MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 3789–3805 doi: 10.1111/j.1365-294X.2010.04698.x

Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae)

ANDREW J. ECKERT,*† ANDREW D. BOWER,‡ SANTIAGO C. GONZÁLEZ-MARTÍNEZ,§ JILL L. WEGRZYN,¶ GRAHAM COOP*† and DAVID B. NEALE†¶**

**Section of Evolution and Ecology, University of California—Davis, Davis, CA 95616, USA, †Center for Population Biology, University of California—Davis, Davis, CA 95616, USA, ‡USDA Forest Service, Pacific Northwest Research Station, Corvallis, OR 97331, USA, §Department of Forest Systems and Resources, Forest Research Institute, CIFOR-INIA, Madrid 28040, Spain, ¶Department of Plant Sciences, University of California—Davis, Davis, CA 95616, USA, **Institute of Forest Genetics, USDA Forest Service, Davis, CA 95616, USA*

Abstract

Genetic variation is often arrayed in latitudinal or altitudinal clines, reflecting either adaptation along environmental gradients, migratory routes, or both. For forest trees, climate is one of the most important drivers of adaptive phenotypic traits. Correlations of single and multilocus genotypes with environmental gradients have been identified for a variety of forest trees. These correlations are interpreted normally as evidence of natural selection. Here, we use a genome-wide dataset of single nucleotide polymorphisms (SNPs) typed from 1730 loci in 682 loblolly pine (*Pinus taeda* L.) trees sampled from 54 local populations covering the full-range of the species to examine allelic correlations to five multivariate measures of climate. Applications of a Bayesian generalized linear mixed model, where the climate variable was a fixed effect and an estimated variance–covariance matrix controlled random effects due to shared population history, identified several well-supported SNPs associating to principal components corresponding to geography, temperature, growing degree-days, precipitation and aridity. Functional annotation of those genes with putative orthologs in Arabidopsis revealed a diverse set of abiotic stress response genes ranging from transmembrane proteins to proteins involved in sugar metabolism. Many of these SNPs also had large allele frequency differences among populations (F_{ST} = 0.10–0.35). These results illustrate a first step towards a ecosystem perspective of population genomics for non-model organisms, but also highlight the need for further integration of the methodologies employed in spatial statistics, population genetics and climate modeling during scans for signatures of natural selection from genomic data.

Keywords: adaptation, ecological genomics, environmental gradients, *Pinus taeda*, population structure, single nucleotide polymorphisms

Received 8 January 2010; revision received 12 April 2010; accepted 22 April 2010

Introduction

Forest trees illustrate clear phenotypic adaptations to environmental gradients at multiple spatial scales (Morgenstern 1996). An extensive history of provenance, common garden and genecological studies has established the highly polygenic basis of these adaptive traits (Langlet 1971; Namkoong 1979). Applications of population and quantitative genetic methodologies to a variety of forest tree species have further elucidated many of the functional genes underlying complex adaptive phenotypes (Neale & Savolainen 2004; González-Martínez *et al.* 2006; Neale 2007; Savolainen & Pyhäjärvi 2007; Savolainen *et al.* 2007; Neale & Ingvarsson 2008; Grattapaglia *et al.* 2009). The further link to the correspondence between genetic and environmental associations; however, is missing from most of these studies

Correspondence: David B. Neale, Fax: +1 530 754 9366; E-mail: dbneale@ucdavis.edu

(but see Eveno *et al.* 2008; Namroud *et al.* 2008; Eckert *et al.* 2009a; Richardson *et al.* 2009). This is despite a multitude of genecological studies establishing the genetic basis for climate-related traits in forests trees (cf. Rehfeldt 1989, 1990; St Clair *et al.* 2005; St Clair 2006). The application of high-throughput sequencing and genotyping technologies, therefore, has generated renewed interest in scanning the functional gene space of these species for loci that are significantly associated with environmental variation.

Geographical variation for environmental variables creates gradients across the ranges of species, allowing natural selection to drive the geographic distribution of traits (Linhart & Grant 1996). Thus, populations distributed across strong environmental gradients are expected to exhibit clinal patterns of gene frequencies for those loci under divergent selection. Climate is one of the major environmental drivers of adaptive traits for forest trees (cf. Richardson *et al.* 2009). Most studies of temperate and boreal trees identify temperature as one of the most important factors influencing local adaptation (Rehfeldt *et al.* 1999; Aitken & Hannerz 2001; Howe *et al.* 2003); however, other climatic variables (e.g. precipitation and aridity) are related to a number of quantitative traits (Rehfeldt *et al.* 2002; Beaulieu *et al.* 2004; Wei *et al.* 2004; St Clair *et al.* 2005). For example, traits related to cold adaptation (i.e. budset and growth cessation) often vary along temperature gradients associated with elevational and latitudinal clines (Aitken & Hannerz 2001). Survival and growth traits (i.e. growth initiation, height growth, and biomass partitioning) are also sensitive to different precipitation and aridity regimes, as well as to temperature (Rehfeldt *et al.* 1999; St Clair *et al.* 2005; Bower & Aitken 2008).

For tree species at the southern edge of the temperate latitudes, such as loblolly pine (*Pinus taeda* L.), climate is also one of the main drivers of local adaptation (Schmidtling 2001). In this species and other pines (cf. Alía *et al.* 1997; Aranda *et al.* 2010), phenotypic responses are mediated by climate and affect life-history strategies such as drought-avoidance tactics that can take the form of slower growth, different biomass allocation or higher water-use efficiency depending on the environment of origin. In addition, the range of loblolly pine is likely limited by climate (to the north by low temperature, to the west by low rainfall), and the average yearly minimum temperatures for the place of origin affect its growth and survival in plantations across its native range (Schmidtling 2001). Genotype-by-environment interactions are also notable in this species and likely have a climatic origin (Sierra-Lucero *et al.* 2002, 2003). The correlations of climate with provenance performance of loblolly pine, as well as the large-scale genomic resources available for this species, its distribution across 370 000 km^2 of climatically diverse environments in the southeastern USA and the multitude of association genetic studies identifying the genes underlying several quantitative traits, make it an ideal system to examine correlations between climate and allele frequencies.

The standard approach to detect loci underlying adaptive phenotypic responses to climate is environmental association analysis, where genetic variation is correlated to climate variables (Vasemägi & Primmer 2005; Holderegger *et al.* 2008). The geographical basis of both climate and genetic variation; however, confounds the interpretation of this form of analysis as patterns of gene flow and genetic drift can also lead to the formation of gene frequency clines at neutral loci (Kimura & Maruyama 1971; Vasemägi 2006). Thus, the methods used to detect correlations between climate and genetic variation need to take into account background levels of population structure, if these correlations are used to identify genes affected by natural selection along climatic gradients (e.g. Felsenstein 2002). This form of correction is employed routinely in genetic association methods where genotypes are correlated to phenotypes (Yu *et al.* 2006).

Here, we use environmental association analysis to search for correlations between climate variables and single nucleotide polymorphisms (SNPs) genotyped across a range-wide sample of loblolly pine populations while accounting for neutral levels of population structure. Specifically, we use SNPs located in 1730 functional genes in combination with geographic information system (GIS) derived climate variables to find those SNPs most likely associated to multivariate measures of climate. In doing so, we construct a list of candidate genes underlying climatic responses for loblolly pine, and highlight the need for further integration of the methodologies employed in spatial statistics, population genetics and climate modelling during scans for signatures of natural selection from population genomic data.

Materials and methods

Focal species

Loblolly pine is distributed throughout the southeastern USA, ranging from Texas to Delaware. Patterns of diversity at isozyme and nuclear microsatellite loci illustrate moderate genetic differentiation between populations located to the east and west of the Mississippi River Valley, as well as increased levels of admixture for trees located on the Gulf Coast Plain and putative population contraction in the western most populations (Wells *et al.* 1991; Schmidtling *et al.* 1999; Al-Rabab'ah & Williams 2002, 2004; Xu *et al.* 2008). These patterns conform to a general phylogeographical pattern named the Mississippi River discontinuity (Soltis *et al.* 2006). The structure of this discontinuity is also consistent with a hypothesized dual Pleistocene refugial model proposed for this species (cf. Schmidtling 2003), which has also been used to explain differential growth abilities, disease resistance and concentrations of secondary metabolites among families located across this discontinuity (Wells & Wakeley 1966; Squillace & Wells 1981). A recent examination of diversity across 23 nuclear microsatellites and 3059 SNPs found evidence for three genetic clusters (WMC, western Mississippi cluster; GCC, Gulf coast cluster; ACC, Atlantic coast cluster) defined largely by the Mississippi River discontinuity (Eckert *et al.* 2010).

Sample collection

Needle tissue was collected from 682 trees sampled across the natural range of loblolly pine (Fig. 1). These samples are a geographical subset of those analysed previously for association (González-Martínez *et al.* 2007) and population genetic studies (Eckert *et al.* 2010) for loblolly pine. Total genomic DNA was isolated from each sample at the USDA National Forest Genetics Laboratory (NFGEL, Placerville, CA, USA) using DNeasy® Plant Kits (Qiagen, Valencia, CA, USA) following the manufacturer's protocol. Trees are geo-referenced by county ($n = 54$), meaning that this is the smallest spatial unit able to be assigned to individual trees, with an average number of sampled trees (± one standard deviation) per county of 13 ± 13 (range: 5–71). We defined local populations to be counties for all further analyses.

Climate data

Climate data were obtained from the PRISM Group (Oregon State University, http://www.prismclimate.org), which included raster files for monthly and annual minimum and maximum temperature and precipitation based on climate normals for the period 1971–2000. All variables had an 800 m × 800 m cell size and are spatially autocorrelated. Temperature (°C) and precipitation (mm), rounded to two significant digits, were multiplied by 100 in these raster files. Elevation data were obtained from an ESRI 30 arc-second grid file available at the WorldClim website (http://www.worldclim.org/current). These data have a resolution of approximately 1 km^2. Monthly and annual temperature measures were obtained by averaging minimum and maximum estimates. We also classify sets of months using seasons defined as: winter (December–February), spring (March–May), summer (June–August) and fall (September–November).

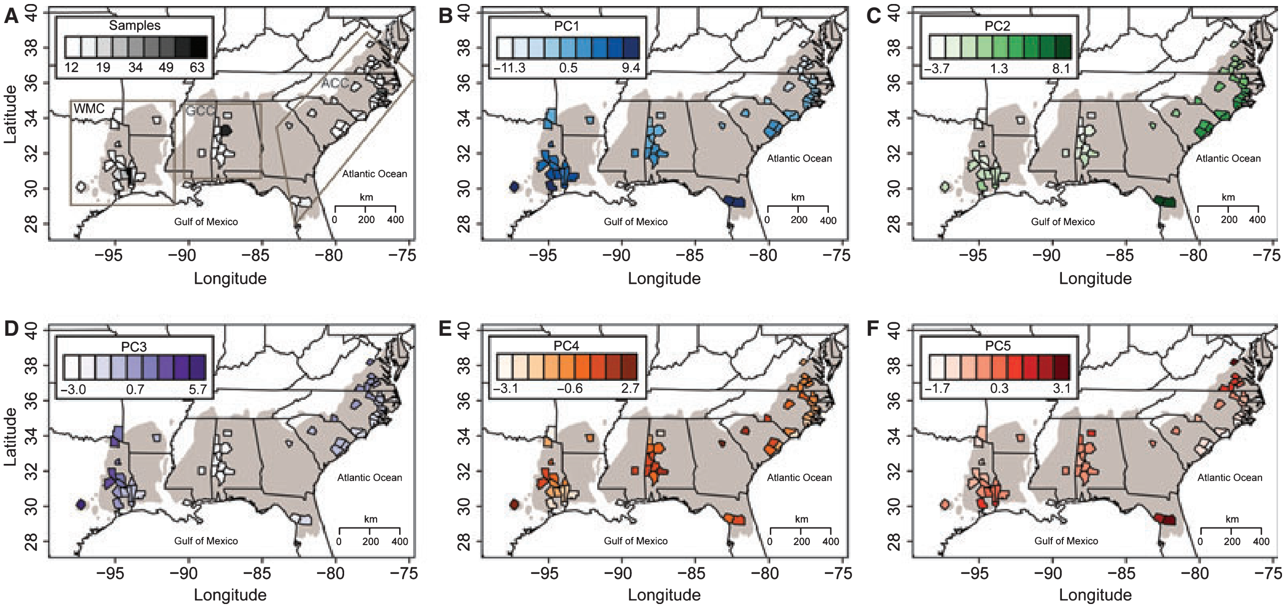

Fig. 1 Spatial patterns for multivariate measures of climate across the range of loblolly pine (*Pinus taeda* L.) illustrate correlations with geography and genetic structure. The shaded area is the range of loblolly pine according to Little (1971). Each polygon is a county in which multiple trees were sampled and genotyped for 1730 SNPs. Colours designate ranges for the listed values, which are the number of sampled trees (A) and geoclimatic PCs 1–5 (B–F). Boxes in panel A enclose the three genetic clusters defined previously for loblolly pine—WMC: west of the Mississippi cluster, GCC: Gulf Coast cluster and ACC: Atlantic Coast cluster.

Climate variables and elevation were summarized for each county using the zonal statistics function in ArcGIS Spatial Analyst (ESRI, Redland, CA, USA). This yielded the minimum, maximum, range, mean and standard deviation of all raster file cells contained within each county. Latitude and longitude for each county were determined using the centroid of each county polygon. County boundaries for each state were obtained from the US Census Bureau 2008 TIGER/line shapefile database (http://www.census.gov/geo/www/tiger/tgrshp2008/tgrshp2008.html), while the range map for loblolly pine was obtained from the US Geological Survey (http://esp.cr.usgs.gov/data/atlas/little/). The distribution for loblolly pine in this map is based on that given by Little (1971).

Two derived climate variables were constructed using spatial and climate data. First, accumulated growing degree-days above 5 °C (GDD5) was calculated on a monthly basis following Rehfeldt (2006). This calculation ignores daily variations around the monthly mean, but we assumed that biases would be consistent across the range of the study. Second, an aridity index (AI) was calculated on a monthly basis following Eckert *et al.* (2010). This index represents the ratio of precipitation to potential evapotranspiration (PET) as estimated using the method of Thornthwaite (1948). Calculations for PET utilized the equations given by Forsythe *et al.* (1995), which relate latitude to daylength.

We used principal components analysis (PCA) to reduce dimensionality for the 60 monthly climate variables, latitude and longitude. The data were centred and scaled prior to PCA. Principal components (PCs) with eigenvalues greater than 1.0 were kept for further analysis. Correlations of the original variables with the first two PCs were visualized using biplots (Gabriel 1971). All analysis was conducted using the PRCOMP function and the BPCA package available in the R computing environment (R Development Core Team 2007).

SNP data

We selected one SNP each from 1730 unique expressed sequence tag (EST) contigs in loblolly pine. This assumes implicitly that SNPs within different EST contigs are independent (i.e. unlinked). The identification of SNPs was carried out previously and was accomplished through the resequencing of 7508 unique EST contigs in a discovery panel of 18 loblolly pine haploid megagametophytes collected from a range-wide sample of trees. Here, contigs were selected for further analysis based upon success of SNP genotyping and coverage of the full linkage map for loblolly pine (TG accession: TG091; http://dendrome.ucdavis.edu/cmap/). Further information concerning the annotation of EST contigs, SNP discovery and SNP genotyping is located in the online Supporting information (see also Eckert *et al.* 2010). When multiple SNPs were available within the same contig, we selected one at random, with the stipulations that the minor allele frequency (MAF; i.e. the allele with frequency < 0.50 across all sampled trees) had to be greater than 2.5% and the absolute value of Wright's inbreeding coefficient (F_{IS}) had to be < 0.25 (see below).

Genotype data for each SNP were obtained using the Infinium platform (Illumina, San Diego, CA, USA) following the manufacturer's instructions. The first application of this platform in forest trees is described fully elsewhere (Eckert *et al.* 2010). Briefly, this assay begins with whole-genome amplification followed by fragmentation for each sample. These products are hybridized directly overnight to array-bound probe sequences. Allele-specific, single-base extension was carried out and followed by fluorescent staining, scanning and analysis. All processing and staining steps were carried out manually, at different times and at different facilities (UC Davis Genome Center and Illumina). Arrays were imaged on a Bead Array reader and analysed using BeadStudio v. 3.1.3.0 (Illumina). All genotype calls were standardized to the top strand for consistency. To ensure similar data quality across data generated at different facilities, we compared distributions for two Illumina-designed quality metrics (GenCall_{50} and Call Frequency, cf. Eckert *et al.* 2009b) between facilities using Mann–Whitney *U*-tests in the R computing environment. The metrics chosen for comparison reflect the overall ability of a particular SNP to be genotyped, where larger values reflect better quality.

For each SNP we calculated the following quantities: observed (H_{OBS}) and expected (H_{EXP}) heterozygosity, MAF, F_{IS} and hierarchical fixation indices for populations (i.e. counties) nested within regions (WMC, western Mississippi cluster; GCC, Gulf coast cluster; ACC, Atlantic coast cluster). Minor allele frequencies were defined using the entire dataset, so that the MAF within counties can exceed 0.50. The significance of the mean F_{IS} and mean hierarchical fixation indices across SNPs was assessed through use of 99% bootstrap ($n = 10\,000$ replicates sampled over SNPs) confidence intervals (CIs). Lastly, we examined allelic correlations (r^2) to validate our assumption that SNPs located in different gene fragments were unlinked. Analyses were conducted using the GENETICS and HIERFSTAT (Goudet 2005) packages in R (R Development Core Team 2007).

Environmental association analysis

Geography confounds inferences of natural selection from correlations between environmental variables and allele frequencies, and can lead to an excess of false positives. This is due primarily to the influence that geography has on gene flow, as well as environmental variation. We used Bayesian geographical analysis, therefore, to search for correlations between allele frequencies and climate variables after correcting for background levels of population structure and differences in sample size (Hancock *et al.* 2008; Coop *et al.* 2010). This approach is similar conceptually to that described by Felsenstein (2002) for within species comparative methods and to that presented by Yu *et al.* (2006) for association mapping within structured populations.

Our null model is that the observed frequency at the *l*th locus, across populations, represents a binomial sample from the unobserved population frequencies. These population frequencies randomly deviate away from a global frequency (ε_l) due to genetic drift, and populations covary in their deviation due to shared ancestry and gene flow. To model this we assume that the transformed population allele frequencies of the *l*th SNP follow a multivariate normal distribution:

$$P(\theta_l|\Omega, \epsilon_l) \sim N(\epsilon_l, \epsilon_l(1-\epsilon_l)\Omega,$$

where θ_l is the transformed allele frequency at SNP l across populations (i.e. counties), and Ω is the covariance matrix which is shared across all loci (Coop *et al.* 2010). Priors are placed on the parameters (Ω and ε_l) and Markov chain Monte Carlo (MCMC) is used to explore the posterior distributions of Ω and the other parameters (θ_l and ε_l) as described in Coop *et al.* (2010). Multiple runs of the MCMC algorithm were carried out to check for convergence. This approach makes the assumption that Ω fully captures the effects of shared history and patterns of gene flow among populations being examined. We chose this to be our null model and estimated Ω using all 1730 SNPs, which conservatively assumes that loci under strong diversifying selection are relatively rare in the genome of loblolly pine.

The correlation between the allele frequency at each SNP and a specific climate variable (Y) was assessed by allowing the environmental variable to be a fixed effect on the mean of the multivariate normal distribution of transformed population allele frequencies:

$$P(\theta_l|\Omega, \epsilon_l, \beta) \sim N(\epsilon_l + \beta Y, \epsilon_l(1-\epsilon_l)\Omega),$$

where β is the linear effect of the environmental variable Y (Coop *et al.* 2010). For the test of correlation we used a single draw of Ω, estimated under the null model from the data across the 1730 SNPs, as the variation among posterior estimates of Ω was small. The posterior probability of the alternative model was calculated by integrating over a uniform prior on β and the other parameters (θ_l and ε_l). As a measure of the support for the correlation we calculated a Bayes factor (BF), the ratio of posterior probability under the alternative (M_1, β allowed to vary) to that under the null (M_0, $\beta = 0$) model. Further details can be found in Coop *et al.* (2010). For each climate variable and SNP, therefore, we have an estimate of the BF, which can be interpreted as how much more likely the alternative is to the null model.

Results

Climate data

Variation within counties for most climate variables had coefficients of variation (CV) less than 10%, with a mean of 2% across all 60 monthly variables. This was likely related to the geographical size of counties, which ranged from 714 to 4927 km^2 (mean: 2179 km^2). Mean elevation was less than 250 m for all 54 sampled counties. Variation of elevation within counties was larger than for climate variables (CV: 10–80%), but had a mean range of 469 m across counties. In general, counties with the largest CV for elevation were also those with the lowest mean elevations.

Strong correlations existed between geography and climate and among climate variables (Fig. S1, Supporting information). Latitude was strongly and negatively correlated with temperature variables (Pearson's $r < -0.86$), as well as GDD5 (Pearson's $r < -0.93$), while longitude was most strongly correlated with summer maximum temperatures ($r = -0.92$) and GDD5 (Pearson's $r = -0.84$). Spring, summer and fall GDD5 were strongly correlated with each other and with the temperature variables (Pearson's $r > 0.90$), but correlations were weaker for winter GDD5 ($0.60 <$ Pearson's $r < 0.82$). Summer precipitation and aridity were highly correlated (Pearson's $r = 0.97$), as were winter and spring precipitation (Pearson's $r = 0.88$), but correlations of precipitation and aridity with other variables were moderate, with highest of those being with longitude.

As expected from the strong correlations among variables, the top five PCs captured most (i.e. 96%) of the overall variance for the 60 climate variables, latitude and longitude (Table 1, Fig. S2, Supporting information). Removal of latitude and longitude from the PCA resulted in similar PC scores and factor loadings for the climate data (Pearson's $r \leq -0.999$), illustrating that the inclusion of geography had little to no effect on the PCA. These five PCs also had eigenvalues greater than one. We refer to these as geoclimatic PCs for the remainder of the manuscript. The top two PCs were

Table 1 Multivariate measures of climate (i.e. geoclimatic PCs) across the range of loblolly pine

PC*	Eigenvalue	PVE†	Description‡
1	37.54	60.56	latitude, longitude, temperature, GDD5, winter aridity
2	12.41	20.02	longitude, spring-fall aridity, precipitation
3	5.50	8.87	winter and summer precipitation, summer aridity
4	2.65	4.27	spring and fall precipitation and aridity
5	1.48	2.39	winter aridity and GDD5

*Principal component.

†Percent variance explained.

‡Seasons are defined in the Materials and methods. Factor loadings are given in Table S1 and Fig. S3 (Supporting information). GDD5, growing degree-days above 5 ºC.

strongly correlated to geography, and, especially for PC1, many of the 60 monthly climate variables. This is evidenced by strong geographical trends for these PCs (Figs 1 and 2), which together accounted for 80.58% of the total variance. Latitude and monthly estimates for minimum and maximum temperature and GDD5 were strongly and positively loaded onto PC1, while longitude, spring and summer precipitation and summer and fall aridity were loaded on PC2 (Table S1 and Fig. S3, Supporting information). Temperature and GDD5 were almost exclusively loaded on PC1 and the remaining PCs were loaded by combinations of precipitation and aridity, with the exception of PC5, which was loaded by winter GDD5.

SNP data

Overall quality for the SNP genotype data was high, with a median call frequency of 0.975 (i.e. a median of 0.025 missing data per SNP) and a median GC_{50} score of 0.80. Given that the data were generated at differing times and at different facilities, we checked for differences in quality between facilities using Mann–Whitney *U*-tests. Significant differences for each metric were found for SNP data generated at different facilities ($P < 1.0 \times 10^{-6}$). These differences, while statistically significant, were small in magnitude. For example, the 2.5% and 97.5% quantiles for the call frequency differed by less than 3% between facilities. A similar trend was apparent for the GC_{50} score, with differences less than 5%. We, therefore, combined all data together to yield a dataset consisting of one SNP selected from 1730 unique genes genotyped for 682 trees distributed across 54 local populations nested within three regions (Table 2).

Sampled genes were distributed across the 12 linkage groups comprising the consensus linkage map for loblolly pine, with 53–88 genes sampled per linkage group. Only 807 of the 1730 genes are located to a position on the genetic map due to non-segregation or marker uninformativeness within the single cross used to construct the map. Approximately 75% of the 1730 genes are functionally annotated based on BLAST comparisons to sequenced plant genomes and EST libraries. Of those, 61% have gene ontology (GO) classifications corresponding to specific molecular functions or involvement with known biological processes (Figs S4 and S5, Supporting information). The majority of the genes,

Table 2 A summary of diversity and population structure estimates for 1730 SNPs genotyped for a range-wide sample of loblolly pine (*Pinus taeda* L.)

	All	WMC	GCC	ACC
Trees	682	247	156	279
Populations*	54	17	11	26
Trees/population	13 (13)	15 (16)	14 (17)	11 (7)
Missing†	0.022 (0.019)	0.027 (0.034)	0.014 (0.017)	0.023 (0.019)
MAF	0.190 (0.135)	0.173 (0.139)	0.186 (0.132)	0.185 (0.131)
H_{OBS}	0.261 (0.143)	0.249 (0.167)	0.266 (0.147)	0.269 (0.150)
H_{EXP}	0.271 (0.148)	0.253 (0.167)	0.271 (0.148)	0.272 (0.150)
F_{IS}	**0.033** (0.057)	**0.017** (0.088)	**0.018** (0.089)	**0.009** (0.072)
F_{ST}‡	**0.043** (0.033)	**0.023** (0.030)	**0.008** (0.025)	**0.012** (0.022)
Fixed§	–	17	1	2

Listed for each measure are totals or averages (standard deviations). Values in bold have 99% bootstrap (n = 10 000 replicates) CIs excluding zero. ACC, Atlantic Coast cluster; GCC, Gulf Coast cluster; WMC, west of the Mississippi cluster; MAF, minor allele frequency; H_{OBS}, observed heterozygosity; H_{EXP}, expected heterozygosity; F_{IS}, Wright's inbreeding coefficient; F_{ST}, fixation index for populations within regions.

*Populations are defined to be counties.

†The average amount of missing data per SNP.

‡Fixation index of populations within regions. Each region was analysed separately.

§The number of SNPs that are fixed for the region.

however, have limited annotation information composed only of sequence similarity to Sitka spruce (*Picea sitchensis* (Bong.) Carr.) ESTs.

Expected and observed heterozygosity across SNPs were within the ranges of those reported previously for conifers and differed on average, with means of $H_{EXP} = 0.271$ and $H_{OBS} = 0.261$, respectively. This translated into a multilocus estimate of F_{IS} that was significantly positive (F_{IS}: 0.033, 99% CI: 0.028–0.036). Estimates of r^2 confirmed the assumption of independence among SNPs, with only 28 pairs of SNPs, out of 1.496×10^6 unique pairwise comparisons, having r^2 values greater than 0.80. The multilocus average (± one standard deviation) of r^2 was 0.00327 ± 0.00504. Of the 28 pairs of SNPs with $r^2 > 0.80$, 17 pairs were within 0.5 cm of each other on the loblolly pine genetic map and three were from different amplicons located at the same genetic locus, while eight were unmapped. All SNPs were kept for further analysis.

Genetic structure was apparent among counties and for counties nested within regions. The minor allele frequencies across SNPs varied significantly among counties, with a multilocus F_{ST} of 0.043 (99% CI: 0.041–0.046). Further hierarchical structure accounted for the bulk of this differentiation, where regions accounted for approximately 70% of the variance in MAF among counties. This translated into a multilocus estimate for the hierarchical fixation index among regions (F_{RT}) of 0.029 (99% CI: 0.027–0.031). Regional structure was also correlated with climate variables, latitude and longitude (Fig. 2). However, counties within regions still exhibited significant genetic (Table 2) and climatic (Fig. 2) structuring, with those counties located to the west of the Mississippi River displaying the greatest magnitude of genetic differentiation. Regions also varied for expected and observed heterozygosity, MAF, F_{IS} and the number of fixed polymorphisms. With respect to the latter two, F_{IS} was significantly positive for all three regions, which likely reflects the further genetic structure among and within counties nested in regions. This was also apparent for the number of fixed polymorphisms (Table 2). Further summaries of genetic diversity and population structure are given in Fig. S6 (Supporting information).

Environmental association analysis

Elements of the correlation matrix across the 54 sampled populations ranged from 0.83 to 0.91, with CV among four draws from the posterior distribution for Ω

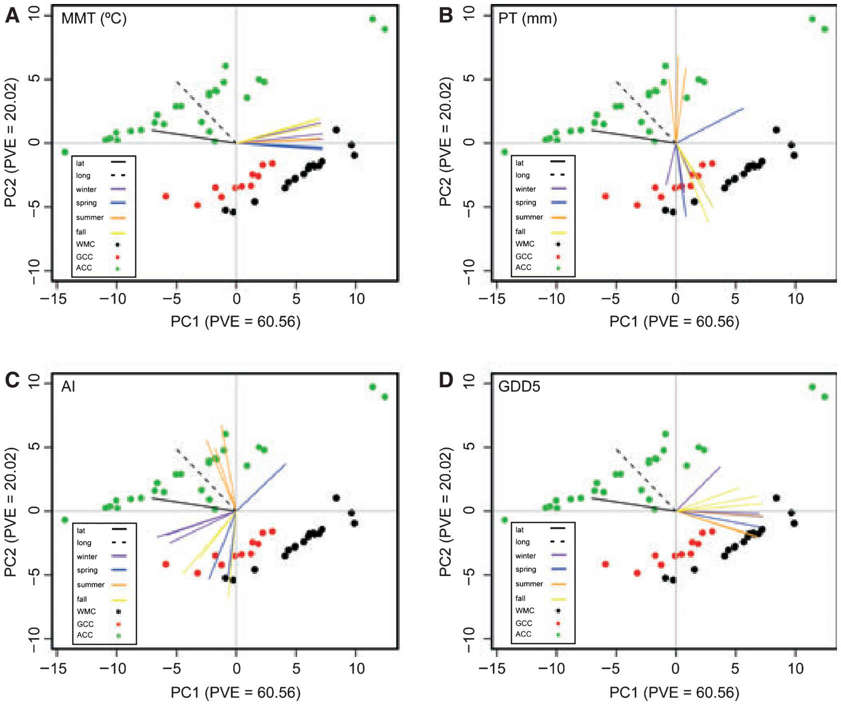

Fig. 2 The confounding between geography, genetic structure and climate variables for the first two PCs is apparent in a biplot parsed by climate variable groups. Climate variables are coloured by season in each panel, while points are coloured according to the three genetic clusters identified by Eckert *et al.* (2010) —WMC: west of the Mississippi cluster, GCC: Gulf Coast cluster and ACC: Atlantic Coast cluster. (A). Biplot for latitude, longitude and monthly minimum and maximum temperatures (MMT). (B) Biplot for monthly for latitude, longitude and monthly precipitation (PT). (C) Biplot for latitude, longitude and monthly AI. (D) Biplot for monthly for latitude, longitude and GDD5.

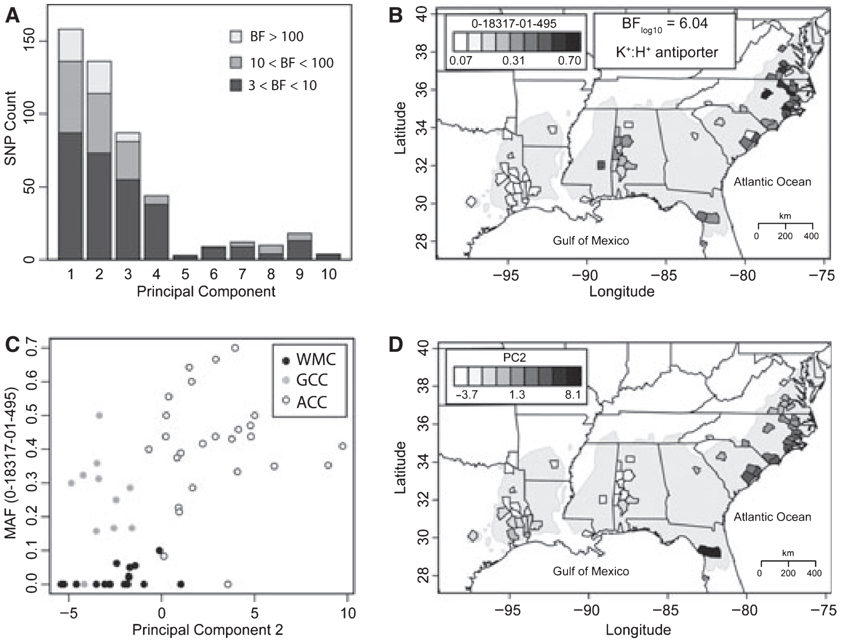

Fig. 3 Application of Bayesian geographical analysis discovered 441 supported (BF > 3) climate associations representing 394 unique SNPs. (A) The number of supported SNPs per geoclimatic PC grouped by category of support. (B). The geographical distribution of the MAF at a SNP located in locus 0–18 317. The BF for the SNP and the annotation information for this locus are given in the inlaid box. (C). A plot of the MAF at SNP 0-18317-01-495 and PC2. Colours denote the three regions (i.e. genetic clusters) identified previously for loblolly pine—WMC: west of the Mississippi cluster, GCC: Gulf Coast cluster and ACC: Atlantic Coast cluster. (D) The geographical distribution for the scores on PC2. This component is largely composed of spring and summer precipitation, summer and fall aridity and longitude.

less than 5% for most elements (Figs S7 and S8, Supporting information). Hierarchical clustering of row and columns of this matrix detected three groups corresponding largely to those identified previously (results not shown; cf. Eckert *et al.* 2010). The WMC region was the most distinct from the other two and the GCC region was nested within a group containing the southwestern most populations of the ACC region.

A total of 394, 118 and 48 SNPs had BFs greater than three, 10 or 100 for at least one geoclimatic PC. These correspond to categories of positive, strong and very strong support according to the Jeffrey's scale for BFs (Jeffreys 1961), though we note that noise not captured by our model can inflate BFs making such interpretations tentative. The vast majority of these SNPs had their highest support for PC1, PC2 and PC3 (Fig. 3A), with the largest BFs observed for PC1 and PC2. These three PCs had 48 SNPs with BFs over 100 (Tables 4, 5 and S2, Supporting information), with only two SNPs being shared across the first 3 PCs (0-12845-02-451 and 2-1087-01-86). Approximately two-thirds of these 48 SNPs were located in genes with sequence similarity to known genes in Arabidopsis (Tables 4 and 5), and of those approximately 40% were nonsynonymous, or linked to nonsynonymous, polymorphisms. The remaining SNPs were largely located in synonymous sites or 3′-untranslated regions (3′-UTR). Twenty-eight (58%) of the 48 SNPs with BFs > 100 are mapped onto the existing linkage map for loblolly pine (Fig. 4). These SNPs are distributed across the genome, with linkage group eight containing the largest number of these loci (n = 5). Many of the 48 SNPs had strong regional differences in allele frequencies (Fig. S9, Supporting information). This was most apparent for SNPs correlated with PC2 and PC3, which are loaded primarily with longitude, precipitation and aridity.

Many of the strongly supported SNPs were located in genes that are prime targets, based on their inferred functions, for a role in adaptation along the gradient described by the PC. For example, a SNP in the 0-18387 locus was fixed primarily in the WMC region (Figs 3B and C) and was associated with higher summer and fall aridity (Fig. 3D). This locus encodes a K$^+$:H$^+$ antiporter, a class of genes often implicated in drought and salt tolerance (Table 3). Several well-supported SNPs, however, resided in genes encoding unknown or hypothetical proteins (Tables 4 and 5).

Discussion

Temperature and precipitation drive large-scale distribution patterns for forest trees (Morgenstern 1996). It is

Table 3 A large fraction (~23%) of the examined SNPs (n = 1730) had moderate to strong support for associations with geoclimatic PCs. A total of 481 correlations representing 394 unique SNPs had BFs greater than 3.0 for PC1 through PC5. Listed are the top three SNPs per PC along with functional annotations derived from tBLASTx analysis of the loblolly pine mRNA (i.e. EST contig) used for design of PCR primers for SNP discovery against the refseq RNA database for Arabidopsis

		BF*†			Exemplar loci		
PC	Description	0.5	1.0	2.0	EST contig‡	BF*	Annotation§
1	Latitude, longitude, temperature, GDD5, winter aridity	87	49	22	0–12 076$^{\mathrm{ns}}$	3.32	Hypothetical protein (At1g01500)
					2–3592$^{\mathrm{nc}}$	3.30	Hypothetical protein (At4g24090)
					0–12 452$^{\mathrm{nc}}$	3.29	Ca^{+2} dependent kinase (At1g05410)
2	Longitude, spring-fall aridity, precipitation	73	41	22	0–18 317$^{\mathrm{syn}}$	6.04	K^+:H^+ antiporter (At2g19600)
					0–8922$^{\mathrm{nc}}$	5.28	TIFY domain-containing protein (At4g32570)
					2–4102$^{\mathrm{syn}}$	5.10	BAG protein (At5g07220)
3	Winter and summer precipitation, summer aridity	55	26	6	CL2115Contig1$^{\mathrm{nc}}$	3.71	Dehydratase like protein (At1g76150)
					0–6445$^{\mathrm{nc}}$	3.37	Hypothetical protein (At3g12650)
					2–10 309$^{\mathrm{nc}}$	2.95	Thioredoxin-related protein (XM_002283585)*
4	Spring and fall precipitation and aridity	38	6	0	UMN-4618$^{\mathrm{nc}}$	1.64	PTAC2 like protein (At1g74850)
					2–10 235$^{\mathrm{nc}}$	1.56	Hypothetical protein (At4g10430)
					CL1905Contig1$^{\mathrm{nc}}$	1.37	LIM transcription factor (At1g10200)
5	Winter aridity and GDD5	2	1	0	CL1054Contig1$^{\mathrm{nc}}$	1.18	Hypothetical protein (AM432844)*
					0–6409$^{\mathrm{ns}}$	0.53	PPR protein (At4g02750)
					CL2339Contig1$^{\mathrm{nc}}$	0.48	Histone 2A protein (At1g54690)

*Bayes factors (BFs) are given on a $\log_{10}$ scale.

†Listed are counts of the number of SNPs with BFs ($\log_{10}$) in the range 0.5–1.0, 1.0–2.0 and 2.0+ for each PC.

‡Genotyped SNPs within these EST contigs are labelled as nonsynonymous (ns), synonymous (syn) or noncoding (nc) using superscripts. Superscripts that are underlined denote those SNPs in linkage disequilibrium with a nonsynonymous polymorphism in the SNP discovery panel.

§The gene model from Arabidopsis is listed in parentheses. Asterisks denote loci annotated using gene models from grapevine (Vitis).

expected that much of the signal of adaptation among forest tree populations, therefore, should occur along these gradients. The genetic basis of adaptive phenotypes related to climate has been established for a multitude of tree species using common gardens. The identification of the genes that underlie these adaptive traits, however, has remained largely elusive, yet emerging association genetic studies (cf. González-Martínez *et al.* 2008) have begun to identify handfuls of such loci. Here, we present an approach that uses environmental data associated directly to wild collected samples in combination with genomic data to make inferences about those loci most correlated with specific climatic gradients. The strongest signals (i.e. BFs > 100) came from 48 SNPs correlated to PCs describing overall temperature, GDD5 and winter aridity (PC1), overall aridity and precipitation (PC2) and precipitation and aridity during the summer and winter (PC3).

Despite substantial advances in the development of genomic resources for several crops (e.g. rice, barley, tomato, wheat) and wild herbs (e.g. Arabidopsis and its relatives), our understanding of the molecular basis of plant adaptation is still incomplete (Alonso-Blanco *et al.* 2009). About half of these potentially adaptive SNPs were identified in genes that code for proteins with unknown function in the Arabidopsis genome or that lack sequence similarity to genes in model plants, although much of this discrepancy could also be due to SNP discovery occurring primarily in 3′-UTR thus limiting sequence similarity between Arabidopsis and pine. The tree growth habit presents life history and physiological particularities that could explain a different response to selection pressure (Niklas 1997). Trees have long been noticed for rapid local adaptation in heterogeneous environments despite large amounts of gene flow and slow rates of evolution (Petit & Hampe 2006 and references therein). Yet trees differ little from herbs at the molecular level (Groover 2005). Differences at the gene and expression levels, therefore, may have relevant developmental consequences and be important to functional responses by trees (Quesada *et al.* 2008). In this context, the comparative analysis of the recently sequenced poplar genome (Tuskan *et al.* 2006), as well as others from woody long-lived perennials such as grapevine (Jaillon *et al.* 2007), is a first step towards a better understanding of the differences in response to environmental variables among life forms at the molecular level.

Table 4 Summary of SNPs (n = 22) with very strong support (BF > 100) for PC one. Dashes indicate that the EST contig had no significant similarity to known gene models in Arabidopsis

SNP*	EST Contig†	BF‡	AT locus§	E-value¶	Annotation	F_{CT} ᶠ	F_{RT} ᶠ
0-12076-01-311	0–12076^{ns}	3.32	At1g01500	8E-05	Hypothetical protein	0.17	0.13
2-3592-01-261	2–3592^{nc}	3.30	At4g24090	1E-39	Hypothetical protein	0.14	0.13
0-12452-03-87	0–12452^{nc}	3.30	At1g05410	4E-09	Ca^{+2} dependent kinase	0.15	0.13
0-17238-01-294	0–17238^{nc}	2.94	At3g28050	3E-20	Nodulin MtN21 family protein	0.21	0.20
2-10235-03-158	2–10235^{nc}	2.82	At4g10430	1E-17	Hypothetical protein	0.22	0.20
0-17251-01-147	0–17251	2.69	–	–	–	0.13	0.07
2-3444-01-348	2–3444^{syn}	2.60	At5g51030	1E-27	Short-chain dehydrogenase/reductase	0.11	0.11
2-10488-01-373	2–10488^{nc}	2.42	At4g10500	1E-36	2OG-Fe(II) oxidoreductase	0.12	0.11
0-12845-02-451	0–12845	2.41	–	–	–	0.19	0.17
UMN-1598-02-647	UMN-1598^{nc}	2.36	At1g35910	1E-49	Trehalose-6-phosphate phosphatase	0.10	0.09
CL305Contig1-05-251	CL305Contig1^{syn}	2.34	At1g48030	0.0	Mitochondrial lipoamide dehydrogenase	0.06	0.03
0-11969-01-142	0–11969	2.33	–	–	–	0.19	0.17
UMN-6523-01-130	UMN-6523	2.32	–	–	–	0.27	0.25
2-1087-01-86	2–1087^{nc}	2.23	At1g04860	2E-15	Ubiquitin-specific protease	0.26	0.26
CL763Contig1-06-141	CL763Contig1^{ns}	2.19	At1g24620	9E-38	Polcalcin	0.09	0.07
UMN-6124-03-373	UMN-6124^{ns}	2.19	At5g53900	1E-41	Hypothetical protein	0.14	0.11
0-13732-01-60	0–13732	2.15	–	–	–	0.12	0.10
0-14570-02-234	0–14570	2.14	–	–	–	0.06	0.04
CL1376Contig1-04-223	CL1376Contig1	2.13	–	–	–	0.15	0.14
0-10663-01-214	0–10663	2.10	–	–	–	0.07	0.03
0-12683-01-213	0–12683^{ns}	2.10	At3g55060	9E-27	Hypothetical protein	0.13	0.13
0-7535-02-83	0–7535^{ns}	2.01	At5g46060	3E-35	Hypothetical protein	0.17	0.15

*Genotyped SNPs are labelled as nonsynonymous (ns), synonymous (syn) or noncoding (nc) using superscripts. Superscripts that are underlined denote those SNPs in linkage disequilibrium with a nonsynonymous polymorphism in the SNP discovery panel. SNPs located in genes without functional annotations are not labelled with respect to these categories.

†The locus in loblolly pine defined by clustering of ESTs into unique contigs for which PCR primers were designed for SNP discovery.

‡BF on a $\log_{10}$ scale.

§The gene model from Arabidopsis.

¶E-value from tBLASTx analysis of the EST contigs against the refseq RNA database for Arabidopsis.

ᶠValues of hierarchical fixation indices estimated using variance components (F_{CT}: among counties; F_{RT}: among regions). The ratio of F_{RT} to F_{CT} gives the amount of among population variance in allele frequencies accounted for by the regions.

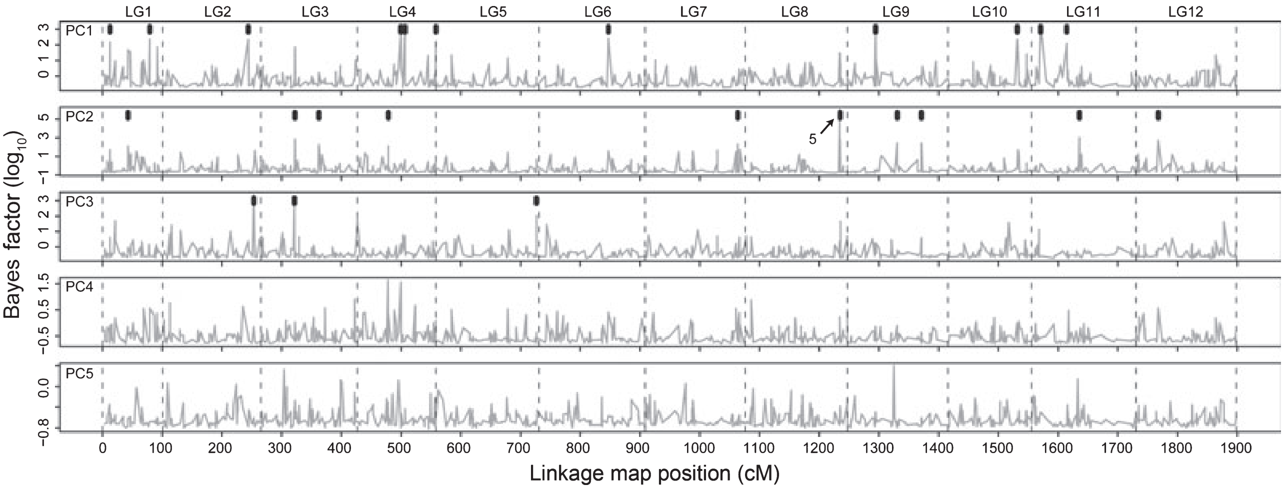

Fig. 4 Markers identified as putatively involved with local adaptation (BFs) are distributed across the genome of loblolly pine. Illustrated are plots of marker position (n = 807 mapped SNPs) in cumulative centiMorgans (cM) versus BFs ($\log_{10}$) for PC1 through PC5. Black dots mark those loci with BFs > 100. Vertical dashed lines demarcate the 12 linkage groups. Lines were smoothed using the loess function available in the R computing environment.

Table 5 Summary of SNPs (n = 22) with very strong support (BF > 100) for PC two. Dashes indicate that the EST contig had no significant similarity to known gene models in Arabidopsis

SNP*	EST Contig†	BF‡	AT locus§	E-value¶	Annotation	F_{CT}**	F_{RT}**
0-18317-01-495	0–18317^{syn}	6.04	At2g19600	8E-13	K$^+$:H$^+$ antiporter	0.22	0.19
0-8922-01-655	0–8922^{nc}	5.28	At4g32570	1E-11	TIFY domain-containing protein	0.23	0.21
2-4102-01-756	2–4102^{syn}	5.10	At5g07220	2E-51	BAG protein	0.16	0.12
UMN-CL194 Contig1-04-130	UMN-CL194 Contig1^{syn}	4.49	At1g47970	1E-04	Hypothetical protein	0.23	0.21
0-6427-02-341	0–6427	4.40	–	–	–	0.22	0.20
0-7881-01-382	0–7881	3.01	–	–	–	0.10	0.07
0-9092-01-302	0–9092^{ns}	2.85	At5g18120	8E-14	Thioredoxin-like protein	0.14	0.13
0-11531-01-379	0–11531^{ns}	2.76	At2g18500	2E-23	Ovate family protein	0.23	0.21
0-12845-02-451	0–12845	2.75	–	–	–	0.19	0.17
0-881-01-114	0–881	2.73	–	–	–	0.08	0.06
UMN-4764-02-149	UMN-4764	2.68	–	–	–	0.07	0.06
0-13762-01-190	0–13762^{nc}	2.46	At3g07130	9E-89	Purple acid phosphatase	0.14	0.11
UMN-7192-01-175	UMN-7192^{ns}	2.43	At1g55240	2E-23	Hypothetical protein	0.06	0.04
0-8304-02-414	0–8304^{ns}	2.40	At5g57480	2E-49	AAA-type ATPase	0.05	0.03
CL3131Contig1-04-127	CL3131 Contig1^{syn}	2.39	At3g18560	4E-08	Hypothetical protein	0.02	0.02
0-6963-01-204	0–6963^{nc}	2.32	At3g33520	3E-03	Actin-related protein	0.08	0.06
CL66Contig4-04-149	CL66 Contig4^{nc}	2.30	At1g71695	5E-107	Peroxidase	0.07	0.06
0-7044-01-319	0–7044	2.28	–	–	–	0.11	0.10
0-14826-01-190	0–14826^{syn}	2.25	At5g51270	4E-24	Protein kinase family protein	0.12	0.09
0-7001-01-143	0–7001^{ns}	2.13	At3g02280	4E-66	Flavodoxin family protein	0.15	0.14
0-9340-01-203	0–9340^{syn}	2.13	At2g20190	4E-31	CLIP-associated protein	0.05	0.03
2-1087-01-86	2–1087^{nc}	2.10	At1g04860	2E-15	Ubiquitin-specific protease	0.25	0.25

*Genotyped SNPs are labelled as nonsynonymous (ns), synonymous (syn) or noncoding (nc) using superscripts. Superscripts that are underlined denote those SNPs in linkage disequilibrium with a nonsynonymous polymorphism in the SNP discovery panel. SNPs located in genes without functional annotations are not labelled with respect to these categories.

†The locus in loblolly pine defined by clustering of ESTs into unique contigs for which PCR primers were designed for SNP discovery.

‡BF on a $\log_{10}$ scale.

§The gene model from Arabidopsis.

¶E-value from tBLASTx analysis of the EST contigs against the refseq RNA database for Arabidopsis.

**Values of hierarchical fixation indices estimated using variance components (F_{CT}: among counties; F_{RT}: among regions). The ratio of F_{RT} to F_{CT} gives the amount of among population variance in allele frequencies accounted for by the regions.

Many of the potentially adaptive SNPs are related to different mechanisms of plant response to abiotic stress, although some of them belong to large gene families where only a few members are well studied in model organisms. This is the case of calcium-dependent protein kinases (CPDKs, 0-12452-03-87), which are key in signal transduction including osmotic stress signaling (Bartels & Sunkar 2005), short-chain dehydrogenase/reductases (SDRs, 2-3444-01-348) and TIFY domain-containing proteins (0-8922-01-655). Proteins with TIFY domains, for example, respond to a wide range of abiotic stresses in rice, including drought, salinity and low temperature, but many of the diverse biologic processes in which these proteins are involved are still unknown (Ye *et al.* 2009). Trees have different strategies to either avoid or tolerate drought and salt stresses (cf. Newton *et al.* 1991), which results in complex patterns at the expression level (cf. Lorenz *et al.* 2005 for loblolly pine). In our study, we have identified potentially adaptive genes representing a wide range of physiological processes, including oxidative stress (e.g. oxidoreductases, 2-10488-01-373; peroxidases, CL66Contig4-04-149; thioredoxin-like proteins, 0-9092-01-302), cell membrane related (e.g. nodulin MtN21 family protein, 0-17238-01-294; K$^+$:H$^+$ antiporter, 0-18317-01-495; fasciclin-like arabinogalactan, 2-3236-01-225) and sugar metabolism (e.g. trehalose-6-phosphate phosphatases, UMN-1598-02-647). Trehalose-6-phosphate phosphatases (TPPs) catalyse the biosynthesis of trehalose-6-phosphate (T-6-P), a precursor of trehalose that is involved in the regulation of sugar metabolism (Eastmond *et al.* 2003). TPP genes are differentially regulated in response to a variety of abiotic stresses and transgenic plants over-expressing trehalose display increased levels of drought, salt and cold tolerance (Iordachescu & Imai 2008). Accumulation of sugars and compatible solutes is one of the most widespread responses to osmotic stress for both plants and animals (Bartels & Sunkar 2005). Interestingly, the only SNP located in a gene with known gene function that was correlated with both PC1 (latitude, temperature and GDD5) and PC2 (longitude, precipitation and aridity) comes from an ubiquitin-specific protease (2-1087-01-86). Ubiquitins are small, highly conserved regulatory proteins that are very abundant in eukaryotes. Their main function is to label proteins for proteasomal degradation. Increased protein degradation in response to environmental stress has been observed in plants as a way to eliminate damaged proteins or to mobilize nitrogen.

Geography can create genetic structure that is correlated to environmental gradients solely through neutral processes such as barriers to gene flow, distance effects (Manel *et al.* 2003; Vasemägi 2006; Storfer *et al.* 2007; Guillot 2009). This is in addition to historical demographic processes that shape population-level patterns of diversity (see Soto *et al.* 2010). Neutral genetic structure affects every locus in the genome, yet a realization of this structure at any particular locus varies wildly. To identify loci that have been subject to natural selection a wide range of tests have been developed (e.g. Nielsen 2005; Zhai *et al.* 2009); two of the most commonly employed are environmental association (Vasemägi & Primmer 2005) and F_{ST} outlier analyses (Storz 2005). Environmental gradients are defined a priori in the environmental association approach, and so loci that show strong associations with these gradients, once population structure has been accounted for, represent good candidates for the functional responses to these gradients. Indeed many of the well-supported SNPs were located within genes encoding proteins likely affecting abiotic stress responses.

In contrast, methods such as F_{ST} outlier analysis are agnostic about the selection pressures and merely require one to define sets of populations (Manel *et al.* 2009; Nosil *et al.* 2009). However, outliers identified using F_{ST} are not necessarily the best candidates for a particular selection pressure, and the interpretation of the selection pressure acting on these outliers is necessarily *post hoc*. For example, Coop *et al.* (2009) showed that the upper tail of pairwise estimates of F_{ST} for human populations is enriched for nonsynonymous polymorphisms suggesting the action of natural selection. The genes containing these SNPs, however, represented a diversity of physiologically and environmentally important genes presuming adaptive along many different gradients. This was shown also for loblolly pine by comparing lists of loci identified as F_{ST} outliers with those composed of loci associated to aridity after corrections for population structure (Eckert *et al.* 2010). No overlap was found between these lists, suggesting that selection along aridity gradients was not occurring at the evolutionary scale of ancestral groups (i.e. genetic clusters), but rather at the scale of local populations. Yet, an assessment of F_{ST} outliers would have identified aridity as one of the major environmental correlates of extreme F_{ST} values, because this measure varies along the same axis as neutral patterns of population structure.

Surprising in our results, because of the rapid decay of linkage disequilibrium in forest trees (Neale & Savolainen 2004), is the set of five SNPs located on linkage group eight that are mapped on top of each other (i.e. zero cM distances), all of which exhibit high degrees of linkage disequilibrium and which all have large values of F_{ST} and extreme BFs for PC2. Such patterns are consistent with linked selection across genomic regions with variable recombination rates, as well as genomic islands of differentiation (Nosil *et al.* 2009). This highlights the need to better understand linkage disequilibrium and rates of recombination in forest trees (see Manel *et al.* 2010a, Keinan & Reich 2010) at multiple genomic scales, especially since many of the strongest associations identified here came from SNPs located in noncoding regions of genes (Tables 4 and 5).

Our application of Bayesian geographical analysis highlights the strengths and weaknesses of using approaches that correct for neutral patterns of population structure. Many false positives are avoided by using corrections for population structure. With respect to Bayesian geographical analysis, power has been shown to be high and many of the true correlations reside in the upper extreme of the distribution of BFs (Coop *et al.* 2010). Corrections for neutral patterns of population structure, however, are conservative when gradients of population structure covary with environmental gradients. Thus, strong candidates identified via measures that do not adjust for population structure should not be treated as uninteresting, just interpreted carefully. The SNP located in locus 2-7808-01 is one such example that displays highly structured allele frequencies and thus a large degree of differentiation among populations (F_{CT} = 0.23), yet had a BFs below one for all five geoclimatic PCs.

We stress also that the significance of a correlation at any SNP depends on the extent to which population structure has been corrected and the form of climatic effects that were modelled. Even methods that correct for population structure assess significance under some model, which is surely only an approximation to reality. We recommend that future studies take an empirical approach and compare the signals found at a set of carefully matched control loci to those at candidate genes (cf. Hancock *et al.* 2008), or, when possible, validate signals along independent clines (Holderegger *et al.* 2008; e.g. Turner *et al.* 2008). We assumed, moreover, linear effects of climate in our Bayesian model, yet nonlinear effects could result near range margins for many species as well as from complicated source-sink dynamics of gene flow (Kirkpatrick & Barton 1997; Savolainen *et al.* 2007).

Forest trees have long been recognized for pervasive patterns of local adaptation (Morgenstern 1996; see also discussion by Savolainen *et al.* 2007). Here, we are working with spatial units much larger than those typically employed in genecological studies (but see St Clair *et al.* 2005; St Clair 2006), thus establishing the question of how local is local when it comes to adaptation (cf. Holderegger *et al.* 2008; Manel *et al.* 2010b). Much of the previous work correlating genetic markers to environmental and climatic variables in forest trees has occurred at refined spatial scales (e.g. Gram & Sork 2001). Our samples span the entire 370 000 km^2 range of loblolly pine and were aggregated into populations with spatial areas comprising approximately 0.5% of this geographical extent on average. Thus, our results apply only at this spatial scale, and consideration of this point makes clear why many of the highly associated SNPs have large allele frequency differences among populations and regions. A similar result was noted by Manel *et al.* (2010b) in an alpine herb after partitioning spatial effects using principal coordinate analysis of neighbour matrices, where molecular adaptation was inferred at local and regional scales.

Natural selection operates at finer spatial scales than addressed here (Epperson 2000). The data at hand, however, are not well suited to addressing questions at finer spatial scales (see Anderson *et al.* 2010). Much of this is due to the type of geo-referencing for our samples, as well as their clumped distribution across the range of loblolly pine. The degree to which genetic data will reflect the action of natural selection, moreover, is affected by numerous factors including the strength of selection relative to the magnitude and spatial patterning of gene flow (García-Ramos & Kirkpatrick 1997). Given the high levels of pollen dispersal for this species (Williams 2010), as well as the observation that county labels within regions account for approximately 30% of the allele frequency differences across our samples on average, our choice of populations may be a decent first approximation to evolutionary relevant units. Other methods employed at finer spatial scales would of course be invaluable to dissecting the genetic basis of adaptation in forest trees (see Sork *et al.* 2010; Manel *et al.* 2010b). This would, however, require more intensive sampling of trees across the landscape, fine-scale climate data and better coverage of the functional and regulatory gene space during polymorphism discovery than we have here (e.g. Storz *et al.* 2007). Indeed, with respect to the latter, a recent study of full gene sequences in population samples of spruce, as opposed to genotyped SNPs identified in small gene fragments, resulted in much stronger signals of selection than observed typically in forest trees (Namroud *et al.* 2010).

A pluralistic approach using a range of methods might be best suited for inferences of natural selection from molecular genetic data, especially when applied to non-model organisms and matched with appropriate study designs and research questions (cf. Manel *et al.* 2009; Eckert *et al.* 2010). While there is a need to better understand the strengths and weaknesses of the quantitative methods employed during scans for local adaptation from genetic data, there is an equally important need to understand what to do once plausible adaptive genetic variation is identified. Correlations should be validated through functional assays, transcriptomic profiling and quantitative genetic dissection. With respect to the latter, correlation analyses such as employed here are aptly suited to integration with association genetic and genecological studies that identify the genetic basis of ecologically relevant phenotypes. The need for appropriate theory to understand the spatial aspects of both adaptive genetic variation and environmental variation is paramount (cf. Sokal *et al.* 1989), especially when viewed as a tool to integrate disparate inferential approaches such as those used here and molecular population genetic approaches reliant on the site-frequency spectrum (see also Manel *et al.* 2003, 2009). The demands of climate modelling and GIS theory will also be equally demanding (cf. Wang *et al.* 2006). Results such as these will also increasingly have large practical consequences, as the focus of conservation and applied population genetics broadens to include empirical studies of putatively adaptive genetic diversity, as well as genomic patterns of diversity and gene flow (Luikart *et al.* 2003; Holderegger *et al.* 2006; Sork & Smouse 2006).

Advancements in high-throughput sequencing and genotyping technologies (reviewed by Gilad *et al.* 2009) have enabled the transition of landscape studies away from patterns of neutral genetic diversity to studying putatively adaptive genes across complex environmental gradients for non-model organisms. This transition is enabled by the sheer amount of genomic data made available by these technologies that can cover many genes, and in some cases the entire genome, of the organisms under study. While more decisive evidence for diversifying selection may come from quantitative genetics (cf. Holderegger *et al.* 2006) and functional studies (but see Nielsen 2009), landscape genomics offers great promise in two respects. First, the link between genetic variation and environmental variation is directly assessed. Second, loci identified in analyses such as these are prime candidates for further experimentation and meta-analysis, where sets of 'interesting' loci are intersected among landscape, population and association genomic studies. Here, we have identified several sets of SNPs, including a core set of 48 well-supported SNPs, consistent with diversifying selection along climate gradients for loblolly pine. This represents a first step towards understanding the molecular basis of ecologically relevant genetic variation for this species and forest trees in general, and when combined with emerging association genetic and traditional genecological studies, offers a way to prioritize gene conservation efforts in the face of climate change.

Acknowledgements

The authors would like to thank 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 W. Patrick Cumbie and Barry Goldfarb for providing geographic information for sampled trees and Aslam Mohamed for creating the linkage map. Comments from four anonymous reviewers much improved the manuscript. This work was supported by the National Science Foundation [grant number DBI-0501763].

References

Aitken SN, Hannerz M (2001) Genecology and gene resource management strategies for conifer cold hardiness. In:*Conifer Cold Hardiness* (eds Bigras FJ, Colombo SJ). pp. 23–53, Kluwer Academic Publishers, The Netherlands.

Alía R, Moro J, Denis JB (1997) Performance of *Pinus pinaster* Ait. provenances in Spain: interpretation of the genotype-environment interaction. *Canadian Journal of Forest Research*, **27**, 1548–1559.

Alonso-Blanco C, Aarts MGM, Bentsink L *et al.* (2009) What has natural variation taught us about plant development, physiology and adaptation? *The Plant Cell*, **21**, 1877–1896.

Al-Rabab'ah MA, Williams CG (2002) Population dynamics of *Pinus taeda* L. based on nuclear microsatellites. *Forest Ecology and Management*, **163**, 263–271.

Al-Rabab'ah MA, Williams CG (2004) An ancient bottleneck in the Lost Pines of central Texas. *Molecular Ecology*, **13**, 1075–1084.

Anderson CD, Epperson BK, Fortin M-J *et al.* (2010) Considering spatial and temporal scale in landscape-genetic studies of gene flow. *Molecular Ecology*, **19**, 3565–3575.

Aranda I, Alía R, Ortega U *et al.* (2010) Intra-specific variability in biomass partitioning and carbon isotopic discrimination under moderate drought stress in seedlings from four *Pinus pinaster* populations. *Tree Genetics and Genomes* **6**, 169–178.

Bartels D, Sunkar R (2005) Drought and salt tolerance in plants. *Critical Reviews in Plant Sciences*, **24**, 23–58.

Beaulieu J, Perron M, Bousquet J (2004) Multivariate pattern of adaptive genetic variation and seed source transfer in *Picea mariana*. *Canadian Journal of Forest Research*, **34**, 531–545.

Bower AD, Aitken SN (2008) Ecological genetics and seed transfer guidelines for *Pinus albicaulis* (Pinaceae). *American Journal of Botany*, **95**, 66–76.

Coop G, Pickrell JK, Novembre J *et al.* (2009) The role of geography in human adaptation. *PLoS Genetics*, **5**, e1000500.

Coop G, Witonsky D, Di Rienzo A *et al.* (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, doi:10.1534/genetics.110.114819.

Eastmond PJ, Li Y, Graham IA (2003) Is trehalose-6-phosphate a regulator of sugar metabolism in plants? *Journal of Experimental Botany*, **54**, 533–537.

Eckert AJ, Bower AD, Wegrzyn JL *et al.* (2009a) Association genetics of coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. *Genetics*, **182**, 1289–1302.

Eckert AJ, Pande B, Ersoz ES *et al.* (2009b) High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). *Tree Genetics and Genomes*, **5**, 225–234.

Eckert AJ, van Heerwaarden J, Wegrzyn JL *et al.* (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*, doi:10.1534/genetics.110.115543.

Epperson BK (2000) *Geographical Genetics*. Princeton University Press, Princeton, New Jersey.

Eveno E, Collada C, Guevara MA *et al.* (2008) Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Molecular Biology and Evolution*, **25**, 417–437.

Felsenstein J (2002) Contrasts for a within-species comparative method. In: *Modern Developments in Theoretical Population Genetics: The Legacy of Gustave Malecot* (eds Slatkin M, Veuille M), pp. 118–129. Oxford University Press, Oxford.

Forsythe WC, Rykiel Jr EJ, Stahl RS *et al.* (1995) A model comparison for daylength as a function of latitude and day of year. *Ecological Modelling*, **80**, 87–95.

Gabriel KR (1971) The biplot graphical display of matrices with application to principal component analysis. *Biometrika*, **58**, 453–467.

García-Ramos G, Kirkpatrick M (1997) Genetic models of adaptation and gene flow in peripheral populations. *Evolution*, **51**, 21–28.

Gilad Y, Pritchard JK, Thornton K (2009) High-throughput (a.k.a. ''next-gen'') sequencing. *Trends in Genetics*, **25**, 463–471.

González-Martínez SC, Krutovsky KV, Neale DB (2006) Forest tree population genomics and adaptive evolution. *New Phytologist*, **170**, 227–238.

González-Martínez SC, Wheeler NC, Ersoz ES *et al.* (2007) Association genetics in *Pinus taeda* L. I. Wood property traits. *Genetics*, **175**, 399–409.

González-Martínez SC, Huber D, Ersoz ES *et al.* (2008) Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. *Heredity*, **101**, 19–26.

Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical *F*-statistics. *Molecular Ecology Notes*, **5**, 184–186.

Gram WK, Sork VL (2001) Association between environmental and genetic heterogeneity in forest tree populations. *Ecology*, **82**, 2012–2021.

Grattapaglia D, Plomion C, Kirst M *et al.* (2009) Genomics of growth traits in forest trees. *Current Opinion in Plant Biology*, **12**, 148–156.

Groover AT (2005) What genes make a tree a tree? *Trends in Plant Science*, **10**, 210–214.

Guillot G. (2009) On the inference of spatial genetic structure from population genetics data. *Bioinformatics*, **25**, 1796–1801.

Hancock AM, Witonsky DB, Gordon AS *et al.* (2008) Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics*, **4**, e32.

Holderegger R, Kamm U, Gugerli F (2006) Adaptative vs. neutral genetic diversity: implications for landscape genetics. *Landscape Ecology*, **21**, 797–807.

Holderegger R, Herrmann D, Poncet B *et al.* (2008) Land ahead: using genome scans to identify molecular markers of adaptive relevance. *Plant Ecology and Diversity*, **1**, 273–283.

Howe GT, Aitken SN, Neale DB *et al.* (2003) From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. *Canadian Journal of Botany*, **81**, 1247–1266.

Iordachescu M, Imai R (2008) Trehalose biosynthesis in response to abiotic stresses. *Journal of Integrative Plant Biology*, **50**, 1223–1229.

Jaillon O, Aury JM, Noel B *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Jeffreys H (1961) *The Theory of Probability*, 3rd edn. Oxford University Press, Oxford.

Keinan A, Reich D (2010) Human population differentiation is strongly correlated with local recombination rate. *PLoS Genetics*, **6**, e1000886.

Kimura M, Maruyama T (1971) Pattern of neutral polymorphism in a geographically structured population. *Genetical Research*, **18**, 125–131.

Kirkpatrick M, Barton NH (1997) Evolution of a species' range. *American Naturalist*, **150**, 1–23.

Langlet O (1971) Two hundred years of genecology. *Taxon*, **20**, 653–722.

Linhart YB, Grant MC (1996) Evolutionary significance of local genetic differentiation in plants. *Annual Review of Ecology and Systematics*, **27**, 237–277.

Little Jr EL (1971) *Atlas of United States Trees, Vol. 1, Conifers and Important Hardwoods*. U.S. Department of Agriculture Miscellaneous Publication 1146, Washington.

Lorenz WW, Sun F, Liang C *et al.* (2005) Water-stress-responsive genes in loblolly pine (*Pinus taeda*) roots identified by analyses of expressed sequence tag libraries. *Tree Physiology*, **26**, 1–16.

Luikart G, England PR, Tallman D *et al.* (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Manel S, Schwartz MK, Luikart G *et al.* (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, **18**, 189–197.

Manel S, Conord C, Despres L (2009) Genome scan to assess the respective roles of host-plant and environmental constraints on the adaptation of a widespread insect. *BMC Evolutionary Biology*, **9**, 288.

Manel S, Joost S, Epperson BK *et al.* (2010a) Perspectives on the use of landscape genetics to detect genetic variation in the field. *Molecular Ecology*, **19**, 3760–3772.

Manel S, Poncet BN, Legendre P, Gugerli F, Holderegger R (2010b) Common factors drive genetic variation of adaptive relevance at different spatial scales in *Arabis alpina*. *Molecular Ecology*, **19**, 3824–3835.

Morgenstern EK (1996) *Geographic Variation in Forest Trees*. UBC Press, Vancouver, BC.

Namkoong G (1979) *Introduction to Quantitative Genetics in Forestry*. USDA Forest Service Tech. Bull. No. 1588, Washington.

Namroud M-C, Beaulieu J, Juge N *et al.* (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology*, **17**, 3599–3613.

Namroud M-C, Guillet-Claude C, Mackay J *et al.* (2010) Molecular evolution of regulatory genes in spruces from different species and continents: Heterogeneous patterns of linkage disequilibrium and selection but correlated recent demographic changes. *Journal of Molecular Evolution*, **70**, 371–386.

Neale DB (2007) Genomics to tree breeding and forest health. *Current Opinion in Genetics and Development*, **17**, 1–6.

Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. *Current Opinion in Plant Biology*, **11**, 1–7.

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. *Trends in Plant Science*, **9**, 325–330.

Newton RJ, Funkhouser EA, Fong F *et al.* (1991) Molecular and physiological genetics of drought tolerance in forest trees. *Forest Ecology and Management*, **43**, 225–250.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Nielsen R (2009) Adaptionism – 30 years after Gould and Lewontin. *Evolution*, **63**, 2487–2490.

Niklas KJ (1997) *The Evolutionary Biology of Plants*. University of Chicago Press, Chicago.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annual Review of Ecology and Systematics*, **37**, 187–214.

Quesada T, Li Z, Dervinis C *et al.* (2008) Comparative analysis of the transcriptomes of *Populus trichocarpa* and *Arabidopsis thaliana* suggests extensive evolution of gene expression regulation in angiosperms. *New Phytologist*, **180**, 408–420.

R Development Core Team, 2007 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org.

Rehfeldt GE (1989) Ecological adaptations in Douglas fir (*Pseudotsuga menziesii* var. *glauca*): a synthesis. *Forest Ecology and Management*, **28**, 203–215.

Rehfeldt GE (1990) Genetic differentiation among populations of *Pinus ponderosa* from the upper Colorado River Basin. *Botanical Gazette*, **151**, 125–137.

Rehfeldt GE (2006) *A Spline Climate Model for the Western United States*. USDA Forest Service Rocky Mountain

Research Station General Technical Report 165, Fort Collins, CO.

Rehfeldt GE, Ying CC, Spittlehouse DL *et al.* (1999) Genetic responses to climate in *Pinus contorta*: niche breadth, climate change, and reforestation. *Ecological Monographs*, **69**, 375–407.

Rehfeldt GE, Tchebakova NM, Parfenova YI *et al.* (2002) Intraspecific responses to climate in *Pinus sylvestris*. *Global Change Biology*, **8**, 912–929.

Richardson BA, Rehfeldt GE, Kim M-S (2009) Congruent climate-related genecological response from molecular markers and quantitative traits for western white pine (*Pinus monticola*). *International Journal of Plant Sciences*, **170**, 1120–1131.

Savolainen O, Pyhäjärvi T (2007) Genomic diversity in forest trees. *Current Opinion in Plant Biology*, **10**, 162–167.

Savolainen O, Pyhäjärvi T, Knürr T (2007) Gene flow and local adaptation in forest trees. *Annual Review of Ecology, Evolution and Systematics*, **38**, 595–619.

Schmidtling RC (2001) *Southern Pine Seed Sources*. USDA Forest Service Southern Research Station General Technical Report 44, Asheville, NC.

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

Schmidtling RC, Carroll E, LaFarge T (1999) Allozyme diversity of selected and natural loblolly pine populations. *Silvae Genetica*, **48**, 35–45.

Sierra-Lucero V, Huber DL, McKeand SE *et al.* (2002) Performance differences and genetic parameters for four coastal provenances of loblolly pine in the southeastern United States. *Forest Science*, **48**, 732–742.

Sierra-Lucero V, McKeand SE, Huber DL *et al.* (2003) Genotype-by-environment interaction and deployment considerations for families from Florida provenances of loblolly pine. *Forest Genetics*, **10**, 85–92.

Sokal RR, Jacquez GM, Wooten MC (1989) Spatial autocorrelation analysis of migration and selection. *Genetics*, **121**, 845–855.

Soltis DE, Morris AB, McLachlan JS *et al.* (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261–4293.

Sork VL, Smouse PE (2006) Genetic analysis of landscape connectivity in tree populations. *Landscape Ecology*, **21**, 821–836.

Sork VL, Davis FW, Westfall R *et al.* (2010) Gene movement and genetic association with regional climate gradients in California valley oak (*Quercus lobata* Nee) in the face of climate change. *Molecular Ecology*, **19**, 3806–3823.

Soto A, Robledo-Arnuncio JJ, González-Martínez SC *et al.* (2010) Climatic niche and neutral genetic diversity of the six Iberian pine species: a retrospective and prospective view. *Molecular Ecology*, **19**, 1396–1409.

Squillace AE, Wells OO (1981) Geographic variation of monoterpenes in cortical oleoresin of loblolly pine. *Silvae Genetica*, **30**, 127–135.

St Clair JB (2006) Genetic variation in fall cold hardiness in coastal Douglas-fir in western Oregon and Washington. *Canadian Journal of Botany*, **84**, 1110–1121.

St Clair JB, Mandel NL, Vance-Borland KW (2005) Genecology of Douglas fir in western Oregon and Washington. *Annals of Botany*, **96**, 1199–1214.

Storfer A, Murphy MA, Evans JS *et al.* (2007) Putting the 'landscape' in landscape genetics. *Heredity*, **98**, 128–142.

Storz JF (2005) Using genome scans of DNA variability to infer the genetic basis of adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Storz JF, Sabatino SJ, Hoffmann FG *et al.* (2007) The molecular basis of high-altitude adaptation in deer mice. *PLoS Genetics*, **3**, e45.

Thornthwaite CW (1948) An approach toward a rational classification of climate. *Geographical Review*, **38**, 55–94.

Turner TL, Levine MT, Eckert ML *et al.* (2008) Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. *Genetics*, **179**, 1–19.

Tuskan GA, DiFazio S, Jansson S *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Vasemägi A (2006) The adaptive hypothesis of clinal variation revisited: single-locus clines as a result of spatially restricted gene flow. *Genetics*, **173**, 2411–2414.

Vasemägi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Wang T, Hamann A, Spittlehouse DL *et al.* (2006) Development of scale-free climate data for western Canada for use in resource management. *International Journal of Climatology*, **26**, 383–397.

Wei R-P, Han SD, Dhir NK *et al.* (2004) Population variation in growth and 15-year-old shoot elongation along geographic and climatic gradients in black spruce in Alberta. *Canadian Journal of Forest Research*, **34**, 1691–1702.

Wells OO, Wakeley PC (1966) Geographic variation in survival, growth and fusiform infection of planted loblolly pine. *Forest Science Monographs*, **11**, 1–40.

Wells OO, Switzer GL, Schmidtling RC (1991) Geographic variation in Mississippi loblolly pine and sweetgum. *Silvae Genetica*, **40**, 105–118.

Williams CG (2010). Long-distance pine pollen still germinates after meso-scale dispersal. *American Journal of Botany*, **97**, 846–855.

Xu S, Tauer CG, Nelson CD (2008) Genetic diversity within and among populations of shortleaf pine (*Pinus echinata* Mill.) and loblolly pine (*Pinus taeda* L.). *Tree Genetics and Genomes*, **4**, 859–868.

Ye H, Du H, Tang N *et al.* (2009) Identification and expression profiling analysis of TIFY family genes involved in stress and phytohormone responses in rice. *Plant Molecular Biology*, **71**, 291–305.

Yu J, Pressoir G, Briggs WH *et al.* (2006) A unified mixed-model method for association mapping accounting for multiple levels of relatedness. *Nature Genetics*, **38**, 203–208.

Zhai W, Nielsen R, Slatkin M (2009) An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Molecular Biology and Evolution*, **26**, 273–283.

Andrew Eckert is a population biologist with interests in population, ecological evolutionary and landscape genomics of forest trees and the native plants of western North America, Andrew Bower is a geneticist for the USDA Forest Service, with research interests including population and ecological genetics of forest trees and native plants. Santiago González-Martínez has broad interests in population genetics and genomics of forest trees Mediterranean plant endemics. Jill Wegrzyn currently works in the field of bioinformatics with research interests in scientific databases, machine-learning, functional genomics, and proteomics. Graham Coop is a population and evolutionary geneticist interested in adaption and selection, the evolution of recombination rates and model development and application for understanding population histories. David Neale is a geneticist interested in the discovery and understanding of functional genetic variation underlying adaptive traits in forest trees.

Supporting information

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

Fig. S1 Visualization of the correlation matrix for 62 climate and geographical variables reveals strong correlations among variables. Labels on the bottom axis denote groups of climate variables, which are listed on the right vertical axis. Single monthly names refer to the aridity index. The image was generated using the HEATMAP function in the base distribution of R (R Development Core Team 2007).

Fig. S2 Scree plots summarizing the application of PCA on 62 climate and geographical variables. (A). Scree plot based upon eigenvalues. (B). Scree plot based on PVE.

Fig. S3 Loadings for PCs 1–10 for 62 climate and geographical variables. Types of variables are differentiated by colour (black: latitude and longitude; blue: minimum monthly temperatures; red: maximum monthly temperatures; orange: monthly aridity index; green: monthly precipitation; yellow: growing-degree days above 5 °C). The PVE is listed for each PC.

Fig. S4 A summary of GO categories nested under the term molecular function for all 1730 genes. (A) The fraction of 1730 genes assigned a term nested within the molecular function category. (B). The distribution of level three GO terms for those genes assigned a GO term for a molecular function.

Fig. S5 A summary of GO categories nested under the term biological process for all 1730 genes. (A) The fraction of 1730 genes assigned a term nested within the molecular function category. (B). The distribution of level three GO terms for those genes assigned a GO term for a biological process.

Fig. S6 Distributions of diversity metrics for the 1730 SNPs genotyped using the Illumina Infinium platform are consistent with expectations for conifers and previous applications of this technology (Eckert *et al.* 2010). The distributions shown in panels A, B and D were smoothed using the DENSITY function in the base distribution of R (R Development Core Team 2007). (A). The distributions of expected (*H*EXP) and observed (*H*OBS) heterozygosity. Vertical dashed lines denote averages. (B). The distribution of Wright's inbreeding coefficient (*F*IS). The vertical solid line denotes the average, while the vertical dashed red lines define the 99% bootstrap confidence interval (*n* = 10 000 replicates) for *F*IS. (C). The allele frequency spectrum for the minor allele visualized as a set of 50 bins. The vertical red line denotes the minimum frequency threshold of 2.5% used for inclusion of SNPs. (D). The distributions of hierarchical fixation indices across loci reveal moderate differentiation among 54 populations across the genome of loblolly pine (*F*RT: hierarchical fixation index for regions relative to total; *F*CR: hierarchical fixation index for counties nested within regions; *F*CT: hierarchical fixation index for counties relative to the total). Further information is available in the Materials and Methods section. Vertical dashed lines denote averages.

Fig. S7 The correlation matrix estimated from a single draw from the posterior distribution of covariance matrices estimated using MCMC reveals broad patterns of population structure consistent with three genetic groups. Draws from the posterior distribution of Ω were similar, with coefficients of variation typically < 5% for most elements (Fig. S8). Colours denote correlations among counties reordered with hierarchical clustering (Euclidean distance and Ward's linkage method). The colour scale ranges from yellow (higher correlations) to red (lower correlations). Labels refer to the genetic clusters identified previously (WMC: west of the Mississippi cluster, GCC: Gulf Coast cluster and ACC: Atlantic Coast cluster).

Fig. S8 Variation among draws of the variance-covariance matrix from its posterior distribution was small. Most elements of this matrix had a coefficient of variation less than 5%. Colours denote values ranging from a maximum of 6.9% (yellow) to a minimum of 0.001% (red).

Fig. S9 Relationship between climate variation and minor allele frequencies for the top two SNPs per climate PC as determined using Bayesian geographical analysis. Colours denote regional populations (black = WMC, green = GCC, red = ACC). Regional populations are defined in the Material and methods.

Table S1 Factor loadings for the top five PCs resulting from PCA on 62 geographical and climate variables

Table S2 Summary of SNPs (*n* = 6) with very strong support (BF > 100) for PC3. This PC is comprised largely of winter and summer precipitation and summer aridity. Dashes indicate that the EST contig had no significant similarity to known gene models in *Arabidopsis*

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.