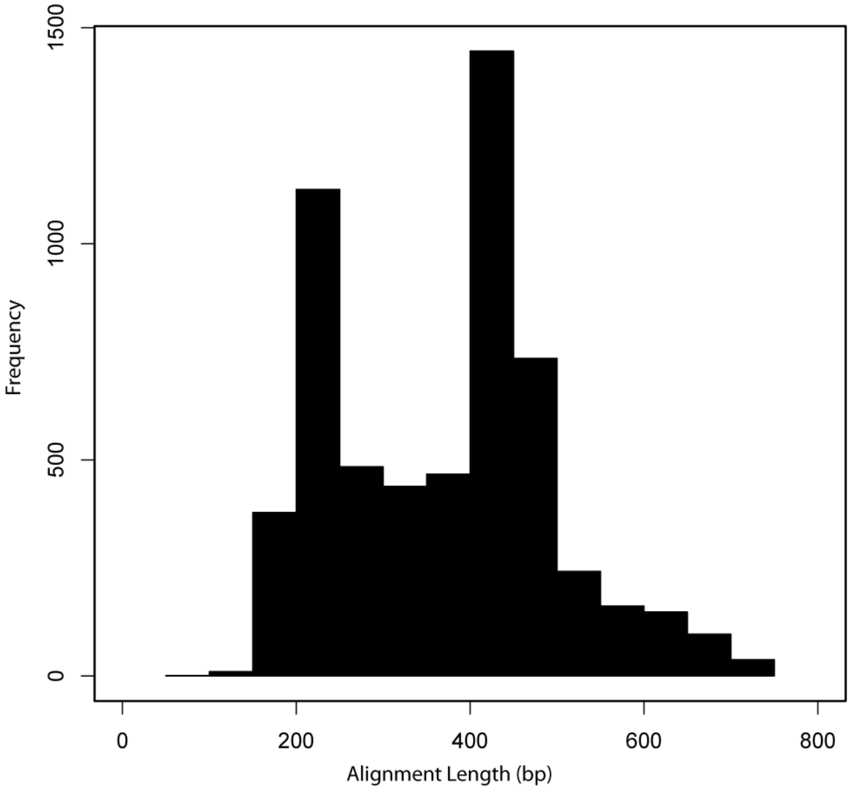

**Figure S3** The distribution of the number of aligned sites across amplicons is strongly bimodal.

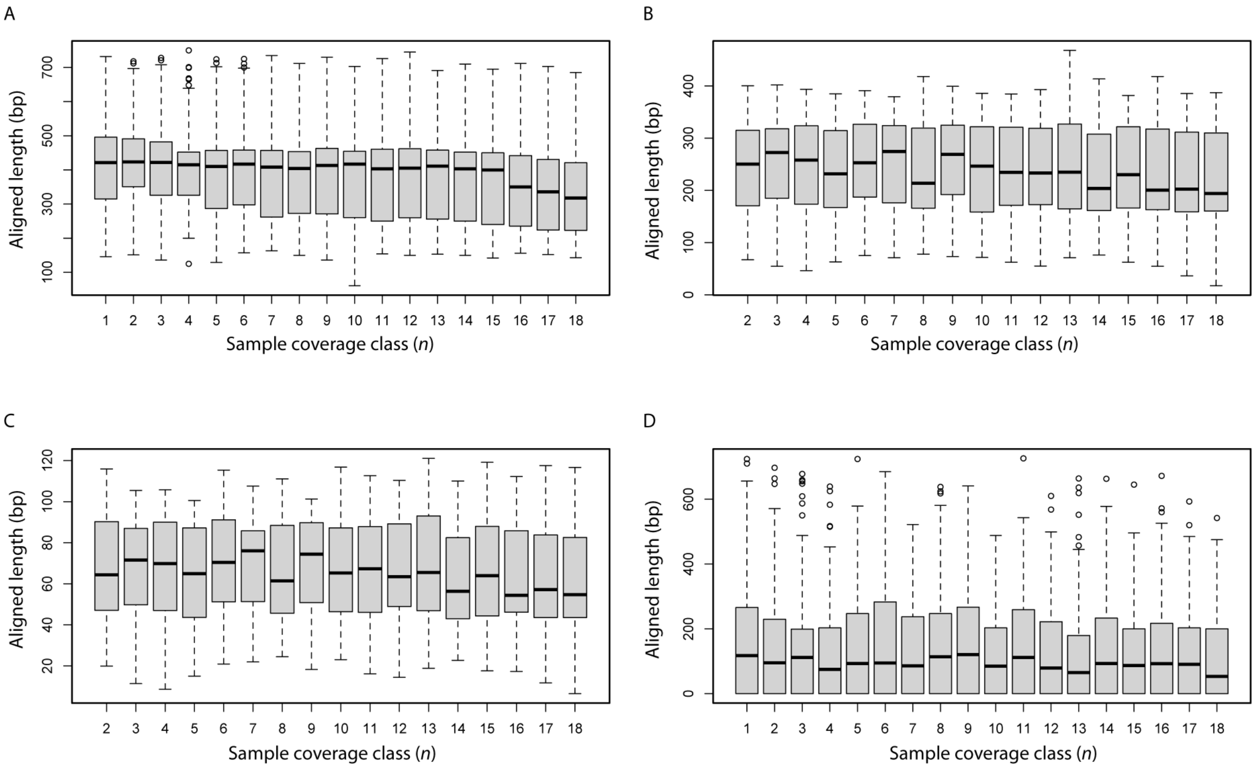

**Figure S4** Alignment length (bp) varied significantly among sample coverage classes for all (A; Kruskal-Wallis test: $\chi^2 = 274.87$, *df* = 17, *P* < 2.2e-16), nonsynonymous (B; Kruskal-Wallis test: $\chi^2 = 41.71$, *df* = 16, *P* = 0.00044), and synonymous (C; Kruskal-Wallis test: $\chi^2 = 40.09$, *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.

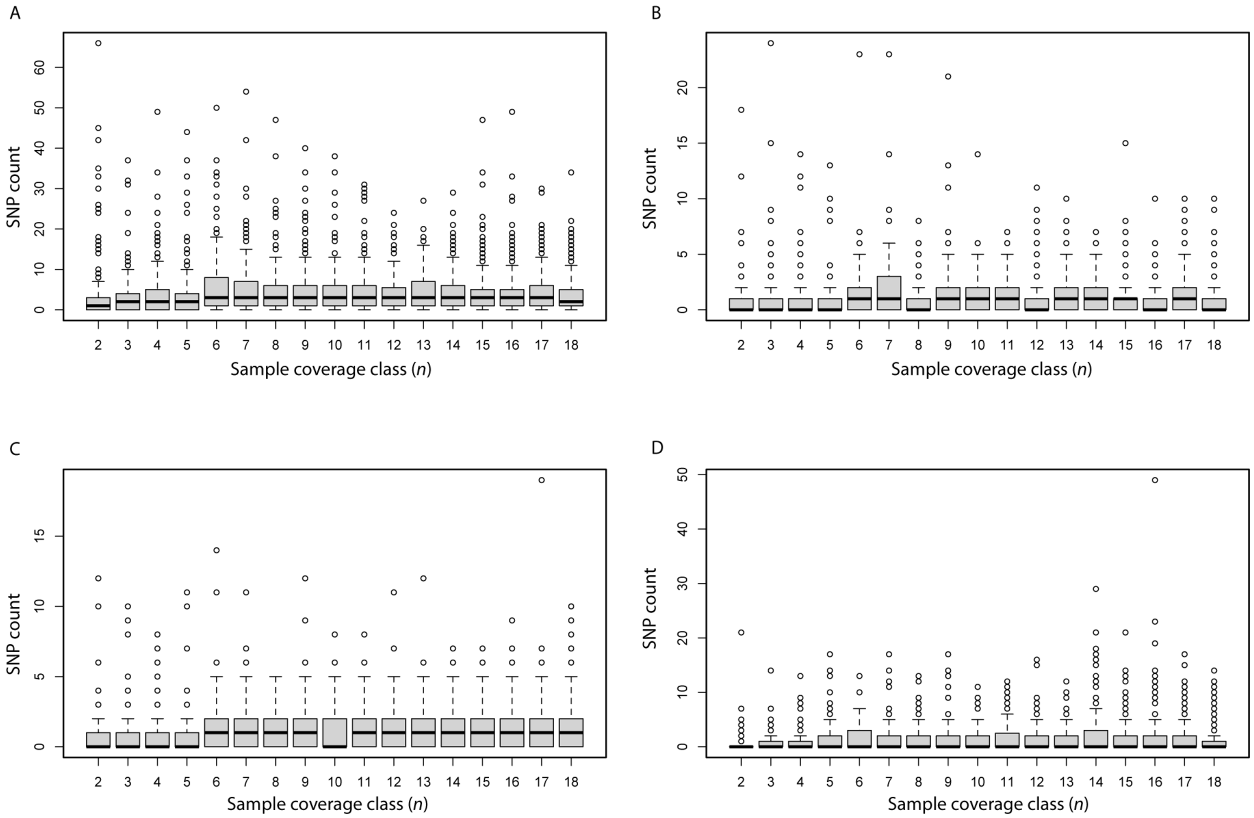

**Figure S5** The number of SNPs varied significantly among sample coverage classes for all (Kruskal-Wallis test: $\chi^2$ = 174.31, *df* = 16, *P* < 2.2e-16), nonsynonymous (Kruskal-Wallis test: $\chi^2$ = 174.31, *df* = 16, *P* = 0.00035), synonymous (Kruskal-Wallis test: $\chi^2$ = 72.05, *df* = 16, *P* = 4.3e-09) and noncoding (Kruskal-Wallis test: $\chi^2$ = 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.

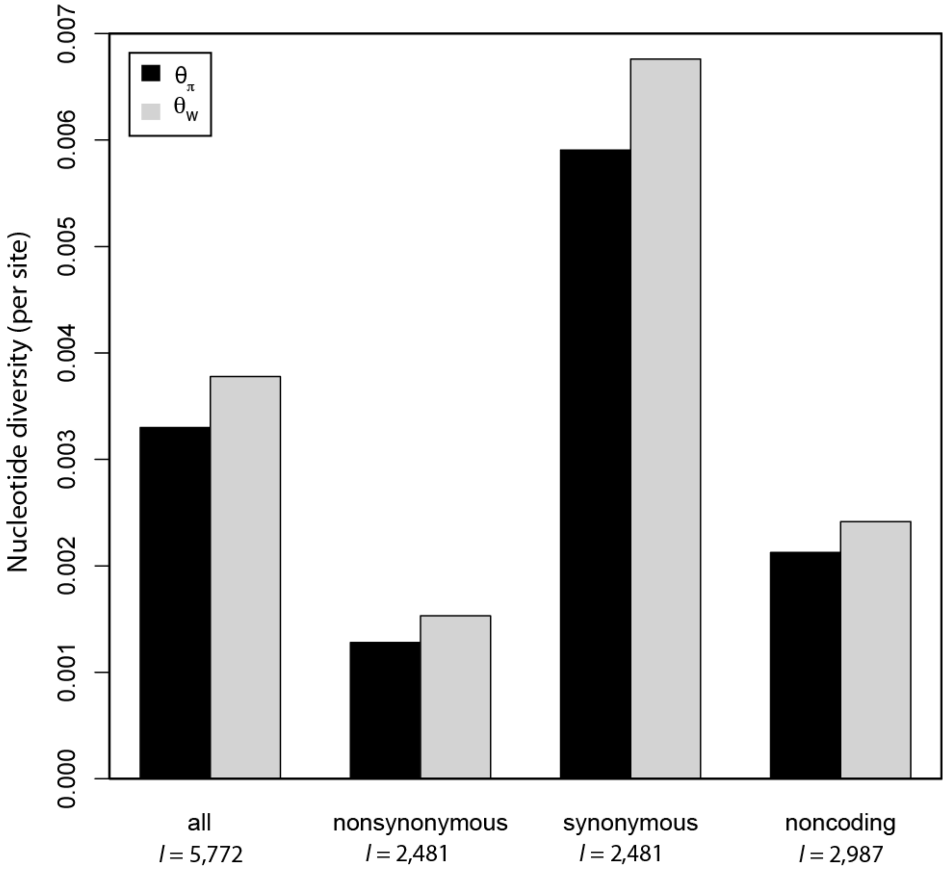

**Figure S6** Average nucleotide diversity for all and annotated amplicons (*l* = number of loci or amplicons). Averages are weighted averages using coverage classes as the weights.

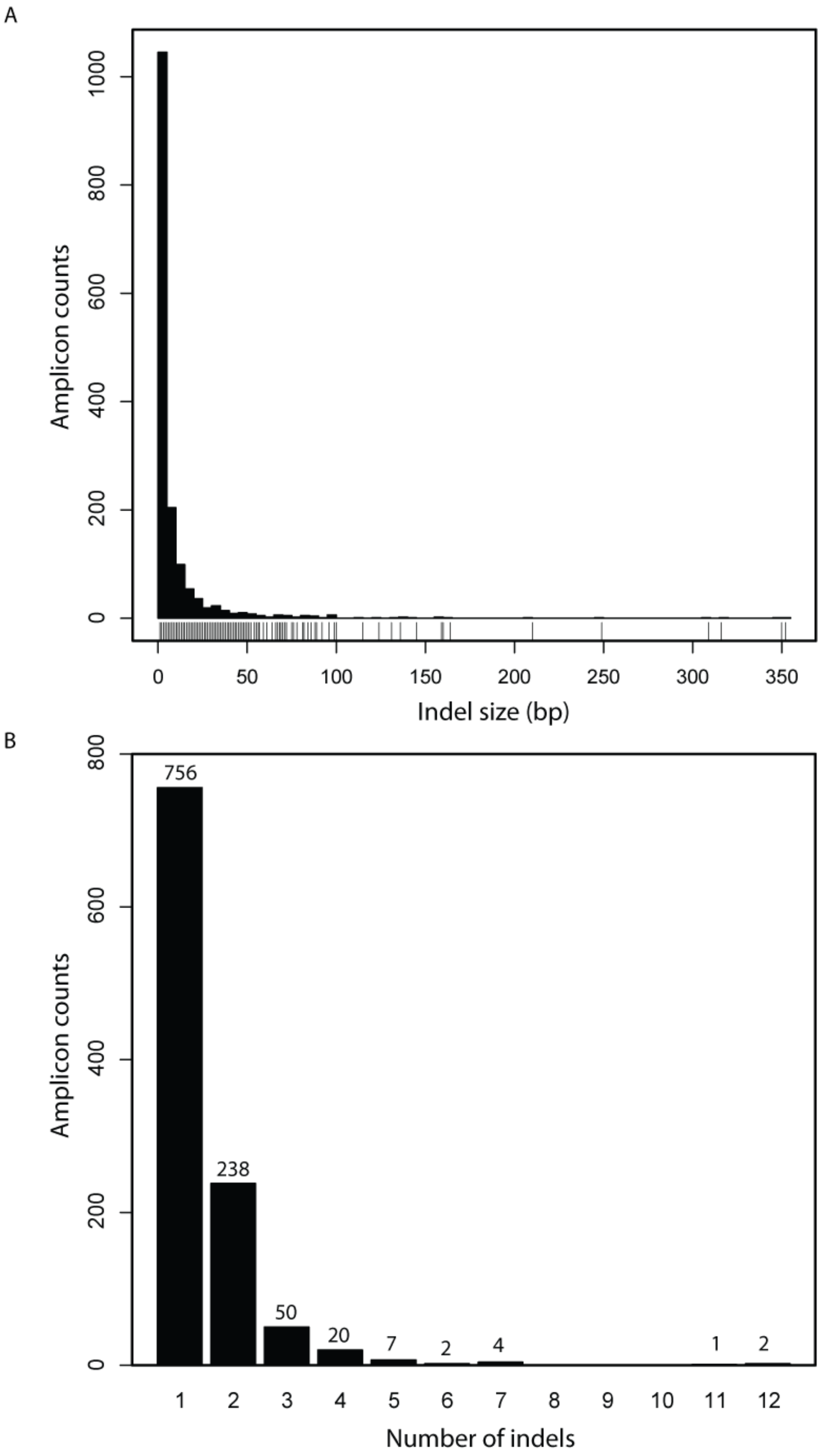

**Figure S7** Frequency distribution for indel size (A) and the number of indels per amplicon (B). The rug plot in panel A identifies bins in the histogram that are difficult to differentiate.

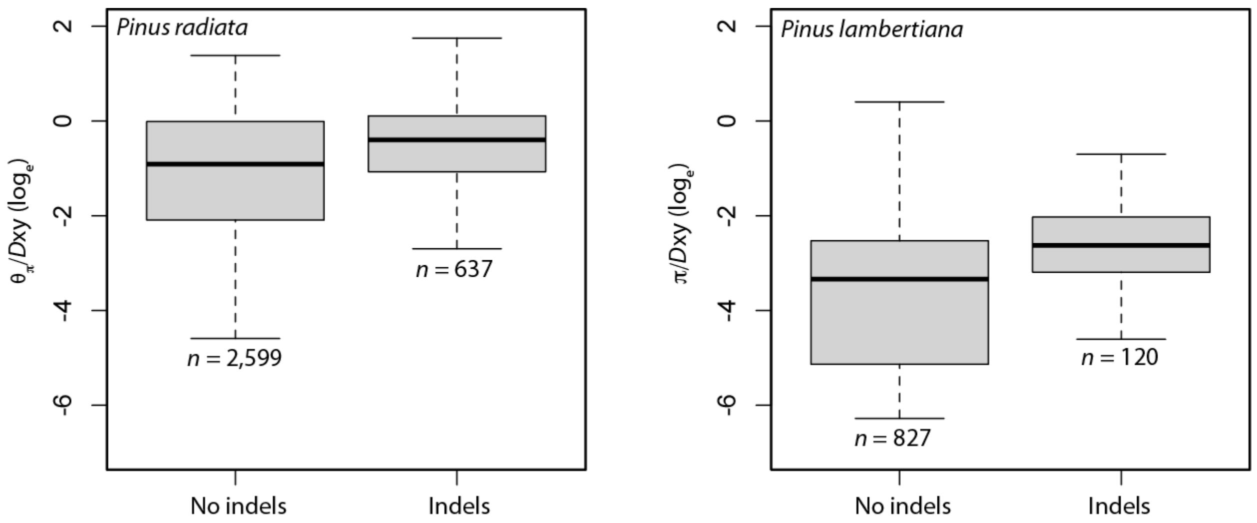

**Figure S8** Nucleotide diversity scaled by divergence differs between classes of amplicons defined based on the presence of at least one indel. The patterns are the same for divergence relative to *Pinus radiata* or *P. lambertiana*. Note that values are on a log-scale (base *e*). Sample sizes for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

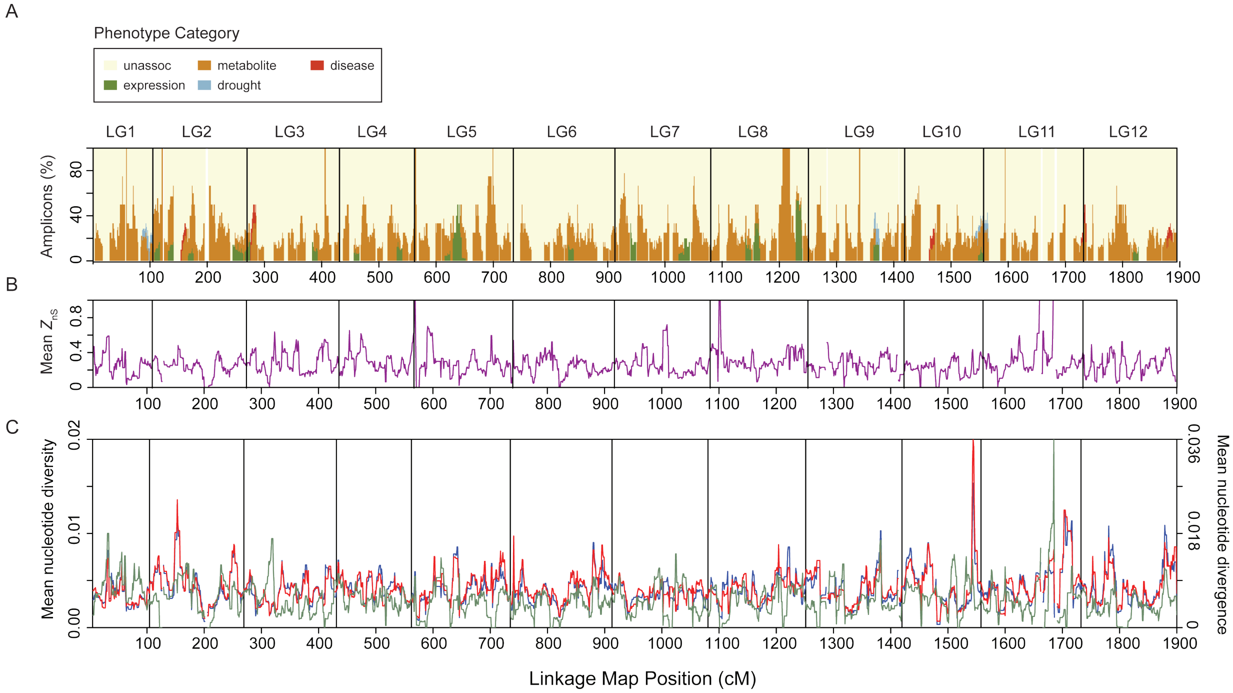

**Figure S9** Distributions of gene categories, linkage disequilibrium, nucleotide diversity and nucleotide divergence across the consensus linkage map of loblolly pine based on sliding windows (5 cM size in steps of 2 cM). (A) Stacked bar plot of gene categories across the consensus linkage map. (B) Intragenic linkage disequilibrium, as assessed using Kelly's $Z_{nS}$ statistic, across the consensus linkage map. (C) Nucleotide diversity ($\theta_{\pi}$ = red, $\theta_{W}$ = 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.

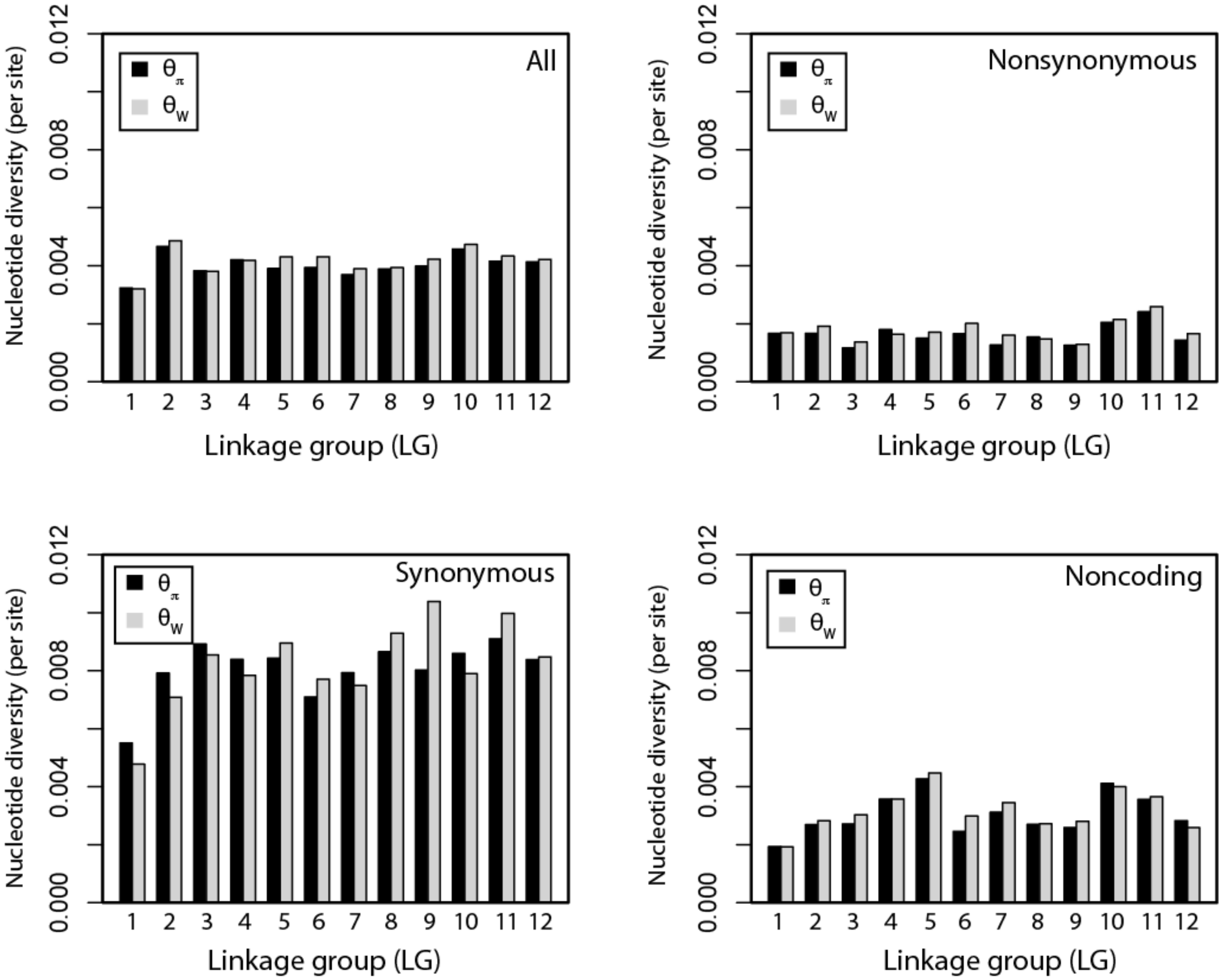

**Figure S10** Average nucleotide diversity across linkage groups. Averages are weighted averages using coverage classes as the weights.

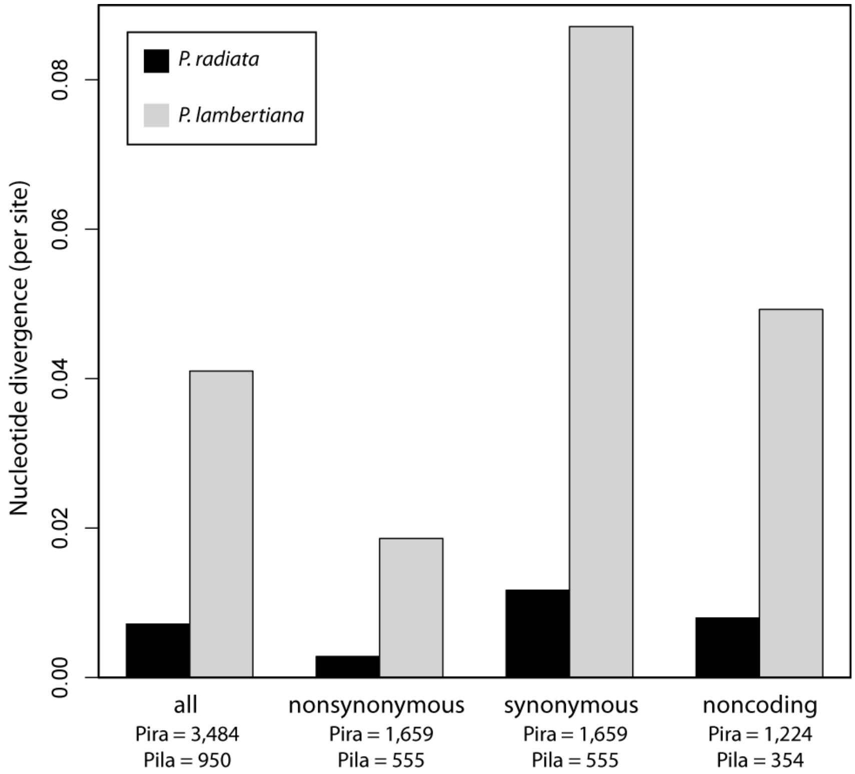

**Figure S11** Average nucleotide divergence ($D_{xy}$) for all and annotated amplicons for each outgroup (*l* = number of loci or amplicons). Averages are weighted averages using coverage classes as the weights. Pila, *Pinus lambertiana*; Pira, *Pinus radiata*.

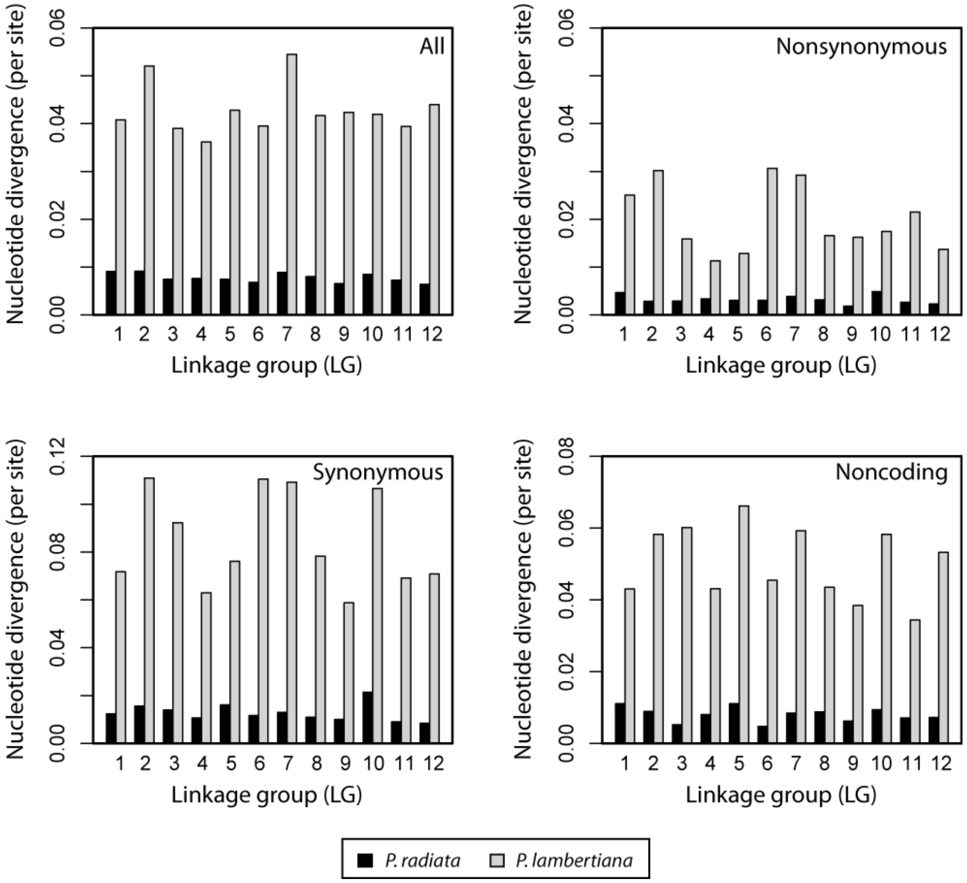

**Figure S12** Average nucleotide divergence ($D_{xy}$) for all and annotated amplicons across linkage groups. Averages are weighted averages using coverage classes as the weights.

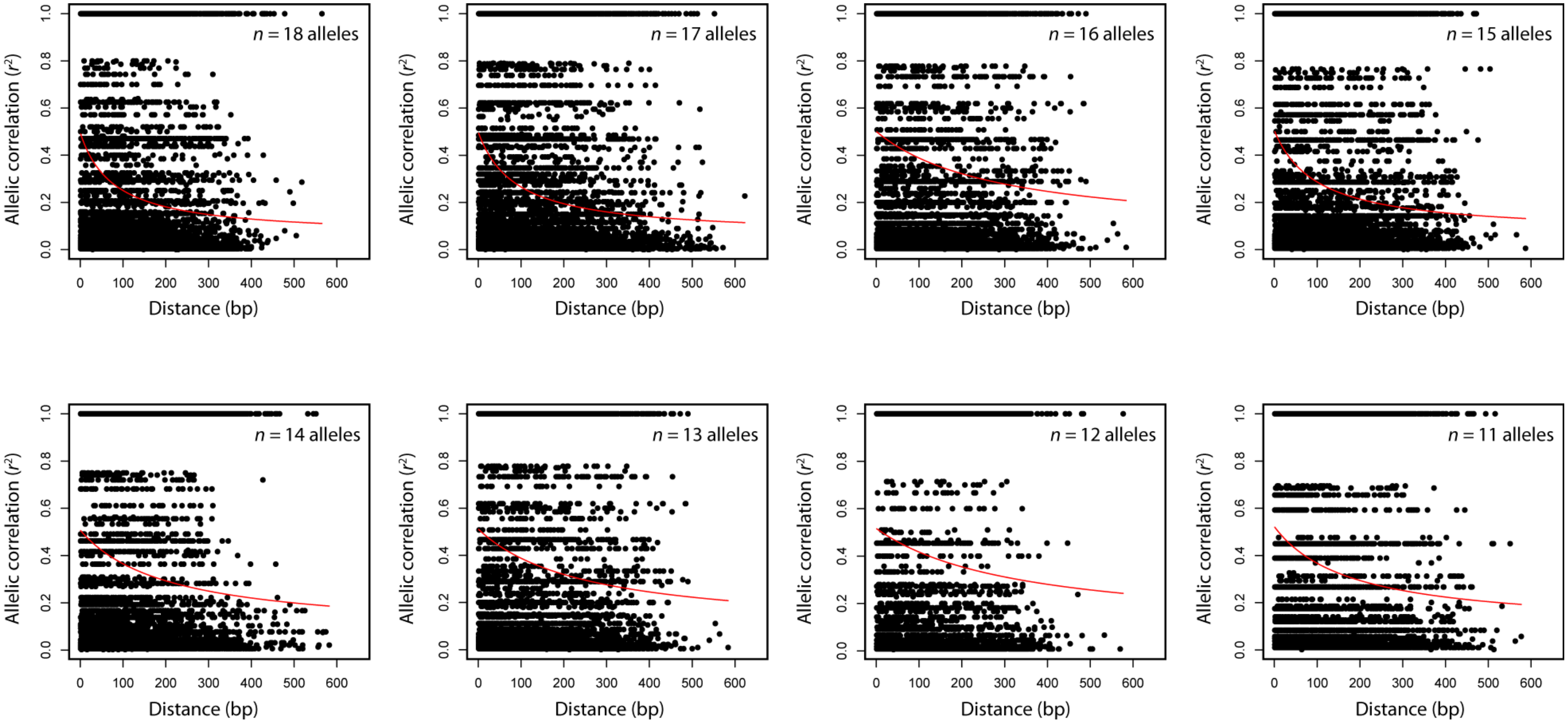

**Figure S13** decay of linkage disequilibrium, as measured using pairwise allelic correlations ($r^2$), with physical distance (bp) across coverage classes as estimated with data including singletons. Red lines give the expected value of $r^2$ following Remington *et al.* (2001). Intragenic pairs of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

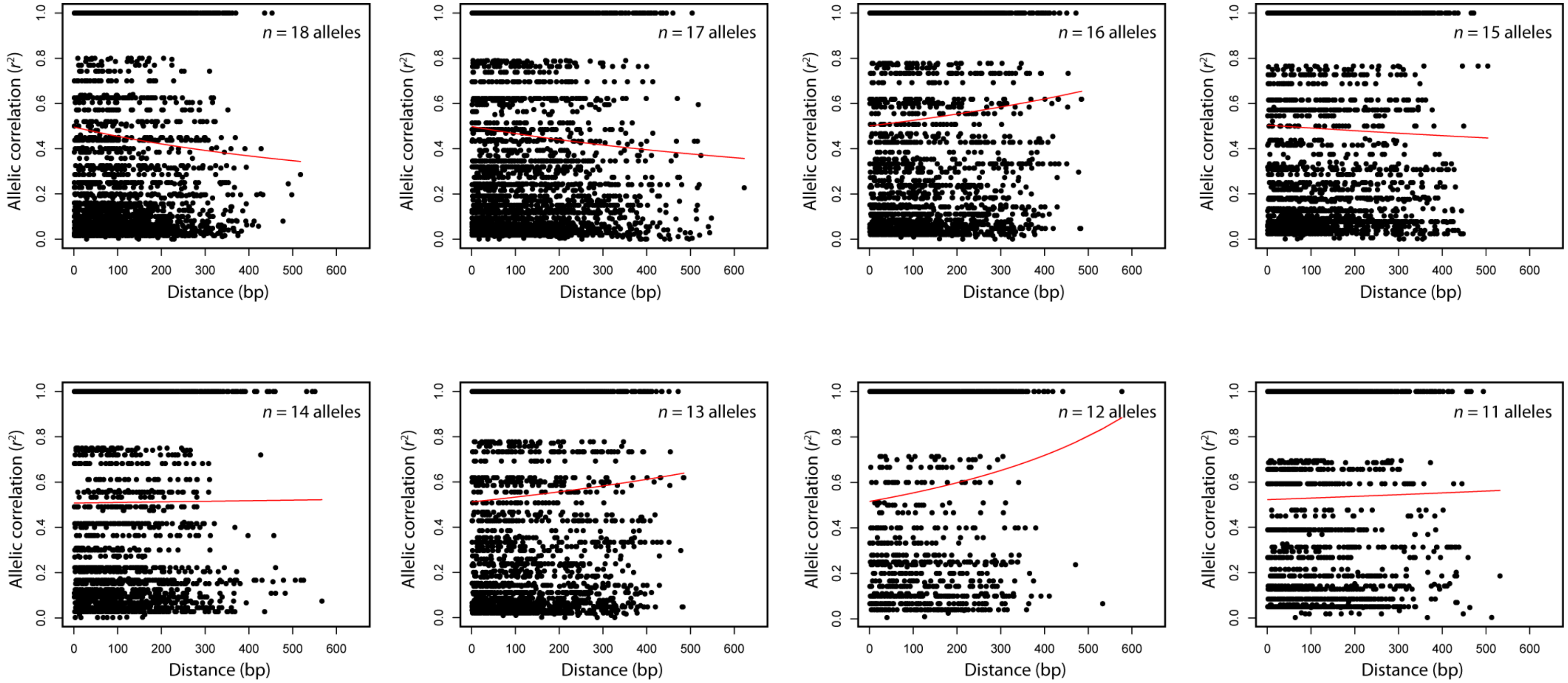

**Figure S14** decay of linkage disequilibrium, as measured using pairwise allelic correlations ($r^2$), with physical distance (bp) across coverage classes as estimated with data excluding singletons. Red lines give the expected value of $r^2$ following Remington *et al.* (2001). Intragenic pairs of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

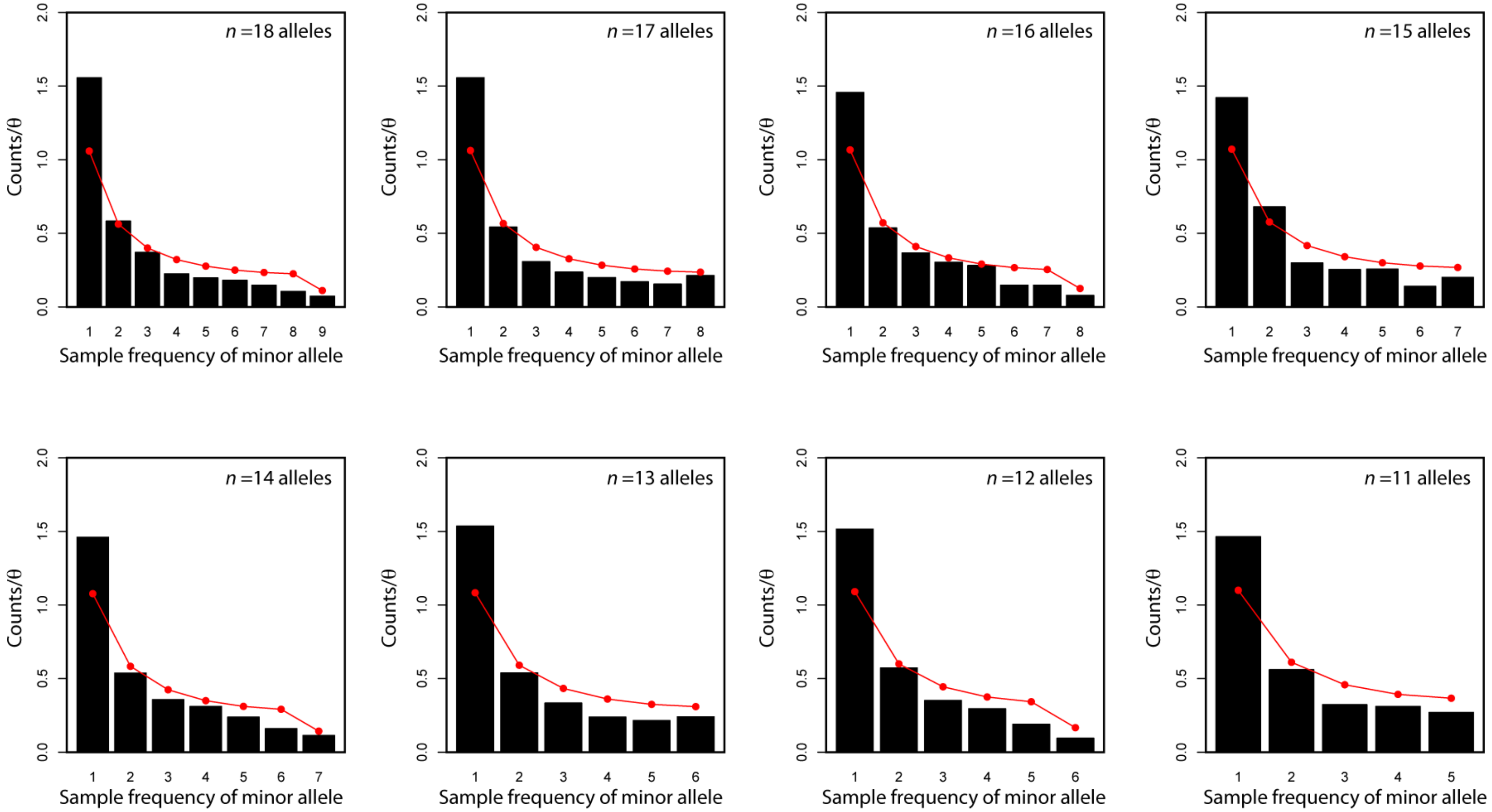

**Figure S15** Observed (black bars) and expected (red points and lines) folded site-frequency spectra for sample coverage classes from 11 to 18 sampled alleles. Counts were standardized by θ for ease of comparison across sample coverage classes. Goodness-of-fit tests to the expected distribution reveal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

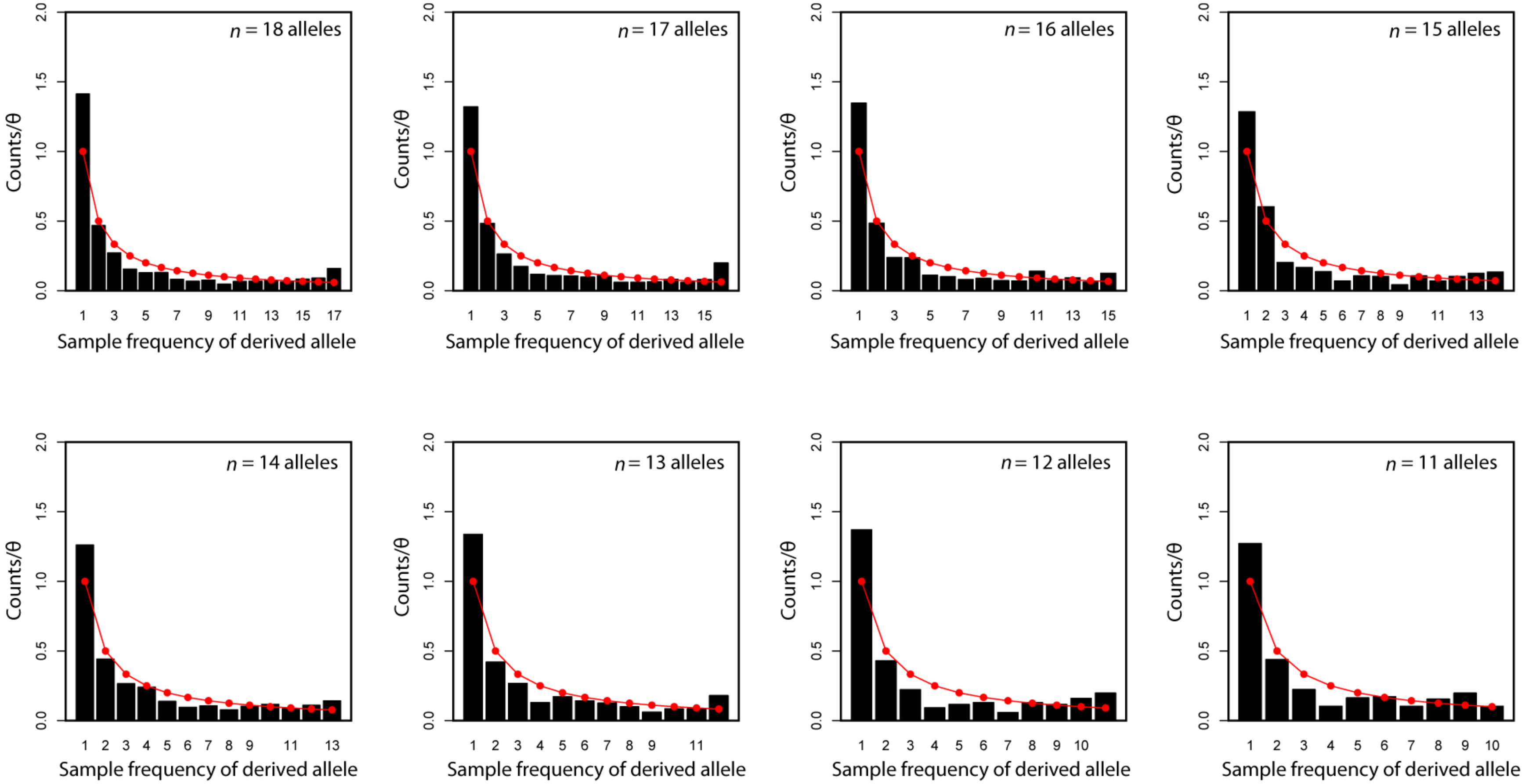

**Figure S16** Unfolded site-frequency spectra across sample coverage classes using *Pinus radiata* as the outgroup. Counts were standardized by θ for ease of comparison across sample coverage classes. Goodness-of-fit tests to the expected distribution reveal that all spectra deviate from neutral expectations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

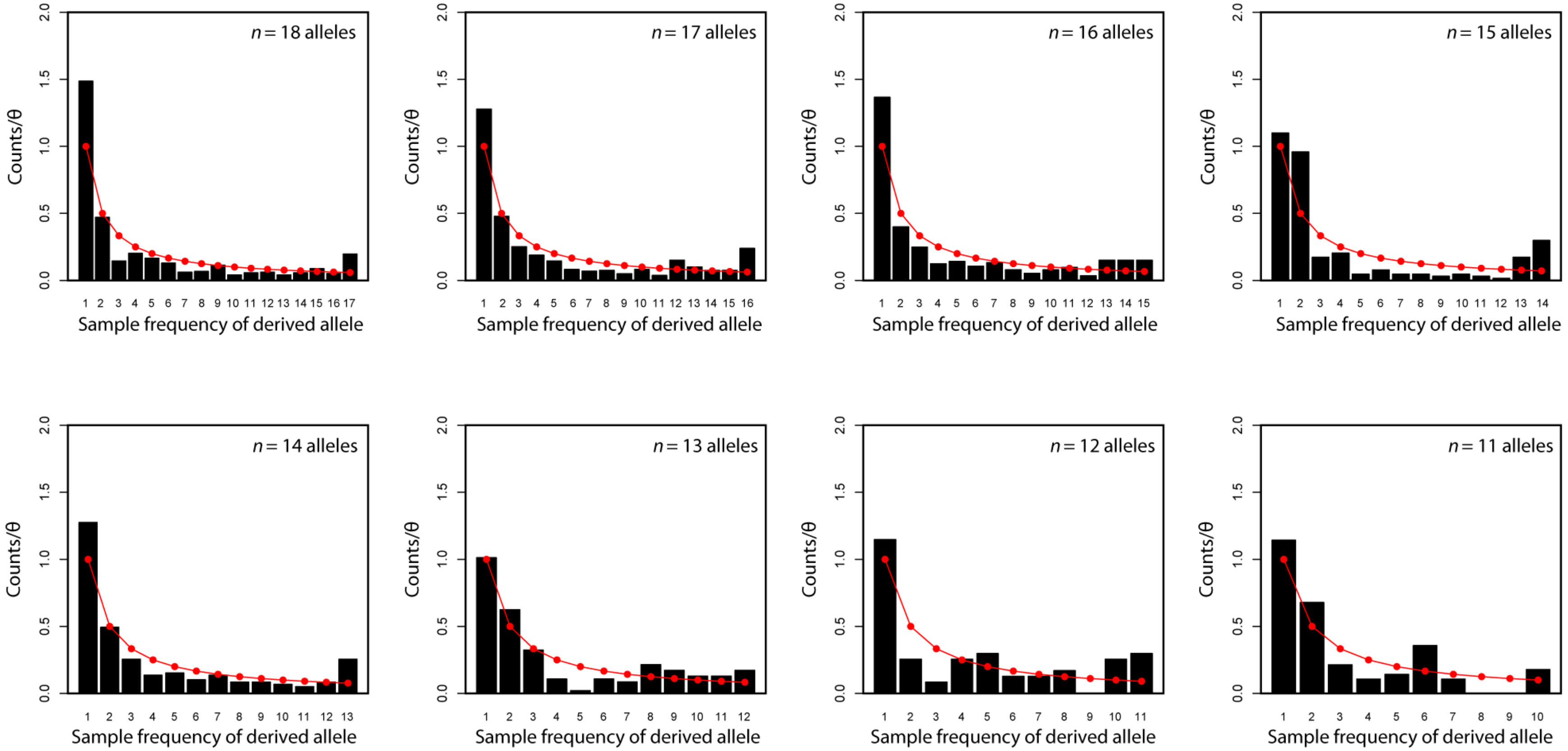

**Figure S17** Unfolded site-frequency spectra across sample coverage classes using *Pinus lambertiana* as the outgroup. Counts were standardized by $\theta$ for ease of comparison across sample coverage classes. Goodness-of-fit tests to the expected distribution reveal that all spectra deviate from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

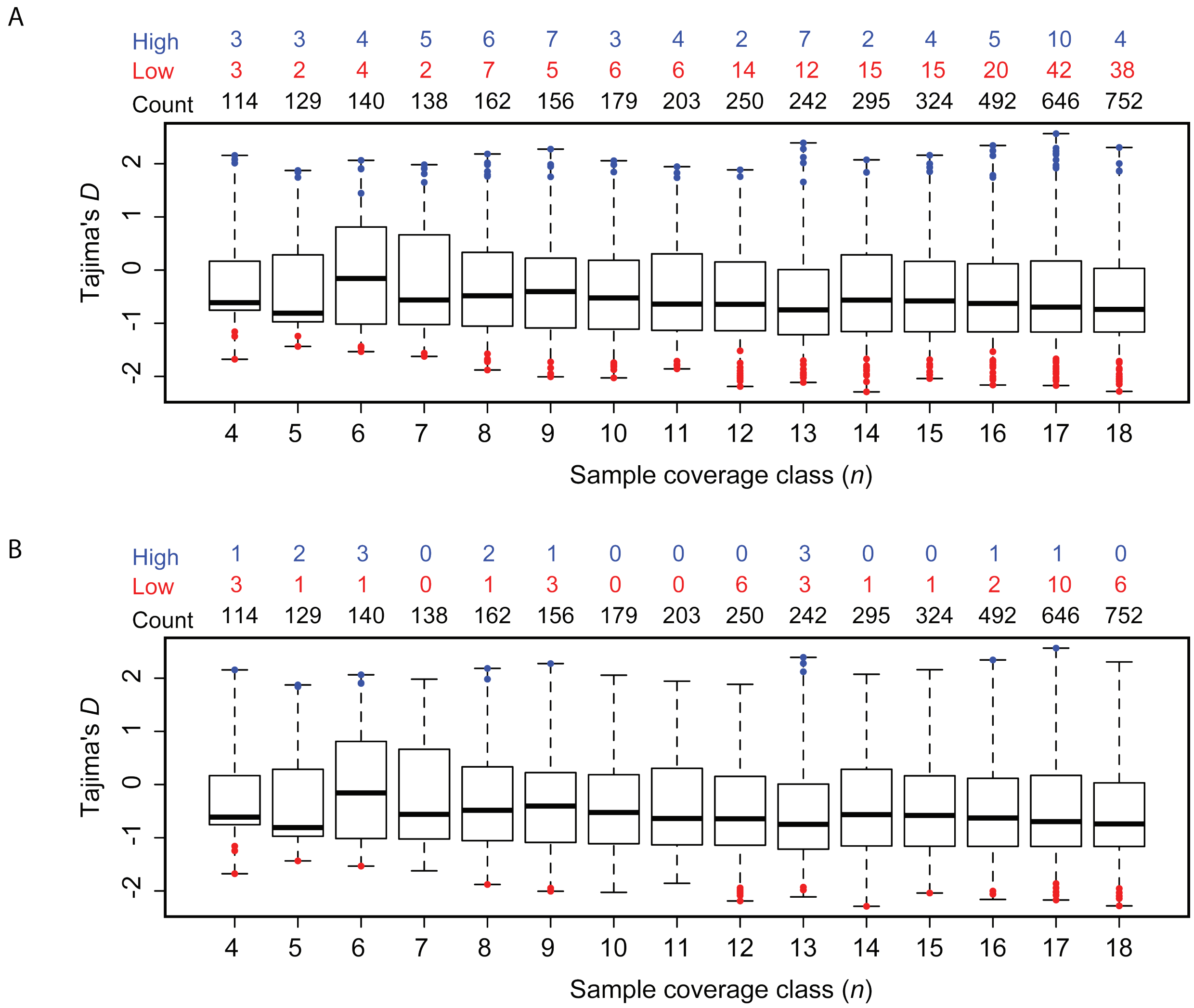

**Figure S18** Summary of the folded site-frequency spectrum using Tajima's *D*. Boxplots give the observed distributions across amplicons, while colored points denote outliers (red = lower tail, blue = upper tail) at the *P* = 0.05 level. Whiskers extend to the extremes of the observed data. (A) The 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.

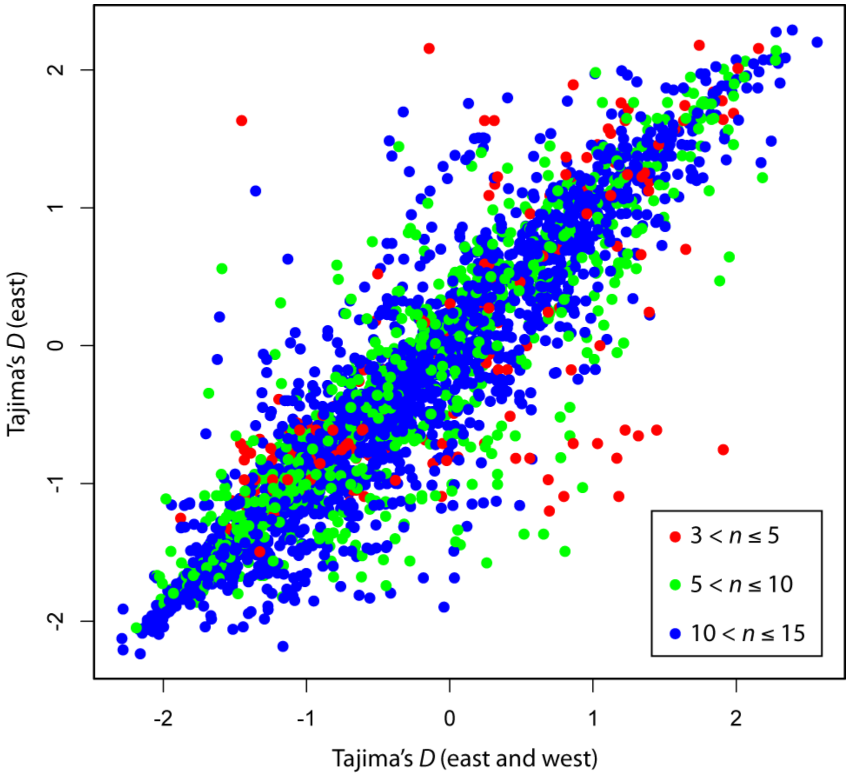

**Figure S19** Summary statistics of the site-frequency spectrum are correlated between the full sample set and the sample set trimmed to just samples obtained from east of the Mississippi River. Colors denote bins of sample coverage classes. The overall correlation structure does not differ among the three classes of sample coverage (ANCOVA: $F$ = 1.27, $df_1$ = 2, $df_2$ = 3,129, $P$ = 0.28, $P_{\mathrm{perm}}$ = 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.

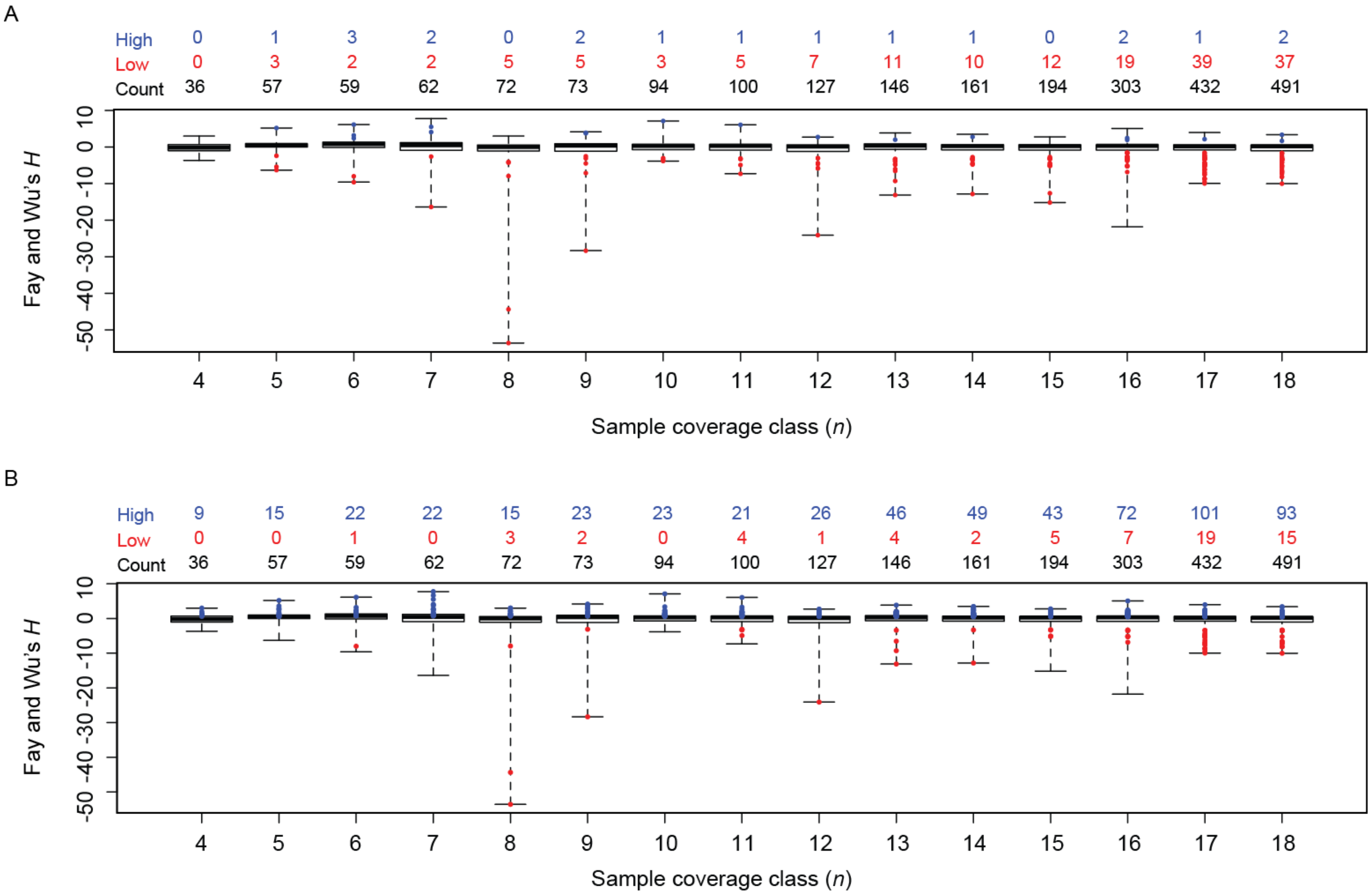

**Figure S20** Summary of Fay and Wu's *H* by coverage class for two neutral models – the standard neutral model (SNM) and the three-epoch model from Ersöz *et al.* (2010). Boxplots give the observed distributions across amplicons, while colored points denote outliers (red = lower tail, blue = upper 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.

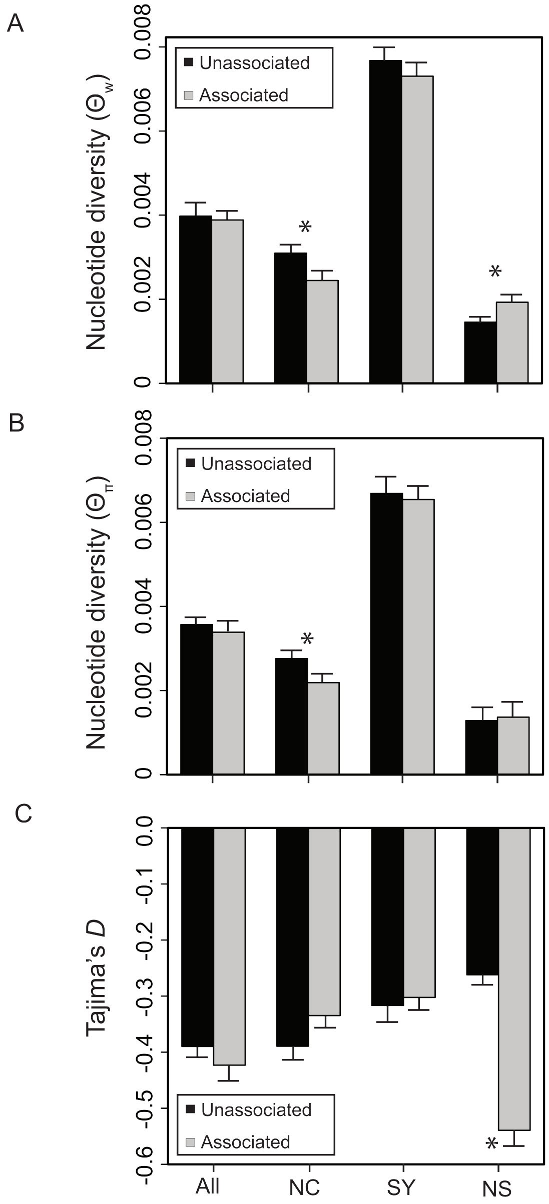

**Figure S21** Summary of differences in the site-frequency spectra for amplicons associated to at least one phenotype and those unassociated to a phenotype. Amplicons associated to at least one phenotype have too many rare variants at nonsynonymous sites, whereas they have too few rare variants at noncoding 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.