Heredity (2010) 104, 3–14

www.nature.com/hdy

ORIGINAL ARTICLE

New insights on the speciation history and nucleotide diversity of three boreal spruce species and a Tertiary relict

J Chen[1], T Källman[1], N Gyllenstrand[2] and M Lascoux
Program in Evolutionary Functional Genomics, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

In all, 10 nuclear loci were re-sequenced in four spruce species. Three of the species are boreal species with very large natural ranges: *Picea mariana* and *P. glauca* are North American, and *P. abies*, is Eurasian. The fourth species, *P. breweriana*, is a Tertiary relict from Northern California, with a very small natural range. Although the boreal species population sizes have fluctuated through the Ice Ages, *P. breweriana* is believed to have had a rather stable population size through the Quaternary. Indeed, the average Tajima's *D* was close to zero in this species and negative in the three boreal ones. Reflecting differences in current population sizes, nucleotide diversity was an order of magnitude lower in *P. breweriana* than in the boreal species. This is in contrast to the similar and high levels of heterozygosity observed in previous studies at allozyme loci across species. As the species have very different histories and effective population sizes, selection at allozyme loci rather than demography appears to be a better explanation for this discrepancy. Parameters of Isolation-with-Migration (IM) models were also estimated for pairs of species. Shared polymorphisms were extensive and fixed polymorphisms few. Divergence times were much shorter than those previously reported. There was also evidence of historical gene flow between *P. abies* and *P. glauca*. The latter was more closely related to *P. abies* than to its sympatric relative *P. mariana*. This last result suggests that North American and Eurasian species might have been geographically much closer in the recent past than they are today.
Heredity (2010) **104,** 3–14; doi:10.1038/hdy.2009.88; published online 29 July 2009

Keywords: picea; nucleotide diversity; divergence time; migration; speciation

Introduction

Classically, the history of species has been approached on two time-scales. Phylogeneticists and systematicists have focused on inferring events on deep times (millions of years) and have mostly been aiming at retrieving the topology and branch lengths of the genealogy of the group of species they were interested in. In contrast, population geneticists have been mostly interested in estimating the relative magnitude of the forces that shaped species current genetic variation: the main conceptual tool in this endeavour has been the coalescent process, which describes the genealogy of a sample of alleles taken within a given population or species. The time scale is set here by the time to the most recent common ancestor (TMRCA) of the sample. Assuming a standard coalescent the expected value of the TMRCA is equal to $4N_e(1–1/n)$, where N_e is the effective population size and n is the number of sequences in the sample (Wakeley, 2008). In some cases, though certainly not in all, the effective population size and generation times are such that the TMRCA is much smaller than the time of the most recent speciation event. For example, in humans, the TMRCA for nuclear genes is on a scale of 100 000 years, which is much less than the speciation time between humans and chimpanzees (c. 5-million years). However, this does not imply that the process modelled by the coalescent, namely random genetic drift, stops playing a role beyond the TMRCA. Actually, recent genomic studies have largely confirmed its importance for recent speciation events (for example, Patterson *et al*., 2006). More generally, Hudson and Coyne (2002) have shown that under a simple allopatric model of speciation with no selection it will take approximately 9–12 N_e generations for the genealogies of $>95\%$ of the loci to be reciprocally monophyletic.

This last result sheds a new light on results recently obtained in spruce species, which are going to be the object of the present investigation. By studying polymorphism and divergence at three nuclear genes among three spruce species (*Picea glauca*, *P. mariana* and *P. abies*) Bouillé and Bousquet (2005) observed a large number of shared polymorphisms. At first glance this may seem surprising, as the divergence between the studied species has been estimated to be around 13–20-million years ago. However, taking the upper limit of the divergence time and assuming an effective population size of 100 000 and an average generation time of 50 years (Bouillé and Bousquet, 2005),

Correspondence: M Lascoux, Program in Evolutionary Functional Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 75326 Uppsala, Sweden.
E-mail: martin.Lascoux@ebc.uu.se
[1]*These authors contributed equally to this work.*
[2]*Current address: Department of Plant Biology and Forest Genetics, Swedish University of Agricultural Sciences, Box 7080, Uppsala 75007, Sweden.*
Received 12 February 2009; revised 24 April 2009; accepted 21 May 2009; published online 29 July 2009

the number of generations since divergence is a mere 4 N_e, which is <9–12 N_e, and therefore shared polymorphisms are indeed expected to be common among these species, even under an allopatric model of speciation.

Shared polymorphisms can be caused by persistence in both descendant species of ancestral polymorphisms or reflect historical or current gene flow. The presence of both shared and fixed polymorphisms among spruce species means that we can take advantage of new approaches to estimate parameters of speciation, namely divergence time, effective sizes of current and ancestral population and migration rates among species. In this study we estimated nucleotide diversity at 10 nuclear loci in four spruce species and used these data to estimate parameters of speciation. The four species were the three species studied by Bouillé and Bousquet (2005), namely *P. glauca*, *P. mariana* and *P. abies* and *P. breweriana*, a Tertiary relict with a limited range confined to Northern California. This is in stark contrast to *P. glauca*, *P. mariana* and *P. abies*, all of whom have continent-wide distributions: *P. glauca* and *P. mariana* are widely distributed across Canada and Alaska from the Atlantic to the Pacific Coast, while *P. abies* is found across Northern and Central Europe from Norway to the Urals Mountains (Figure 1). There are also other differences between *P. breweriana* and its northern relatives that make the comparison among them particularly interesting. First, although there is mounting evidence that the population sizes of *P. abies*, as well as other higher latitude species have been strongly affected by climate changes during the Quaternary (Heuertz *et al*., 2006; Ingvarsson, 2008; Pyhäjärvi *et al*., 2007), *P. breweriana* is believed to have had a fairly stable population size during the same period (Ledig *et al*., 2005 and references therein). How will this be reflected in the estimates of current and ancestral effective population sizes? Second, expected levels of heterozygosity estimated from allozyme studies in different Picea species are surprisingly stable and fall between 0.1 and 0.3 in a majority of studies (Ledig *et al*., 1997). For example, in comparable surveys of genetic variation of allozymes in *P. abies* (Lagercrantz and Ryman, 1990) and *P. breweriana* (Ledig *et al*., 2005) values of the expected heterozygosity were 0.12 in both species, even though the population sizes of the two species today differ by several orders of magnitude. Heterozygosity at allozyme loci has often been used as a proxy for effective population size (for example, Bazin *et al*., 2006). A strong difference in effective population sizes estimated from nucleotide diversity would put this practice into question, at least for species such as spruces. Bouillé and Bousquet (2005) dataset was very limited and they did not carry out any formal estimation of demographic parameters. For example, they used the fact that *P. mariana* and *P. glauca* do not cross naturally today, as well as the pattern of divergence in chloroplast DNA to argue that shared polymorphisms are likely to be ancestral. Although this may well be the case, caution is warranted here, as recent studies suggest that current mating pattern may not be a good proxy for ancient ones (for example, Slotte *et al*., 2008). Finally the presence of shared polymorphisms between *P. abies* and its North American relatives, in particular *P. glauca*, if confirmed on a larger dataset, could also lead to new insights on the history of spruce species. Although the species' range today is quite disjunct, the detection of large amount of spruce pollen in ice cores from Greenland indicates that boreal coniferous forest developed there some 400 000 years ago during a warm interval (de Vernal and Hillaire-Marcel, 2008). Hence the different species might have been geographically closer than today and gene flow might have occurred among them.

Materials and methods

Species sampling and DNA sequencing

Between 20 and 94 individuals of each species were sampled from their natural distribution range (Supplementary File S1). Haploid DNA was extracted from individual megagametophytes using Qiagen DNeasy plant mini kit (Hilden, Germany). PCR amplifications were done using the proofreading enzyme Phusion (Finnzyme, Espoo, Finland) and successful amplifications were treated with ExoSAP-IT (USB, Cleveland, OH, USA) before sequencing. Details regarding origin, extension times and annealing temperature as well as primer sequences for individual loci can be found in Supplementary Table S2. Sequence data were collected with

Figure 1 Natural distribution ranges of the four studied spruce species.

either a MegaBace 1000 using ET-terminator (GE Healthcare, Piscataway, NJ, USA) or an ABI3730 with BigDye v3.1 (Applied Biosystems, Foster City, CA, USA). Sequence data were base-called, assembled and edited with the PHRED, PHRAP and CONSED software suite keeping only high quality data and, furthermore, manually checking all putative polymorphic sites (Ewing and Green, 1998; Ewing *et al*., 1998; Gordon *et al*., 1998). Sequence data for *P. abies* was partly obtained from another study on population genetics of *P. abies* (Heuertz *et al*., 2006).

Sequence data analysis

All obtained sequences were manually aligned and assigned to coding and non-coding regions by aligning genomic sequences against their corresponding mRNA sequence. Standard population genetic parameters for individual loci were calculated either in DNA sequence polymorphism (DNAsp) version 4.5 (Rozas *et al*., 2003) or with the programme compute (Thornton, 2003). As neutrality is one of the major assumptions of the Isolation-with-Migration (IM) model we also tested departure-from-neutrality using the Hudson–Kreitman–Aguadé framework and the maximum likelihood implementation of this test statistics with the programme mlHKA (Wright and Charlesworth, 2004). Recent studies have shown that the Hudson–Kreitman–Aguadé test is one of the most powerful tests to detect positive selection using polymorphism data (Innan and Kim, 2008; Zhai, Nielsen and Slatkin, 2009).

The population recombination parameter (ρ) was estimated with LDhat 2.1 (McVean *et al*., 2002). The range of estimated θ and ρ values from individual loci was later used to define previous uniform ranges in multilocus estimates of ρ and θ with the software Rhothetapost (Haddrill *et al*., 2005).

Estimation of IM parameters

Speciation parameters were estimated under an IM model using the software MIMAR (Becquet and Przeworski, 2007). MIMAR implements a Markov Chain Monte Carlo method to estimate speciation parameters from a slightly different version of Wakeley and Hey (1997) summary statistics. Briefly, segregating sites (S) are classified into four categories, S_1, S_2, S_s and S_f. For each locus, S_1 and S_2 are the number of polymorphic sites unique to samples 1 and 2 (S_1 and S_2, respectively), S_s is the number of sites with shared alleles between the two samples, and S_f is the number of sites with fixed alleles in either samples. MIMAR also differs from previous implementations of the IM model by taking recombination into account but, otherwise, relies on the same set of major assumptions. Namely, loci are assumed to be neutral, and the two species under study are more closely related than any of them is to a third species. In particular, this means that there should not be other unsampled populations exchanging genes with the sampled populations or their ancestor. Clearly this is a demanding assumption. Finally, as pointed by the authors MIMAR may not provide precise estimates unless datasets have both shared and fixed alleles between the two samples.

When assigning polymorphisms to the four categories we excluded indels and sites with missing data and used outgroup sequences to infer the derived allele frequency in the species i, f_i. *Pinus taeda* was the only outgroup in analyses involving *P. breweriana*, and both *Pinus taeda* and *P. breweriana* were used as outgroups (Ran *et al*., 2006) to minimize the error in inferring the ancestral states in pairwise analyses of the remaining three spruce species. S_1, S_2, S_s and S_f were then calculated according to the MIMAR manual supplied by Becquet and Przeworski (2007). Specifically, for polymorphic sites, if $0< f_i \leqslant 1$ in each species, the allele is shared; if $f_i=0$, $f_j=1$, $i\neq j$, the allele is fixed in the sample j; and if $f_i=0$, and $0< f_j <1$, $i\neq j$, the allele is specific to sample j.

The IM model used in MIMAR is defined by 6 parameters: the population split time in generations, T_{gen}; three population mutation rates $\theta_1=4N_{e1}\mu$, $\theta_2=4N_{e2}\mu$ and $\theta_A=4N_{eA}\mu$, where N_e is the effective population size of the two descendant populations and of the ancestral population and μ is the mutation rate per base pair; and the migration rates $M_{12}=4N_{e1}m_{12}$ and $M_{21}=4N_{e1}m_{21}$. It is to be noted that migration rates are scaled by the effective size of population 1. The Markov Chain Monte Carlo was run at least 7×10^6 steps, 1×10^6 burning steps and sampled every 50 steps. The previous distributions and the variance of the normal kernel distributions were optimized after several test runs. This step was crucial to obtain good mixing and thereby, a good acceptance rate. The population recombination rate per base pair is defined as $\rho=4N_er$, where N_e is the effective population size and r is the recombination rate per base pair per generation. We ignored gene conversions, and took the average value of multilocus estimates of ρ as the recombination rate between two species in the analysis. The number of Ancestral Recombination Graphs per locus was set to $X=100$.

Two demographic models were implemented. First, we used a standard neutral model in each pair of species, thereafter called the standard IM model. Second, because the average Tajima's D values were negative in the three boreal species, we added population growth for those while retaining a standard neutral model for *P. breweriana* that did not show any clear evidence of departure from the standard neutral model. More complex demographic models such as bottlenecks were not considered as the MIMAR model is already extremely computer intensive and growth models gave a satisfactory fit (see below). To assess convergence and the mixing of Markov Chain Monte Carlo, we carried out two independent runs and monitored the performance of MIMAR using code provided with MIMAR (For all the parameters and model settings, please refer to Supplementary File S3).

To examine whether our models captured the main features of the history of the species studied here, we used a goodness-of-fit test. We generated simulated datasets for parameters sampled from the posterior distribution estimated by MIMAR, and compared the observed values of a number of statistics: S_1, S_2, S_s, S_f, π, F_{st} and Tajima's D to the distribution of the summary statistics of the simulated data in accordance with Becquet and Przeworski (2007). The P-value for each statistics was calculated for both standard IM model and growth model.

Phylogeny

We used BEST 2.0 (Liu and Pearl, 2007), a Bayesian hierarchical model built in MrBayes (Ronquist and

Huelsenbeck, 2003) to estimate simultaneously gene trees and the species tree. We sampled randomly one individual in each species and used *Pinus taeda* as an outgoup. The same 10 unlinked genes as in the MIMAR analysis were used and an inverse γ substitution model was chosen. 100×10^6 steps and a total of eight chains were carried out to obtain convergence of the Markov Chain Monte Carlo. The species tree was compared with that inferred by the estimates of divergence time given by MIMAR.

Results

Polymorphism

Sequence data from 10 loci were collected from four different spruce species, three North American ones, *P. breweriana*, *P. glauca* and *P. mariana* and one European relative, *P. abies*. The average number of megagametophytes/species was 35. In total just over 8000 bp of aligned sequence was obtained and 210, 26, 119 and 153 segregating sites were identified in *P. abies*, *P. breweriana*, *P. glauca* and *P. mariana* respectively. Estimates of standard population genetic statistics for the four different species are summarized in Figure 2 (details for individual loci are given in Supplementary Table S4). The nucleotide diversity was an order of magnitude lower in *P. breweriana* than in the other three species that all harboured similar levels of variation. In all four species the majority of loci showed negative Tajima's *D* values, even though a vast majority of them were non-significant and for *P. breweriana* the observed values were close to zero (Figure 2). The locus Sb29 showed a deviating pattern with positive values of Tajima's *D* in all species but *P. breweriana*. In *P. abies* and *P. mariana* all loci were polymorphic, whereas GI was monomorphic in

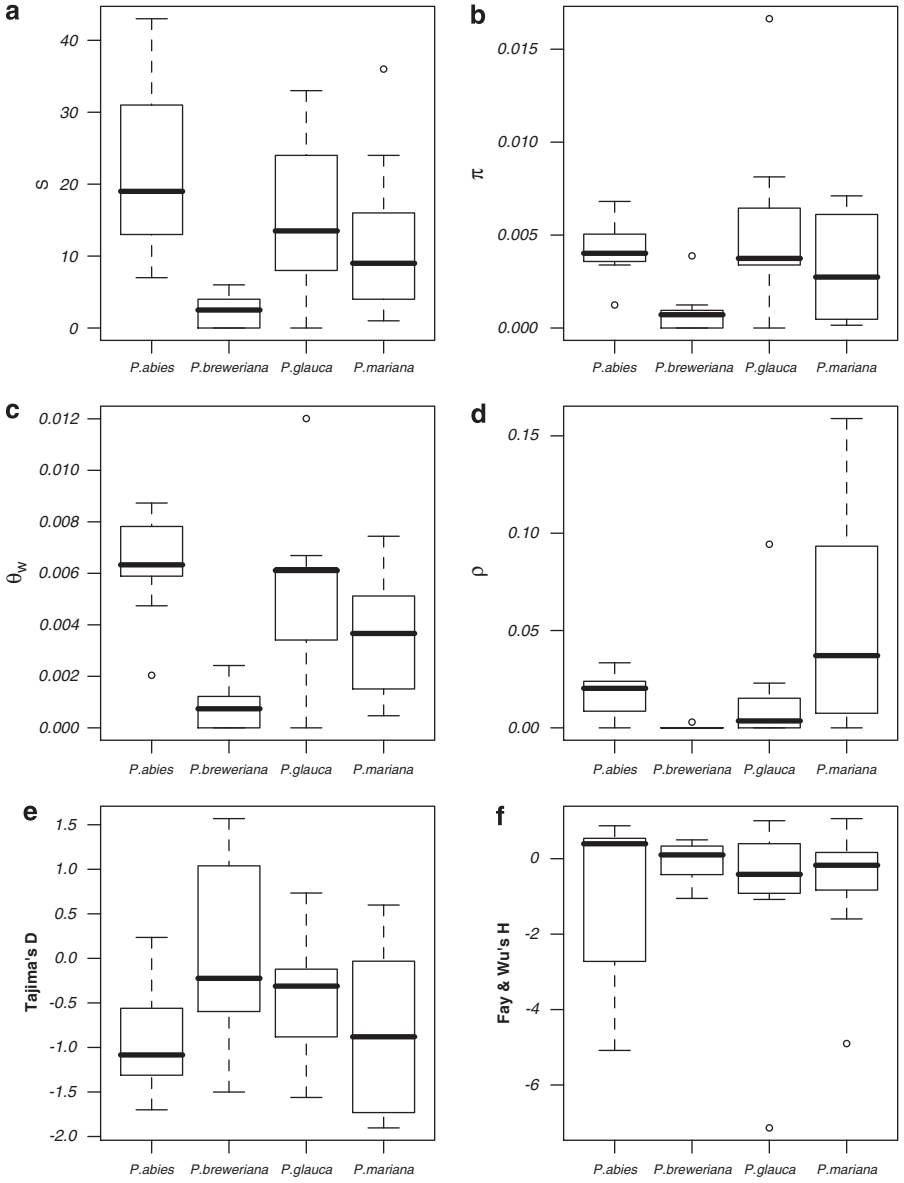

Figure 2 Polymorphism patterns among the four spruce species. Bars represent the median, boxes the interquartile range and whiskers extend out to 1.5-times the interquartile range. (**a**) Number of segregating sites, (**b**) Pairwise nucleotide diversity, (**c**) Watterson's θ, (**d**) Population recombination rate (ρ), (**e**) Tajima's *D*, and (**f**) Fay and Wu's H.

P. glauca and loci GI, Sb16 and se1364 were fixed in *P. breweriana*. All four species had highest diversity at synonymous sites and the non-synonymous diversity was, averaged over loci, less than a third of the silent diversity (Supplementary Table S4). Finally, only ZTL deviated from neutral patterns when using multilocus Hudson–Kreitman–Aguadé tests (data not shown).

Thirty-seven indel variants between 1 and 43-bp long were detected among and within species. Only three of the indels were due to fixed length differences between species. Seven of the indel variants were found in more than one species and the remaining ones were specific to a given species. With the exception of a microsatellite shared between *P. breweriana* and *P. mariana* all identified indels in *P. breweriana* were private. *P. glauca* and *P. abies* had the highest proportions of shared indels with four indels in common. As expected most of the single nucleotide polymorphisms and indel variants were found in the non-coding regions.

Speciation parameters

The segregating sites were classified into the four categories, S_1, S_2, S_s and S_f. Ten segregating sites were shared by *P. abies*, *P. glauca* and *P. mariana*. Unexpectedly, considering their current geographic distribution, *P. abies* and *P. glauca* shared 42 polymorphic sites, more than twice the number shared between *P. glauca* and *P. mariana*, whose distribution ranges today almost completely overlap. Three loci ZTL, COL2 and FT3 contributed most to the shared polymorphisms, whereas the fixed differences between species mainly came from FT3 and GI. It is interesting to note that in contrast to other genes, GI showed a proportionally large number of fixed differences between species, but no shared variation (Supplementary Table S3).

Standard IM model

Figures 3 and 4 show the marginal posterior distribution for parameters of the standard IM model. By assuming a mutation rate $\mu = 1.0 \times 10^{-8}$ per site per generation and an average generation time $t = 50$ years (Bouillé and Bousquet (2005)), we estimated the effective population size N_1, N_2, N_A, the divergence time T and the migration rates M_{12} and M_{21}, in pairwise comparisons. Because of the uncertainty attached to mutation rate and generation time estimates we also provide estimates for $\mu = 2.0 \times 10^{-8}$ and generation time of 25 or 50 years (Supplementary File S5).

Encouragingly, the estimates of the effective population size are consistent across the six pairwise analyses, suggesting that the standard IM model does capture some general features of the population histories (Becquet and Przeworski, 2007). *P. abies* has the largest effective population size estimates of the four species, around 136 000–159 000. Although *P. glauca* and *P. mariana* are largely sympatric today, the estimated effective population sizes of *P. mariana* (around 84 700–97 000) is just 2/3 that of *P. glauca* (121 000–133 000). The estimate of effective population size of *P. breweriana* was an order of magnitude smaller (11 600–12 900). It is interesting to note that in contrast to what has been observed in many previous studies (see Becquet, 2007 and references therein), the ancestral effective population sizes, with values ranging from 92 000–125 000, was smaller, or of the same order of magnitude, than the inferred effective population sizes of the three boreal species. This is in agreement with the expectations, as these populations have obviously gone through strong population growth since their inception.

The different pairwise analyses indicate that *P. breweriana* diverged first from the common ancestor of the four spruce species around 13–17 Mya and that *P. mariana* diverged from the other two boreal species around 12–15 Mya. Finally, Figure 3b shows that split of *P. abies* and *P. glauca* occurred nearly 8.5 Mya. Migration rates estimates were generally low, with the notable exception of the estimate of gene flow from *P. glauca* to *P. abies*, which had a value greater than 1.

Goodness-of-fit test for the standard IM model

To assess the goodness-of-fit of the standard IM model we compared four summary statistics produced by coalescent simulations based on parameter values randomly sampled from the parameter posterior distributions with the corresponding values of the summary statistics from the actual data. The standard IM model gives a reasonably good fit for S_1, S_2, S_s and S_f, π and F_{st}, but the fit for Tajima's D tends to be poor (Figure 5, Supplementary File S6). The negative mean values of Tajima's D could be explained by the presence of selection, and/or departure from constant population size. As Tajima's D values were consistently negative across loci, but individual loci did not statistically depart from neutrality, a demographic explanation seems more likely (Heuertz *et al*., 2006).

Growth model

As Tajima's D was strongly negative for the three boreal species and close to zero in *P. breweriana*, we also estimated speciation parameters using an IM model, in which the boreal species were subjected to a decrease in population size followed by population expansion and the *P. breweriana* population size was kept constant. The present effective population size estimates from the growth models were almost twice as large as the estimates obtained under the standard IM model (Figures 6 and 7 and Supplementary File S5). However, estimates of the ancestral effective population size were not affected by the change of model. In the two North American species *P. glauca* and *P. mariana* growth was initiated 2–2.6 Mya, but the latter grew much faster. *P. abies* population expansion was more recent, ca 1.85–2 Mya with a similar growth rate to that of *P. mariana*. The estimates of the divergence time were half of those obtained under the standard IM model and the genealogical relationships between the four species were not as clear. The species split around 6–7.8 Mya, except *P. glauca* and *P. breweriana* that diverged 13 Mya. The amount of gene flow between species pair increased, but the directions remained the same as under the standard IM model. Like in standard IM model the amount of gene flow from *P. glauca* to *P. abies* was notably high, being close to 4.4 (Supplementary File S5).

Goodness-of-fit test for the growth model

In most cases the growth model improved the goodness-of-fit for all summary statistics and in particular for Tajima's D, which was the one showing poorest fit under

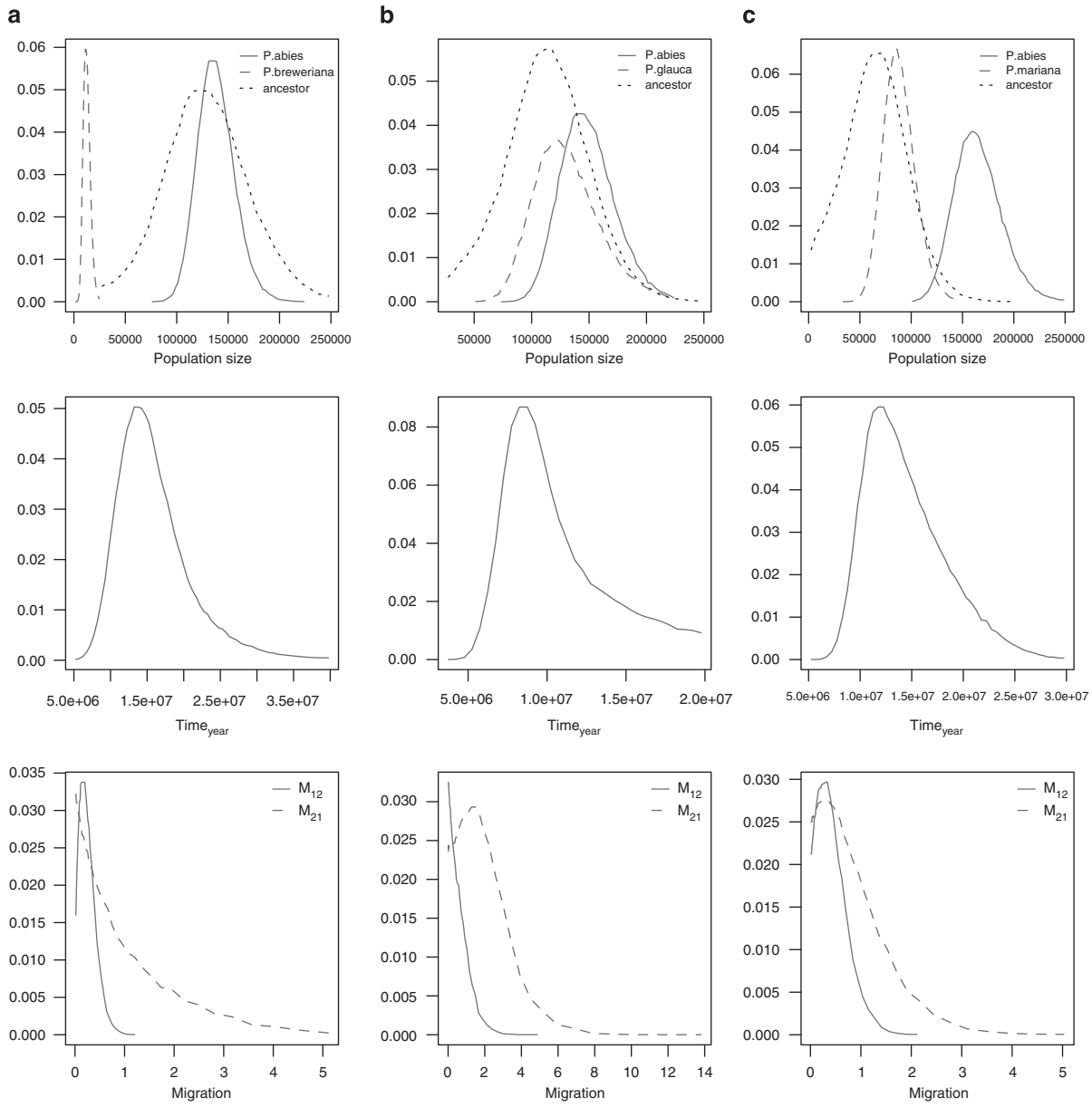

Figure 3 The smoothed marginal posterior distributions of estimated parameters under the standard Isolation-with-Migration (IM) model estimated by MIMAR ($\mu = 1.0 \times 10^{-8}$ per site per generation, 50 years per generation). (**a**) *P. abies* × *P. breweriana*, (**b**) *P. abies* × *P. glauca* and (**c**). *P. abies* × *P.mariana*.

the standard IM model (Figure 5). The two exceptions were S_f for *P. abies* and *P. glauca*, and S_f and π_2 for *P. abies* and *P. mariana* (Supplementary File S6). The poor fit of S_f ($P < 0.01$) could be because of too small number of fixed polymorphisms between *P. abies* and the two North American species (Supplementary Table S3), whereas poor fit of π_2 for *P.abies* and *P. mariana* ($P < 0.05$) suggests that a simple growth model may not capture every aspect of the species history.

Phylogeny

The species tree given by BEST supported the relationship inferred by the estimates of divergence time given by MIMAR (Figure 8). *P. breweriana* was at the basal position of the four species we analyzed, and *P. abies* and *P. glauca* clustered together.

Discussion

The main findings of the present study are the following. First, nucleotide variation was in general relatively low, considering the large distribution range of the boreal species and the notoriously high levels of heterozygosity at allozyme loci in conifer species. Nucleotide variation was also much lower in *P. breweriana* than in the other three species and its frequency spectrum was closer to standard neutral expectations. All three boreal species, on the other hand, had patterns similar to those observed in earlier studies of conifer species (*P. glauca*, Bouillé and Bousquet, 2005; *P. abies*, Heuertz *et al*., 2006; and *Pinus sylvestris*, Pyhäjärvi *et al*., 2007), likely reflecting recent population growth and/or past bottleneck(s). Second, shared polymorphisms were extensive and fixed sites almost entirely confined to the pairwise comparisons

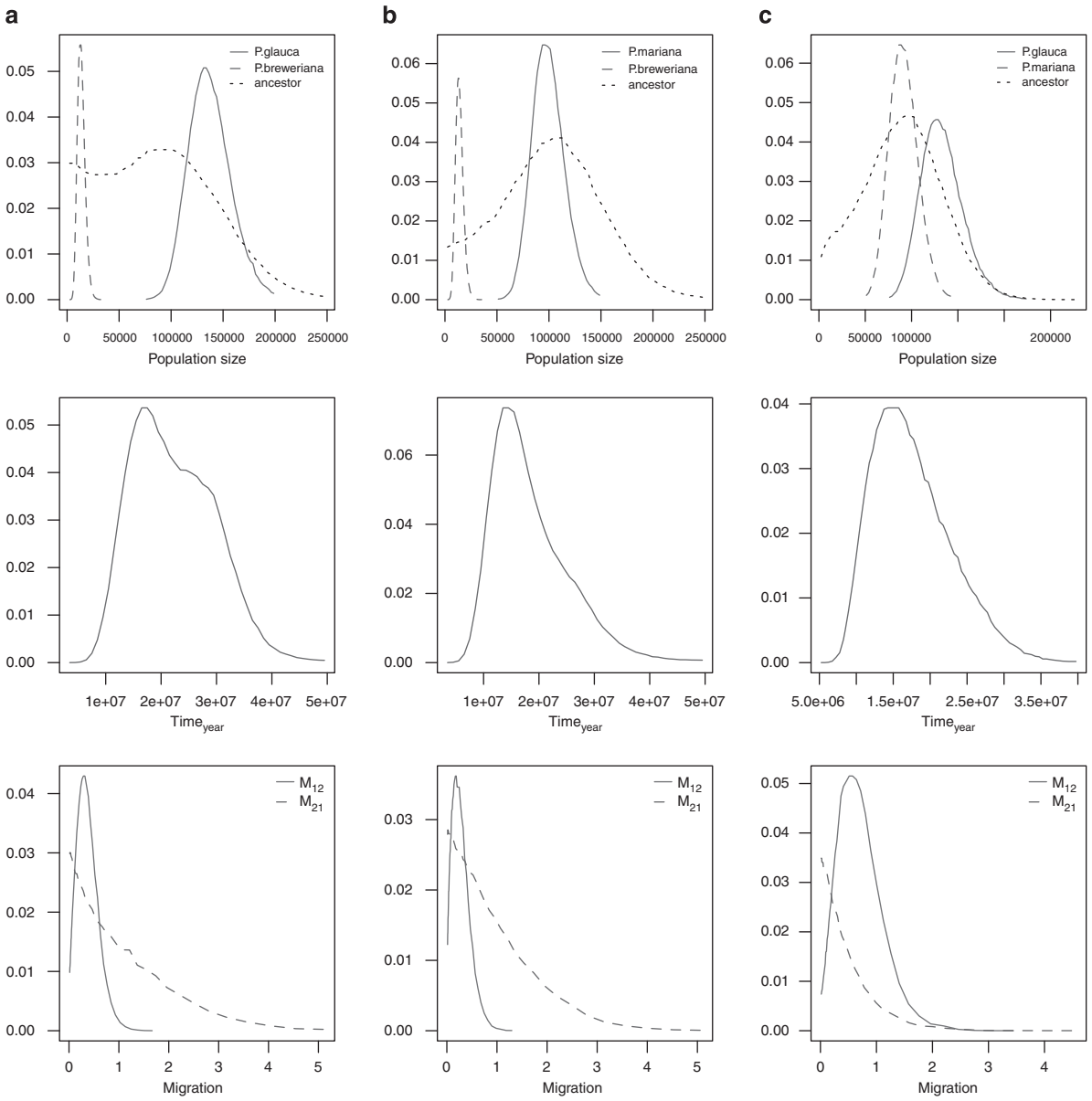

Figure 4 The smoothed marginal posterior distributions of estimated parameters under the standard Isolation-with-Migration (IM) model estimated by MIMAR ($\mu = 1.0 \times 10^{-8}$ per site per generation, 50 years per generation). (**a**) *P. glauca* × *P. breweriana*, (**b**) *P. mariana* × *P. breweriana* and (**c**). *P. glauca* × *P. mariana*.

involving *P. breweriana*. Estimates of effective populations sizes tended to be larger or of the same order of magnitude than previously published ones (Bouillé and Bousquet, 2005; Pyhäjärvi *et al*., 2007) except in *P. breweriana*, and divergence time much shorter than previous estimates (Bouillé and Bousquet, 2005). In particular, the divergence time between *P. abies* and the two North American boreal species *P. mariana* and *P. glauca* was shorter than the divergence time between the latter two, and there was even evidence for gene flow between *P. abies* and *P. glauca*. Among the species in this study and under all models, the effective population size of *P. abies* was the largest. This is also at variance with earlier species comparison (Bouillé and Bousquet, 2005), where *P. abies* had the smallest effective population size. This later result may simply indicate that estimates based on too few loci and too few sampled individuals are not reliable.

Polymorphism at allozyme and nucleotide levels

In contrast to the similar and high levels of polymorphisms observed at allozyme loci among the four spruce species studied here (Lagercrantz and Ryman, 1990; Ledig *et al*., 1997 and references therein, 2005), nucleotide variation was an order of magnitude lower in *P. breweriana* than in its boreal relatives. This low value is in agreement with the fossil record and with a mean Tajima's *D* value close to zero, both of which lent support to a scenario where *P. breweriana* retained a small population size since the Tertiary and through recent times. A similar discrepancy between levels of polymorphism at allozyme and nucleotide level has already been noted, in particular in nucleotide polymorphism surveys in *P. abies* (Heuertz *et al*., 2006) and *Pinus sylvestris* (Pyhäjärvi *et al*., 2007). In the latter case a direct comparison was done through genotyping and resequen-

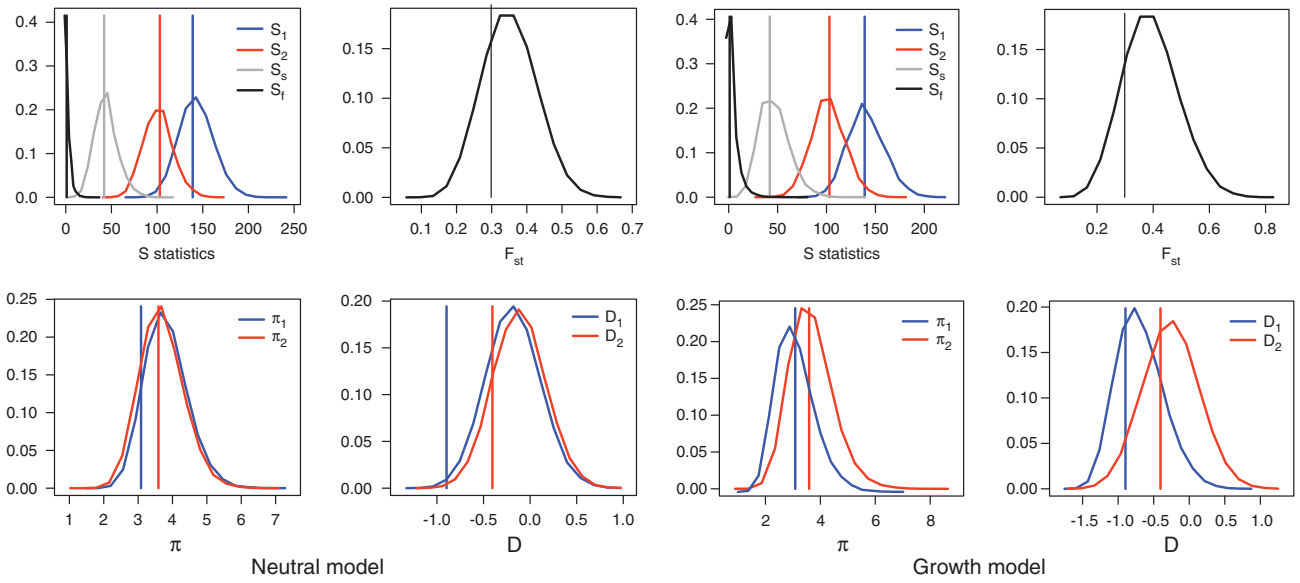

Figure 5 An example of goodness-of-fit plots from *P. abies* × *P. glauca* under both the neutral and growth models.

cing of six allozyme loci (Pyhäjärvi, 2008). Mutation and recombination rates were not higher at allozyme loci than at other loci. The haplotype distribution at allozymes was suggestive of recent partial sweeps or balancing selection, though more complex demographic scenarios than those tested could perhaps have also resulted in the observed pattern. As the same discrepancy seems to be observed across many species with different demographic histories, such as *P. breweriana* and the three boreal species studied here, selection at allozyme loci seems a more likely explanation than demographics.

Speciation in spruce species and biogeographical implications

There are a number of underlying assumptions to the IM models, in general, and to MIMAR, in particular, that may not always be met in this study, and could lead to biases in demographic estimates. First, one of the main assumptions is selective neutrality, which might be problematic when using coding sequence. None of the loci in this study show significant departure from neutrality when using Tajima's *D* or Fay and Wu's H. A single locus, ZTL, deviated from neutral patterns when using multilocus Hudson–Kreitman–Aguadé tests. However, additional MIMAR runs excluding this locus did not alter the parameter estimates significantly (data not shown). It decreased the estimates for a small amount of gene flow between *P. abies* and *P. breweriana*, which had a peak near to zero and could be an artefact of the Bayesian implementation (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004). However, the estimate of gene flow from *P. glauca* to *P. abies* was hardly influenced even though ZTL was excluded (1.12 under standard IM model and 4.32 under growth model). Second, IM models further assume a two-species system, where there is no gene flow with any other species. This is clearly not the case in this study, where at least the four included species seem to have exchanged genes with each other and with species not considered here: gene flow between *P. glauca* and *P. engelmanii*, and between *P. mariana* and *P. rubens* has been documented (for example, Perron and Bousquet, 1997; Major *et al*., 2005), and *P. abies* and *P. obovata* have a large hybrid zone centred around the Ob river (Tollefsrud *et al*., 2008). However, because *P. engelmanii* can be considered as a subspecies of *P. glauca* (Rajora and Dancik, 2000) and *P. rubens* a derivative of *P. mariana* (Perron *et al*., 2000), putative gene flow between them and the species studied here might have a minor effect on the level of polymorphism of the latter. The *P. abies* populations included in the present study all belong to the western part of the range, and therefore putative gene flow from *P. obovata* is likely to be limited or absent. Of course, shared polymorphisms with other spruce species exist, in particular between *P. abies* and other Eurasian species (our own unpublished data) and we cannot rule out that ancient gene flow from other species did inflate our estimates of the current effective population sizes.

One of the most striking results is the low divergence and the presence of unidirectional gene flow between *P. glauca* and *P. abies*. This, of course, seems surprising at first glance as both species are today found on different continents. Furthermore, it is also at variance with (i) chloroplast DNA phylogenies that place *P. abies* and *P. glauca* in distinct lineages (Ran *et al*., 2006), and (ii) the difficulty to cross those species (Mikkola, 1969). When *P. glauca* pollen is used to fertilize *P. abies* female cones the pollen tubes penetrate through the nucellar cap and in some cases eggs are fertilized. The embryos, however, die at different stages. It is interesting to note that the pollen tubes of *P. abies* penetrate more effectively into the nucellus of *P. glauca* than the pollen tubes in the reciprocal crosses, whereas our results would have led us to expect the opposite. On the other hand, in agreement with our results that did not detect any gene flow between these species, sterility of the cross *P. mariana* × *P. glauca* is more strongly established and seems to be due to incompatibility: the growth of the pollen tube is halted midway through the nucellar cap or sooner. So, in summary, even if crosses between *P. abies*

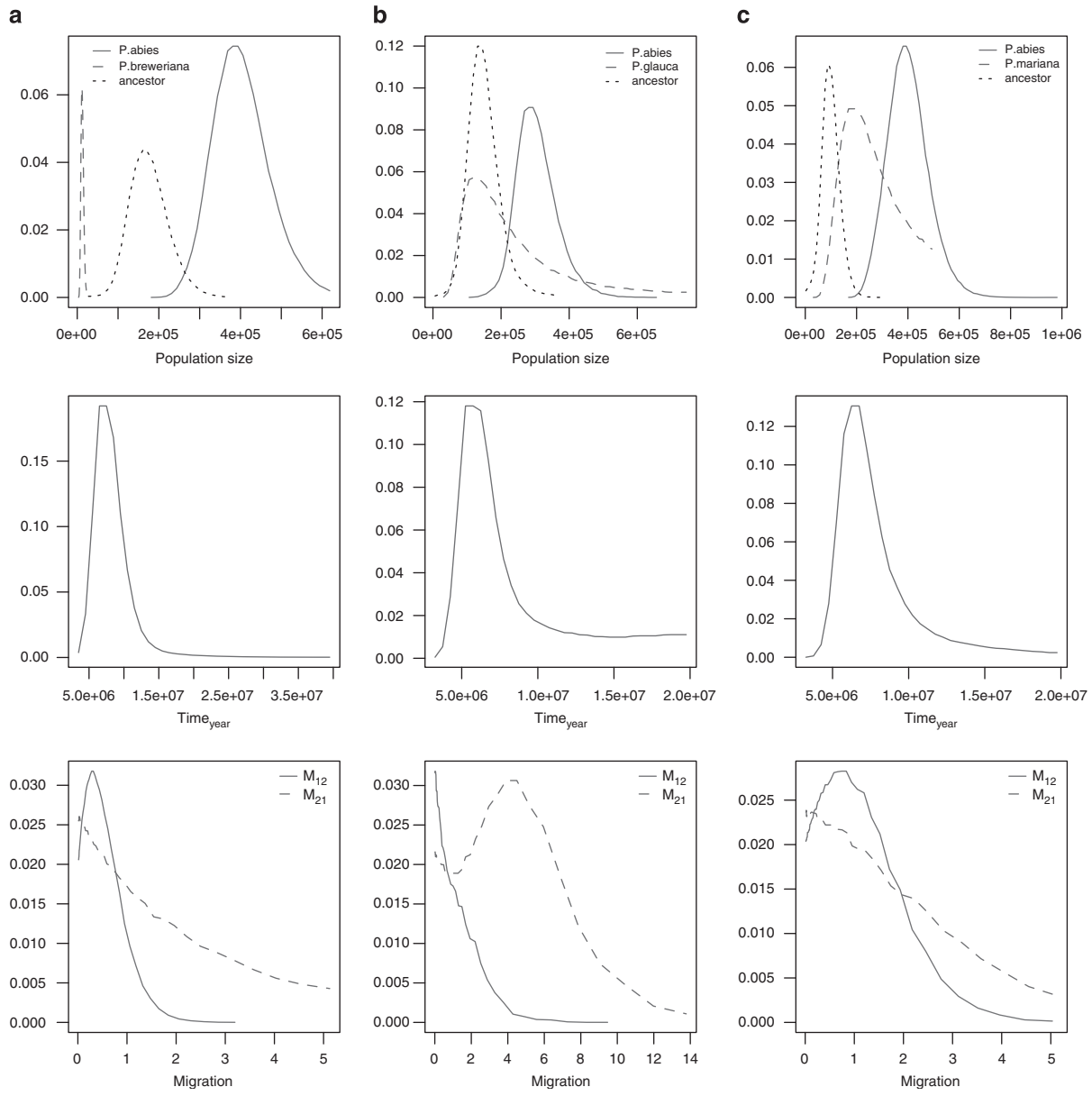

Figure 6 The smoothed marginal posterior distributions of estimated parameters under a growth model estimated by MIMAR ($\mu = 1.0 \times 10^{-8}$ per site per generation, 50 years per generation). (**a**) *P. abies* × *P. breweriana*, (**b**) *P. abies* × *P. glauca* and (**c**) *P. abies* × *P. mariana*.

and *P. glauca* are today difficult, if not impossible, they might well have been easier just after the two species diverged. Also, it is possible that *P. obovata*, which is closely related to *P. abies* and is geographically closer to the northern part of the range of *P. glauca*, acted as a bridge between the two species.

The estimates of divergence time between the four species show that *P. breweriana* separates from the others, whereas *P. glauca* was more closely related to *P. abies* than to *P. mariana*. Phylogenetic relationship based on organelle DNA could be unreliable because of higher rate of random genetic drift rate and complete linkage among loci, making mitochondrial DNA in essence equivalent to a single locus (Hudson and Coyne, 2002). Thus, we randomly picked out one individual in each species and constructed a phylogenetic tree based on multi-locus nuclear DNA using BEST programme (Liu and Pearl, 2007). The topology of the resulting tree was congruent with the relationships deduced from MIMAR estimates of divergence times. BEST, in contrast to MIMAR, assumes no recombination within locus and no migration between species. Ignoring migration amounts to assume that all shared polymorphisms are ancestral and will, therefore, lead to the overestimation of the closeness of relationship among species such as *P. abies* and *P. glauca* that have a high number of shared polymorphisms. Ignoring recombination will make trees look more star-like (Schierup and Hein, 2000). In this case none of these departures seem to have had an overwhelming effect.

Conclusion

To our knowledge this is one of the first multi-species survey of nucleotide diversity in conifers and the first study to provide estimates of IM parameters. It confirms

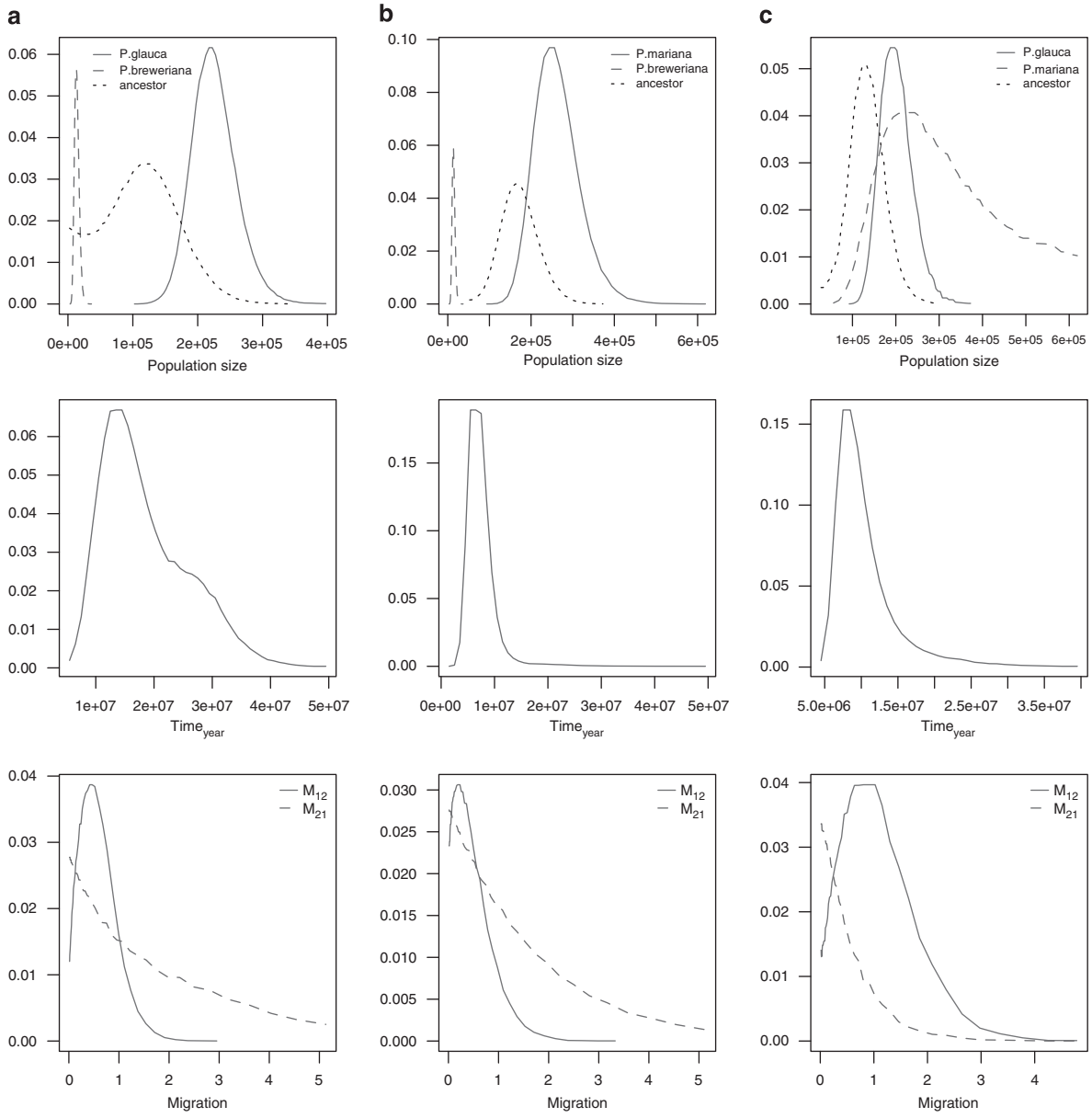

Figure 7 The smoothed marginal posterior distributions under a growth model estimated by MIMAR ($\mu = 1.0 \times 10^{-8}$ per site per generation, 50 years per generation). (**a**) *P. glauca* × *P. breweriana*, (**b**) *P. mariana* × *P. breweriana* and (**c**) *P. glauca* × *P. mariana*

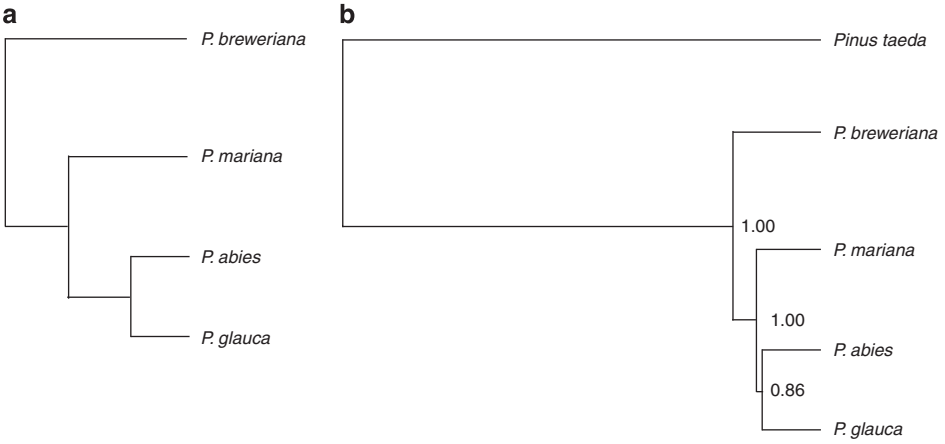

Figure 8 Phylogenetic relationships among the four studied spruce species. (**a**) Deduced from the divergence times estimated by MIMAR. (**b**) The phylogenetic tree and posterior probabilities for the topology as inferred with the programme BEST.

the relatively low nucleotide diversity previously observed in conifers (e.g. Heuertz *et al*., 2006) and the extent of shared polymorphisms, and highlights the recent divergence and diversity of histories and effective population sizes among species of this important group of forest trees. The availability of well described group of species with contrasted histories and ecology will certainly prove to be very useful when assessing the relative roles played by historical contingencies and general adaptive mechanisms in trees response to environmental changes.

Acknowledgements

The research was funded by grants from the Carl Tryggers Foundation, the Philip-Sörensen Foundation and the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) to Martin Lascoux. The EVOLTREE network of excellence financed Niclas Gyllenstrand. Thomas Källman received support from the Nilsson-Ehle Foundation. We thank Céline Becquet for kindly answering questions about MIMAR, and Jing Huang for help with labwork. We also want to thank Thomas Ledig and Jean Bousquet for providing us with seeds from the North American spruce species.

References

Bazin E, Glémin S, Galtier N (2006). Population size does not influence mitochondrial genetic diversity in animals. *Science* **312**: 570–572.

Becquet C (2007). Population genetic approaches to the study of speciation. PhD thesis, Department of Human Genetics, University of Chicago.

Becquet C, Przeworski M (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Res* **17**: 1505–1519.

Bouillé M, Bousquet J (2005). Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer Picea (Pinaceae): Implications for the long-term maintenance of genetic diversity in trees. *Am J Bot* **92**: 63–73.

De Vernal A, Hillaire-Marcel C (2008). Natural variability of Greenland climate, vegetation, and ice volume during the past million years. *Science* **320**: 1622–1625.

Ewing B, Green P (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**: 186–194.

Ewing B, Hillier L, Wendl MC, Green P (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175–185.

Haddrill P, Thornton KR, Charlesworth B, Andolfatto P (2005). Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res* **15**: 790–799.

Heuertz M, De Paoli E, Källman T, Larsson H, Jurman I, Morgante M *et al*. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway Spruce [*Picea abies* (L.) Karst]. *Genetics* **174**: 2095–2105.

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.

Hudson RR, Coyne JA (2002). Mathematical consequences of the genealogical species concept. *Evolution* **56**: 1557–1565.

Gordon D, Abajian C, Green P (1998). Consed: A graphical tool for sequence finishing. *Genome Res* **8**: 195–202.

Ingvarsson PK (2008). Multilocus patterns of nucleotide polymorphism and the demographic history of Populus tremula. *Genetics* **180**: 329–340.

Innan H, Kim Y (2008). Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics* **179**: 1713–1719.

Lagercrantz U, Ryman N (1990). Genetic structure of Norway spruce (Picea abies): Concordance of morphological and allozymic variation. *Evolution* **44**: 38–53.

Ledig FT, Hodgskiss P, Johnson D (2005). Genetic diversity, genetic structure, and mating system of Brewer spruce (Pinaceae), a relict of the Arcto-Tertiary forest. *Am J Bot* **92**: 1975–1986.

Ledig FT, Jacob-Cervantes V, Hodgkiss PD, Equiluz-Piedra T (1997). Recent evolution and divergence among populations of a rare Mexican endemic, Chihuahua spruce, following Holocene climatic warming. *Evolution* **51**: 1815–1827.

Liu L, Pearl DK (2007). Species trees from gene tree: Reconstructing Bayesian posterior distribution of a species phylogeny using estimated gene tree distributions. *Syst Biol* **56**: 504–514.

Major JE, Mosseler A, Johnsen KH, Rajora BP, Barci DC, Kim KH *et al*. (2005). Reproductive barriers and hybridity in two spruces, *Picea rubens*, *Picea mariana*, sympatric in eastern North America. *Can J Bot* **83**: 163–175.

McVean G, Awadalla P, Fearnhead P (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**: 1231–1241.

Mikkola L (1969). Observations on interspecific sterility in *Picea*. *Ann Bot Fennici* **6**: 285–339.

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

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006). Genetic evidence for complex speciation of humans and chimpanzee. *Nature* **441**: 1103–1108.

Perron M, Bousquet J (1997). Natural hybridization between black spruce and red spruce. *Mol Ecol* **6**: 725–734.

Perron M, Perry DJ, Andalo C, Bousquet J (2000). Evidence from sequence-tagged-site markers of a recent progenitor-derivative species pair in conifers. *Proc Natl Acad Sci USA* **97**: 11331–11336.

Pyhäjärvi T (2008). Roles of demography and natural selection in molecular evolution of trees, focus on *Pinus sylvestris*. PhD thesis, University of Oulu, Finland.

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

Rajora O, Dancik BP (2000). Population genetic variation, structure, and evolution in Engelmann spruce, white spruce, and their natural hybrid complex in Alberta. *Can J Bot* **78**: 768–780.

Ran JH, Wei XX, Wang XQ (2006). Molecular phylogeny and biogeography of Picea (Pinaceae): Implications for phylogeographical studies using cytoplasmic haplotypes. *Mol Phylogenet Evol* **41**: 405–419.

Ronquist F, Huelsenbeck JP (2003). MRBAYES: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.

Schierup MH, Hein J (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**: 879–891.

Slotte T, Huang H, Lascoux M, Ceplitis A (2008). Polyploid speciation did not confer instant reproductive isolation in Capsella (Brassicaceae). *Mol Biol Evol* **25**: 1472–1481.

Thornton K (2003). Libsequence: A C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**: 2325–2327.

Tollefsrud MM, Kissling R, Gugerli F, Johnsen O, Skrøppa T, Cheddadi R *et al*. (2008). Genetic consequences of glacial survival and postglacial colonization in Norway spruce: Combined analysis of mitochondrial DNA and fossil pollen. *Mol Ecol* **17**: 4134–4150.

Wakeley J (2008). *Coalescent theory—An introduction*. Roberts and Company Publishers.

Wakeley J, Hey J (1997). Estimating ancestral population parameters. *Genetics* **145**: 847–855.

Wright SJ, Charlesworth B (2004). The HKA test revisited: A maximum likelihood ratio test of the standard neutral model. *Genetics* **168**: 1071–1076.

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

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)