Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold-hardiness in coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*)


*Section of Evolution and Ecology, University of California at Davis, Davis, CA 95616, USA
§Center for Population Biology, University of California at Davis, Davis, CA 95616, USA
**Department of Plant Sciences, University of California at Davis, Davis, CA 95616, USA
§§Institute of Forest Genetics, Pacific Southwest Research Station, U.S. Department of Agriculture Forest Service, Placerville, CA 95667, USA
***Department of Ecosystem Science and Management, Texas A&M University, College Station, TX 77843, USA
Running Title: Positive selection at cold-hardiness related candidate genes

Key Words: coalescent, cold-hardiness, coastal Douglas-fir, demographic inference, selective sweeps

Author for correspondence:

David B. Neale
Department of Plant Sciences
Mail Stop 6
University of California, Davis
Davis, CA 95616
(530) 754-8431
dbneale@ucdavis.edu
Abstract

Forest trees exhibit remarkable adaptations to their environments. The genetic basis for phenotypic adaptation to climatic gradients has been established through a long history of common garden, provenance and genecological studies. The identities of genes underlying these traits, however, have remained elusive, and thus so have the patterns of adaptive molecular diversity in forest tree genomes. Here, we report an analysis of diversity and divergence for a 121 cold-hardiness candidate gene set in coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*). Application of several different tests for neutrality, including those that incorporated demographic models, revealed signatures of selection consistent with selective sweeps at 3 to 8 loci, depending upon the severity of a bottleneck event and the method used to detect selection. Given the high levels of recombination, these candidate genes are likely to be closely linked to the target of selection if not the genes themselves. Putative homologs in Arabidopsis act primarily to stabilize the plasma membrane and protect against denaturation of proteins at freezing temperatures. These results indicate that surveys of nucleotide diversity and divergence, when framed within the context of further association mapping experiments, will come full circle with respect to their utility to the dissection of complex phenotypic traits into their genetic components.
INTRODUCTION

A long history of common garden experiments has shown that forest trees are adapted phenotypically to local and regional climates (MORGENSTERN 1996). Latitudinal clines for phenological traits have been documented across a diverse set of temperate forest tree species ranging from Norway spruce (*Picea abies* (L.) Karst.) to European aspen (*Populus tremula* L.). The frequent nature of these patterns, as well as the high heritabilities and polygenic nature for many of these traits (HOGE *et al.* 2003), suggest that forest tree genomes should be rife with signatures of positive natural selection.

Early efforts to detect these signatures using molecular markers were largely unsuccessful (MCKAY AND LATTA 2002). One of the limiting factors for these efforts was the genomic distribution, availability and nonfunctional nature of the genetic markers that were utilized (MERKLE AND ADAMS 1987). The development of high-throughput allele-specific marker detection for large sets of functional genes has revolutionized the search for adaptive genetic variation within conifers (GONZALEZ-MARTINEZ *et al.* 2006). Surveys of nucleotide diversity across a diverse set of conifer species have highlighted common genomic attributes, such as the rapid decay of linkage disequilibrium (LD), and have documented the first tangible evidence of natural selection in forest tree genomes (SAVOLAINEN AND PYHÄJÄRVI 2007; EVENO *et al.* 2008; PALMÉ *et al.* 2008, 2009).

Here, we take a candidate gene approach to the identification of genes underlying cold-hardiness phenotypes in coastal Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco var. *menziesii*). A 40 year history of common garden experiments, provenance trials and geneecological studies have documented the genetic basis of growth, germination and bud phenology traits, as well as the steep clines in genetic

The structure of these clines is consistent with the hypothesis that diversifying selection (Mckay and Latta 2002) or local selective sweeps (Slatkin and Wiehe 1998; Santiago and Caballero 2005) should characterize most of the signatures of positive natural selection in the genome of coastal Douglas-fir. If this is the case, we expect to observe signatures of elevated levels of intermediate frequency polymorphisms in the site frequency spectrum for those genes that underlie these clinal traits.

While the phenotypes involved with adaptation to cold temperatures are well known, the identities of the genes underlying these traits have remained elusive.

Several quantitative trait loci (QTLs) have been identified (Jermstad et al. 2001a,b, 2003) and a handful of candidate genes have been shown to collocate with these QTLs (Wheeler et al. 2005), but the vast majority of candidate genes putatively involved with cold-hardiness remain to be investigated. The genes underlying physiological responses to cold, however, are well known in Arabidopsis thaliana L. (Thomashow 1999). Similar types of genes are thought to underlie cold-hardiness related traits in conifers (Yakovlev et al. 2005; Holliday et al. 2008). An initial survey of nucleotide diversity at 18 candidate genes in coastal Douglas-fir, however, found only limited departures from neutrality (Krutovsky and Neale 2005). The vast majority of candidate genes, as well as inference of demographic influences upon genomic patterns of nucleotide diversity thus remain to be investigated.

Mounting evidence indicates that patterns of positive selection are pervasive throughout the genomes of Drosophila (Begun et al. 2007), Arabidopsis (Clark et al.
2007) and humans (Voight et al. 2006). Evidence on a genome-scale is only beginning
to emerge for conifers (Namroud et al. 2008). We used an expanded set of candidate
genes to address three goals: (i) describe patterns of polymorphism and divergence
within a set of 121 candidate genes for cold-hardiness, (ii) identify those genes with
patterns of polymorphism and divergence inconsistent with neutral expectations and (iii)
discuss the types and prevalence of selection across a conifer genome.

MATERIALS AND METHODS

Plant material and DNA isolation: A sample of 24 unrelated trees obtained
from six regions located across Washington and Oregon was assembled as a diversity
panel (Table S1) for the discovery of single nucleotide polymorphisms (SNPs). These
regions are situated along temperature, precipitation and geneecological gradients
identified previously for coastal Douglas-fir in Oregon and Washington (St. Clair et al.
macrocarpa [Vasey] Mayr) collected from the Angeles National Forest was used as an
outgroup. Total genomic DNA was isolated from haploid megagametophyte tissues
dissected out from multiple open-pollinated seeds collected from each tree using a
DNeasy® 96 Plant Kit (Qiagen, Valencia, CA). All DNA extractions and quantifications
were performed at the USDA National Forest Genetics Electrophoresis Laboratory
(Placerville, CA USA).

Choice of candidate genes: Candidate genes with a putative role in conferring
tolerance to cold temperatures were selected according to three criteria: (i) genes found
to collocate with QTLs for cold-hardiness in Douglas-fir (Wheeler et al. 2005), (ii) genes
with physiological roles in cold tolerance response and (iii) genes showing differential
expression in microarray studies in *Arabidopsis thaliana* L. (FOWLER AND THOMASHOW 2002; LEE et al. 2005). Standard BLAST tools were used to mine the Douglas-fir expressed sequence tags (EST) collections housed at Virginia Polytechnic University (http://staff.vbi.vt.edu/estap) for orthologous sequences to selected candidate genes. Those collections represent four EST libraries, three of which are cold hardiness libraries derived from one year old seedlings subjected to fall, winter and spring ambient air conditions in Corvallis, Oregon during 2003. A threshold of 1E^-10 was used to define homology during BLAST analyses. In total, 553 putative homologs were selected for primer design, of which 378 were able to have primers designed successfully. A total of 121 of those 378 genes were selected based on diversity of function and sequence-based primer validation from a single Douglas-fir sample (Tables S2 and S3).

**PCR, DNA sequencing and sequence analysis:** Sequence data for each candidate gene were obtained using standard PCR and Sanger sequencing protocols (Table S4). The Pine Alignment and SNP Identification Pipeline (PineSAP) was used to assemble, call bases, align and identify SNPs from the 5,220 sequence reads (WEGRZYN et al. 2009). The final alignments for each candidate gene were generated using a minimum Phred quality score of 20 for base calls associated with SNPs. All alignments were further trimmed, edited and validated using CodonCode Aligner v. 2.0.4 (CodonCode, Dedham, MA). Gene annotation, data summaries and laboratory protocols are given at our website (http://dendrome.ucdavis.edu/dfgp).

**SNP genotyping and linkage mapping analyses:** A subset of the 933 discovered SNPs (n = 384) was chosen to validate polymorphism identifications made by PineSAP and to investigate the distribution of those polymorphisms across existing
linkage maps of Douglas-fir through the genotyping of 192 progeny from a three-
generation outbred pedigree (cf. JERMSTAD et al. 1998) using Illumina’s GoldenGate™
SNP genotyping platform. This high-throughput platform works well for the large and
complex genomes of conifers (Pavy et al. 2008). Multiplexed genotyping was conducted
at the DNA Technologies Core facility located at the University of California, Davis
Genome Center. Signal intensities were quantified and matched to specific alleles using
the BeadStudio v 3.1.14 software (Illumina). A minimum GenCall50 (GC50) score of 0.20
was chosen for inclusion of SNP loci in the final data set and genotypic clusters were
edited manually as needed (Eckert et al. 2008).

The existing linkage maps for coastal Douglas-fir are based largely on
anonymous RFLP markers. Here, we used 189 fully informative RFLP markers, three
isozyme markers and the 77 segregating SNPs to infer a sex-averaged genetic linkage
map using JoinMap v. 1.4 (Stam 1993). All SNP loci were assessed for conformity to
Mendelian segregation ratios prior to linkage mapping using goodness-of-fit $\chi^2$
tests. We
removed SNPs from the same candidate gene that mapped to less than 0.5 cM from
each other prior to producing the final genetic map. Generation of linkage maps within
this pedigree is described in detail elsewhere (Jermstad et al. 1998, 2001a,b, 2003;
Krutovsky et al. 2004). In brief, we used a LOD threshold of 4.0 to assign markers to
linkage groups followed by a threshold of 0.10 to order markers within those inferred
linkage groups. Recombination frequencies were converted to map distances in cM
using the Kosambi mapping function. Linkage map results are deposited and freely
available at the forest tree CMAP website (http://dendrome.ucdavis.edu/cmap/;
accession no. TG05).
Linkage disequilibrium analyses: Linkage disequilibrium (LD) between pairs of sites within each candidate gene and between pairs of sites in different candidate genes mapped to the same linkage group was evaluated by using DNASP v. 4.50.3 (ROZAS et al. 2003). We used the squared allelic correlation coefficient ($r^2$) between pairs of parsimony informative sites to estimate LD. We used Fisher exact tests to calculate $P$-values under the null hypothesis that LD estimates at least as extreme as those observed could have arisen under linkage equilibrium. The decay of LD with distance in base pairs (bp) between sites within the same candidate gene was evaluated with nonlinear regression using the Gauss-Newton algorithm as implemented in the *nls* function of R (HILL AND WEIR 1988; REMINGTON et al. 2001).

Population structure and historical demography: The extent of structure among populations of coastal Douglas-fir observed at chloroplast and nuclear microsatellites, as well as allozymes, is low and consistent with weak effects of isolation by distance (LI AND ADAMS 1989; VIARD et al. 2001; KRUTOVSKY et al. 2009). We analyzed patterns of differentiation for all 115 polymorphic candidate genes using pairwise measures of divergence ($D_{xy}$) among the six populations and related this measure to geographical distance using Mantel tests.

To assess the effects of historical demography, especially given the weak patterns of population structure, on the ability to identify candidate genes under selection, we used a likelihood-free Markov chain Monte Carlo (MCMC) approach to estimate parameters under an instantaneous growth model (MARJORAM et al. 2001; MARJORAM AND TAVARÉ 2006). This model is specified by two parameters, the ratio ($f$) of the current effective population size ($N_{ec}$) to the effective population size prior to the
growth event \((N_{ep})\) and the time in units of \(4N_{ec}\) generations at which the growth event began \((\tau)\). We assigned uniform priors to both parameters, used the average Tajima’s \(D\) over the 18 loci analyzed by KRUTOVSKY AND NEALE (2005) as the metric \((D = -0.248)\) and specified a rejection threshold \((\epsilon)\) of 0.05.

Parameters were estimated with Markov chains that were run for \(1.2 \times 10^6\) steps with the first \(0.2 \times 10^6\) steps being discarded as burn-in and a thinning interval of 100 steps. The resulting 10,000 samples for each parameter were taken as an approximation of their unknown posterior distributions. The joint posterior distribution of \(f\) and \(\tau\) was smoothed using the kernel smoothing functions available in \(R\). Convergence was assessed through three independent MCMC runs by analyzing similarities between the resulting approximate posterior distributions and using effective sample size (ESS) estimates calculated using the CODA ver. 0.13-2 package in \(R\).

To understand the effects of more complicated demographic scenarios on our abilities to detect loci under selection, we also analyzed two bottleneck models, where we set the time of the bottleneck to occur at 10,000 \((\text{BIM}_10; \tau = 0.001)\) or 100,000 \((\text{BIM}_100; \tau = 0.015)\) years ago with bottleneck sizes that were 1% of the current population size. The duration of the bottleneck was set to 0.005 coalescent time units, and the estimate of \(\Theta\) upon which the simulations were conditioned was adjusted so that the average simulated value of \(\theta_{\pi}\) per locus was equal to the observed value for each of the 18 loci used to fit the model. Model fit was assessed by simulating 10,000 sets of 18 loci for which we calculated the average and standard deviation of Tajima’s \(D\). We used these simulated sets of data to differentiate among models (instantaneous growth or the two bottleneck models) using rejection sampling \((\epsilon = 0.05)\) for both \(D\) and
the standard deviation of $D$) and the relationship between acceptance rates and Bayes factors ($BF = \frac{P(S|BIM)}{P(S|IGM)}$, where $P(S|Model)$ is given by the acceptance rate and represents the probability of the observed statistics given the model). The observed values of average $D$ and the standard deviation of $D$ were -0.248 and 0.837, respectively (KRUTOVSKY AND NEALE 2005). Following JEFFREYS (1998), we considered values greater than $10^{-1/2}$ for the BF as “substantial” evidence for a given model (cf. ROSS-IBARRA et al. 2009). This scheme conditions all comparisons on the estimates of $\tau$ and $f$ from the instantaneous growth model (i.e. these were not estimated again) given added demographic effects (e.g. bottlenecks).

**Nucleotide diversity and neutrality tests:** Several estimators of $\Theta = 4Neu$, where $Ne$ is the effective population size and $u$ is the mutation rate, were used to assess the magnitude of nucleotide diversity across candidate genes. A likelihood ratio test was performed subsequently to test the null hypothesis that $\Theta$ is constant across loci (HUDSON 1991). The number of haplotypes ($h$) and haplotypic diversity ($Hd$) were calculated for comparison with diversity estimators based on the site frequency spectrum. Pairwise divergence ($D_{xy}$) for each candidate gene was calculated by comparing the sequences of coastal Douglas-fir to a single sequence from bigcone Douglas-fir. We assumed for all calculations that our data conformed to the infinite sites model. Sites violating these assumptions, or those with missing data, were ignored.

A suite of summary statistics was used to detect departures of the site frequency spectrum from neutral expectations within each candidate gene. The significance of each statistic was determined under three different null models using 10,000 coalescent simulations conditional on the observed value of $\theta_\pi$. when simulating data under the
demographic models, we adjusted the value of $\Theta$ so that the median value of the simulated $\theta_n$ was equal to the observed data. This was done to account for the fact that $\theta_n$ is only an unbiased estimator of $\Theta$ under genetic drift. The positive false discovery rate (FDR) method was used to correct for multiple testing (STOREY 2003). We chose to focus on positive selection, because many of the tests we employed are most powerful in differentiating positive from background selection (ZENG et al. 2006, 2007a,b). We employed the DHEW test to search further for departures from neutrality. This test uses the combination of Tajima’s $D$, Fay and Wu’s normalized $H$ and the Ewens-Watterson test of neutrality to detect positive selection. The rejection region for this test was determined through 50,000 coalescent simulations conditional on the observed value of $\theta_n$. We used a nominal threshold of $P = 0.0005$ in the calculation of this region for each locus to account for the fact that we are performing 115 independent tests.

Lastly, we used polymorphism-divergence based tests to search for candidate genes with reduced levels of polymorphism relative to divergence. A maximum-likelihood implementation of the Hudson-Kreitman-Aguade test (HKA) was carried out with MLHKA ver. 2.0 (WRIGHT AND CHARLESWORTH 2004). We implemented a likelihood-ratio test by choosing randomly 20 candidate genes exhibiting neutral site frequency spectra and three candidate genes exhibiting reduced levels of polymorphism relative to divergence. These three genes were chosen because they had reduced levels of polymorphism relative to divergence ($\theta_n/D_{xy} < 0.15$). The likelihood function was evaluated and parameters estimated under both models using MCMC with $1.0 \times 10^8$ steps. Statistical significance was determined by comparing twice the difference between the log-likelihood scores to a $\chi^2$ distribution with three degrees of freedom,
where in model one it was assumed all 23 loci to be neutral and in model 2 it was assumed that the three focal genes were non-neutral. Three independent MCMC simulations were used to check for convergence of the likelihood score and parameter estimates.

We also computed the ratio of nonsynonymous (Ka) to synonymous (Ks) divergence for each gene using DNASP and correlated the Ka/Ks ratio to the polymorphism/divergence ratio using simple linear regression in R. This ratio is the average of each pairwise estimate of Ka/Ks of the coastal Douglas-fir samples to the single bigcone Douglas-fir. We used bootstrapping (n = 10,000 replicates) across samples to construct 95% and 99% confidence intervals for our measure of Ka/Ks.

**RESULTS**

**Patterns of nucleotide diversity and divergence:** A total of 59,173 bp of aligned sequence data, of which 932 bp had missing, ambiguous or indel data, were generated through the partial resequencing of 121 candidate genes (GenBank accession nos.: EU864545-EU867209). The average length of these gene fragments was 489 bp (± 204 bp). These genes are distributed throughout the genome of coastal Douglas-fir (Fig. S1), with 64 of the 77 segregating SNPs being placed on the existing linkage map. These 64 SNPs represent 42 candidate genes, 36 from this study and six from KRUTOVSKY AND NEALE (2005). Sequence data were obtained for 21 of the 24 samples per gene on average. Of the 121 candidate genes, 115 were polymorphic yielding a total of 933 SNPs. The number of polymorphisms varied greatly across genes with an average of eight SNPs per locus (range: 0-29). The majority of SNPs were silent (n = 732), with 478 of those being located in noncoding regions. The remaining 201
SNPs were nonsynonymous, with an average of two nonsynonymous SNPs per gene (range: 0-12). The average SNP density ranged from a maximum of one SNP per 43 bp for silent sites to a minimum of one SNP per 112 bp for nonsynonymous sites.

Average estimates of nucleotide diversity were on the order of $4.0 \times 10^{-3}$ per site, with silent sites having the greatest ($\theta_W = 0.00777$, $\theta_n = 0.00756$) and nonsynonymous sites the smallest ($\theta_W = 0.00214$, $\theta_n = 0.00200$) average estimates of diversity (Fig. 1, Tables 1 and S5). Diversity at synonymous ($\theta_{ns}$) and silent ($\theta_{nsil}$) sites was approximately four-fold greater than at nonsynonymous sites ($\theta_{na}$), with most genes exhibiting a $\theta_{na}/\theta_{ns}$ ratio less than one ($n = 77$). Six genes, however, had $\theta_{na}/\theta_{ns}$ ratios greater than one, with values ranging from 1.13 to 3.09 (Table S2).

Nucleotide diversity was also heterogeneous across candidate genes. A likelihood ratio test indicated that $\Theta$ significantly varied across genes when considering all sites ($\chi^2 = 156.59$, $df = 114$, $P = 0.005$), with the same result occurring for different categories of sites (silent: $\chi^2 = 189.11$, $df = 111$, $P < 0.001$; nonsynonymous: $\chi^2 = 95.67$, $df = 71$, $P = 0.027$). In general, $\theta_W$ was larger than $\theta_n$ illustrating an excess of rare SNP alleles relative to expectations under neutrality.

All candidate genes exhibited an excess of divergence relative to diversity, with the average divergence estimate approximately two-fold greater than the average diversity estimate (Fig. 1, Tables 1 and S6). The magnitudes of divergence also differed by category of sites. Divergence at synonymous (Ks) and silent (Ksil) sites was greater than at nonsynonymous (Ka) sites, with 76 genes exhibiting a Ka/Ks ratio less than one or zero. However, 18 genes had a Ka/Ks ratio that was greater than one, with values
ranging from 1.01 to 14.48. Two of these 18 genes had bootstrap confidence intervals excluding one (Table 1). The remaining 27 genes had undefined values for Ka/Ks.

**Patterns of intra- and intergenic linkage disequilibrium:** The number of haplotypes at each polymorphic locus varied from two \((H_d = 0.087)\) to 20 \((H_d = 0.991)\), with an average of five \((\bar{H}_d = 0.585)\) per locus. The estimated minimum number of recombination events \((R_M)\) was positively correlated with the number of haplotypes and the number of segregating sites (Kendall’s \(\tau > 0.419, P < 0.001\) ), whereas divergence for all, silent, synonymous and nonsynonymous sites were uncorrelated with \(R_M\).

Nucleotide diversity at synonymous and nonsynonymous sites was also uncorrelated with \(R_M\).

The relationship of the squared allelic correlation coefficient \((r^2)\) with physical distance illustrated that LD decayed approximately 50% over 1,100 bp, with \(r^2\) dropping to < 0.25 by 900 bp (Fig. 2A). Using Fisher’s exact tests, the value of \(r^2\) for 866 of the 2,837 intragenic site pairs remained significant after a Bonferroni correction. Significant intergenic LD was also apparent for candidate genes that were located proximally on linkage groups. For example, linkage group 17 has two candidate genes whose polymorphisms are almost in complete LD despite being ~30 cM apart, while genes farther apart show little LD (Fig. 2B).

**Historical demography:** The instantaneous growth model (IGM) yielded well-defined approximate posterior probability distributions for the parameters of interest. Replicated runs of the MCMC sampler converged to similar posterior probability distributions, as well as to roughly similar point estimates of the parameters (Fig. S2, Table S7). Parameter estimates for this model are consistent with population growth.
producing a modest increase in effective population size \((f = 0.60-0.73)\) beginning approximately 0.039 to 0.045 coalescent time units ago. These translate into minimum age estimates for the onset of growth of approximately 270,000-310,000 years ago if we assume a mutation rate on the order of \(1.0 \times 10^{-9}\) mutations/bp/year (WILLYARD et al. 2006), a generation time of 50 years, and a genome-wide estimate for \(\Theta\) of 5.92.

We also fit the data to two extreme bottleneck models, where the effective population size during the bottleneck was 5% of the current size (Fig. 3). All other parameters were set to the mean values across the MCMC replications (Table S7). When this event was set to occur at 10,000 years ago, the Bayes factor with respect to the instantaneous growth model was 0.15. When the bottleneck, however, was set to be 100,000 years ago the Bayes factor was 2.5. Following JEFFREYS (1998), the IGM fits the data substantially better than the BIM_10 (BF = 6.67, i.e. \(1/0.15\)), but the IGM and BIM_100 models are not differentiated substantially by the data (BF = 2.5). We focus subsequently on the two best models for purpose of neutrality testing.

**Neutrality tests:** Six candidate genes exhibited clear patterns of polymorphism and divergence consistent with natural selection. The results from univariate tests are given in Tables S8 and S9, as well as are \(P\)-values under null models corresponding to Wright-Fisher mating, the instantaneous growth model and the extreme bottleneck model set to occur at 100,000 years ago. On average, the summary statistics reflected the excess of low frequency SNP alleles \((D = -0.142)\) and the slight excess of high-frequency (Fay and Wu’s \(H = -0.106\)) derived SNP alleles.

A compound test based on \(D, H\) and the Ewens-Watterson haplotype test \((DHEW)\) identified four genes as having patterns of diversity consistent with selective
sweeps (Table 2, Fig. S3). Products from all four genes function as structural
components of the plasma membrane or provide protection against temperature-
dependent denaturation of proteins. Inclusion of demographic models (IGM and
BIM_100) into the simulations caused one of these loci (CN634517.1) to become non-
significant.

Three genes exhibited a significant excess of divergence relative to diversity
using a maximum-likelihood implementation of the HKA test ($\chi^2 = 12.60, \text{df} = 3, P =
0.006$). These excesses ranged from a two- to three-fold reduction in diversity relative to
divergence (Table S10, Fig. S3), which was also apparent across different categories of
sites (Fig. S4). All three loci function in generalized stress responses at the levels of
protein folding and transcriptional regulation, with locus Pm_CL908Contig1 also being
significant for the DHEW test. A weak, yet statistically significant relationship, also
existed between $K_a/K_s$ and $\theta_\pi/D_{xy}$ ($F_{1,91} = 4.79, P = 0.031, r^2 = 0.05$). The driver behind
this result was three loci with extreme values of $K_a/K_s$ (i.e. $> 5$, cf. Table 1), all of which
had $\theta_\pi/D_{xy}$ less than 0.15 (Table 2). Removal of these three loci changed the
relationship to be non-significant.

Several additional genes exhibited patterns of polymorphism consistent with
balancing or diversifying selection at a nominal significance level of $P = 0.05$ depending
upon which null model and summary statistic was used (Tables S8 and S9). None of
these were significant using the DHEW test. For three of these five genes, a significant
and positive relationship existed between geographical distance and pairwise $D_{xy}$
among populations (Fig. S5). This pattern is suggestive of diversifying selection along
environmental gradients, especially given that only five candidate genes had a
significant value of Mantel’s $r$ and three of those five had positive values of $D$ that were nominally significant. Two of these genes remained significant at the FDR threshold of $Q = 0.10$ for Tajima’s test when an extreme bottleneck was included as part of the null model. These loci, CN636901.1 and Pm_CL1400Contig1, encode proteins that are involved with abiotic stress response and cell wall architecture, respectively.

**DISCUSSION**

**Nucleotide diversity and linkage disequilibrium:** Nucleotide diversity in coastal Douglas-fir is similar, but slightly higher, to that observed in other conifer species (SAVOLAINEN AND PYHÄJÄRVI 2007). The average nucleotide diversities for all, silent and nonsynonymous sites were similar to those reported previously (KRUTOVSKY AND NEALE 2005). Magnitudes of intragenic LD were moderate as compared to those reported for conifers. Our estimate of recombination ($\rho/\Theta = 0.86$) was similar to that reported in loblolly pine ($\rho/\Theta = 0.29$; BROWN et al. 2004), but almost an order of magnitude less than that reported in Norway spruce ($\rho/\Theta = 1.23-4.05$; HEUERTZ et al. 2006) and the central populations of Scots pine ($\rho/\Theta = 3.80-5.80$; PYHÄJÄRVI et al. 2007).

Interlocus LD was higher than expected based on results from loblolly pine and the previous study in coastal Douglas-fir (BROWN et al. 2004; KRUTOVSKY AND NEALE 2005). While genomic variation in recombination rates has been documented across plants (GAUT et al. 2007) that could reflect repetitive regions of the genome where recombination is suppressed, several additional processes such as natural selection, mutation rate variation, population structure or extreme bottlenecks can also account for these patterns. This is important given that interlocus LD has dramatic effects on the
amount and pattern of genetic diversity and population differentiation, as well as the
response to and efficacy of natural selection, across the genome of an organism (Le
Corre and Kremer 2003).

Patterns of selection: Most genes exhibited an excess of nucleotide diversity at
synonymous relative to nonsynonymous sites, suggesting the widespread occurrence of
across different categories of sites were consistent with this pattern as well (Fig. S6).
These patterns have been noted extensively in forest trees and have driven the
hypothesis that detectable patterns of positive selection in conifers are rare. This
hypothesis has been confirmed in a recent examination of genome-wide patterns of
population differentiation in white spruce (Namroud et al. 2008; but see Eveno et al.
2008). Further examination using phylogenetic approaches, however, has detected
patterns consistent with directional and purifying selection (Palmé et al. 2008, 2009),
suggesting that consideration of relevant time scales is important to the search for
selection.

We concur with this hypothesis showing that approximately 5% of the genes
exhibited patterns consistent with positive selection. Our estimate is likely to be a lower
bound due to low statistical power and the conservative assumptions that we made, as
well as to fact that causal variants may exhibit patterns of polymorphism that are
indistinguishable from neutral variants due strictly to the quantitative genetic
mechanisms underlying the trait (Le Corre and Kremer 2003). Also of consideration
are other models of selective sweeps incorporating selection on a standing crop of
neutral variants (cf. Innan and Kim 2004). In those models, linked neutral variants are
less likely to show the traditional signs of selective sweeps. Contrary to expectations based on patterns of low levels of intragenic LD, the strongest signals came from those genes where patterns were consistent with selective sweeps. The vast majority of candidate genes showed little evidence of elevated levels of intermediate frequency SNP alleles as expected under diversifying selection and recurrent selective sweep models. Under the bottleneck scenarios, however, two genes had significantly positive values of Tajima’s $D$.

It is likely that these patterns represent relatively recent and single, not recurrent, selective sweep events. A genome-wide phenomenon of recurrent selective sweeps would produce average values of $D$ and $H$ that were negative and positive, respectively (Przeworski 2002). Our genes, however, are not random genomic samples, and 42 of the 121 candidate genes exhibit this predicted pattern producing average values of $D = -0.83$ and $H = 0.55$. A conclusion to rule out all occurrences of recurrent sweep models (Gillespie 2001) within conifer genomes, therefore, is premature, especially given our ignorance about large-scale genomic patterns of LD, the spatial distribution of genes across chromosomes and the historical demography of these lineages.

The simplicity of our demographic models may not account for all of the confounding effects of demography on tests for selection. It also is possible that we have missed evidence supportive of other demographic models due to our limited sampling with respect to the range of coastal Douglas-fir. Expansion from bottlenecks could produce patterns consistent with signatures of single and recurrent selective sweeps depending on the timing and severity of the bottleneck. This is the most common form of demographic model fit to European forest trees (Heuertz et al. 2006;
The severity of the bottleneck would have to be extreme to produce genome-wide patterns mirroring single selective sweeps. Incorporation of an extreme bottleneck (99% reduction of $N_{eo}$) occurring 100,000 years ago into the null model caused one of the four genes significant previously in the DHEW test to become non-significant. It remains significant, however, at a nominal threshold of $P = 0.001$.

Paleobotanical and geological evidence indicate that Pleistocene glaciation in North America differed dramatically from that in Europe making an extreme bottleneck model unlikely for coastal Douglas-fir. The spatial extent of continental glaciers was smaller and vegetation was not compressed to the same degree, especially on the west coast (GRAHAM 1999). The most parsimonious demographic model for coastal Douglas-fir is one of population expansion. Inclusion of a bottleneck following the instantaneous expansion, however, did provide a marginally better fit to the data (Fig. 3). The data used to fit test the demographic models did not have outgroup information, and further work, therefore, should examine rigorously a broad range of demographic models (cf. INVGVARSSON 2008).

**Putative targets of selection:** The most significant departures from neutrality came from genes that encoded proteins that are structural components of or associated with cellular membranes or those that function as or regulate chaperones. Locus Pm_CL908Contig1 contains a GRAM domain, which characterizes membrane-associated proteins such as glucosyltransferases (DOERKS et al. 2000; CARO et al. 2007). The Arabidopsis homolog of locus ES420171.1 is a multi-spanning G-protein coupled receptor that stabilizes the plasma membrane in response to freezing temperatures (BRETON et al. 2003). Both CN634517.1 and Pm_CL61Contig1 have
homologs whose primary functions are to act as or in the regulation of chaperones (ROMANO et al. 2004). Locus Pm_CL922Contig1 encodes a thaumatin-like protein, whose homolog in Arabidopsis has gene expression induced by fungal infection (HU AND REDDY 1997). While the link between cold-hardiness and disease resistance is speculative, it is tempting to correlate overall health with the ability to grow and tolerate cold efficiently. Several of these genes are also responsive to abscisic acid (ABA).

Controversy has surrounded the role of ABA during cold acclimation in Arabidopsis, however, due to its general effect on plant growth in response to stress (GILMOUR AND THOMASHOW 1991).

Limitations and implications: There are several limitations to our analyses. We used a single sequence from a recently diverged outgroup so that misspecification of ancestral states is possible. This would reduce greatly the power of the HKA test employed here and significantly affect estimation of $H$ (cf. BAUDRY AND DEPAULIS 2003). This also justifies our avoidance of McDonald-Kreitman type analyses. Background selection could have produced signatures of selective sweeps, but not with respect to $H$ (HUDSON AND KAPLAN 1995). Lastly, we did not incorporate effects of population structure into our hypothesis testing. The effects of undetected structure on inference of selection from summary statistics are tied closely to sampling design and the strength of substructure (PRZEWORSKI 2002; STÄDLER et al. 2009), and thus conclusions regarding the conservative or liberal nature of our estimates are premature.

Selective sweeps, while not likely the dominant type of selection across the genome of coastal Douglas-fir, are nevertheless important. Here, we have identified several genes whose patterns of diversity and divergence are inconsistent with the
standard neutral model, as well as simple demographic scenarios involving instantaneous growth and bottlenecks. Depending upon demographic assumptions, we have identified 3 to 8 such loci. These genes are prime targets for further association genetic and functional studies, thus beginning to bring molecular population and quantitative genetic approaches full circle with respect to the dissection of complex traits.

ACKNOWLEDGMENTS

The authors would like to thank F. Thomas Ledig for contributing bigcone Douglas-fir seeds, Valerie Hipkins and her staff at NFGEL for performing the DNA extractions, Katie Tsang and Jacqueline Silva for helping to obtain sequence data and Jeff Ross-Ibarra for helpful discussion about demographic inference. The manuscript was much improved by comments from two anonymous reviewers. Funding for this project was made available through a USDA NRI Plant Genome grant (04-712-0084).

REFERENCES


INGVARSSON, P., 2008. Multilocus patterns of nucleotide polymorphism and the


PAVY, N., B. PELGAS, S. BEAUSEIGLE, S. BLAIS, F. GAGNON et al., 2008 Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: Application to the large and unsequenced genomes of white and black spruce. BMC Genomics 9: 21.


PYHÄJÄRVI, T., M. R. GARCÍA-GIL, T. KNÜRR, M. MIKKONEN, W. WACHOWIAK et al., 2007 Demographic history has influenced nucleotide diversity in European Pinus sylvestris populations. Genetics 177: 1713–1724.


**Table 1.** Measures of diversity and divergence across 121 candidate genes putatively affecting cold-hardiness in Douglas-fir. Values in bold represent totals for the number of segregating sites ($S$) and averages for measures of diversity ($\theta_W$, $\theta_\pi$) and divergence ($D_{xy}$). Numbers in parentheses are averages ± one standard deviation for $S$ and standard deviations for $\theta_W$, $\theta_\pi$ and $D_{xy}$. Estimates of diversity and divergence are scaled to length of the aligned data ignoring sites with missing data or insertion-deletion polymorphisms (indels).

<table>
<thead>
<tr>
<th></th>
<th>$S$</th>
<th>$\theta_W$</th>
<th>$\theta_\pi$</th>
<th>$D_{xy}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>933</td>
<td>0.00450</td>
<td>0.00435</td>
<td>0.00852</td>
</tr>
<tr>
<td></td>
<td>(8 ± 7)</td>
<td>(0.00331)</td>
<td>(0.00358)</td>
<td>(0.00619)</td>
</tr>
<tr>
<td>Silent</td>
<td>732</td>
<td>0.00777</td>
<td>0.00756</td>
<td>0.01296</td>
</tr>
<tr>
<td></td>
<td>(6 ± 6)</td>
<td>(0.00677)</td>
<td>(0.00765)</td>
<td>(0.01059)</td>
</tr>
<tr>
<td>Synonymous</td>
<td>254</td>
<td>0.00791</td>
<td>0.00760</td>
<td>0.01413</td>
</tr>
<tr>
<td></td>
<td>(2 ± 2)</td>
<td>(0.00821)</td>
<td>(0.00901)</td>
<td>(0.01498)</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>201</td>
<td>0.00214</td>
<td>0.00200</td>
<td>0.00580</td>
</tr>
<tr>
<td></td>
<td>(2 ± 2)</td>
<td>(0.00281)</td>
<td>(0.00297)</td>
<td>(0.00806)</td>
</tr>
</tbody>
</table>
Table 2. A list of candidate genes putatively affected by positive natural selection.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Gene Product</th>
<th>Result(^{a})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compound DHEW test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pm_CL908Contig1</td>
<td>GRAM-containing/ABA-responsive protein</td>
<td>P(_D) = 0.001, P(<em>H) &lt; 0.001, P(</em>{EW}) = 0.080</td>
</tr>
<tr>
<td>ES420171.1</td>
<td>cold-regulated plasma membrane protein</td>
<td>P(_D) = 0.009, P(<em>H) = 0.050, P(</em>{EW}) = 0.035</td>
</tr>
<tr>
<td>ES420250.1</td>
<td>dehydrin-like protein</td>
<td>P(_D) = 0.072, P(<em>H) = 0.083, P(</em>{EW}) = 0.042</td>
</tr>
<tr>
<td>CN634517.1</td>
<td>lumenal binding protein</td>
<td>P(_D) = 0.034, P(<em>H) = 0.148, P(</em>{EW}) = 0.076</td>
</tr>
<tr>
<td><strong>Polymorphism-to-divergence</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pm_CL61Contig1</td>
<td>cyclosporin A-binding protein</td>
<td>k = 0.32</td>
</tr>
<tr>
<td>Pm_CL908Contig1</td>
<td>GRAM-containing/ABA-responsive protein</td>
<td>k = 0.58</td>
</tr>
<tr>
<td>CN634517.1</td>
<td>transcription regulation protein</td>
<td>k = 0.41</td>
</tr>
<tr>
<td><strong>Synonymous-to-nonsynonymous divergence</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pm_CL922Contig1</td>
<td>thaumatin-like protein</td>
<td>Ka/Ks = 14.48** (, \theta_\pi/D_{xy} = 0.087)</td>
</tr>
<tr>
<td>CN634677.1</td>
<td>LRR receptor-like protein kinase</td>
<td>Ka/Ks = 10.78*, (\theta_\pi/D_{xy} = 0.066)</td>
</tr>
</tbody>
</table>

\(^{a}\)Results for the DHEW test are given as P-values (P) for each of the component tests (\(D =\) Tajima’s \(D\); \(H =\) Fay and Wu’s \(H\); \(EW =\) Ewens-Watterson test) comprising the joint test. Values for the EW test are one minus the left tailed probabilities (cf. ZENG et al. 2007b) Listed are P-values for drift within a constant size population. Loci significant when demographic models are included in the simulations are bolded. For polymorphism-to-divergence tests, parameter estimates for a maximum likelihood implementation of the HKA test are listed. Estimates of \(k\) are from a nested model where all three putative targets of selection are allowed to have free parameters. The parameter \(k\) specifies the level of
elevation ($k > 1$) or reduction ($k < 1$) in diversity relative to divergence. Asterisks denote significance differences from 1.0 for the Ka/Ks ratio using bootstrapping. **$P < 0.01$, *$P < 0.05$ (cf. MATERIALS AND METHODS).
**FIGURE LEGENDS**

**FIGURE 1.** The distribution of diversity and divergence across different categories of sites. Values are given on a per site scale. Divergence is measured against bigcone Douglas-fir. Vertical lines extend to 1.5 times the inter-quartile range. S, synonymous sites; Sil, silent sites; NS, nonsynonymous sites; Total, all sites.

**FIGURE 2.** Patterns of intra- and interlocus linkage disequilibrium (LD) for coastal Douglas-fir. (A) Intragenic patterns of LD using only parsimony informative sites. The fitted lines (black) are the expectation of $r^2$ (plus a 95% confidence interval for $C$) using the nonlinear least squares estimate of $C = 4N_e r$, where $N_e$ is the effective population size and $r$ is the recombination rate. The thick black line is that given for the decay of LD in a previous study of coastal Douglas-fir (KRUTOVSKY AND NEALE 2005). (B) Patterns of intergenic LD along linkage group 17. Dark lines delineate comparisons within and among candidate genes, with letters indicating placement of the candidate genes within the matrix.

**FIGURE 3.** The distribution of and historical demographic inferences for coastal Douglas-fir. (A) The distribution of coastal Douglas-fir in Oregon and Washington. Locations for samples are indicated with filled circles and grouped into local populations. (B) Patterns of population structure in coastal Douglas-fir. Data and results are from KRUTOVSKY et al. (2009), who used 1283 samples and 31 allozyme and microsatellite loci. Illustrated is the smallest value of $K$ judged as optimal using the $\Delta K$ method (EVANNO et al. 2005) for
categories corresponding to individuals (upper) and populations (lower). (C)

Demographic inferences for coastal Douglas-fir. The solid line represents the average value of Tajima’s $D$ across the 18 loci used to fit the model (cf. KRUTOVSKY AND NEALE 2005), whereas the dashed line denotes the mean of the simulated distribution under each model (IGM, instantaneous growth model; BIM_10, bottleneck 10,000 years ago; BIM_100, bottleneck model 100,000 years ago). Listed for the BIM models are Bayes factors relative to the IGM.
A

Squared allelic correlation coefficient ($r^2$)

Distance (bp)

4N_e r = 0.00376
95% CI: 0.00308-0.00452

B

<table>
<thead>
<tr>
<th>Locus</th>
<th>cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pm_CL1868Contig1 (A)</td>
<td>0.0</td>
</tr>
<tr>
<td>Pm_CL1994Contig1 (B)</td>
<td>26.7</td>
</tr>
<tr>
<td>rfpPsmelFG_0462_a</td>
<td>41.2</td>
</tr>
<tr>
<td>CN636795.1 (C)</td>
<td>55.3</td>
</tr>
<tr>
<td>sSPcDFE028B10110 (D)</td>
<td>57.5</td>
</tr>
<tr>
<td>Pm_CL1692Contig1 (E)</td>
<td>63.7</td>
</tr>
</tbody>
</table>

$\rho^2$