MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 3504–3522 doi: 10.1111/mec.12830

Climatic adaptation and ecological divergence between two closely related pine species in Southeast China

YONGFENG ZHOU,*† LIRUI ZHANG,*‡ JIANQUAN LIU,* GUILI WU* and OUTI SAVOLAINEN†§

*State Key Laboratory of Grassland Agro-Ecosystem, School of Life Science, Lanzhou University, Lanzhou 730000, Gansu, China, †Plant Genetics Group, Department of Biology, University of Oulu, 90014 Oulu, Finland, ‡Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland, §Biocenter Oulu, University of Oulu, 90014 Oulu, Finland

Abstract

Climate is one of the most important drivers for adaptive evolution in forest trees. Climatic selection contributes greatly to local adaptation and intraspecific differentiation, but this kind of selection could also have promoted interspecific divergence through ecological speciation. To test this hypothesis, we examined intra- and interspecific genetic variation at 25 climate-related candidate genes and 12 reference loci in two closely related pine species, *Pinus massoniana* Lamb. and *Pinus hwangshanensis* Hisa, using population genetic and landscape genetic approaches. These two species occur in Southeast China but have contrasting ecological preferences in terms of several environmental variables, notably altitude, although hybrids form where their distributions overlap. One or more robust tests detected signals of recent and/or ancient selection at two-thirds (17) of the 25 candidate genes, at varying evolutionary timescales, but only three of the 12 reference loci. The signals of recent selection were species specific, but signals of ancient selection were mostly shared by the two species likely because of the shared evolutionary history. F_{ST} outlier analysis identified six SNPs in five climate-related candidate genes under divergent selection between the two species. In addition, a total of 24 candidate SNPs representing nine candidate genes showed significant correlation with altitudinal divergence in the two species based on the covariance matrix of population history derived from reference SNPs. Genetic differentiation between these two species was higher at the candidate genes than at the reference loci. Moreover, analysis using the isolation-with-migration model indicated that gene flow between the species has been more restricted for climate-related candidate genes than the reference loci, in both directions. Taken together, our results suggest that species-specific and divergent climatic selection at the candidate genes might have counteracted interspecific gene flow and played a key role in the ecological divergence of these two closely related pine species.

Keywords: climate change, divergent selection, ecological speciation, gene flow, landscape genetics, pine, population genetics

Received 21 October 2013; revision received 9 May 2014; accepted 21 May 2014

Introduction

Climate is one of the most important drivers for adaptive evolution in forest trees (Aitken *et al.* 2008; Richardson *et al.* 2009; Sork *et al.* 2010; Alberto *et al.* 2013). Climatic selection is known to contribute strongly to local adaptation and intraspecific differentiation (e.g. Morgenstern 1996; Savolainen *et al.* 2007; Olson *et al.* 2013), but this kind of selection could also have promoted interspecific divergence through ecological speciation (Rundle & Nosil 2005; Schluter 2009; Schluter & Conte 2009; Keller & Seehausen 2012; Nosil 2012). Because of the shared evolutionary history, ancient

Correspondence: Dr. Jianquan Liu, Fax: +86 931 8914288; E-mail: liujq@lzu.edu.cn

climatic selection might be shared between closely related species. Most recent climatic selection, in turn, might be species specific, because forest trees often have the highest fitness in their own environments (i.e. local adaptation, Savolainen *et al.* 2007), and different species often occupy different climate niches (i.e. niche divergence, Hua & Wiens 2013). At the stage of incipient speciation between diverging populations, selection for adaptive differentiation has also been shown to counteract the homogenizing effects of gene flow (Mayr 1963; Slatkin 1987; Lenormand 2002; Savolainen *et al.* 2007; Sousa *et al.* 2013) by decreasing the survival and reproductive success of maladapted immigrants from nonlocal diverged populations (Gavrilets & Cruzan 1998). Such diversifying selective forces might not affect all loci, but be limited to loci experiencing local selection, whereas neutral loci would be affected only through linkage disequilibrium with the selected loci (Nosil *et al.* 2009). Such selection could give rise to lower estimated migration for these genes than for reference genes in models assuming neutrality (Sousa *et al.* 2013). The effects of the genes underlying these processes may gradually extend to the genome through linkage disequilibrium, to form reproductive incompatibility between species (Wu 2001; Wu & Ting 2004; Feder *et al.* 2012).

Identifying candidate genes underlying genetic differences associated with climate variations can help explain how species have adapted to their past and present climate conditions and allow us to predict how they will respond to future climate change (Hoffman & Sgro 2011; Franks & Hoffmann 2012). Long-term effects of repeated selection should be detectable by comparing divergence between species and diversity within species, while signals of more recent selection should be detectable by tests based on F_{ST} outliers, site frequency spectrum (SFS) or linkage disequilibrium (LD) analyses (Nielsen 2005; Hohenlohe *et al.* 2010). Recent local selection along environmental gradients can be detected by examining covariance between allele frequency and climate variables (e.g. Coop *et al.* 2010). Patterns of diversity and divergence can be predicted for simple null models, and estimates for specific genes can be compared to expectations derived from such models (Fisher 1930; Wright 1930; Hedrick 2011). However, as demographic effects also cause deviations from the null models, detecting selection requires additional tools, such as comparing a set of candidate genes potentially related to the trait of interest to a set of reference genes, likely not related to the trait (e.g. Chen *et al.* 2012; Keller *et al.* 2012; Kujala & Savolainen 2012).

Evidence of climatic adaptation has been widely detected in forest trees. Correlations have been detected in diverse tree species between climatic gradients and variation of genes linked to adaptive traits, for example bud set and cold adaptation (Holliday *et al.* 2010; Ma *et al.* 2010; Chen *et al.* 2012; Olson *et al.* 2013), and growth traits associated with precipitation, aridity and temperature (Bower & Aitken 2008; Eckert *et al.* 2010; Hall *et al.* 2011; Keller *et al.* 2011; Prunier *et al.* 2011, 2012; Mosca *et al.* 2012). In addition, selection signals have been detected at climate-related candidate genes in numerous tree species, even if particular phenotypes have not been examined (e.g. Gonzalez-Martinez *et al.* 2006; Pyhäjärvi *et al.* 2007; Eckert *et al.* 2009; Grivet *et al.* 2009; Wachowiak *et al.* 2009). However, it is less clear whether climatic adaptation is also an important driver of tree speciation.

This study focuses on a pair of closely related pine species, *Pinus massoniana* and *Pinus hwangshanensis*. All previous phylogenetic analyses have placed them in the same section and subsection of the genus *Pinus* (Wang *et al.* 1999; Gernandt *et al.* 2005), but they differ with respect to morphological characteristics and timber anatomy (Wu 1980; Xing *et al.* 1992). They are distributed across Southeast China and have different elevations (Xing *et al.* 1992; Luo *et al.* 2001). Molecular phylogenetic analyses including closely related species suggested the two pine species diverged about 3–5 million years ago, MYA (Leslie *et al.* 2012). *Pinus massoniana* tends to occur at elevations below 900 m at the base of the mountains, while *P. hwangshanensis* is mainly present at altitudes above 700 m to the treeline (Fu *et al.* 1999; Luo *et al.* 2001; Li *et al.* 2010a). Spontaneous hybrids are frequently observed at intermediate altitudes, but they have substantially lower (<50%) seed germination rates than both parents (Li *et al.* 2010a, 2012a). The two species share common ancestral polymorphisms in their mitochondrial DNA, but between-species divergence of the chloroplast DNA is much higher, indicating ongoing speciation (Zhou *et al.* 2010). Such a pair of closely related species with overlapping distributions offers valuable opportunities for detecting signals linked to climatic adaptation and examining interactions between selection and gene flow during ecological divergence.

We used multiple complementary approaches to detect genes that have been targeted by selection over three different evolutionary timescales: within each species (recent selection), between the two species (PM-PH) and in the shared lineage leading to *P. massoniana* and *P. hwangshanensis* (i.e. ancient selection between these species and an outgroup, *Pinus koraiensis*, PM-PK and PH-PK). Specifically, we aimed to address the following three questions. (i) Are the same genes targeted by recent selection within each species, or do different loci contribute to recent climatic adaptation? (ii) Did climate-related candidate genes contribute to the species divergence—do they have higher levels of interspecific genetic differentiation and more restricted gene flow between species than reference loci, as expected? At the same timescale, do

climate-related candidate SNPs show signals of divergent selection or significant correlation with altitudinal divergence between these two species? (iii) Is there evidence of ancient climatic selection in the lineage between PM-PH and *P. koraiensis*?

Materials and methods

Plant samples and DNA isolation

We collected samples from 11 and 15 natural sites spanning the ranges of *Pinus hwangshanensis* and *Pinus massoniana*, respectively (Table S1, Supporting information), including eight parapatric populations (populations of the two species present on the same mountain, spanning different but overlapping altitudinal ranges, see Li *et al.* 2010a,b), three allopatric populations of *P. hwangshanensis* and seven allopatric populations of *P. massoniana*. At each parapatric site, *P. hwangshanensis* was sampled at a higher altitude than *P. massoniana* (Table S1, Supporting information). On average, *P. massoniana* and *P. hwangshanensis* samples were collected at altitudes of 1143 and 606 m, respectively (Table S1, Supporting information). As gene flow between the two species might complicate our further analyses, such as selection tests using polymorphic information within species, we avoided sampling hybrids by collecting samples from trees located at least 100 m higher or lower than the margins of the contact zones. Cones were collected from different trees separated by at least 100 m. Seeds from each mother tree were kept in individual paper bags and stored at 4 °C in a dry environment. We randomly selected four individuals per site, so in total we sampled 44 and 60 *P. hwangshanensis* and *P. massoniana* individuals, respectively, and pooled the samples for most analyses, although the populations are somewhat differentiated (Chiang *et al.* 2006; Zhou *et al.* 2010; Ge *et al.* 2012; Y. Zhou, L. Duvaux, L. Zhang, O. Savolainen & J. Liu, in preparation). We discuss the impact of this where relevant. Two individuals of *Pinus koraiensis* (subgenus Strobus) from northeast China (Mohe, Heilongjiang, 53°03′N, 122°22′E) were sampled for use as an outgroup.

Genomic DNA was extracted from haploid megagametophytes of germinated seeds using a QIAGEN DNeasy Plant Mini Kits (QIAGEN, Inc., Valencia, CA, USA), with polyvinylpyrrolidone (PVP) added to the buffer (final concentration, 1%).

Loci studied

We selected a set of 25 candidate genes (Table S2, Supporting information) that had been previously implicated in climatic adaptation (cold hardiness and/or drought tolerance) in pines (Gonzalez-Martinez *et al.* 2006; Eveno *et al.* 2008; Wachowiak *et al.* 2009; Grivet *et al.* 2011). All of these genes were first identified by analyses of gene expression patterns in plants under cold and drought stress (e.g. Joosen *et al.* 2006; Lorenz *et al.* 2006). Similarly, a set of 12 loci (Dvornyk *et al.* 2002; Brown *et al.* 2004; Ma *et al.* 2006; Wachowiak *et al.* 2011; Ren *et al.* 2012, Table S2, Supporting information) assumed to be neutral were selected for use as references.

PCR, sequencing, sequence alignment and annotation

Primers were designed using the PRIMER3 software (http://frodo.wi.mit.edu/primer3/) and available genomic sequences of conifers. Primer sequences were listed in Table S2 (Supporting information). Target loci were amplified using a Gene Amp PCR system 9700 DNA Thermal Cycler (Applied Biosystems, Foster City, CA, USA) and 25 μL PCR mixtures containing 10 ng haploid template DNA, 50 mM Tris-HCl, 1.5 mM MgCl2, 250 mg/mL bovine serum albumin (BSA), 0.5 mM dNTPs, 2.0 mM of each primer and 0.75 U of Taq polymerase. The thermal profile consisted of primary denaturation at 95 °C for 6 min, followed by 32 cycles of 30 s at 95 °C, 45 s at a primer-specific annealing temperature (Table S2, Supporting information) and 1 min or 1 min 30 s at 72 °C, with a final extension of 6 min at 72 °C. Amplification products were then purified using a TIANquick Midi Purification Kit (Tiangen, Beijing, China). Sequencing reactions were performed with the forward and reverse PCR primers for all amplicons, using an ABI Prism BigDye Terminator Cycle Sequencing Kit, version 3.1 and an ABI3130xl Genetic Analyzer (Applied Biosystems) at Lanzhou University or ABI3730xl Genetic Analyzer at the Beijing Genome Institute (BGI). Singletons were verified by resequencing reamplified fragments from the same megagametophyte. Only sequences with clear single peaks were retained.

We aligned DNA sequences using MUSCLE (Edgar 2004) implemented in MEGA 5.0 (Tamura *et al.* 2011). All putative polymorphic sites were subsequently confirmed by visual inspection of the chromatograms. Coding and noncoding regions (introns and untranslated regions) were annotated by BLAST searches of the National Center for Biotechnology Information database (http://www.ncbi.nlm.nih.gov/).

Population genetic analyses

To measure genetic diversity within each species, we determined the number of segregating sites (S) and nucleotide diversity statistics (π, Nei 1987; θ_w, Watterson 1975) for all sites, silent sites and nonsynonymous sites, and the number of haplotypes (K) and haplotypic diversity (H_d, Nei 1987) for all sites. Average within-population diversity (π_{aver}) and total diversity (π_T) were

compared in both species. As the divergence between the two species was low, we estimated *F* statistics hierarchically, both between species (F_{CT}) and among populations within species (F_{ST}). Grouping of populations based on population structure (Fig. S2, Supporting information) and biogeographic analyses (Chiang *et al.* 2006; Zhou *et al.* 2010; Ge *et al.* 2012; Y. Zhou, L. Duvaux, L. Zhang, O. Savolainen & J. Liu, in preparation) resulted in the identification of five and seven populations (roughly by province, Table S1, Supporting information) for *P. hwangshanensis* and *P. massoniana*, respectively, with 8–12 sampled individuals in each population. We calculated nucleotide divergence per site for nonsynonymous sites (K_a) and silent sites (K_s). We also calculated the ratio of replacement to silent polymorphism (π_a/π_s) in each species, the ratio of replacement divergence to synonymous divergence (K_a/K_s) between the two species and between each species and the outgroup. All statistics were computed for each locus using DNASP v5 (Librado & Rozas 2009). Neighbour-joining (NJ) trees for all the loci examined were also constructed for all haplotypes of both species under the Hasegawa–Kishino–Yano (HKY) mutation model using Geneious 5.6 (http://www.geneious.com/).

To investigate the effects of selection and population history on LD, the level of LD was measured as the correlation coefficient r^2 (Hill & Robertson 1968) using parsimony-informative sites. Indels and sites with three nucleotide variants were excluded from the analysis. Under the mutation-drift-equilibrium model, the decay of LD with physical distance was estimated using nonlinear regression of r^2 between polymorphic sites and the distance (in base pairs) between sites (Hill & Robertson 1968). The nonlinear least-squares (NLS) estimate of ρ ($\rho = 4N_e c$, where N_e is effective population size and c is the recombination rate) between adjacent sites was fitted using the NLS function implemented in the R 2.15.2 statistical package (R Core Team 2013).

Population structure analysis

To examine genetic structure in each species, we used a Bayesian clustering approach using STRUCTURE v.2.3.3 (Hubisz *et al.* 2009), which assigns individuals (with admixture allowed) to a predetermined number (*K*) of clusters. We ran six replicates for each value of *K* from 1 to 8 (100 000 burn-in cycles followed by 1 000 000 cycles of data collection) and identified the most likely number of genetic clusters representing the data, according to the ΔK statistic (Evanno *et al.* 2005).

Selection tests

Signals of natural selection were detected from three different evolutionary timescales: within each species (recent selection), between the two species (which diverged about 3 million years ago, MYA, Table 5) and in the lineages leading to these two species (i.e. PM-PK and PH-PK), by comparing them to the outgroup (which diverged 45–85 MYA, Willyard *et al.* 2007).

To detect recent selection within species, we compared the patterns of sequence variation to the neutral equilibrium model using a range of statistics for each species. The SFS-based statistic the Tajima's *D* (Tajima 1989) were computed using DNASP v5 (Librado & Rozas 2009). The site- and haplotype-frequency spectra based on the compound DHEW tests (Zeng *et al.* 2007) were conducted using scripts provided by Dr. Kai Zeng. Because the two species hybridize, we also estimated the likelihood that natural selection has occurred at individual loci using the recently developed maximum frequency of derived mutations (MFDM) test (Li 2011). The MFDM test compares the size of each basal branch in the tree and is robust to the effect of admixtures when migrant detectors (MDs) are employed. As the migrant lineage is expected to first coalesce with the lineages of the MDs before coalescing with any others (Li 2011), we used the three most frequent haplotypes in one species as MDs in the other to identify the unbalanced trees caused by interspecific introgression rather than selection. This analysis was conducted for each locus of each species and for PM-PH excluding indels. Two loci (*ptlim-2* and *erd3*) in *P. hwangshanensis* were excluded from the analysis because of the small sample size.

To examine local adaptation to the seasonal climate within species and ecological divergence between the two species driven by climate selection at different altitudes (one of the most important differentiating ecological factors between the two species), we tested for correlations between allele frequencies and altitudes under a Bayesian generalized linear mixed model (BAYENV, Coop *et al.* 2010; Gunther & Coop 2013). This method tests for covariance between the candidate SNPs frequencies and environmental or geographic variables (e.g. altitude) that exceed the expected covariance estimated using the reference SNPs. The covariance matrix of population differences Ω captures the pattern of allele frequency variance among populations as expected under genetic drift. We conducted BAYENV analyses for *P. massoniana* and for the combined data set (PM-PH), but not for *P. hwangshanensis* (as too few populations were sampled). A total of 159 SNPs for *P. massoniana* and 250 SNPs for PM-PH from the reference loci were used to characterize the neutral covariance of the population history, with five runs of a 500 000-step Monte Carlo Markov Chain (MCMC) in the BAYENV 2.0 program (Gunther & Coop 2013). The mean covariance matrix over runs was used as the covariance matrix in the following analyses. We then tested for covariance between altitudes and

the population-specific allele counts at candidate SNPs while using the reference SNPs—derived Ω as a covariate to control for population history. For a given SNP, BAYENV 2.0 compares a null model of covariance of SNP frequency with population history to an alternative model that includes a covariance of environmental or geographic variables. A Bayes factor (BF) is calculated as the ratio of the posterior probabilities under the alternative and null models. Because linear models are sometimes sensitive to outliers, the rank-based nonparametric statistic, Spearman's ρ, was used to avoid including spurious correlations. We analysed each SNP individually and determined the distributions of posterior odds ratio (PO) for the candidate SNPs (326 for *P. massoniana*, 634 for the PM-PH) by 10 runs with 500 000 MCMC steps with different random seeds. The results were averaged across the 10 runs. The top 5% BFs with the top 5% Spearman's ρ values were then interpreted as showing strong support for clinal adaptation along the altitude gradients in *P. massoniana* or for climatic selection driving or maintaining ecological divergence in PM-PH, as suggested by the manual.

To detect SNPs that might be affected by divergent selection in the two pine species, we used the BAYESCAN v2.1 program (Foll & Gaggiotti 2008). Loci that have been under local selection are expected to show a higher F_{ST} values than neutral loci (e.g. Beaumont & Nichols 1996; Foll & Gaggiotti 2008). Briefly, BayeScan tests for signals of local adaptation in multilocus data by separately modelling a population-specific effect β based on the island model of demographic history, and a locus-specific effect α, that is sensitive to the strength of selection and is retained when the locus-specific component is necessary to explain the observed pattern of diversity. Bayes factors and posterior probabilities were calculated to indicate how much more likely the model with selection is compared to the neutral model. We ran BAYESCAN v2.1 analyses separately on the reference SNPs and candidate SNPs under identical run conditions (10 pilot runs with a burn-in of 50 000 iterations, followed by 100 000 output iterations with a thinning interval of 10, resulting in 10 000 iterations for posterior estimation). To minimize numbers of false positives, outliers were identified at the 5% significance level of posterior probability, corrected by the false discovery rate.

To examine longer term selection, we used the multilocus Hudson–Kreitman–Aguade test (HKA, Hudson *et al.* 1987) to assess the fit of the data to the neutral equilibrium model between these two species and between each species and the outgroup, using the program HKA (http://genfaculty.rutgers.edu/hey/software#HKA) with 10 000 simulations. The MLHKA test (maximum-likelihood HKA test, Wright & Charlesworth 2004), an extension of the HKA test, was then used to identify genes or groups of genes that likely have been subject to selection by comparing the neutral model and the model with hypothesized selection at specific loci (Wright & Charlesworth 2004). The selection parameter K measures the degree of diversity within species compared to divergence between species. $K \ll 1$ suggests selective sweeps, reflected in the increase of divergence relative to polymorphism, while $K \gg 1$ suggests balancing selection, maintaining nucleotide variation (Wright & Charlesworth 2004). K estimates were calculated for all loci with different seed numbers and starting values for the divergence time parameter (T) in the MLHKA program (http://labs.eeb.utoronto.ca/wright/Stephen_I._Wright/Programs.html) with 1 000 000 MCMC cycles.

Finally, to detect selection in the PM-PH, PM-PK and PH-PK lineages, we conducted McDonald–Kreitman (McDonald & Kreitman 1991) and McDonald–Kreitman Poisson Random Field (MKPRF, Bustamante *et al.* 2002) tests. The MK test (McDonald & Kreitman 1991) is based on a comparison of synonymous (or silent) and nonsynonymous (replacement) variation within species and divergence between species. The population selection coefficient ($\gamma = 2N_eS$, where N_e is the effective population size and S is the selection coefficient) was also estimated using the counts of fixed and polymorphic SNPs in the MKPRF software (Bustamante *et al.* 2002). The MKPRF is based on a hierarchical Bayesian model that can describe not only the strength and direction of selection at an individual locus, but also an overall trend for selection across a group of loci. MKPRF obtains samples from the posterior distribution of the parameters using a MCMC. Ten independent chains were run for 10 000 'burn-in' iterations and then sampled every 10 iterations for a total of 100 000 samples for each chain. The Gelman Rubin statistic (based on the variation between chains) for each parameter in all cases was close to 1.0, indicating that the chains converged well. We summarized the selection coefficients for each locus using the mean and 95% credible interval for both candidate genes and reference loci.

Isolation-with-migration analyses

To estimate migration rates and splitting times for candidate genes and reference loci, we used the isolation-with-migration (IM) model (Nielsen & Wakeley 2001; Hey & Nielsen 2004; Hey 2010). The mutation rate per site per year ($\mu = K_s/2T$) was estimated based on divergence (K_s) across reference loci and divergence time (T) of 45–85 million years ago (MYA, Willyard *et al.* 2007) between these two species (subgenus *Pinus*) and the outgroup *P. koraiensis* (subgenus *Strobus*). Repeated preliminary runs with different seeds were conducted with 100 000 steps and 50 000 burn-in steps in the IMA2 program (Hey

2010) to capture the prior estimates. Subsequently, long runs with 5 000 000 steps and 500 000 burn-in steps were conducted twice for both candidate genes and reference loci. We recorded peaks in the posterior distributions and the 95% highest posterior density (HPD) interval for the parameters. Similar posterior distributions were obtained from each simulation, and average values of the demographic parameters were finally calculated.

Results

Genetic similarity between P. massoniana and P. hwangshanensis

The overall estimate of silent divergence (K_s = 0.013) was low, indicating that divergence between the two species and adaptation to different habitats occurred recently. In accordance with the weak divergence, the taxa shared about 50% of the segregating sites, and only five fixed sites were observed at the candidate genes (Table 1). In addition, the two species shared at least one haplotype at most locus, except at four candidate genes (*aqua-MIP*, *dhn1*, *GI* and *Glu*; Table S4 and Fig. S3, Supporting informations).

Table 1 Summary statistics for nucleotide variation within and between *Pinus massoniana* and *Pinus hwangshanensis*

	P. massoniana		*P. hwangshanensis*	
	Candidate	Reference	Candidate	Reference
L	16 888	7262	16 888	7262
SNPs (silent)	321 (234)	159 (137)	480 (348)	202 (185)
θ_w	0.0045	0.0041	0.0069	0.0059
H_d	0.64	0.58	0.77	0.80
	PM-PH (candidate genes)		PM-PH (reference loci)	
SS	118		36	
S1	214		78	
S2	273		114	
SF	5		0	
K_s	0.013		0.012	

L, total analysed length in base pairs; *SNPs* (silent), single nucleotide polymorphisms for total and (silent sites); θ_w, Watterson's nucleotide diversity (Watterson 1975); H_d, haplotype diversity; SS, segregating sites shared by PM and PH; S1, exclusive segregating sites in *P. massoniana*; S2, exclusive segregating sites in *P. hwangshanensis*; SF, fixed segregating sites between PM and PH; K_s, silent nucleotide divergence.

Effective population size, nucleotide diversity and linkage disequilibrium

The length of the 37 sequenced loci varied from 304 to 1246 bp, amounting to a total of 16 888 bp and 7262 bp for all candidate genes and reference loci, respectively (Table 1). *Pinus massoniana* has a wider geographic distribution than *Pinus hwangshanensis* (Zhou *et al.* 2010), but it exhibited much lower silent nucleotide diversity (P = 0.025, Fig. 1a and Table S3, Supporting information). The average within-population estimates of nucleotide diversity were similar to the species-wide estimates for both species (π_{aver} = 0.0030, π_T = 0.0039 in *P. massoniana* and π_{aver} = 0.0054, π_T = 0.0058 in *P. hwangshanensis*, Table S3, Supporting information).

The extent of LD is also expected to be related to effective population size (and possibly to selection). The short sequences and relatively low samples sizes gave rise to very large confidence intervals of ρ ($4N_ec$). The point estimate of recombination parameter ρ for reference loci of *P. hwangshanensis* (0.0196) was about four times higher than that of *P. massoniana* (0.0048). The higher estimate for *P. hwangshanensis* was consistent with the higher θ ($4N_e\mu$) estimate for *P. hwangshanensis* and thus possibly higher effective population size in *P. hwangshanensis* (Table 2). LD within genes decayed within about 200 bp, as in many other outcrossing coniferous species (Pyhäjärvi *et al.* 2007; Chen *et al.* 2009; Li *et al.* 2010b, 2012b; Kujala & Savolainen 2012; Fig. S1, Supporting information).

Signals of selection at three evolutionary scales

Recent selection within each species was identified as departures from the standard neutral equilibrium model using Tajima's *D*, DHEW tests and MFDM tests, which use slightly different aspects of the frequency spectrum. In *P. massoniana*, two candidate genes (*agp4*, *aqua-MIP*) were consistently identified by all the tests as having been subject to positive selection. In addition, some other loci, including eight candidate genes (*a3ip2*, *ccoaomt*, *dhn7*, *dhn9*, *GI*, *Glu*, *PHYO* and *Pod*) and two reference loci (*c3h* and *LHCA4*), showed departures from the null model according to at least one of the more robust tests (the DHEW and/or MFDM test; Table 4 and Table S9, Supporting information). In *P. hwangshanensis*, Tajima's *D* was significantly more negative at candidate genes than at reference loci (P = 0.040, Fig. 1b), suggesting that candidate genes might have been influenced by selective sweeps or purifying selection. Six candidate genes (*aqua-MIP*, *comt*, *dhn1*, *dhn7*, *Glu* and *PHYO*) and one reference locus (*PAL*) were identified as having been subject to selection according to the robust DHEW and/or MFDM tests (Table 4 and

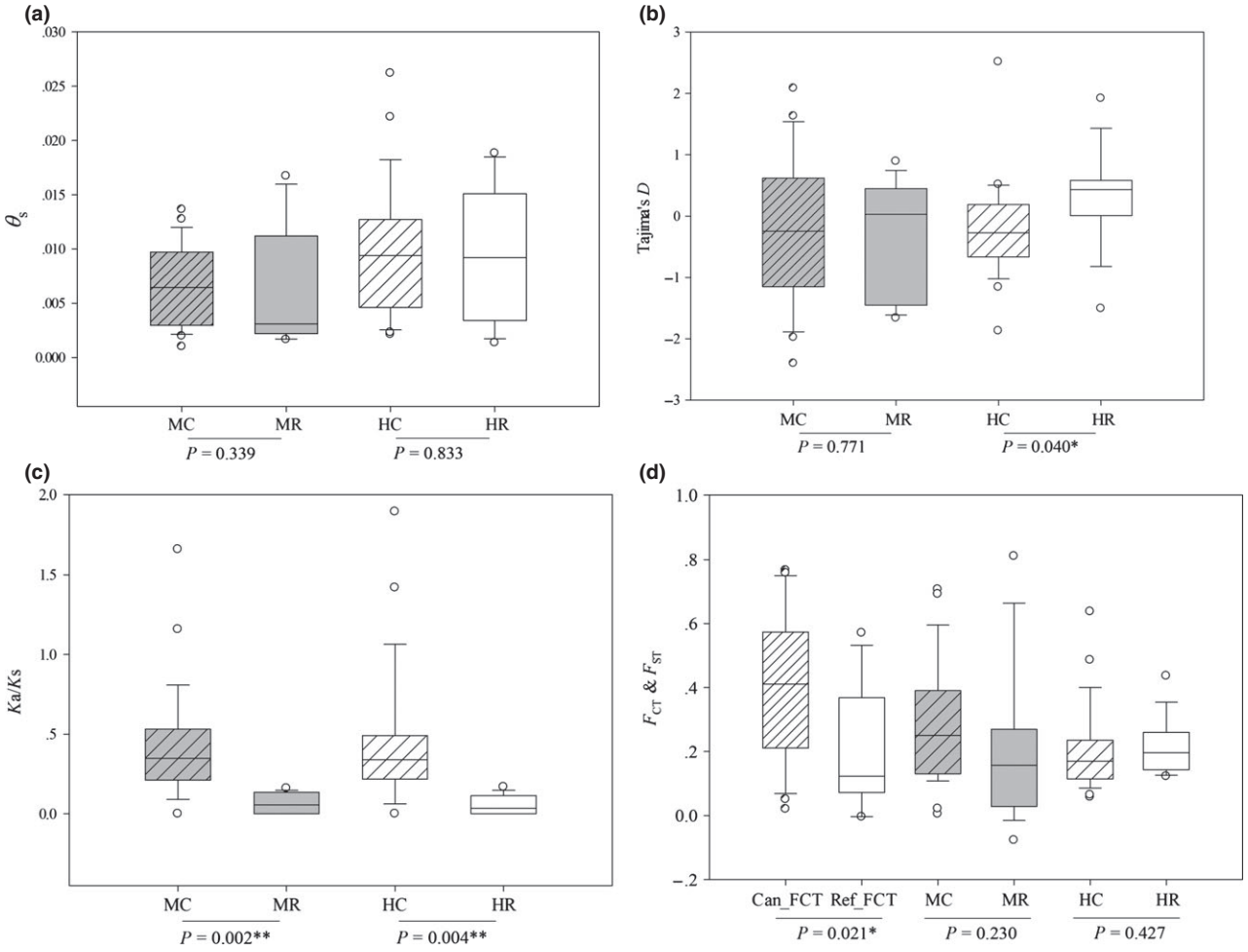

Fig. 1 Box plots of summary statistics for candidate genes and reference loci of *Pinus massoniana* and *Pinus hwangshanensis*. (a) Silent nucleotide diversity (θ_S), (b) Tajima's *D*, (c) ratios of nonsynonymous to synonymous nucleotide substitution rates (K_a/K_s), obtained using *Pinus koraiensis* as an outgroup, and (d) genetic differentiation among populations within species (F_{ST}) and between species (F_{CT}). Bars represent the median, while the bottom and top of each box represent the 25% and 75% percentiles, respectively, and the whiskers extend to 1.5 times the interquartile range. Dots are outliers. *$P < 0.05$, **$P < 0.01$. MC, MR, HC and HR are abbreviations for *P. massoniana* candidate genes, *P. massoniana* reference loci, *P. hwangshanensis* candidate genes and *P. hwangshanensis* reference loci, respectively.

Table 2 Descriptive statistics for population nucleotide recombination ($4N_ec$) of candidate genes and reference loci in the two species

Loci	Rm (95% CI)	ρ (95% CI)	θ (95% CI)	ρ/θ
Pinus massoniana				
Candidate genes	3.16 (0.04 17.46)	0.0455 (0.0003, 0.2941)	0.0045 (0.0008, 0.0107)	10.11
Reference loci	1.92 (0.01, 7.21)	0.0048 (0.0001, 0.1238)	0.0041 (0.0009, 0.0184)	1.17
Pinus hwangshanensis				
Candidate genes	3.08 (0.03, 8.74)	0.0613 (0.0005, 0.1175)	0.0069 (0.0018, 0.0165)	8.88
Reference loci	2.42 (0.01, 6.82)	0.0196 (0.0022, 0.2497)	0.0059 (0.0011, 0.0142)	3.32

Table S9, Supporting information). Among the genes detected under recent selection, eight and three were specific to *P. massoniana* and *P. hwangshanensis*, respectively. Four (*aqua-MIP*, *dhn7*, *Glu and PHYO*) were shared by the two species, but the SNPs at these genes detected by the MFDM test differed between them (Table 4 and Table S9, Supporting information).

We used BAYENV 2.0 (Gunther & Coop 2013) to detect the locus-specific covariance between altitudes and candidate SNPs against the reference SNPs derived

population-specific covariance in *P. massoniana* and in the combined data set (PM-PH). As suggested by the manual, we used both the Bayes factor and the Spearman's ρ to evaluate the results. In *P. massoniana*, we detected four SNPs from three candidate genes (two of them from *Glu*) showing substantial evidence (BF > 3.2, Jeffreys 1961) of correlation with altitude (Table S8, Supporting information). In the combined analysis, the PM-PH, the top 30 BFs (BF > 4.4, substantial evidence for the alternative hypothesis, Jeffreys 1961) along with the top 5% Spearman's ρ values were interpreted as showing strong support for ecological divergence along the altitudinal gradients. Note that these correlations were detected after taking into account the population structure derived from the reference loci. We identified 24 candidate SNPs (including four nonsynonymous SNPs, Table 3) representing nine of the 25 studied candidate genes (Table 3), among which altitudinal associations were especially strong for the *Glu* gene (putative glucan-endo-1, 3-beta-glucosidase precursor; nine significant associations, belonging to two LD groups with $r^2 > 0.2$ and four representing the top four BFs; Table 3). All these candidate genes with SNPs associated with altitudinal divergence in PM-PH were found to have been subject to recent selection within the species, by at least one robust test (Tables 3 and 4). Many of these environmental correlations were due to the fact that the candidate loci were more highly diverged between species along the altitude than the reference loci (Fig. 2). The same loci seem to have experienced selection for altitudinal divergence between the two species.

Table 3 The top 30 Bayes factors with the top 5% Spearman's ρ values for correlations between allele frequency and altitude in PM-PH

Rank	SNPs ID	Bayes factors	Spearman's ρ	Annotations
1	*Glu*_637	17.271**	−0.605**	Int.
2	*Glu*_342	16.853**	−0.611**	SS
3	*Glu*_132	16.722**	0.637**	SS
4	*Glu*_368	15.169**	−0.594*	NSS Ala-Val
5	*MdhA*_432	11.278**	0.671**	Int.
6	*aqua-MIP*_378	9.574*	0.540*	Int.
7	*Glu*_813	8.935*	0.594*	Int.
8	*Pod*_164	8.774*	−0.546*	NSS Ala-Ser
9	*aqua-MIP*_116	7.982*	0.532*	NSS Pro-Leu
10	*GI*_345	6.592*	0.424*	SS
11	*MdhA*_378	6.575*	−0.504*	Int.
12	*Glu*_628	6.538*	0.574*	Int.
13	*dhn1*_996	6.322*	0.440*	Int.
14	*MdhA*_153	6.216*	−0.489*	Int.
15	*lp3-1*_235	5.934*	0.510*	Int.
16	*aqua-MIP*_167	5.713*	0.428*	NSS Pro-Leu
17	*dhn1*_678	5.599*	0.404	SS
18	*Pod*_323	5.570*	−0.496*	Int
19	*Glu*_838	5.377*	−0.481*	Int
20	*PHYO*_114	5.356*	−0.420	Int.
21	*hlhl*_140	4.877*	0.532*	Int.
22	*MdhA*_64	4.767*	0.547*	Int.
23	*hlhl*_230	4.741*	0.533*	Int.
24	*Glu*_948	4.474	−0.407	Int.

Bayes factors: 3.2–10, substantial evidence; 10–100, strong evidence; Int., Intron; SS, synonymous sites; NSS, Nonsynonymous sites.
*$P < 0.05$; **$P < 0.01$; two-tailed *t*-tests.

F_{ST} outliers implemented in the program BAYESCAN were used to detect loci under divergent selection between the two species. The distribution of F_{ST} across candidate SNPs was similar to that across reference SNPs (Fig. 3). BAYESCAN identified six outliers [*aqua-MIP*_116, *dhn1*_723, *GI* (239, 345, $r^2 = 1$), *Glu*_716 and *agp4*_201] belonging to five candidate genes (all of them were detected under recent selection within species, Table S9, Supporting information) at the 5% significance level (corrected by the false discovery rate, Fig. 3). However, no outliers at reference loci were detected (Fig. 3). These six candidate SNPs identified by BAYESCAN, except *agp4*_201 (BF = 0.85), were either significantly associated with altitudes or highly linked to the SNPs associated with altitudes in the BAYENV 2.0 analyses in the PM-PH. All the four candidate genes showed signals of recent selection within both species harboured at least one SNPs under divergent selection between PM and PH. The hierarchical model of F_{ST} (Excoffier *et al.* 2009) would in principle fit better to our data (De Mita *et al.* 2013), but we aimed to detect signals of divergent selection between the two species here. Population structure analyses showed that populations nested within species with small amount of individuals harboured admixed ancestry (Y. Zhou, L. Duvaux, L. Zhang, O. Savolainen & J. Liu, in preparation).

Over longer evolutionary timescales, in the PH-PM comparison, the K_a/K_s ratio was about three times higher for the candidate genes than for the reference loci (0.412 and 0.152, respectively; Table S5, Supporting information), but such comparisons have low power between closely related species. The HKA test revealed significant deviations from the neutral model ($P < 0.01$, Table S6, Supporting information). Two of the candidate genes detected by the MFDM test (*dhn1* and *LEA*) have been subject to balancing selection ($K \gg 1$, Table 4) according to the MLHKA test. Because of recent divergence and the rarity of fixed replacement sites between the two closely related pines, the MK type tests had low power.

In the PM-PK and PH-PK comparisons, the K_a/K_s ratio was about four times higher for candidate genes

Table 4 Intra- and Interspecific differentiation and selection tests for the two closely related pines at each locus

	Pinus massoniana						*Pinus hwangshanensis*						
Locus	D_T	F_{ST}	DHEW *P* value	MFDM *P* value	MFDM (PM-PH)	MLHKA's *K* (PM-PK)	D_T	F_{ST}	DHEW *P* value	MFDM *P* value	MLHKA's *K* (PM-PH)	MLHKA's *K* (PH-PK)	F_{CT}
Climate-related candidate genes													
a3ip2	−1.68	0.045	0.052	0.010*	0.005**	0.652	0.136	0.046	0.453	0.326	0.829	0.557	0.572
agp4	−2.403**	0.021	0.000***	0.017*	0.012*	2.103	0.208	0.027	0.835	0.476	1.37	4.407**	0.523
aqua-MIP	−1.979*	0.077	0.009**	0.017*	0.217	0.988	−0.249	0.021	0.205	0.008**	0.856	0.831	0.75
agp-like	0.901	0.112	0.652	0.204	0.079	0.892	0.504	0.072	0.633	0.186	1.483	1.462	0.199
araH	0.895	0.062	0.974	1	1	2.315	−0.116	0.207	0.874	1	0.658	2.232	0.02
araR	2.084*	0.684	0.924	0.87	0.759	1.355	−0.412	0.036	0.117	0.381	1.557	1.506	0.215
ccoaomt	−1.152	0.103	0.014*	0.102	0.008**	0.173**	0.502	0.079	0.907	0.093	3.198	0.615	0.427
comt	0.217	0.145	0.385	0.102	0.059	0.326*	−1.025	0.049	0.021*	0.093	0.815	0.202**	0.086
dhn1	−1.889*	0.036	0.182	1	0.002**	0.556	−0.964	0.009	0.129	0.002**	6.707***	3.404**	0.522
dhn2	0.529	0.603	0.67	0.542	0.317	1.693	−0.305	0.043	0.96	1	0.959	1.575	0.31
dhn7	−1.443	0.208	0.041*	0.047*	0.006**	2.573	−0.269	0.008	0.231	0.01*	0.917	2.24	0.479
dhn9	−0.1001	0.035	0.215	0.007**	0.1	1.764	−0.583	0.048	0.496	0.233	1.324	2.813*	0.579
erd3	1.517	0.145	1	0.383	1	1.618	−0.497	0.007	0.714	n. a.	0.681	0.505	0.347
GI	−0.927	0.112	0.022*	0.136	0.297	0.355	−0.044	0.206	0.254	0.512	1.759	0.476	0.766
Glu	−0.731	0.236	0.068	0.008**	0.594	0.948	−0.994	0.204	0.057	0.005**	1.072	0.762	0.759
hlhl	−1.151	0.144	0.176	0.136	0.772	0.625	0.517	0.12	0.638	0.372	0.639	0.502	0.71
Ino3	−0.043	0.308	0.315	0.271	0.099	0.186*	0.183	0.236	0.342	0.279	2.686	0.508	0.365
LEA	1.539	0.123	0.798	0.678	0.002**	0.499	−0.76	0.375	0.142	0.326	2.807*	1.576	0.14
lp3-1	−0.333	0.284	0.446	0.746	1	3.315*	−0.092	0.101	0.813	1	0.719	2.049	0.068
MdhA	0.423	0.309	0.652	0.339	0.97	1.207	−0.624	−0.025	0.439	0.186	1.108	1.595	0.57
PHYO	−0.246	0.116	0.399	0.025*	0.025*	0.240**	−1.872*	0.056	0.007	0.05*	1.169	0.277*	0.736
Pod	0.186	0.065	0.359	0.025*	0.316	0.718	−1.16	0.036	0.231	0.698	2.094	1.851	0.411
pp2c	−1.143	0.064	0.07	0.746	1	1	−0.636	−0.014	0.603	1	1.284	1.045	0.33
ppap12	1.63	0.154	0.932	0.78	0.653	3.964*	2.515*	0.116	0.986	0.512	0.656	2.487	0.051
RD21A	−0.35	0.067	0.306	0.169	0.099	3.740**	0.304	0.256	0.505	0.419	0.7	0.758	0.26
Average	−0.07	0.17					−0.277	0.093					0.408
Reference loci													
PtIFG2009	−0.318	0.028	0.552	0.979	0.067	1.151	−0.526	0.206	0.321	0.14	1.791	2.38	0.087
c3h	−1.591	0.083	0.003**	0.169	0.218	1.41	0.598	0.152	0.732	0.279	1.157	1.572	0.15
c4 h2	0.391	0.743	0.463	0.746	0.97	2.92	1.219	0.206	0.659	0.558	1.833	0.719	0.445
CesA2	−1.444	0	0.119	1	1	0.83	0.465	0.211	0.779	1	0.948	0.908	0.292
GapCp	−1.466	0.216	0.089	0.136	0.673	0.704	0.571	0.184	0.673	0.233	0.806	0.541	0.571
LHCA4	−1.665	0.084	0.119	0.006**	0.002**	0.652	−0.163	0.205	0.655	0.14	1.19	1.094	0.516
nir	0.39	0.132	0.437	0.407	0.356	1.333	0.237	0.248	0.362	0.14	0.79	1.217	0.078
PAL	0.511	0.086	0.509	0.136	0.059	0.435	−1.51	0.128	0.026*	0.14	2.13	0.801	0.096
pho	−0.749	0.193	0.089	0.576	1	0.754	0.402	0.063	0.817	1	0.947	0.47	0.279
ptlim-1	0.683	0.164	0.585	0.059	0.059	0.542	1.919*	0.254	0.903	1	0.6	0.235	0.067
ptlim-2	0.892	0.142	0.949	1	1	4.842	0.184	0.176	0.904	n. a.	1.343	2.995*	−0.004
WD40	0.374	0.195	0.413	0.102	0.158	3.765**	0.488	0.147	0.398	0.233	0.7	2.738*	−0.002
Average	−0.333	0.172					0.179	0.182					0.214

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

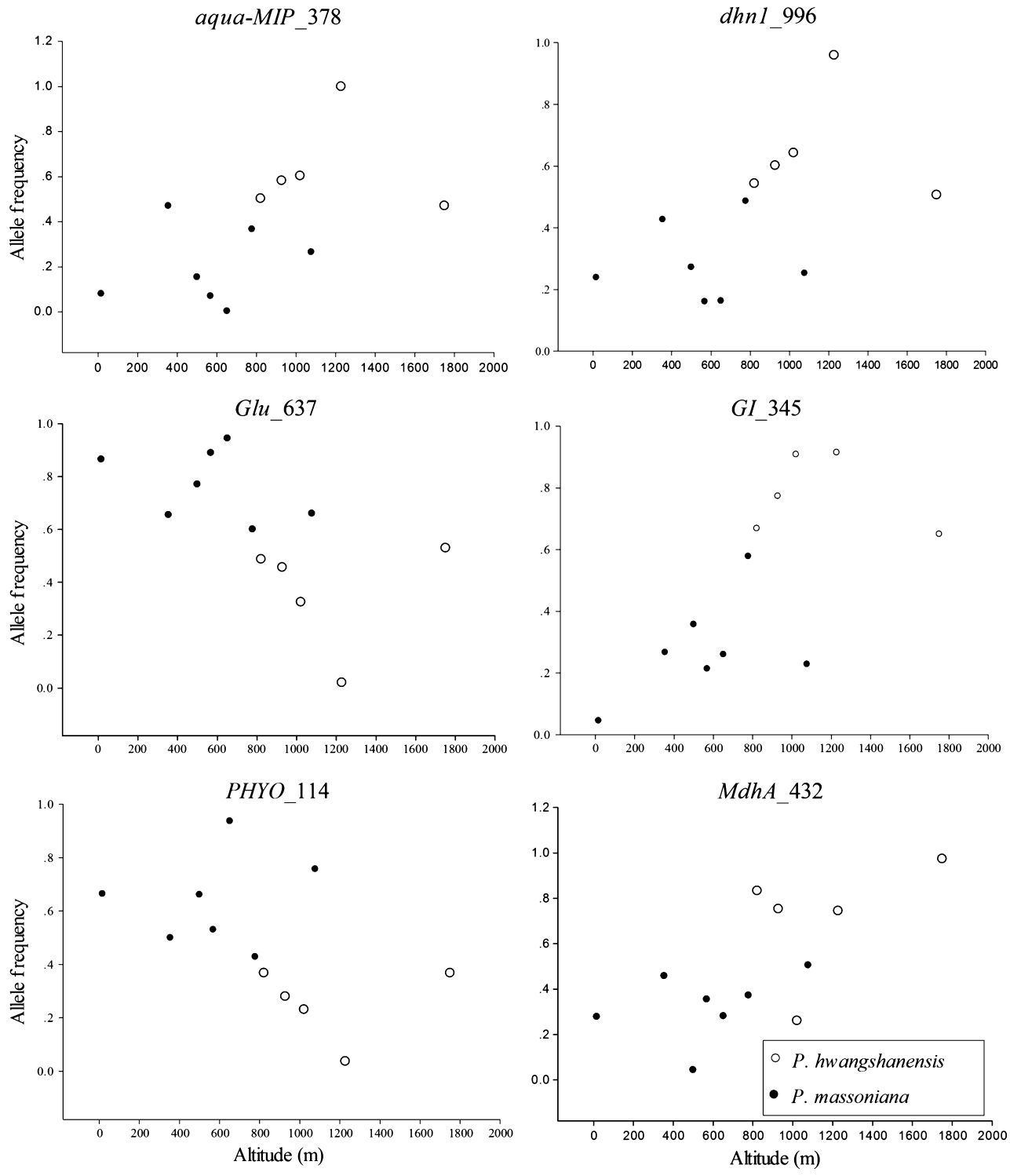

Fig. 2 Examples of candidate SNPs showing significant regression in transformed allele frequencies with altitude.

than for reference loci (Fig. 1c, Table S5, Supporting information), and the HKA test suggested that the ratios deviated significantly from null expectations in both cases ($P < 0.001$, Table S6, Supporting information). We then examined the specific genes that had been subject to selection using the MLHKA test. In the PM-PK comparison, four candidate genes (*ccoaomt*, *comt*, *Ino3* and *PHYO*) showed significantly lower ratios of polymorphism within species to divergence between species (i.e. MLHKA's $K \ll 1$), likely because of

selective sweeps. For PH-PK, two candidate genes (*comt* and *PHYO*) were suggested to have been influenced by selective sweeps. In addition, the MLHKA test suggested three candidate genes and one reference locus in the PM-PK lineage, and three candidate genes and two reference loci in the PH-PK lineage had been influenced by balancing selection, of which one locus was shared by both lineages (Table 4).

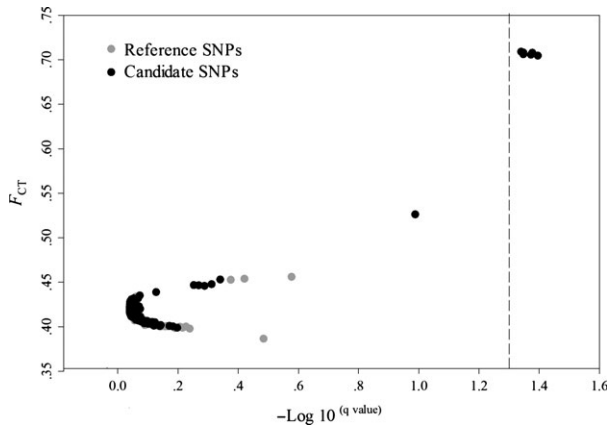

Fig. 3 F_{ST} outliers detected using BAYESCAN.

In the PM-PK and PH-PK comparisons, the MKPRF test, which jointly analyses multiple loci, detected multiple candidate genes with excess polymorphic or fixed replacement sites. Most of them were shared by PM-PK and PH-PK as many changes must have occurred during their shared history after divergence from *Pinus koraiensis*. None of the reference loci in the two comparisons exhibited deviation from the neutral expectations (Fig. 4a, b).

We distinguished different kinds of selection by jointly using of the HKA tests, the MK tests and their extensions. All the candidate loci detected under selection by the MLHKA test were also detected as undergoing similar

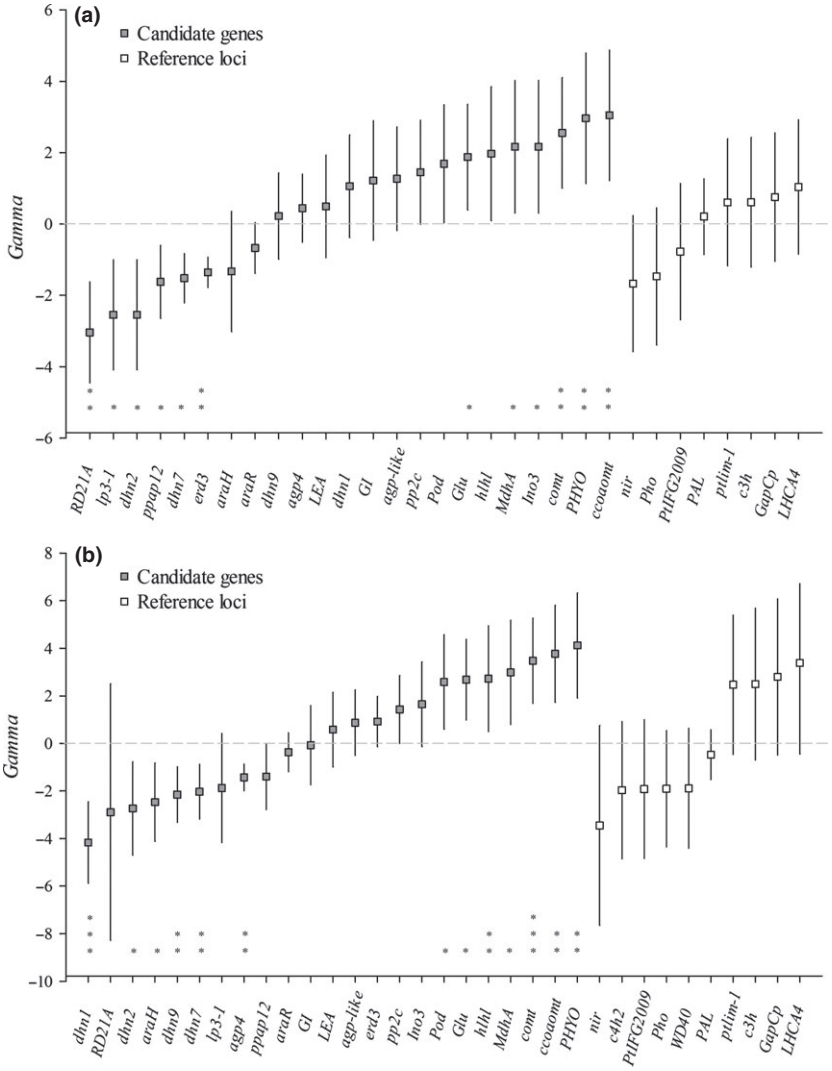

Fig. 4 Results of McDonald–Kreitman Poisson Random Field (MKPRF) analysis: means and 95% credible intervals of posterior distributions for selection coefficients (γ) at each examined locus for (a) *Pinus massoniana* and (b) *Pinus hwangshanensis*, $*P < 0.05$; $**P < 0.01$ and $***P < 0.001$.

kind of selection by the MKPRF test in both PM and PH lineages (the MLHKA and MKPRF tests are not independent). Taken together, for both the PM-PK and PH-PK lineages, we found positive selection or selective sweeps associated with *ccoaomt*, *comt*, *Glu*, *MdhA* and *PHYO*. Lineage-specific positive selection in the branch leading to PM was found in *Ino3*, in the lineage leading to PH at *Hlh1* and *Pod*. Balancing selection in both lineages was detected for *dhn2* and *dhn7*, and lineage-specific balancing selection was detected at four loci (Table 4 and Fig. 4).

Intraspecific and interspecific differentiation

We expected that candidate loci may have diverged before reference loci because of selection for adaptation to the species' different habitats. As the divergence between species was very low (K_s = 0.013), we also used *F* statistics to describe the differentiation. Between the two species, interspecific genetic differentiation (F_{CT}) was significantly higher for the candidate genes (F_{CT} = 0.408 ± 0.239) than for the reference loci (F_{CT} = 0.214 ± 0.203, P = 0.021, Fig. 1d), even though the levels of diversity were similar (Fig. 1a). Within both species, the levels of intraspecific genetic differentiation (F_{ST}) among populations for candidate genes were similar to those of reference loci for the two species (Fig. 1d). Gene exchange at candidate genes might be counteracted by local selection against maladapted immigrants from nonlocal species.

Isolation-with-migration model

Isolation-with-migration model was used to estimate the divergence time and migration rates. Our repeated simulation runs using the IMA2 program resulted in unambiguous marginal posterior probability distributions for the demographic parameters for both sets of candidate and reference loci. The estimate of the effective population size (N_e) for *P. hwangshanensis* was about twice of that for *P. massoniana* (Table 5), in accordance with higher estimates of genetic diversity and recombination for *P. hwangshanensis*. The detected effective migration rate provided clear evidence for rejecting the isolation model. Gene flow seems to have occurred symmetrically in both directions between these two species (Table 5). In addition, we estimated that the species diverged about three MYA based on all reference loci data (Table 5 and Fig. 5). However, estimates of gene flow in both directions based on the climate-related candidate gene data were much lower (Table 5 and Fig. 5). The IMa program was originally designed for analysing neutral makers, but we interpret these results in the light of later developments by Sousa *et al.* (2013), who showed that some patterns of selection may result in lower migration estimates at the targeted loci.

Discussion

We compared intra- and interspecific genetic variation at 25 climate-related candidate genes and 12 presumed reference loci in two closely related pine species with overlapping distributions but different altitudinal preferences. We found that 17 of the 25 climate-related candidate genes were subject to positive or negative selection according to one or more of the robust tests in the two species at three different evolutionary timescales. Signals indicating ancient selection were mostly shared, but signals indicating recent selection were species specific. BAYESCAN detected six SNPs from candidate genes (but none from reference genes) under divergent selection between the two species. In addition, a total of 24 SNPs from nine candidate genes showed significant covariance with altitudinal divergence between the two species beyond the population structure derived from the reference SNPs. Genetic differentiation between the two species was significantly higher in the climate-related candidate genes than in the reference loci. Our

Table 5 Maximum-likelihood estimates (MLE) and 95% highest posterior density (HPD) intervals of demographic parameters from pairwise IMa2 multilocus analyses for candidate genes and reference loci

Value	θ_1	θ_2	θ_A	$m_1 > 2$	$m_2 > 1$	T	N_1	N_2	N_A	T (year)	M_1	M_2
Reference loci												
HiPt	1.745	3.135	0.875	1.538	0.884	1.105	58 167	104 500	29 167	2 946 667	1.342	1.385
Mean	1.796	3.213	1.130	1.602	1.039	1.117	59 867	107 100	37 667	2 978 667	1.437	1.669
HPD95Lo	1.235	2.335	0.015	0.686	0.317	0.560	41 167	77 833	500	1 492 000	0.423	0.370
HPD95Hi	2.385	4.145	2.345	2.558	1.879	2.150	79 500	138 167	78 167	5 733 333	3.050	3.894
Climate-related candidate genes												
HiPt	3.685	5.855	2.625	0.406	0.264	1.357	122 833	195 167	87 500	3 618 667	0.75	0.77
Mean	3.709	5.902	2.710	0.422	0.274	1.400	123 633	196 733	90 333	3 733 333	0.78	0.81
HPD95Lo	3.125	5.025	1.715	0.276	0.164	1.058	104 167	167 500	57 167	2 821 333	0.43	0.41
HPD95Hi	4.315	6.805	3.755	0.579	0.391	1.769	143 833	226 833	125 167	4 717 333	1.25	1.33

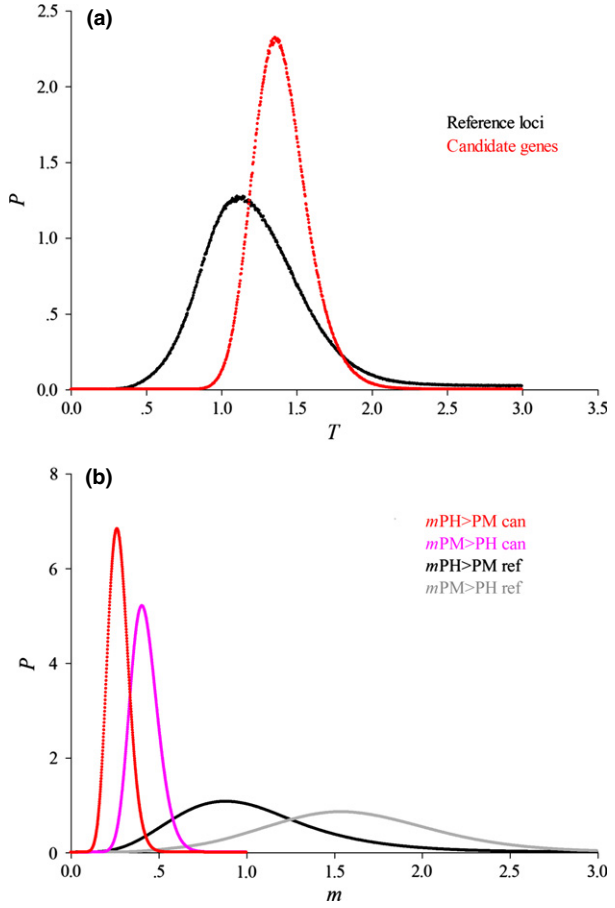

Fig. 5 Marginal posterior distributions of (a) divergence time (*T*) and (b) gene flow (*m*), under the IM model between *Pinus massoniana* and *Pinus hwangshanensis* for candidate genes and reference loci.

further simulations, based on an IM model, suggested that levels of migration between the two species have been lower for the climate-related candidate genes than for the reference loci in both directions. The findings indicate that the observed genetic pattern is probably due to climatic adaptation driving ecological divergence between the two species. Species-specific climatic selection within each species and divergent climatic selection between the two species might restrict interspecific gene flow by preventing the spread of locally adapted alleles to the other species, playing an important role in initiating speciation of the closely related pines. Further, as these loci were more diverged than the reference loci, climatic selection may still be ongoing.

Selection at climate-related candidate genes

The estimated K_a/K_s ratios between the two pine species, and between them and the outgroup, were significantly lower for the reference loci than for the candidate genes (Fig. 1c), suggesting that more relaxed purifying selection or positive selection has acted on the latter.

Eight climate-related candidate genes were identified as having been subject to ancient and/or recent selection by at least two tests at various evolutionary timescales. These were *PHYO*, *GI*, *dhn1*, *dhn7*, *ccoaomt*, *agp4*, *aqua-MIP* and *Glu*. All these eight genes have been found to be highly correlated with the climatic adaptation of plants, and evidence of selection acting on them has been previously detected in population genetic analyses of other tree or herbaceous species. First, *PHYO* (phytochrome A, *PHYA* in *Arabidopsis thaliana*) homologues in model plants have been found to be involved in growth responses to temperature variations, although this gene is particularly important in the photoperiodic pathway (Howe *et al.* 1996; Smith 2000; Garcia-Gil *et al.* 2003). Population genetic variation at homologous loci has been found to be correlated with the bud set and cold tolerance of several forest trees (Chen *et al.* 2012; Holliday *et al.* 2012). Recently, positive selection was found to be strongly acting at *PHYA* in a northern population in the perennial *Arabidopsis lyrata* (Toivainen *et al.* 2014). Second, another photoperiod-related gene, the *Gigantea* (*GI*) has been found to be subject to selection in spruce populations collected from locations with different temperature regimes (Chen *et al.* 2012). Third, the *dehydrin* genes (*dhn*1 and *dhn*7 in the present study) have been found to be highly expressed in response to any type of stress that causes dehydration at the cellular level, including cold, drought and salinity (Close 1997; Yakovlev *et al.* 2008; Velasco-Conde *et al.* 2012). Selection signals have been detected for members of this gene family in numerous widely distributed conifer species (Cato *et al.* 2006; Eveno *et al.* 2008; Grivet *et al.* 2009; Palme *et al.* 2009; Wachowiak *et al.* 2009; Kujala & Savolainen 2012). Fourth, *agp4* homologues play important roles in cell wall formation, while *aqua-MIP* and *ccoaomt* genes putatively participate in controlling the water content of cells (Cruz *et al.* 1992; Dubos *et al.* 2003). Homologues of *agp4*, *aqua-MIP* and *ccoaomt* have been found to be subject to selection in several pine species (Gonzalez-Martinez *et al.* 2006; Eveno *et al.* 2008; Grivet *et al.* 2009). Finally, nine SNPs at *Glu* (putative glucan-endo-1, 3-beta-glucosidase precursor) showed significant covariance with altitude. *Glu* is upregulated under osmotic stress in plants (Dubos *et al.* 2003) and participates in control of osmotic balance during dehydration mediated by adjustments of sugar metabolism (Eveno *et al.* 2008). Genome-wide association analyses in *Medicago truncatula* also showed that allele frequency at *Glu* was significantly correlated with climatic gradients (Yoder *et al.* 2014).

Three of the 12 reference genes (with no previous record of selection) showed signs of selection in this study (*LHCA4*, *c3h* and *PAL*, Table 4). Both *PAL* (Fukasawa-Akada *et al.* 1996) and *c3h* (Wei *et al.* 2006) are reportedly upregulated under cold treatment in herbs and may have played important roles during cold adaptation of the two pines studied here. As our reference sequences were from coding areas, we were more likely to detect some selection than if we had chosen noncoding reference sequences (e.g. Ometto *et al.* 2006). Further functional evidence is needed to confirm suspected roles of these genes in the focal species and their relation to climatic adaptation (see Pavlidis *et al.* 2012b). Some selection at the reference genes would make our comparisons between the two groups conservative (Hahn 2008).

The data set shows some overall evidence of deviation from the SNE, which can influence some SFS-based tests (Garrigan *et al.* 2010), but the DHEW (Zeng *et al.* 2007) and MFDM (Li 2011) tests provide more robust evidence. As gene flow between the pair of species also influences the detection of recent selection, we used the MFDM test and its migrant detection capacity to eliminate this effect (Li 2011). In the tests for ancient selection, we expect that gene flow between the species would slow down or prevent the process of fixation of mutations. In this sense, we would expect our tests for positive selection to be conservative. For the selection tests, we combined the data across populations. When populations are differentiated, this might lower the probability of detecting selection in MK tests, because balanced polymorphism possible due to local adaptation would reduce the level of divergence to polymorphism. However, the multilocus MKPRF approach is much more powerful than the individual locus-based MK test (Eilertson *et al.* 2012). Further, HKA tests can be sensitive to population subdivision (Ingvarsson 2004). The HKA test also may also be influenced by recent bottlenecks, which may contribute to excess of polymorphisms at some loci (Wright & Charlesworth 2004). However, Tajima's *D* did not deviate significantly from the null hypothesis at reference loci in these two species (Fig. 1b). In addition, demographic histories of the two pines were effectively modelled by approximate Bayesian computation (ABC), suggesting that neither the bottleneck model nor the expansion model had a higher probability than the SNE (Y. Zhou, L. Duvaux, L. Zhang, O. Savolainen & J. Liu, in preparation).

More than half of the climate-related candidate genes were found to be subject to positive or balancing selection. This is consistent with expectations as climatic selection is likely to affect conifer populations frequently and strongly, for several reasons. First, conifers often have large current and historical population sizes and thus may be subject to efficient selection ($\gamma = 2N_e s$, e.g., Gossmann *et al.* 2010; in eukaryotes). Second, many conifers have been repeatedly affected by rounds of climate change. Third, unlike annual plants, the seeds of forest trees often colonize new sites substantial distances from their parent populations, thus exposing them to greater environmental selective pressures. Furthermore, current methods may not efficiently detect selection based on polygenic traits (Pavlidis *et al.* 2012a), which are involved in much climatic adaptation (Howe *et al.* 2003; Savolainen 2011; Savolainen *et al.* 2011; Alberto *et al.* 2013; Yoder *et al.* 2014), and selection may have often acted on standing variation instead of new mutations. It should be more difficult to detect effects of selection in such cases (Hermisson & Pennings 2005; Savolainen *et al.* 2013).

Ecological divergence with gene flow

Intraspecific gene flow is important for the cohesion of species, and the rapid spread of advantageous alleles among populations (together with associated hitchhiking events) may be more important than the slower homogenizing spread of neutral alleles (Slatkin 1976; Rieseberg & Burke 2001; Rieseberg *et al.* 2003; Morjan & Rieseberg 2004). Consequently, intraspecific genetic differentiation (F_{ST}) among populations is generally lower at major loci underlying selected phenotypic traits than at neutral loci (Morjan & Rieseberg 2004). However, we also found elevated interspecific genetic differentiation (F_{CT}) and clearer species boundaries in phylogenetic trees (Fig. S3, Supporting information) associated with climate-related candidate genes. This analysis is subject to false positives due to the hierarchical population structure (Excoffier *et al.* 2009), but we expect that the comparison of candidate and reference genes protects against this problem. We found clear evidence that different loci or SNPs have been targeted by selection in the two focal pine species. In particular, we identified six outliers from five candidate genes showing significantly higher than expected interspecific differentiation (Fig. 3). This is consistent with the hypothesis that divergent selection drives genetic differentiation by targeting specific loci, and hitchhiking loci (Fisher 1930; Barton 2000), eventually leading to the formation of reproductive barriers between closely related species linked by interspecific gene flow, that is, ecological speciation (Schluter 2000; Rundle & Nosil 2005).

Genes may act as 'leaders' or 'followers' during speciation (Schwander and Leimar 2011). The leading genes are often related to 'magic' traits, that is, traits that are subject to divergent selection and contribute to reproductive isolation, such as ecologically important traits in plants (Servedio *et al.* 2011). According to Wu's

genic view of speciation (Wu 2001), interspecific flow of genes associated with magic traits ('leaders') is restricted first, while the flow of other genes ('followers') initially persists. However, the gene flow of followers may gradually decline as a result of linkage to the leaders (divergent hitchhiking) and eventually cease throughout the whole genome (Wu 2001; Wu & Ting 2004; Feder *et al.* 2012). Our comparisons of interspecific divergence between sets of loci clearly support these predictions, as we observed weaker migration at climate-related candidate genes than at reference loci (Table 5 and Fig. 5). Thus, selection for climatic adaptation may have counteracted interspecific gene flow, thereby promoting divergence led by genes underlying related traits (Feder *et al.* 2012).

Five SNPs, from four candidate genes, were fixed between the species (Table 1). Population genetic analysis suggested that all these four genes with fixed sites were under recent divergent selection within species and divergent selection between species (Table 4), but few of them showed signals of ancient selection, as expected (Table 4 and Fig. 4). We also found that the five fixed SNPs were either associated with altitude or closely linked to SNPs showing significant altitudinal associations (Table 3). Thus, local selection for ecological traits such as altitudinal preference and divergence between the two pines might have occurred simultaneously. This hypothesis is consistent with suggestions that most extant conifers, especially those of the Northern Hemisphere, originated during the climatic oscillations following the late Neogene. Analyses of marine-core records have shown that the global climate was drier, cooler and more variable during this period; thus, the climatic oscillations between about 3.0 and 2.5 MYA mark the onset of the Northern Hemisphere glaciation (Raymo 1994; Tiedemann *et al.* 1994; Zachos *et al.* 2001). This period also coincides with the estimated ages of the two closely related pines (between 2.5 and 4.0 MYA, Table 5) and other coniferous species of the Northern Hemisphere, which generally split from closely related species at estimated times ranging between 3.0 and 5.0 MYA (Leslie *et al.* 2012; Mao *et al.* 2012). Our findings indicate that climate changes at that time may have played an important role in initiating speciation of the closely related pines we examined by exerting strong divergent selective pressures on key genes associated with related traits.

Acknowledgements

We thank Dr. Victoria Sork and four anonymous reviewers for their careful reviews and valuable comments. We are grateful for useful discussions with Dr. Remy Petit, Dr. Delphine Grivet, Dr. Tanja Pyhäjärvi and Dr. Sebastian Ramos-Onsins. We thank Dr Bin Tian and Xingmin Tian for their help in collecting samples. This research was supported by grants from the National Key Project for Basic Research (2014CB954100, 2012CB114504), the National Natural Science Foundation of China (30725004) and the '111' collaboration Program. Data analyses were partly conducted on computer clusters in the Finnish IT Center for Science (CSC). Y. Z.'s stay in Finland was supported by the Chinese Scholarship Council (CSC NO. 2010618044), the funding from the Center for International Mobility (CIMO, Finland) and Biocenter Oulu (to O. S.).

References

Aitken SN, Yeaman S, Holliday JA, Wang TL, Curtis-McLane S (2008) Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*, **1**, 95–111.

Alberto F, Aitken S, Alia A *et al.* (2013) Evolutionary response to climate change—evidence from tree populations. *Global Change Biology*, **19**, 1645–1661.

Barton NH (2000) Genetic hitch-hiking. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, **355**, 1553–1562.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **263**, 1619–1626.

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

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 15255–15260.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL (2002) The cost of inbreeding in *Arabidopsis*. *Nature*, **416**, 531–534.

Cato SA, Pot D, Kumar S, Douglas J, Gardner RC, Wilcox PL (2006) Balancing selection in a dehydrin gene associated with increased wood density and decreased radial growth in *Pinus radiata* (Abstract). In: Plant & Animal Genome XIV Conf, San Diego.

Chen J, Kallman T, Gyllenstrand N, Lascoux M (2009) New insights on the speciation history and nucleotide diversity of three boreal spruce species and a Tertiary relict. *Heredity*, **104**, 3–14.

Chen J, Källman T, Ma X *et al.* (2012) Disentangling the roles of history and local selection in shaping clinal variation in allele frequencies and gene expression in Norway spruce (*Picea abies*). *Genetics*, **191**, 865–881.

Chiang YC, Hung KH, Schaal BA *et al.* (2006) Contrasting phylogeographical patterns between mainland and island taxa of the *Pinus luchuensis* complex. *Molecular Ecology*, **15**, 765–779.

Close T (1997) Dehydrins: a commonality in the response of plants to dehydration and low temperature. *Physiologia Plantarum*, **100**, 291–296.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Cruz RT, Jordan WR, Drew MC (1992) Structural changes and associated reduction of hydraulic conductance in roots of *Sorghum bicolour* L. following exposure to water deficit. *Plant Physiology*, **99**, 203–212.

De Mita S, Thuillet A-C, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

Dubos C, Le-Provost G, Pot D *et al.* (2003) Identification, characterization of water-stress-responsive genes in hydroponically grown maritime pine (*Pinus pinaster*) seedlings. *Tree Physiology*, **23**, 169–179.

Dvornyk V, Sirvio A, Mikkonen M, Savolainen O (2002) Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. *Molecular Biology and Evolution*, **19**, 179–188.

Eckert AJ, Wegrzyn JL, Pande B *et al.* (2009) 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*). *Genetics*, **183**, 289–298.

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*, **185**, 969–982.

Edgar RC (2004) MUSCLE, multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Eilertson KE, Booth JG, Bustamante CD (2012) SnIPRE: selection inference using a poisson random effects model. *PLoS Computational Biology*, **8**, e1002806.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

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.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.

Fisher RA (1930) *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford [Variorum edition, Bennett J. H. (Editor), 1999, Oxford University Press, Oxford].

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Franks SJ, Hoffmann AA (2012) Genetics of climate change adaptation. *Annual Review of Genetics*, **46**, 185–208.

Fu LG, Li N, Elias TS, Mill RR (1999) Pinaceae. In: *Flora of China*, Vol. 4 (eds Wu Z, Raven PH), pp. 15–90. Science Press, Beijing. Available from http://www.efloras.org/.

Fukasawa-Akada T, Kung SD, Watson JC (1996) Phenylalanine ammonia-lyase gene structure, expression, and evolution in *Nicotiana*. *Plant Molecular Biology*, **30**, 711–722.

Garcia-Gil MR, Mikkonen M, Savolainen O (2003) Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. *Molecular Ecology*, **12**, 1195–1206.

Garrigan D, Lewontin R, Wakeley J (2010) Measuring the sensitivity of single-locus "neutrality tests" using a direct perturbation approach. *Molecular Biology and Evolution*, **27**, 73–89.

Gavrilets S, Cruzan MB (1998) Neutral gene flow across single locus clines. *Evolution*, **52**, 1277–1284.

Ge XJ, Hsu TW, Hung KH *et al.* (2012) Inferring multiple refugia and phylogeographical patterns in *Pinus massoniana* based on nucleotide sequence variation and DNA fingerprinting. *PLoS ONE*, **7**, e43717.

Gernandt DS, Lopez GG, Garcia SO, Liston A (2005) Phylogeny and classification of *Pinus*. *Taxon*, **54**, 29–42.

Gonzalez-Martinez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB (2006) DNA sequence variation, selection of tag single nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. *Genetics*, **172**, 1915–1926.

Gossmann TI, Song BH, Windsor AJ *et al.* (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, **27**, 1822–1832.

Grivet D, Sebastiani F, Gonzalez-Martinez SC, Vendramin GG (2009) Patterns of polymorphism resulting from long-range colonization in the Mediterranean conifer Aleppo pine. *New Phytologist*, **184**, 1016–1028.

Grivet D, Sebastiani F, Alia R *et al.* (2011) Molecular footprints of local adaptation in two Mediterranean conifers. *Molecular Biology and Evolution*, **28**, 101–116.

Gunther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Hahn MW (2008) Toward a selection theory of molecular evolution. *Evolution*, **62**, 255–265.

Hall D, Ma XF, Ingvarsson PK (2011) Adaptive evolution of the *Populus tremula* photoperiod pathway. *Molecular Ecology*, **20**, 1463–1474.

Hedrick P (2011) *Genetics of Populations*. Jones and Bartlett Publishers, Sudbury, Massachusetts.

Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, **169**, 2335–2352.

Hey J (2010) Isolation with Migration models for more than two populations. *Molecular Biology and Evolution*, **27**, 905–920.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, **38**, 226–231.

Hoffman AA, Sgro CM (2011) Climate change and evolutionary adaptation. *Nature*, **470**, 479–485.

Hohenlohe PA, Phillips PC, Cresko WA (2010) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *International Journal of Plant Sciences*, **171**, 1059–1071.

Holliday JA, Ritland K, Aitken SN (2010) Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). *New Phytologist*, **188**, 501–514.

Holliday JA, Suren H, Aitken SN (2012) Divergent selection and heterogeneous migration rates across the range of Sitka spruce (*Picea sitchensis*). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **279**, 1675–1683.

Howe GT, Gardner G, Hackett WP, Furnier GR (1996) Phytochrome control of short-day-induced bud set in black cottonwood. *Physiologia Plantarum*, **97**, 95–103.

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

Hua X, Wiens JJ (2013) How does climate influence speciation? *The American Naturalist*, **182**, 1–12.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

Ingvarsson PK (2004) Population subdivision and the Hudson-Kreitman-Aguade test: testing for deviations from the neutral model in organelle genomes. *Genetical Research*, **83**, 31–39.

Jeffreys H (1961) *Theory of Probability*. Oxford University Press, Oxford.

Joosen RVL, Lammers M, Balk PA *et al.* (2006) Correlating gene expression to physiological parameters and environmental conditions during cold acclimation of *Pinus sylvestris*, identification of molecular markers using cDNA microarrays. *Tree Physiology*, **26**, 1297–1313.

Keller I, Seehausen O (2012) Thermal adaptation and ecological speciation. *Molecular Ecology*, **21**, 782–799.

Keller SR, Levsen N, Ingvarsson PK, Olson MS, Tiffin P (2011) Local selection across a latitudinal gradient shapes nucleotide diversity in balsam poplar, *Populus balsamifera* L. *Genetics*, **188**, 941–952.

Keller SR, Levsen N, Olson MS, Tiffin P (2012) Local adaptation in the flowering time gene network of balsam poplar, *Populus balsamifera* L. *Molecular Biology and Evolution*, **29**, 3143–3152.

Kujala ST, Savolainen O (2012) Sequence variation patterns along a latitudinal cline in Scots pine (*Pinus sylvestris*): signs of clinal adaptation? *Tree Genetics & Genomes*, **8**, 1451–1467.

Lenormand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, **17**, 183–189.

Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ, Mathews S (2012) Hemisphere-scale differences in conifer evolutionary dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 6217–16221.

Li HP (2011) A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Molecular Biology and Evolution*, **28**, 365–375.

Li Y, Stocks M, Hemmilä S *et al.* (2010a) Demographic histories of four spruce (Picea) species of the Qinghai-Tibetan Plateau and Neighboring areas inferred from multiple nuclear loci. *Molecular Biology and Evolution*, **27**, 1001–1014.

Li SX, Chen Y, Gao HD, Yin TM (2010b) Potential chromosomal introgression barriers revealed by linkage analysis in a hybrid of *Pinus massoniana* and *P. hwangshanensis*. *BMC Plant Biology*, **10**, 37.

Li SX, Tang ZX, Zhang DF, Ye N, Xu CX, Ying TM (2012a) Genome-wide detection of genetic loci triggering uneven descending of gametes from a natural hybrid pine. *Tree Genetics & Genomes*, **8**, 1371–1377.

Li Z, Zou J, Mao K *et al.* (2012b) Population genetic evidence for complex evolutionary histories of four high altitude juniper species in the Qinghai-Tibetan Plateau. *Evolution*, **66**, 831–845.

Librado P, Rozas J (2009) DNASP v5, a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Lorenz WW, Sun F, Liang C *et al.* (2006) Water stress-responsive genes in loblolly pine (*Pinus taeda*) roots identified by analyses of expressed sequence tag libraries. *Tree Physiology*, **26**, 1–16.

Luo SJ, Zou HY, Liang SW (2001) Study on the introgressive hybridization between *P. hwangshanensis* and *P. massoniana*. *Scientia Silvae Sinicae*, **6**, 118–122.

Ma XF, Szmidt AE, Wang XR (2006) Genetic structure and evolutionary history of a diploid hybrid pine *Pinus densata* inferred from the nucleotide variation at seven gene loci. *Molecular Biology and Evolution*, **23**, 807–816.

Ma XF, Hall D, StOnge K, Jansson S, Ingvarsson PK (2010) Genetic differentiation, clinal variation and phenotypic associations with growth cessation across the *Populus tremula* photoperiodic pathway. *Genetics*, **186**, 1033–1044.

Mao KS, Milne RI, Zhang LB *et al.* (2012) Distribution of living Cupressaceae reflects the breakup of Pangea. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 7793–7798.

Mayr E (1963) *Animal Species and Evolution*. Belknap Press of Harvard University Press, Cambridge.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the ADH locus in *Drosophila*. *Nature*, **351**, 652–654.

Morgenstern EK (1996) *Geographic Variation in Forest Trees*. UBC Press, Vancouver, British of Columbia, Canada.

Morjan CL, Rieseberg LH (2004) How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology*, **13**, 1341–1356.

Mosca E, Eckert AJ, Di Pierro EA *et al.* (2012) The geographical and environmental determinants of genetic diversity for four alpine conifers of the European Alps. *Molecular Ecology*, **21**, 5530–5545.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York, New York.

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

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov Chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Nosil P (2012) *Ecological Speciation*. Oxford University Press, Oxford.

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

Olson MS, Levsen N, Soolanayakanahally RY *et al.* (2013) The adaptive potential of *Populus balsamifera* L. to phenology requirements in a warmer global climate. *Molecular Ecology*, **22**, 1214–1230.

Ometto L, De Lorenzo D, Stephan W (2006) Contrasting patterns of sequence divergence and base composition between Drosophila introns and intergenic regions. *Biology Letters*, **2**, 604–607.

Palme A, Pyhajarvi T, Wachowiak W, Savolainen O (2009) Selection on nuclear genes in a *Pinus* phylogeny. *Molecular Biology and Evolution*, **26**, 893–905.

Pavlidis P, Metzler D, Stephan W (2012a) Selective sweeps in multilocus models of quantitative traits. *Genetics*, **192**, 225–239.

Pavlidis P, Jensen JD, Stephan W, Stamatakis A (2012b) A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, **29**, 3237–3248.

Prunier J, Laroche J, Beaulieu J, Bousquet J (2011) Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. *Molecular Ecology*, **20**, 1702–1716.

Prunier J, Gerardi S, Laroche J, Beaulieu J, Bousquet J (2012) Parallel and lineage-specific molecular adaptation to climate in boreal black spruce. *Molecular Ecology*, **21**, 4270–4286.

Pyhäjärvi T, Garcia-Gil MR, Knurr T, Mikkonen M, Wachowiak W, Savolainen O (2007) Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. *Genetics*, **177**, 1713–1724.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Available from http://www.R-project.org/.

Raymo ME (1994) The initiation of Northern Hemisphere glaciation. *Annual Review of Earth and Planetary Sciences*, **22**, 353–383.

Ren GP, Abbott RJ, Zhou YF, Zhang LR, Peng YL, Liu JQ (2012) Genetic divergence, range expansion and possible homoploid hybrid speciation among pine species in Northeast China. *Heredity*, **108**, 552–562.

Richardson BA, Rehfeldt GE, Kim MS (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.

Rieseberg LH, Burke JM (2001) The biological reality of species: gene flow, selection, and collective evolution. *Taxon*, **50**, 47–67.

Rieseberg LH, Widmer A, Arntz MA, Burke JM (2003) Directional selection is the primary cause of phenotypic diversification. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12242–12245.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.

Savolainen O (2011) The genomic basis of local climatic adaptation. *Science*, **334**, 49–50.

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

Savolainen O, Kujala ST, Sokol C *et al*. (2011) Adaptive potential of northernmost tree populations to climate change, with emphasis on Scots pine (*Pinus sylvestris* L.). *Journal of Heredity*, **102**, 526–536.

Savolainen O, Lascoux M, Merila J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.

Schluter D (2000) Ecological character displacement in adaptive radiation. *The American Naturalist*, **156**, S4–S16.

Schluter D (2009) Evidence for ecological speciation and its alternative. *Science*, **323**, 737–741.

Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9955–9962.

Schwander T, Leimar O (2011) Genes as leaders and followers in evolution. *Trends in Ecology & Evolution*, **26**, 143–151.

Servedio MR, Van Doorn GS, Kopp M, Frame AM, Nosil P (2011) Magic traits in speciation: 'magic' but not rare? *Trends in Ecology & Evolution*, **26**, 389–397.

Slatkin M (1976) The rate of spread of an advantageous allele in a subdivided population. In: *Population Genetics and Ecology* (eds Karlin S, Nevo E), pp. 767–780. Academic Press Inc, New York, New York.

Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science*, **236**, 787–792.

Smith H (2000) Phytochromes and light signal perception by plants-an emerging synthesis. *Nature*, **407**, 585–591.

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

Sousa VC, Carneiro M, Ferrand N, Hey J (2013) Identifying loci under selection against gene flow in isolation-with-migration models. *Genetics*, **194**, 211–233.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5, Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**, 2731–2739.

Tiedemann R, Sarnthein M, Shackleton NJ (1994) Astronomical timescale for the Pliocene Atlantic $\delta^{18}O$ and dust flux records of Ocean Drilling Program Site 659. *Paleoceanography*, **9**, 619–638.

Toivainen T, Pyhäjärvi T, Niittyvuopio A, Savolainen O (2014) A recent local sweep at the *PHYA* locus in the Northern European Spiterstulen population of *Arabidopsis lyrata*. *Molecular Ecology*, **23**, 1040–1052.

Velasco-Conde T, Yakovlev I, Majada JP, Aranda I, Johnsen O (2012) Dehydrins in maritime pine (*Pinus pinaster*) and their expression related to drought stress response. *Tree Genetics & Genomes*, **8**, 957–973.

Wachowiak W, Balk PA, Savolainen O (2009) Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold related candidate genes in Scots pine (*Pinus sylvestris* L.). *Tree Genetics & Genomes*, **5**, 117–132.

Wachowiak W, Palmé AE, Savolainen O (2011) Speciation history of three closely related pines *Pinus mugo* (T.), *P. uliginosa* (N.) and *P. sylvestris* (L.). *Molecular Ecology*, **20**, 1729–1743.

Wang XR, Tsumura Y, Yoshimaru H, Nagasaka K, Szmidt AE (1999) Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbcL*, *matK*, *rpl20-rps18* spacer, and *trnV* intron sequences. *American Journal of Botany*, **86**, 1742–1753.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Wei H, Dhanaraj AL, Arora R, Rowland LJ, Fu Y, Sun L (2006) Identification of cold acclimation-responsive *Rhododendron* genes for lipid metabolism, membrane transport and lignin biosynthesis: importance of moderately abundant ESTs in genomic studies. *Plant, Cell and Environment*, **29**, 558–570.

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R (2007) Fossil calibration of molecular divergence infers a moderate

mutation rate and recent radiations for *Pinus*. *Molecular Biology and Evolution*, **24**, 90–101.
Wright S (1930) The Genetical Theory of Natural Selection: a review. *Journal of Heredity*, **21**, 340–356.
Wright SI, Charlesworth B (2004) The HKA test revisited: a Maximum-Likelihood-Ratio test of the standard neutral model. *Genetics*, **168**, 1071–1076.
Wu ZY (1980) *Vegetation in China*. Science Press, Beijing.
Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.
Wu CI, Ting CT (2004) Genes and speciation. *Nature Reviews. Genetics*, **5**, 114–122.
Xing YH, Fang YX, Wu GR (1992) The preliminary study on natural hybridization between *Pinus hwangshanensis* and *P. massoniana* in Dabie Mountain of Anhui Province. *Journal of Anhui Forest Science and Technology*, **4**, 5–9.
Yakovlev IA, Asante DK, Fossdal CG, Partanen J, Junttila O, Johnsen O (2008) Dehydrins expression related to timing of bud burst in Norway spruce. *Planta*, **228**, 459–472.
Yoder JB, Stanton-Geddes J, Zhou P, Briskine R, Young ND, Tiffin P (2014) Genomic signature of adaptation to climate in *Medicago truncatula*. *Genetics*, **196**, 1263–1275.
Zachos J, Pagani M, Sloan L, Thomas E, Billups K (2001) Trends, rhythms, and aberrations in global climate 65 ma to present. *Science*, **292**, 686–693.
Zeng K, Mano SH, Shi SH, Wu CI (2007) Compound tests for the detection of hitchhiking under positive selection. *Molecular Biology and Evolution*, **24**, 1898–1908.
Zhou YF, Abbott RJ, Jiang ZY, Du FK, Milne RI, Liu JQ (2010) Gene flow and species delimitation: a case study of two pine species with overlapping distributions in southeast china. *Evolution*, **64**, 2342–2352.

J.L., Y.Z. and O.S. conceived and designed the study, Y.Z., L.Z. and G.W. performed laboratory experiments, Y.Z. analyzed data and wrote the first draft of the paper, J.L., Y.Z. and O.S. contributed to revisions of the original manuscript.

Data accessibility

The haplotype sequences of each locus reported here were deposited in GenBank under accession numbers KJ921127–KJ921496. Sequence alignment files and inputs for selection tests are available at Dryad (doi:10.5061/dryad.f0q11).

Supporting information

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

Table S1 Geographic origin of the populations used in this study.

Table S2 Putative functions, gene structures, primer sequences, annealing temperatures, GenBank homologues and sources for each amplicon.

Table S3 Summary statistics for nucleotide diversity within and differentiation between *Pinus massoniana* and *Pinus hwangshanensis* at 37 analysed loci.

Table S4 Summary statistics of site frequency spectrum, diversity within and divergence between the two pines.

Table S5 The ratio of nonsynonymous diversity (π_a) to silent diversity (π_s) within PM and PH and the ratio of nonsynonymous nucleotide divergence (K_a) to synonymous nucleotide divergence (K_s) in PM-PH, PM-PK and PH-PK.

Table S6 The HKA tests results for all loci in PM-PH, PM-PK and PH-PK.

Table S7 The MK tests results in PM-PK and PH-PK.

Table S8 The top 5% Bayes factors with Spearman's ρ and Pearson's *r* values for correlations between allele frequency and altitudes in *Pinus massoniana*.

Table S9 Summary results of selection tests at three different evolutionary timescales.

Fig. S1 Scatter plots of the squared correlation coefficient of allele frequencies (r^2) as a function of distance in base pairs between pairs of polymorphic sites in *Pinus massoniana* and *Pinus hwangshanensis* at candidate genes and reference loci, respectively.

Fig. S2 Population structure of (a) *Pinus massoniana* and (b) *Pinus hwangshanensis*. The code numbers for each subpopulations under each figures are linked to Table S1.

Fig. S3 Neighbour-joining (NJ) tree at selection-targeted candidate genes, genes for possible adaptive introgression and selection-free loci in the two closely related species with using *Pinus koraiensis* as an outgroup. Clades for *Pinus massoniana* and *Pinus hwangshanensis* were coloured in blue and red, respectively.