Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae)

Andrew J. Eckert¹,², Joost van Heerwaarden³, Jill L. Wegrzyn³, C. Dana Nelson⁴, Jeffrey Ross-Ibarra³, Santiago C. González-Martínez⁵ and David. B. Neale²,³,⁶,⁷

¹Section of Evolution and Ecology, University of California at Davis, Davis, CA 95616, USA
²Center for Population Biology, University of California at Davis, Davis, CA 95616, USA
³Department of Plant Sciences, University of California at Davis, Davis, CA 95616, USA
⁴Southern Institute of Forest Genetics, USDA Forest Service, Saucier, MS 39574, USA
⁵Department of Forest Systems and Resources, Forest Research Institute, CIFOR-INIA, Madrid 28040, Spain
⁶Institute of Forest Genetics, Pacific Southwest Research Station, U.S. Department of Agriculture, Davis, CA 95616, USA
⁷Corresponding author: David B. Neale, Department of Plant Sciences, University of California at Davis, One Shields Avenue, Davis, CA 95616; fax: (530) 754-9366; phone: (530) 754-8431; e-mail: dbneale@ucdavis.edu

Key words: aridity, environmental association, *Pinus taeda*, population structure, single nucleotide polymorphisms
ABSTRACT

Natural populations of forest trees exhibit striking phenotypic adaptations to diverse environmental gradients, thereby making them appealing subjects for the study of genes underlying ecologically relevant phenotypes. Here, we use a genome-wide dataset of single nucleotide polymorphisms genotyped across 3059 functional genes to study patterns of population structure and identify loci associated with aridity across the natural range of loblolly pine (*Pinus taeda* L.). Overall patterns of population structure, as inferred using principal components and Bayesian cluster analyses, were consistent with three genetic clusters likely resulting from expansions out of Pleistocene refugia located in Mexico and Florida. A novel application of association analysis, which removes the confounding effects of shared ancestry on correlations between genetic and environmental variation, identified five loci correlated to aridity. These loci were primarily involved with abiotic stress response to temperature and drought. A unique set of 24 loci was identified as $F_{ST}$ outliers based on the genetic clusters identified previously and after accounting for expansions out of Pleistocene refugia. These loci were involved with a diversity of physiological processes. Identification of non-overlapping sets of loci highlights the fundamental differences implicit in the use of either method, and suggests a pluralistic, yet complementary, approach to the identification of genes underlying ecologically relevant phenotypes.
INTRODUCTION

Environmental heterogeneity at multiple spatial scales influences the distribution of genetic variation across plant populations. Correlations between genetic variation and environmental gradients have been identified in a variety of plant species (Antonovics and Bradshaw 1970; Antonovics 1971; Hamrick and Allard 1972; Westfall and Conkle 1992; Linhart and Grant 1996; Mitton 1997; Savolainen et al. 2007), and such associations are often interpreted as evidence of natural selection (Allard et al. 1972; Dudley 1996a, 1996b; Gram and Sork 2001; Väsemägi and Primmer 2005; Parisod and Christin 2008). Renewed interest in identifying correlations between environmental and genetic variation has emerged as high-throughput sequencing and genotyping platforms are applied to functional genetic variation within natural populations of non-model species (cf. Joost et al. 2007; Namroud et al. 2008).

Forest trees illustrate clear phenotypic adaptations to environmental gradients at multiple spatial scales (Morgenstern 1996; Savolainen et al. 2007 and references therein). An extensive history of provenance, common garden and geneecological studies has established the highly polygenic basis of these adaptive traits (Langlet 1971; Namkoong 1979). One of the major abiotic stressors for conifers is water availability. Traits related to water-deficit stress (WDS) have a genetic basis for many conifers and variation at these traits is adaptive (Zhang and Marshall 1994; Aitken et al. 1995; Johnsen et al. 1999; Olivas-Garcia et al. 2000; Brendel et al. 2002; González-Martínez et al. 2007; Baltunis et al. 2008). Physiological responses to drought involve many different cellular and molecular pathways (Newton et al. 1991; Ingram and Barrels 1996), and functional and gene expression studies in Arabidopsis
have implicated several gene networks in WDS responses, as well as interactions between drought and cold-hardiness traits (BRAY 2004; SHINOZAKI AND YAMAGUCHI-SHINOZAKI 2000, 2007). Population genetic studies, moreover, have identified functionally diverse genes with skewed site-frequency spectra (GONZÁLEZ-MARTÍNEZ et al. 2006), extreme allele frequency differences across populations (EVENO et al. 2008), altered gene expression (DUBOS AND PLOMION 2003; WATKINSON et al. 2003; YANG AND LOOPSTRA 2005) or significant associations with WDS (GONZÁLEZ-MARTÍNEZ et al. 2008) for multiple pine species distributed across strong precipitation gradients.

This study investigates the potential association between genetic variation at individual loci and aridity gradients for loblolly pine (Pinus taeda L.). The predominant abiotic gradient across the range of loblolly pine is water availability. Water use efficiency is a WDS related trait, and in loblolly pine is heritable, but displays significant non-additive variance as well as strong environmental influences (WARREN et al. 2001) and genotype by environment interactions (BALTUNIS et al. 2008). The canonical approach to searching for environmental associations would be to correlate measures of genetic diversity (e.g. allele frequencies, heterozygosity) to environmental variation related to drought stress (LINHART AND GRANT 1996; MITTON 1997; MITTON et al. 1998; VASEMÄGI AND PRIMMER 2005). One limitation to this approach is that environment is likely to be confounded with geography and, by proxy, overall genetic structure. Correction for population structure is common in association studies that aim at identifying markers contributing to phenotypes (cf. Yu et al. 2006). An obvious solution to the confounding of genetic structure with environmental variation, therefore, is to
apply existing association approaches to environmental data, treating the environment methodologically as a phenotype.

Here, we use two genome-wide marker datasets to address the following questions: (1) What are the patterns of population structure across the range of loblolly pine? (2) What is the degree of confounding between environmental variation and population structure due to geography? (3) Which loci are associated to water availability across the range of loblolly pine? (4) Are loci with strong genotypic correlations to aridity also those with extreme allele frequency differences among populations? In answering these questions, we highlight the need for further integration of environmental and genetic data in genome scans for loci subject to natural selection, as well as the promise of combining complementary approaches for the identification of functionally important genetic variation within natural populations (cf. VASEMÄGI AND PRIMMER 2005).

MATERIALS AND METHODS

Focal Species

Loblolly pine (*Pinus taeda* L.) is distributed throughout the southeastern United States, ranging from Texas to Delaware. Its 370,000 km² range is divided primarily by the Mississippi River Valley, with 60% of the distribution range being located east of the Mississippi River (AL-RABAB’AH AND WILLIAMS 2002). Isozyme and nuclear SSR loci illustrate moderate genetic differentiation between populations located to the east and west of the Mississippi River Valley, as well putative population contraction in the western most populations (WELLS *et al*. 1991; SCHMIDTLING *et al*. 1999; AL-RABAB’AH AND
Williams 2002, 2004; Xu et al. 2008). A review of phylogeographical patterns in unglaciated eastern North America identified six major patterns, of which loblolly pine conforms to the Mississippi River discontinuity (Soltis et al. 2006). The structure of this discontinuity is consistent with a dual Pleistocene refugial model, which has also been used to explain differential growth abilities, disease resistance and concentrations of secondary metabolites among families located across this discontinuity (Wells and Wakeley 1966; Squillace and Wells 1981; Schmidtling 2003).

**Sampling**

Needle tissue was collected from 907 largely unrelated trees sampled across the natural range of loblolly pine (Fig. 1). Seven hundred of these trees are first-generation selections (i.e. trees grown from wild collected seed) with known source localities. These samples are geo-referenced by county ($n_{\text{counties}} = 188$). The average number of sampled trees per county was 4 ± 6 (range: 1 – 67). The remaining 207 are second-generation selections (i.e. trees resulting from crosses between the first generation selections). These trees also comprise two experimental populations currently being used for association mapping – Weyerhaeuser (WeyCo; cf. González-Martínez et al. 2007) and North Carolina State University (NCSU; cf. Cumbie et al. unpublished). Total genomic DNA was isolated from each sample at the USDA National Forest Genetics Laboratory (NFGEL, Placerville, CA) using DNeasy® Plant Kits (Qiagen, Valencia, CA, USA) following the manufacturer’s protocol.

**Marker Discovery and Genotyping**
We utilized two sets of molecular markers. The first set comprises 23 unlinked nuclear simple sequence repeat (SSRs) markers selected from the PtTX marker set for medium to high polymorphism rate and full coverage of the linkage map (Auckland et al. 2002; González-Martínez et al. 2006, 2007). The second set comprises approximately 23,000 single nucleotide polymorphism (SNP) markers, of which we chose 7,216 for genotyping, identified through the resequencing of 7,535 unique expressed sequence tag (EST) contigs in 18 loblolly pine haploid megagametophytes. These SNPs cover the entire linkage map for loblolly pine (TG accession: TG091; http://dendrome.ucdavis.edu/cmap/), with 117 ± 18 SNPs mapped on average per linkage group for a total of 1815 mapped SNPs with an average distance of 1.2 ± 1.1 cM between SNPs. Selection of SNPs for genotyping was based largely on quality scores derived from the original sequence data and not on functional or site annotations. This ensured thorough coverage of the available sequence resource for loblolly pine (cf. http://dendrome.ucdavis.edu/adept2/). Genotyping of SNPs utilizing the Infinium platform (Illumina, San Diego, CA, USA) was carried out at the University of California Davis Genome Center. Arrays were imaged on a Bead Array reader (Illumina) and genotype calling was performed using BeadStudio v. 3.1.3.0 (Illumina). Further information regarding the discovery, annotation and PCR, genotyping and DNA sequencing protocols for both marker types is available in File S1. The complete data are available as Files S2 and S3.

For each marker locus we calculated observed \((H_O)\) and expected \((H_E)\) heterozygosity as well as Wright’s inbreeding coefficient \((F_{IS} = 1 – H_O/H_E)\). Loci with extreme values of \(F_{IS} (|F_{IS}| > 0.25)\) were removed prior to analysis. We used Fisher
exact tests (GUO AND THOMPSON 1992) with Bonferroni corrections to test for Hardy-Weinberg equilibrium for each SNP and SSR marker, respectively. All analyses were preformed using the R environment (R CORE DEVELOPMENT TEAM 2007).

Environmental data

Climate data were gathered from the WORLDCLIM 2.5 minute geographical information system (GIS) layer using Diva-GIS ver. 5.4 (HIJIMANS et al. 2005; available at http://www.diva-gis.org/). Monthly minimum and maximum temperatures, monthly precipitation and 19 bioclimatic variables were obtained from this layer (Table S1). The temperature and precipitation data were used to estimate potential evapotranspiration (PET) with the method of THORNTHWAITE (1948). An aridity index (AI) was defined as the ratio of precipitation to PET (File S1), with this ratio being defined quarterly. Annual quarters were defined starting with January 1st through March 31st as quarter one and are labeled as AI1 through AI4. We focus on aridity because it encapsulates water availability as a function of temperature and precipitation. Thus, the remainder of the paper focuses solely on aridity.

Patterns of population structure

Population genetic structure was analyzed by means of principal component analysis (PCA) on genotypes from individual trees. Briefly, PCA was performed on the normalized $n \times m$ matrix, $M$, of genotypes, where $n$ is the number of trees and $m$ is the number of loci. Similar analyses were conducted for SSR markers using a method that accounts for the dependence among alleles at a locus (VAN HEERWAARDEN et al. 2010).
The eigenvalues corresponding to the principal components (PCs) were inspected to
determine the number of major independent axes of genetic differentiation in the data.
Following outlier removal and reanalysis of $M$, significance of PCs was determined by
comparing the value of each standardized eigenvalue to a Tracy-Widom (TW)
distribution (PATTERSON et al. 2006). Outliers were defined as trees with PC scores
greater than six standard deviations away from the mean for any of the first 10 PCs.
Trees were assigned to discrete genetic clusters based on $K - 1$ significant PCs, where
$K$ is the number of clusters being considered (PASCHOU et al. 2007; VAN HEERWAARDEN
et al. 2010). Specifically, Ward’s hierarchical clustering algorithm was applied to the
matrix of Euclidean distances, calculated from the significant PCs, and the resulting
dendrogram was used to assign individuals to each cluster using the CUTREE function in
R.

For comparison, we also used the program STRUCTURE ver. 2.2 to infer the
number of genetic clusters and membership coefficients within those clusters using the
23 nuclear SSR markers (PRITCHARD et al. 2000; FALUSH et al. 2003). Analysis using the
SNP data was avoided because of lack of convergence among runs likely related to
insufficient run times for the MCMC sampler (data not shown). We varied the number of
genetic clusters ($K$) from one to 12 and for each value ran 50 independent Markov chain
Monte Carlo (MCMC) simulations. Each run was carried out for $1.2 \times 10^7$ steps, with first
$2.0 \times 10^6$ steps being discarded as burn-in. We assumed further that allele frequencies
were correlated among populations and that our data contained admixed trees. The
optimal value of $K$ was determined using the $\Delta K$ method (EVANNO et al. 2005) and by
inspection of the relationship between the log-probability of the data and $K$. Average
admixtorest coefficients in all cases were estimated for each value of $K$ using the 
*LargeKGreedy* algorithm with 1000 random input orders as implemented in the program 
CLUMPP ver. 1.1 (Jakobsson and Rosenberg 2007). These values were visualized 
using bar plots constructed with DISTRUCT ver. 1.1 (Rosenberg 2004).

Associations between population structure, geography and environment were 
studied by fitting a general linear model to each measure of structure (assignment 
probabilities, PCs) with per-county aridity indices and latitude and longitude as 
explanatory variables. The relation between the different environmental variables and 
genetic assignment was visualized by a biplot of the PCA on the environmental 
variables with plotted points colored according to their corresponding genetic cluster.

**Environmental associations and outlier analysis**

Loci associated to environment were identified using a standard association 
mapping approach, substituting aridity for phenotype. Analysis was done separately for 
each variable and each SNP. Each tree was assigned a “phenotype” consisting of its 
corresponding county-level aridity index. Following Price et al. (2006), we used PCA 
analysis to correct for spurious associations due to confounding of ancestry and aridity. 
Briefly, PCA is performed as described above, but excluding the target SNP. Two 
vectors of ancestry-corrected residuals are obtained by multiple linear regression on 
environmental and genotypic values, using the $k$ significant genetic PCs as independent 
variables. Association between genotype and aridity is described by the squared 
correlation $r^2$ between the two vectors. For each SNP, scored in $N$ individuals, the test 
statistic is calculated as $(N-k-1)r^2$, which is approximately $\chi^2$ distributed with one degree
of freedom (PRICE et al. 2006). SNP loci showing the strongest association with different aridity indices were identified by Q-Q analysis of P-values. The magnitude of environmental differences among SNPs was evaluated using a general linear model with environment as a dependent variable and corrected genotypic values as explanatory variables. Differences of environment among SNP genotypes were evaluated using a general linear model with environment as a dependent variable and corrected genotypic values as explanatory variables. Multiple testing was accounted for using the false discovery rate (FDR) method of STOREY AND TIBSHIRANI (2003) with a significance threshold of $Q = 0.05$, although this method, or any multiple test correction method assuming multiple independent tests of the same null hypothesis, will be conservative due to the correlations among environmental variables.

For comparison, we searched for patterns of adaptive differentiation among populations using FDIST2 (BEAUMONT AND NICHOLS 1996; BEAUMONT AND BALDING 2004), with the populations being those identified using PCA or STRUCTURE. Outlier analyses were conducted only for the SNP data, as opposed to the SSR data, because these represent functional variation likely to be targets of natural selection. We used results from clustering of SSRs and SNPs to define populations because the range of loblolly pine is continuous, the number of discrete populations is relatively unknown and studies employing $F_{ST}$ outlier analysis in forest trees justify a priori population definitions with some form of structure analysis (EVENO et al. 2008; NAMROUD et al. 2008). We simulated $1.0 \times 10^6$ loci under two null models using ms (HUDSON 2002) - an island model and a two Pleistocene refugial model (File S1, Fig. S1). The latter model was used to account for effects of historical demography on the null expectation of $F_{ST}$
across loci (cf. EXCOFFIER et al. 2009). An iterative approach was used to adjust the median $F_{ST}$ (WEIR AND COCKERHAM 1984) and $H_E$ until they were within 5% of observed values. A modified form of the msstats software developed with the libsequence C++ library (THORNTON 2003; available from: http://www.rilab.org/code/random_code.html) was used to estimate $F_{ST}$ and $H_E$, while the CPlot and PVJ programs, which are distributed as part of FDIST2, were used to estimate 99% quantiles of the null distribution and $P$-values, respectively. Extreme values of $F_{ST}$ conditional on expected heterozygosity were defined as those that lay above the 99% quantile of the null distribution.

RESULTS

Genotyping summary

To our knowledge, this is the first successful application of the Illumina Infinium genotyping platform to a non-model plant species. Further details concerning conversion rates and quality scores are available in File S1. We selected one SNP that was typed successfully per EST locus for further analysis ($n = 3082$ SNPs). Forty-five of these SNPs deviated significantly from HWE proportions, with 23 also having $F_{IS}$ values $> 0.25$. These 23 were removed prior to further analysis. An ascertainment bias was also apparent for the frequency distribution of the minor allele across all 3059 SNP loci (File S1; Fig. S2). Call rates across the 23 SSR loci averaged 95%, with 18 of the 23 loci deviating significantly from HWE proportions. All 23 SSR loci had values of $F_{IS} < 0.25$ when only first generation selections with known source localities were considered.
Integration of these datasets resulted in 622 trees sampled from 167 of the 188 county locations typed for 3059 SNPs and 23 SSRs. Summaries of these data are located in Table 1, with further descriptions of genotyping results presented in File S1 and Figure S3. These data were used to infer patterns of population structure and to search for multivariate and locus-specific environmental associations.

**Patterns of environmental heterogeneity**

Strong correlations exist among environmental measures, as well as between those measures and geographical location (Fig. S4). In general, sites located along the Gulf Coast Plain have the highest annual precipitation (1650 mm), whereas sites located in the northeast and southwest have the lowest (980 mm). The highest average annual temperatures are observed in the southeast (20.8°C), while sites located in the northeast have the lowest (12.8°C). These trends are apparent for measures of aridity both spatially and temporally (Fig. 1). In general, all sites have water surpluses in the winter, with those located along the Gulf Coast Plain having the largest. Water deficits are apparent in the northeast (spring) and southwest (summer) as the year progresses. By fall, all sites have water surpluses, with those located in the extreme southeast being the driest.

**Patterns of population structure**

Patterns of population structure for loblolly pine are accounted for primarily by the Mississippi River discontinuity. This is apparent for results from PCA and STRUCTURE using both SSRs and SNPs. PCA analysis on the SNP data revealed the presence of
seven significant PCs defining eight genetic clusters. Visual inspection of the
eigenvalues, however, shows the presence of two major PCs explaining 2.4% of the
total variation (56% of the significant variation), which indicates the presence of three
clearly differentiated clusters (Figs. 2 and 3). The remaining five significant clusters lack
a strong geographical basis. These three clusters are largely divided along the
Mississippi River Valley, with a further division of the eastern cluster into Gulf and
Atlantic Coast clusters. Similar results were observed for the SSR markers using PCA,
with two of the three significant PCs clustered across the Mississippi River discontinuity
(Figs. 2 and 3).

The optimal value of $K$ was two as determined by the $\Delta K$ statistic using
STRUCTURE with the SSRs. Inspection of bar plots for the admixture coefficient when
$K = 2$, averaged across the 50 replicated MCMC runs, indicated that each cluster is
geographically based, with one cluster corresponding to trees that are located primarily
west of the Mississippi River and the other to those located to the east (Fig. S5). The
use of $\Delta K$ to choose an optimal value of $K = 2$, however, is difficult, because $\Delta K$ in this
case compares the lack of structure ($K = 1$) to some structure ($K = 2$ and 3). We chose
to use values of three and five for further analysis, because $K = 3$ had the second
largest value of $\Delta K$ and $K = 5$ is where the median value of log-probability of the data
leveled off (Fig. 2). Geographical trends in cluster assignments for these values of $K$
reflected the west-east division, as well as further divisions of the eastern cluster along
a southwest to northeast axis (Fig. 3).

The method of clustering did not dramatically affect the assignment of trees to
clusters or the geographical basis of those clusters (Fig. 3). The use of PCA and
STRUCTURE on the SSR data yielded similar patterns for $K = 3$, with values of $F_{ST}$ calculated for each assignment scheme being significantly correlated ($r^2 = 0.50$, $P < 2.2 \times 10^{-16}$). Some discrepancy between the methods was apparent. For example, PCA placed the trees located in Livingston county, Louisiana, into the eastern population, while STRUCTURE associated them with the western population, which has been noted previously (SCHMIDTLING 2003).

Multilocus population structure was strongly confounded with aridity. Geography was correlated with patterns of population structure and aridity for all clustering methods and choices of $K$ (Tables 2 and S2). The first two PCs from PCA using the SNP data were correlated with geography and aridity. The four aridity indices, longitude and latitude together explained 61% of the variance in PC1 and 39% of the variance in PC2, thus illustrating that the genetic clusters identified previously differ with respect to aridity (Fig. S6). All aridity indices also showed significant correlations with either longitude or latitude, as well (Fig. S4). This was reflected in the fact that aridity and geography by themselves explained much of the variance in PC1 ($r^2 = 0.46$ for all four aridity indices, $P < 2.2 \times 10^{-16}$; $r^2 = 0.51$ for latitude and longitude, $P < 2.2 \times 10^{-16}$) and PC2 ($r^2 = 0.25$, $P < 2.2 \times 10^{-16}$ for all four aridity indices; $r^2 = 0.16$ for latitude and longitude, $P < 2.2 \times 10^{-16}$). Three of the four aridity indices (AI1, AI2 and AI4) were also significantly correlated (min $r^2 = 0.53$, $P < 2.2 \times 10^{-16}$). A visual representation of the correlations between these variables and genetic structure is given in Figure 2. Similar yet weaker correlations were apparent for PCs and STRUCTURE assignment probabilities using the SSR data (Table 2).
Environmental associations and outlier analysis

Association analysis resulted in the identification of five loci with significant correlations to aridity (Table 3, Fig. S7). The strongest associations were between four loci and aridity during the second quarter (AI2), with subsets of these loci also associating with aridity during the first (AI1) and fourth quarters (AI4). Only a single locus was found to associate significantly with aridity during the third quarter (AI3). The four significant loci associated with aridity during quarters one, two and four together explained 9.2% of the variance in AI2 (Fig. 4), 5.6% of the variance in AI1 and 4.7% of the variance in AI4 (15.2%, 9.1%, and 8.1% of the ancestry corrected environment). For comparison, the average SNP explained 0.1% ± 0.2% of the variance in AI2 (0.2% ± 0.3% for ancestry corrected environment). The single locus associated with AI3 explained 2.6% of the variation (3.0% of variation in ancestry corrected environment).

All five significant SNPs were located in loci with high sequence similarity to coding sequences in Arabidopsis that primarily affect abiotic and pathogenic stress responses (see Discussion; Table 3). Two of these five SNPs are mapped to linkage group three. Three of the five SNPs are located in synonymous positions, while the remaining two are located in an intron and a 3’ UTR.

Clusters defined using the SSR markers resulted in few outlier SNPs, patterns indicative of residual within cluster substructure and multilocus values of $F_{ST}$ less than 1% (Fig. 5). Most of these outliers can be accounted for by including a dual Pleistocene refugia model as the null model. This model fit the frequency distribution of the minor allele as well as, if not better than, that of drift alone or an island model (File S1, Fig. S3). Three loci had significantly elevated $F_{ST}$ estimates, however, when $K = 5$. 
A different pattern emerges when the cluster memberships are based on SNP loci. In this case, 24 and 15 loci are significant outliers after accounting for demography when $K = 3$ or $K = 5$, respectively. In both cases, the multilocus values of $F_{ST}$ (0.022 for $K = 3$, 0.016 for $K = 5$) are more similar to those reported previously (Schmidtling 2003). Only seven loci are shared between lists of outlier loci for each value of $K$, however the outlier loci unique to a particular value of $K$ were always located in the upper tail of the distribution for $F_{ST}$ of the alternative value of $K$. We focus on the results when $K = 3$; results for $K = 5$ are shown in Table S3 and Figures S8 and S9. All outlier loci have values of $F_{ST}$ that are approximately six to 12-fold larger than the average $F_{ST}$ across all SNPs (Table 4).

Nineteen of the 24 loci identified as outliers had significant tBLASTx hits to annotated loci in Arabidopsis (Table 4). The remaining five had little sequence similarity to known protein sequence in plants, although high nucleotide sequence similarity was detected in Sitka spruce (*Picea sitchensis* (Bong.) Carr.) EST libraries for four of those five using blastn. Putative functions of gene products in Arabidopsis range from growth regulators to pathogenic stress response. Approximately half of the SNPs located in outlier loci are nonsynonymous point mutations, with the remainder being located largely in synonymous positions. Eleven of the 24 outliers are mapped and span five different linkage groups. Six of those 11 are located in close proximity (< 2 cM) on linkage group eight.

There was no overlap between the loci associated to aridity gradients and those identified as outliers. Loci associated with aridity gradients had values of $F_{ST}$ within the range of the mean value across loci ($F_{ST} = 0 - 0.035$), while $F_{ST}$ outliers showed no
meaningful correlation with aridity gradients (maximum $r^2 = 0.008$). Correlations between $F_{ST}$ and $\chi^2$ statistics for aridity during each quarter, moreover, were non-significant and the percent variance explained was less than 0.5% in all cases. The genomic location of both sets of loci were different, with two of the five loci associated to aridity being located approximately four cM apart on linkage group three and the $F_{ST}$ outliers being spread across linkage groups one, three, six, seven and eight (Fig. 6).

DISCUSSION

Population structure in loblolly pine

This study presents the first genome-wide analysis for population structure of functional genetic variation among natural populations of loblolly pine. Our results are broadly consistent with previous conclusions regarding patterns of population structure for loblolly pine and conifers in general (LEDIG 1998; SCHMIDTLING et al. 1999; AL-RABAB’AH AND WILLIAMS 2002, 2004; SCHMIDTLING 2003; GONZÁLEZ-MARTÍNEZ et al. 2006, 2007; Xu et al. 2008) - genetic structure is weak, primarily accounted for by the Mississippi River discontinuity and is consistent with a dual Pleistocene refugial model. Our results also lay the foundation for performing population structure corrections during association mapping.

We detected further substructure across the range of loblolly pine that was apparent only in the SNP dataset. While this result could possibly be a function of the ascertainment bias for the SNP data, such biases are not expected to affect the relative placement of trees in PC space (McVean 2009). The same three genetic clusters accounted for most of the geographical patterns observed for levels of population
structure across values of $K$ ranging from three to eight (Fig. S10), however, thus justifying our focus on $K = 3$. This is also reflected in the correlations of genetic structure to aridity (Table 2), because two to three clusters accounted for most of the significant correlations. Geographical trends in cluster assignments show clearly that trees in Florida, a putative refugium, are strongly clustered with those along the Atlantic rather than Gulf Coastal Plain and that trees located to the west of the Mississippi River are genetically distinct. This pattern supports expansion from dual refugia, one located in southern Florida and one in southern Texas or Mexico (Schmidting 2003).

Novel to the analyses presented here is the placement of the Gulf Coast trees into a unique genetic cluster (Fig. 3C). This result is consistent with a scenario of expansion with differentiation or admixture. These trees were intermediate to the eastern and western clusters along the first PC, which is consistent with admixture. Testing these hypotheses, however, is complicated by introgression between loblolly pine and closely related sympatric pines (Chen et al. 2004), shared ancestral polymorphism (Syring et al. 2008) and sample size differences among the populations considered (McVean 2009).

**Environmental associations and $F_{ST}$ outliers**

Environmental association analysis identified five genes associated to aridity gradients. The primary functions of gene products encoded by these loci were abiotic and biotic stress responses. All five loci have putative orthologs in Arabidopsis that are responsive to abscisic (ABA) or jasmonic (JA) acid, two plant hormones with well-documented correlations to abiotic stress responses (Glazebrook et al. 2003; Wong et
al. 2006; WASTERNACK 2007; FABRO et al. 2008; MIZUNO AND YAMASHINO 2008). In addition, gene expression for four of the five loci in Arabidopsis or close relatives was repressed at low temperatures (KREPS et al. 2002; VERGNOLLE et al. 2005; WONG et al. 2006), with one locus significantly upregulated in the presence of chitin (LIBAULT et al. 2007) and another upregulated during infection by biotrophic fungi (FABRO et al. 2008). The protein products encoded by these genes also likely affect general osmotic stress responses because drought cold-tolerance responses are often correlated (BECK et al. 2007). These five genes did not have extreme values of $F_{ST}$ for any of the six clustering assignments, with values similar to the mean across all loci in each case.

$F_{ST}$ outliers encoded proteins involved in a wide range of plant processes including response to viral infection (0-13230; NICAISE et al. 2007), cuticular wax biosynthesis (2-1205; COSTAGLIOLI et al. 2005), floral and gametophytic development (2-3591; TAN AND IRISH 2006) and structural components of photosystem complexes (CL1799Contig1, JANSSON 1994). Additional loci have been previously associated with growth (0-14415), leaf nitrogen content (2-1087) and pitch canker resistance (1-3324) for loblolly pine (CUMBIE et al. unpublished; QUESADA et al. unpublished). None of these 24 loci, nor the additional eight $F_{ST}$ outlier loci for $K = 5$, illustrated strong correlations to aridity.

Many of the outlier loci showed striking geographical trends along the major axis of differentiation (southwest to northeast), with some approaching fixation of different alleles at opposite ends of this axis (Fig. S11). One of these was locus CL3949Contig1, which encodes a peptidyl-prolyl cis-trans isomerase affecting protein folding and signal transduction in Arabidopsis (CHOU AND GASSER 1997). A SNP in this locus was an $F_{ST}$
outlier when $K = 5$ (Table S3) and was in the upper tail of the distribution of $F_{ST}$ for $K = 3$. A putative ortholog of this locus in coastal Dougals-fir (Pseudotsuga menziesii (Mirb.) Franco var. menziesii) was also identified as an $F_{ST}$ outlier during a scan across candidate genes for associations to cold-hardiness phenotypes (Eckert et al. 2009).

Trees in northeast part of the range for loblolly pine were almost fixed for the A allele at this SNP, whereas the C allele was at high frequency in the southwest. Genome-wide datasets thus open the door to comparative association analysis across natural populations of conifers, where replication is across evolutionary lineages sampled across similar environmental gradients (cf. Turner et al. 2008).

Comparison between environmental association and $F_{ST}$ outlier approaches

A multitude of correlations between environmental variables and genetic variation has been noted across diverse taxa (Mitton 1997). To our knowledge this is the first study to use the analytical machinery of association analysis to discover environmental associations between functional genetic markers and environmental gradients. This approach identified an independent set of genes relative to the $F_{ST}$ outlier approach, which highlights the fundamental differences between these methods. The association approach used here assesses the effect of natural selection along specific environmental gradients, while $F_{ST}$ outlier methods aim to identify loci influenced by natural selection driving allele frequency differences among populations (i.e. ancestral groups) and are thus agnostic about the environmental gradients driving the extreme values of $F_{ST}$. We highlight the advantages and appropriate uses of each method by
comparing and contrasting the interpretations attributed to significant results for these two approaches (see also LATTA 1998; BARTON 1999; LE CORRE AND KREMER 2003).

Environmental gradients are defined *a priori* in the association approach and correlation to them tested after corrections for population structure. Significant associations to an environmental gradient, therefore, likely represent those polymorphisms underlying functional responses to that gradient. In contrast, *post hoc* interpretations of environmental differences are attributed to the cause of $F_{ST}$ outliers. Geography, however, can create genetic structure that is correlated to environment solely through neutral processes such as barriers to gene flow, distance effects and historical population size and range changes (MANEL et al. 2003; STORFER et al. 2008). As seen here, aridity does differ significantly among the identified genetic clusters and a tempting conclusion to draw about $F_{ST}$ outliers is one related to WDS. This implies that caution should be used when interpreting the biological significance of outlier loci, especially if overall correlations of structure to geography and environmental heterogeneity have not been assessed.

The association approach does not require identification of discrete genetic clusters. As illustrated here, the process of defining populations affects the identification of $F_{ST}$ outliers. Although geographical patterns of cluster membership did not change dramatically between methods, they did become clearer in the full SNP dataset. Differing definitions of populations may, therefore, lead to spurious results even if the island model is a good descriptor of the underlying genetic structure, unless the “islands” are known without error. A better approach in this case would be to define populations based on environment, while accounting for within population substructure,
rather than neutral marker based estimates. This is often done for species with subdivided ranges by organizing population samples along environmental or geographical gradients (cf. Eveno et al. 2008).

The identification of genetic clusters may be difficult or inappropriate for species such as loblolly pine that are distributed continuously across large geographical expanses, and for which paleobotanical (cf. references in Schmidtling 2003) and population genetic evidence suggests historical fluctuations in population size (Waples and Gaggiotti 2006; Guillot 2009). Invocation of the island model in these cases to derive the null distribution of test statistics will result in undesirable statistical behaviors.

Populations of most North American conifers are likely far from demographic equilibrium and acknowledging this may lead to better inferences of selection (Westfall and Millar 2004). As illustrated here, incorporation of demographic models into $F_{ST}$ approaches dramatically reduces the number of outliers detected (Fig. 5). Even under equilibrium, however, misspecification of the hierarchical nature of population structure may increase the false positive rate by nearly an order of magnitude (Excoffier et al. 2009). Results from analyses of $F_{ST}$ outlier analyses that do not consider these scenarios (cf. Namroud et al. 2008) should be viewed with some skepticism.

In contrast to association methods, the $F_{ST}$ outlier approach identifies loci independently of environment. Genome-wide scans for outliers thus enable identification of loci underlying phenotypic responses to diverse environmental gradients correlated to ancestry (but see McKay and Latta 2002; Le Corre and Kremer 2003). In contrast, the association approach has low power when there is complete confounding between environment gradients and axes of ancestry. Thus, lists of outlier loci are
useful because they may allow a reverse engineering of the relevant phenotypes and possibly the environmental gradients on which the phenotypes are selected through comparative genomics, functional experimentation and GIS analysis. This would in principle help clarify the relationship between genotype, adaptive phenotypes and fitness (LUIKART et al. 2003; STORZ 2005; ROSS-IBARRA et al. 2007).

We are not the first to note differences between results obtained using association versus $F_{ST}$ outlier approaches. In an analysis of SNPs associated to human diseases, there was no significant difference in $F_{ST}$ across associated versus randomly chosen genes (LOHMUELLER et al. 2009). This suggests that environment, as opposed to ancestry, plays a significant role in disease risk for many causative polymorphisms in humans. Analogously, natural selection may not be driving large-scale adaptive differences among lineages of loblolly pine, but may instead be selecting genotypes along environmental gradients regardless of ancestry. Selection may thus be operating across different spatial and evolutionary scales in forest trees and our ability to detect it will depend upon the scale of sampling and the method employed.

Alternatively, quantitative genetic theory predicts that clinal variation for phenotypic traits is derived from small frequency differences at many loci (BARTON 1999), with the among population component of linkage disequilibrium across loci being largely responsible for changes in phenotypic means (LE CORRE AND KREMER 2003). In this case, loci underlying a quantitative trait are expected to illustrate only small allele frequency differences among populations, but manifest large phenotypic effects (LATTA 1998, MCKAY AND LATTA 2002; LE CORRE AND KREMER 2003). An observation consistent with this expectation was made previously in European aspen (Populus tremula L.) for
two SNPs associated to bud phenology (INGVARSSON et al. 2008). This is consistent with the low $F_{ST}$ for SNPs associated to aridity and suggests that scans for $F_{ST}$ outliers may miss many loci underlying adaptive phenotypes.

While our approach to identifying ecologically relevant genetic variants has numerous advantages, it is not without its own limitations. Within-county sampling may be more effective for association analyses than the county-level approach taken here, because fine-scale environmental variation may of be of considerable biological relevance (MITTON et al. 1998; PARISOD AND CHRISTIN 2008). Alternatively, allele frequency approaches utilizing local populations have also been shown to be powerful in identifying loci affected by diversifying selection (HANCOCK et al. 2008). Our SNP data, moreover, represent only a fraction of the coding portion of the loblolly pine genome, and further SNP discovery and analysis would enable truly genomic approaches to ecologically relevant genetic variation. Our $F_{ST}$ outlier approach is limited by the assumption that our demographic model reflects the true phylogeographical history of loblolly pine. While model misspecification can lead to erroneous results, we note that our model is likely more biologically plausible than the standard model used in $F_{ST}$ outlier analyses, and in fact explains the observed data as well as, if not better than, drift or the island model alone. Lastly, future approaches using sets of targeted candidate genes, as opposed to genome scans, may be more fruitful in identifying loci correlated with specific environmental gradients.

Conclusions
We identified general patterns of population structure, environmental correlates to those patterns, and 29 SNP loci that were associated to aridity \((n = 5)\) or had extreme values of \(F_{ST} (n = 24)\) for loblolly pine. The five SNP loci associated to aridity were located in genes encoding proteins primarily involved with biotic and abiotic stress response, while the 24 outlier loci were located in genes encoding proteins involved with a diverse set of physiological functions. Increased awareness of the strengths, weaknesses and complementarity of the methods employed during scans for ecologically important genetic variation is needed as truly population genomic data sets emerge for non-model species.

ACKNOWLEDGEMENTS

The authors would like to thank Sedley Josserand, Dennis Deemer, Craig Echt, Valerie Hipkins, Vanessa K. Rashbrook, Charles M. Nicolet, John D. Liechty, Benjamin N. Figueroa and Gabriel G. Rosa for laboratory and bioinformatics support. We would also like to thank Andrew Bower for advice on GIS-derived climate data, W. Patrick Cumbie and Barry Goldfarb for providing geographic information for sampled trees and Aslam Mohamed for creating the linkage map. The manuscript was also much improved by critical and insightful comments made by two anonymous reviewers. This work was supported by the National Science Foundation [grant number DBI-0501763].
REFERENCES


BECK, E. H., S. FETTIG, C. KNake, K. HARTIG, AND T. BHATTARAI, 2007 Specific and
unspecific responses of plants to cold and drought stress. Journal of Biosciences

1626.

BEAUMONT, M. A., AND D. J. BALDING, 2004 Identifying adaptive genetic divergence among
populations from genome scans. Molecular Ecology 13: 969–980.

BRAY, E. A., 2004 Genes commonly regulated by water-deficit stress in Arabidopsis

BRENDEL, O., D. POT, C. PLOMIOn, P. ROZENBERG, AND J. M. GUEHL, 2002 Genetic
parameters and QTL analysis of δ13C and ring width in maritime pine. Plant, Cell
and Environment 25: 945–953

CHEN, J., C. G. TAUER, G. BAI Y. HUANG, M. E. PAYTON, et al., 2004 Bidirectional
introgression between Pinus taeda and Pinus echinata: Evidence from
morphological and molecular data. Canadian Journal of Forest Research 34:
2508–2516.

CHOU, I. T. AND C. S. GASSER, 1997 Characterization of the cyclophilin gene family of
Arabidopsis thaliana and phylogenetic analysis of known cyclophilin proteins.
Plant Molecular Biology 35: 873–892.

COSTAGLIOLI, P., J. JOUBÈS, C. GARCIA, M. STEF, B. ARVEILER, et al., 2005 Profiling
candidate genes involved in wax biosynthesis in Arabidopsis thaliana by


LANGLET, O., 1971 Two hundred years of geneecology. Taxon **20**: 653–722.


MIZUNO, T., AND T. YAMASHINO, 2008 Comparative transcriptome of diurnally oscillating
genes and hormone-responsive genes in Arabidopsis thaliana: Insight into
circadian clock-controlled daily responses to common ambient stresses in plants.

MORGENSTERN, E. K., 1996 Geographic Variation in Forest Trees. UBC Press,
Vancouver, BC.

NAMKOONG, G., 1979 Introduction to Quantitative Genetics in Forestry. USDA Forest

NAMROUD, M-C., J. BEAULIEU, N. JUGE, J. LAROCHE AND J. BOUSQUET, 2008 Scanning the
genome for gene single nucleotide polymorphisms involved in adaptive

NEWTON, R. J., E. A. FUNKHOUSER, F. FONG AND C. G. TAUER, 1991 Molecular and
physiological genetics of drought tolerance in forest species. Forest Ecology and

Coordinated and selective recruitment of eIF4E and eIF4G factors for potyvirus

OLIVAS-GARCIA, J. M., B. M. CREGG, AND T. C. HENNESSEY, 2000 Genotypic variation in
carbon isotope discrimination and gas exchange of ponderosa pine seedlings
under two levels of water stress. Canadian Journal of Forest Research 30: 1581–
1590.


VERGNOLLE, C. M-N. VAULTIER, L. TAConNAT, J-P. RENOU, J-C. KADER, et al., 2005 The cold–induced early activation of phospholipase C and D pathways determines
the response of two distinct clusters of genes in Arabidopsis cell suspensions.

Plant Physiology **139**: 1217–1233.


WESTFALL, R. D., AND C. I. MILLAR, 2004 Genetic consequences of forest population
dynamics influenced by historic climatic variability in the western USA. Forest

WONG, C. E., Y. LI, A. LABBE, D. GUEVARA, P. NUIN, et al., 2006 Transcriptional profiling
implicates novel interactions between abiotic stress and hormonal responses in

XU, S., C. G. TAUER, AND C. D. NELSON, 2008 Genetic diversity within and among
populations of shortleaf pine (*Pinus echinata* Mill.) and loblolly pine (*Pinus taeda*

YANG, S-H., AND C. A. LOOPSTRA, 2005 Seasonal variation in gene expression for loblolly
pines (*Pinus taeda*) from different geographical regions. Tree Physiology 25:
1063–1073.

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

ZHANG, J., AND J. D. MARSHALL, 1994 Population differences in water-use efficiency of
well-watered and water-stressed western larch seedlings. Canadian Journal of
<table>
<thead>
<tr>
<th></th>
<th>SSRs</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of loci</td>
<td>23</td>
<td>3059</td>
</tr>
<tr>
<td>Missing data (%)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>6.5 ± 5.7</td>
<td>2.2 ± 2.6</td>
</tr>
<tr>
<td>Expected Heterozygosity (&lt;i&gt;H&lt;/i&gt;&lt;sub&gt;E&lt;/sub&gt;)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.72 ± 0.19</td>
<td>0.25 ± 0.16</td>
</tr>
<tr>
<td>Fixation Index (&lt;i&gt;F&lt;/i&gt;&lt;sub&gt;IS&lt;/sub&gt;)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.11 ± 0.07</td>
<td>0.02 ± 0.06</td>
</tr>
<tr>
<td>Number of loci in HWE&lt;sup&gt;b&lt;/sup&gt;</td>
<td>23</td>
<td>3037</td>
</tr>
</tbody>
</table>

<sup>a</sup>Numbers are averages (± one standard deviation).

<sup>b</sup>Number of loci that do not deviate from HWE after correcting for multiple testing using a Bonferroni correction.
TABLE 2. Population structure ($K$ = 3), geography and environmental heterogeneity are correlated (Pearson’s $r$) across the range of loblolly pine.

<table>
<thead>
<tr>
<th></th>
<th>SSR-STR$^a$</th>
<th>SSR-PCA$^b$</th>
<th>SNP-PCA$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AP1</td>
<td>AP2</td>
<td>AP3</td>
</tr>
<tr>
<td>Latitude</td>
<td>-0.21***</td>
<td>0.20***</td>
<td>0.02</td>
</tr>
<tr>
<td>Longitude</td>
<td>-0.37***</td>
<td>0.34***</td>
<td>0.04</td>
</tr>
<tr>
<td>AI1</td>
<td>0.05</td>
<td>0.13**</td>
<td>0.13**</td>
</tr>
<tr>
<td>AI2</td>
<td>0.23***</td>
<td>-0.25***</td>
<td>0.03</td>
</tr>
<tr>
<td>AI3</td>
<td>-0.12**</td>
<td>0.12**</td>
<td>0.01</td>
</tr>
<tr>
<td>AI4</td>
<td>0.21***</td>
<td>-0.25***</td>
<td>0.07</td>
</tr>
</tbody>
</table>

$^a$Correlations are with the assignment probabilities (AP) for each cluster derived using STRUCTURE.

$^b$Correlations are with the first two principal components (PCs), which define three clusters.

* $P < 0.05$, ** $P < 0.005$, *** $P < 0.0005$
### TABLE 3. Significant associations between genotypes and aridity gradients (FDR $Q < 0.05$)

<table>
<thead>
<tr>
<th>SNP Locus</th>
<th>Gene Product</th>
<th>At Locus</th>
<th>E-value $^c$</th>
<th>Al $^d$</th>
<th>$r^e$</th>
<th>$P$ $^f$</th>
<th>$Q$ $^f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-16169-01-216$^{syn}$</td>
<td>hexose:hydrogen symporter</td>
<td>At5g26340</td>
<td>3E-118</td>
<td>AI1</td>
<td>0.027</td>
<td><strong>4.79E-08</strong></td>
<td>1.84E-04</td>
</tr>
<tr>
<td>0-2179-01-216$^{nc(utr)}$</td>
<td>photosystem II protein</td>
<td>At2g06520</td>
<td>2E-24</td>
<td>AI1</td>
<td>0.034</td>
<td><strong>6.31E-07</strong></td>
<td>1.81E-03</td>
</tr>
<tr>
<td>0-10240-01-410$^{syn}$</td>
<td>C3HC4-type RING finger</td>
<td>At4g26400</td>
<td>1E-05</td>
<td>AI2</td>
<td>0.039</td>
<td><strong>9.63E-07</strong></td>
<td>2.21E-03</td>
</tr>
<tr>
<td>0-18281-01-345$^{nc(intron)}$</td>
<td>MATE efflux family protein</td>
<td>At4g25640</td>
<td>9E-72</td>
<td>AI2</td>
<td>0.022</td>
<td>8.23E-06</td>
<td>1.35E-02</td>
</tr>
</tbody>
</table>

$^a$nc, noncoding; ns, nonsynonymous; syn, synonymous; utr, untranslated region  
$^b$At Locus refers to the locus tag for *Arabidopsis thaliana*.  
$^c$E-values from tBLASTx analysis of the loblolly pine EST contigs against the NCBI refseq RNA database for Arabidopsis.  
$^d$AI, aridity index  
$^e$r$^2$ is that from a general linear model with environment as the dependent variable and corrected genotypic values as explanatory variables.  
$^f$P-values listed in bold are also significant using a Bonferroni correction. Significance refers to the $(N-k-1)r^2$ statistic of PRICE et al. (2006).
**Table 4. Significant $F_{ST}$ outliers ($P < 0.01$) based on $K = 3$ using a two refugia model have values of $F_{ST}$ six to 12-fold greater than the average across all loci.**

<table>
<thead>
<tr>
<th>SNP Locus*</th>
<th>Gene Product*</th>
<th>At Locus*</th>
<th>E-value*</th>
<th>$F_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1126-02-419&lt;sup&gt;ns&lt;/sup&gt;</td>
<td>O-methyltransferase</td>
<td>At5g54160</td>
<td>1E-17</td>
<td>0.182</td>
</tr>
<tr>
<td>0-11531-01-379&lt;sup&gt;ns&lt;/sup&gt;</td>
<td>ovate family protein</td>
<td>At2g18500</td>
<td>2E-23</td>
<td>0.185</td>
</tr>
<tr>
<td>0-12076-01-310&lt;sup&gt;ns&lt;/sup&gt;</td>
<td>hypothetical protein</td>
<td>At1g01500</td>
<td>7E-05</td>
<td>0.157</td>
</tr>
<tr>
<td>0-13230-02-146&lt;sup&gt;ns&lt;/sup&gt;</td>
<td>eukaryotic translation initiation factor</td>
<td>At3g60240</td>
<td>6E-50</td>
<td>0.144</td>
</tr>
<tr>
<td>0-14415-01-190&lt;sup&gt;ns&lt;/sup&gt;</td>
<td>unknown</td>
<td>------</td>
<td>------</td>
<td>0.209</td>
</tr>
<tr>
<td>0-15241-02-133&lt;sup&gt;ns&lt;/sup&gt;</td>
<td>hypothetical protein</td>
<td>At2g01300</td>
<td>5E-16</td>
<td>0.262</td>
</tr>
<tr>
<td>0-17238-01-290&lt;sup&gt;nc intron&lt;/sup&gt;</td>
<td>MtN21 family protein</td>
<td>At3g28050</td>
<td>3E-20</td>
<td>0.213</td>
</tr>
<tr>
<td>0-2784-01-297&lt;sup&gt;ns&lt;/sup&gt;</td>
<td>nucleoporin family protein</td>
<td>At1g59660</td>
<td>2E-10</td>
<td>0.247</td>
</tr>
<tr>
<td>0-6427-02-341&lt;sup&gt;nc utr&lt;/sup&gt;</td>
<td>unknown</td>
<td>------</td>
<td>------</td>
<td>0.186</td>
</tr>
<tr>
<td>0-7745-01-176&lt;sup&gt;syn&lt;/sup&gt;</td>
<td>unknown</td>
<td>------</td>
<td>------</td>
<td>0.239</td>
</tr>
<tr>
<td>0-8922-01-645&lt;sup&gt;nc intron&lt;/sup&gt;</td>
<td>TIFY domain containing protein</td>
<td>At4g32570</td>
<td>1E-11</td>
<td>0.194</td>
</tr>
<tr>
<td>1-3327-01-113&lt;sup&gt;ns&lt;/sup&gt;</td>
<td>unknown</td>
<td>------</td>
<td>------</td>
<td>0.242</td>
</tr>
<tr>
<td>2-1087-01-86&lt;sup&gt;nc utr&lt;/sup&gt;</td>
<td>ubiquitin-specific protease</td>
<td>At1g04860</td>
<td>2E-15</td>
<td>0.235</td>
</tr>
<tr>
<td>2-1205-02-74&lt;sup&gt;nc utr&lt;/sup&gt;</td>
<td>3-ketoacyl-CoA synthase</td>
<td>At1g68530</td>
<td>2E-54</td>
<td>0.180</td>
</tr>
<tr>
<td>2-2953-01-168&lt;sup&gt;nc utr&lt;/sup&gt;</td>
<td>unknown</td>
<td>------</td>
<td>------</td>
<td>0.176</td>
</tr>
<tr>
<td>2-3591-03-186&lt;sup&gt;nc utr&lt;/sup&gt;</td>
<td>zinc finger-homeodomain protein 2</td>
<td>At4g24660</td>
<td>1E-53</td>
<td>0.174</td>
</tr>
<tr>
<td>2-7619-01-191&lt;sup&gt;syn&lt;/sup&gt;</td>
<td>VHS/GAT domain-containing protein</td>
<td>At5g16880</td>
<td>8E-29</td>
<td>0.168</td>
</tr>
<tr>
<td>2-9190-01-446&lt;sup&gt;syn&lt;/sup&gt;</td>
<td>auxin responsive protein</td>
<td>At4g30080</td>
<td>2E-26</td>
<td>0.182</td>
</tr>
<tr>
<td>CL1569Contig1-03-45&lt;sup&gt;nc utr&lt;/sup&gt;</td>
<td>hypothetical protein</td>
<td>At3g16070</td>
<td>2E-08</td>
<td>0.193</td>
</tr>
<tr>
<td>CL1799Contig1-04-224&lt;sup&gt;nc intron&lt;/sup&gt;</td>
<td>chlorophyll a/b binding protein</td>
<td>At3g54890</td>
<td>1E-112</td>
<td>0.168</td>
</tr>
<tr>
<td>CL872Contig1-03-287&lt;sup&gt;ns&lt;/sup&gt;</td>
<td>unknown</td>
<td>------</td>
<td>------</td>
<td>0.239</td>
</tr>
<tr>
<td>UMN-1592-02-53&lt;sup&gt;ns&lt;/sup&gt;</td>
<td>SWIB/MDM2 domain-containing protein</td>
<td>At4g34290</td>
<td>2E-30</td>
<td>0.172</td>
</tr>
<tr>
<td>UMN-897-01-82&lt;sup&gt;nc utr&lt;/sup&gt;</td>
<td>unknown</td>
<td>------</td>
<td>------</td>
<td>0.253</td>
</tr>
<tr>
<td>UMN-CL194Contig1-04-130&lt;sup&gt;syn&lt;/sup&gt;</td>
<td>unknown</td>
<td>------</td>
<td>------</td>
<td>0.187</td>
</tr>
</tbody>
</table>

*nc, noncoding; ns, nonsynonymous; syn, synonymous, utr, untranslated region

*All loci listed as unknown have putative homologs in Sitka spruce (*Picea sitchensis*) EST libraries.

*Locus tag from *Arabidopsis thaliana* (*At*).

*E-values from tBLASTx analysis of the loblolly pine EST contigs against the NCBI refseq RNA database for Arabidopsis.
Figure Legends

Figure 1. The distribution of loblolly pine, sampling localities (black points) and aridity gradients (color gradients) used to assess patterns of population structure and environmental associations. The distribution of loblolly pine is available as a GIS layer from the US Geological Survey (http://esp.cr.usgs.gov/data/atlas/little/). Aridity gradients were smoothed with kriging using algorithms in the fields package available in R (R CORE DEVELOPMENT TEAM 2007). Aridity gradients are shown by annual quarter (Q1: Jan-Mar, Q2: Apr-Jun, Q3: Jul-Sep, Q4: Oct-Dec).

Figure 2. Patterns of population structure for loblolly pine are correlated with geography and aridity. (A) The relationship between the log-probability of the data and the number of clusters (K) using the SSR data. Arrows denote values of K assumed during further analysis. (B) Assignment probabilities (AP) for trees arranged from west to east assuming K = 3 or K = 5. (C) Scree plots for PCA analysis using SNPs (left) and SSR markers (right). Major eigenvalues are denoted with black circles. (D) A biplot indicating correlations among aridity variables, as well as with population structure. The PCs were derived by PCA on the four aridity indices. Colors denote cluster memberships as determined using PCA on the SNP data.

Figure 3. Cluster assignments are not dramatically affected by choice of clustering method (STRUCTURE and PCA) and marker type (SSRs and SNPs). Illustrated are assignments summarized by county for each method-marker pair used to assess
patterns of population structure for $K = 3$. Each point is a small pie chart denoting the proportion of trees in that county belonging to a specific genetic cluster.

Figure 4. Genotypic associations for four loci with the aridity index during the second quarter (AI2) illustrate small yet significant correlations. Horizontal lines denote the genotypic mean, with diamonds representing the upper and lower bounds of the 95% confidence interval for the mean. Ancestry-corrected environmental indices are plotted on the $y$-axis. For clarity, the SNP genotypes are plotted on the $x$-axis, although ancestry corrected SNP genotypes are used in the association analysis.

Figure 5. The quantity and identity of $F_{ST}$ outliers under island and dual Pleistocene refugia models are affected by cluster assignments. Each point represents a SNP locus ($n = 3059$), with lines denoting the upper 99th percentile of the null distribution (solid: island model; dashed: two refugia model). Each plot represents a different set of cluster assignments based on differing marker sets for $K = 3$. (A) Cluster assignments from STRUCTURE based on SSR markers. (B) Cluster assignments from PCA using SSR markers. (C) Cluster assignments from PCA using SNP markers.

Figure 6. Loci associated to aridity and those identified as $F_{ST}$ outliers are distributed across the linkage map of loblolly pine. Plotted are trend lines across the 1898 cM linkage map for $F_{ST}$ and $\chi^2$ statistics for four aridity indices (AI1, AI2, AI3 and AI4). Red dots denote significant outliers ($n = 11$) or associated SNPs ($n = 2$), while dashed green lines denote the 2.5%, 50% and 97.5% quantiles of the genomic distribution for each
statistic. Three of the five SNPs associated significantly to aridity were not segregating within the cross that was used to construct the linkage map and are thus unmapped.