Demographic histories of four spruce (*Picea*) species of the Qinghai-Tibetan Plateau and neighboring areas inferred from multiple nuclear loci

Yuan Li,[†,1,2] Michael Stocks,[†,2] Sofia Hemmilä,[†,2] Thomas Källman,[†,2] Hongtao Zhu,[1] Yongfeng Zhou,[1] Jun Chen,[2] Jianquan Liu,[*,1] and Martin Lascoux[*,2,3]

[1]Institute of Molecular Ecology, The MOE Key Laboratory of Arid and Grassland Ecology, College of Life Science, Lanzhou University, Lanzhou, Gansu, China

[2]Program in Evolutionary Functional Genomics, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[3]Laboratory of Evolutionary Genomics, CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai, China

†These authors contributed equally to the present study.

***Corresponding authors:** E-mail: ljqdxy@public.xn.qh.cn; martin.lascoux@ebc.uu.se.

Associate editor: Jody Hey

Abstract

Nucleotide variation at 12–16 nuclear loci was studied in three spruce species from the Qinghai-Tibetan Plateau (QTP), *Picea likiangensis*, *P. wilsonii*, and *P. purpurea*, and one species from the Tian Shan mountain range, *P. schrenkiana*. Silent nucleotide diversity was limited in *P. schrenkiana* and high in the three species from the QTP, with values higher than in boreal spruce species, despite their much more restricted distributions compared with that of the boreal species. In contrast to European boreal species that have experienced severe bottlenecks in the past, coalescent-based analysis suggests that DNA polymorphism in the species from the QTP and adjacent areas is compatible with the standard neutral model (*P. likiangensis*, *P. wilsonii*, and *P. schrenkiana*) or with population growth (*P. purpurea*). In order to test if *P. purpurea* is a diploid hybrid of *P. likiangensis* and *P. wilsonii*, we used a combination of approaches, including model-based inference of population structure, isolation-with-migration models, and recent theoretical results on the effect of introgression on the geographic distribution of diversity. In contrast to the three other species, each of which was predominantly assigned to a single cluster in the Structure analysis, *P. purpurea* individuals were scattered over the three main clusters and not, as we had expected, confined to the *P. likiangensis* and *P. wilsonii* clusters. Furthermore, the contribution of *P. schrenkiana* was by far the largest one. In agreement with this, the divergence between *P. purpurea* and *P. schrenkiana* was lower than the divergence of either *P. likiangensis* or *P. wilsonii* from *P. schrenkiana*. These results, together with previous ones showing that *P. purpurea* and *P. wilsonii* share the same haplotypes at both chloroplast and mitochondrial markers, suggest that *P. purpurea* has a complex origin, possibly involving additional species.

Key words: *Picea*, Qinghai Tibetan Plateau, effective population size, divergence time, introgression, speciation.

Introduction

Coniferous species, in general, and spruce species, in particular, are an especially interesting group of species from an evolutionary point of view. The combination of large effective population sizes, long generation times, and introgression have led to a remarkable evolutionary web of species characterized by an exceptional amount of shared polymorphisms and low levels of divergence between species (Syring et al. 2007; Chen et al. 2010; Willyard et al. 2009). Inferring the evolutionary history of conifer species is therefore particularly challenging. In the present study, we combine different analytical approaches to understand the relationships of four closely related Asian spruce (*Picea*) species from the Qinghai-Tibetan Plateau (QTP) and adjacent areas (fig. 1), a complex and fragmented landscape resulting from a tumultuous geological and climatic evolution over the past 50 million years (Ma) (Royden et al. 2008). *Picea wilsonii* Masters has the largest distribution and occurs at intermediate altitudes (1,400–2,800 m above sea level [asl]) on the northeastern edge of the plateau and in small scattered populations as far east as Hebei in Central China. *Picea likiangensis* (Franch.) Pritzel has the southernmost distribution and is found at high altitudes (2,500–4,100 m asl) along the southern limit of the QTP. *Picea likiangensis* has traditionally been subdivided into three main varieties, var. *rubescens* in the north, var. *linzhensis* in the west, and var. *likiangensis* in the southeast. *Picea purpurea* Masters has a smaller distribution range that is nested between those of *P. likiangensis* and *P. wilsonii* at altitudes ranging from 2,600 to 3,800 m asl. Finally, the distribution range of *P. schrenkiana* Fischer et Meyer is presently completely separated from the distributions of the three other species. *Picea schrenkiana* occurs in the Tian Shan mountain range north of the vast and barren Tarim Basin and also extends across the border into Kazakhstan and Kyrgyzstan (Fu et al. 1999) at altitudes ranging from 1,200 to 3,500 m

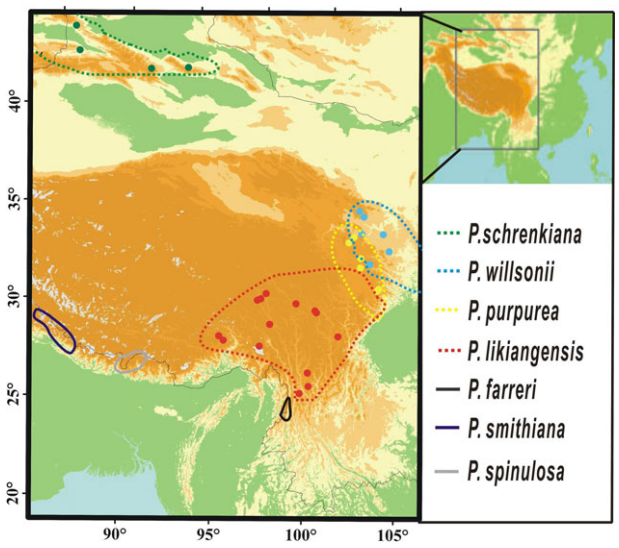

FIG. 1. The locations, on and around the QTP, of each of the sampled populations for the four spruce species studied here, *Picea schrenkiana*, *P. wilsonii*, *P. purpurea*, and *P. likiangensis*, together with their approximate distribution ranges. Also shown are the distribution ranges of species that might have contributed to the ancestry of the four studied species. The inset box shows the location of the studied area in a larger scale map of Asia. The *x* axis gives the longitude and the *y* axis gives the latitude.

asl. Collectively, these four species constitute a representative species complex of the modern geographic distribution of spruce on the QTP and neighboring highlands (Dupont-Nivet et al. 2008). All four species are diploid (Fu et al. 1999), and the phylogenetic relationships among them are unclear. Based on morphological evidence (seed scales of cones and stomatal lines along the leaf abaxial surface), *P. schrenkiana* and *P. wilsonii*, on the one hand, and *P. purpurea* and *P. likiangensis*, on the other hand, are assigned to two different sections (Farjón 1990; Fu et al. 1999). However, chloroplast DNA (cpDNA) sequence data give a different grouping: Although all four species belong to the same clade, together with three other minor species of the same area not considered here (*P. brachytyla*, *P. smithiana*, and *P. farreri*), *P. wilsonii* and *P. purpurea* belong to one subclade and *P. schrenkiana* and *P. likiangensis* are part of another one (Ran et al. 2006). The latter grouping is also supported by an extensive study of both cpDNA and mitochondrial DNA (mtDNA) in *P. wilsonii*, *P. purpurea*, and *P. likiangensis* (Du F, Petit RJ, Liu JQ, unpublished data), and the results further suggest that *P. wilsonii* is monomorphic for both organelle markers and has extensive haplotype sharing with *P. purpurea* at both cpDNA and mtDNA markers.

So far there have been very few studies based on nuclear DNA sequence variation in tree species of the QTP (but see Ma et al. 2006). Most studies have focused on local population structure using allelic variation at allozymes (e.g., Luo et al. 2005), random amplified polymorphic DNA (e.g., Peng et al. 2007), or microsatellite markers (e.g., Wang et al. 2005) or on the phylogeography of individual species using cpDNA or mtDNA (e.g., Meng et al. 2007; Du et al. 2009). Although certainly useful, these latter studies are inherently limited by the uniparental inheritance and lack of recombination of organelle genomes and statistically sound population genetic inferences of deeper time events require the use of independent nuclear loci (Wakeley 2003; Lascoux et al. 2004). We therefore sequenced 12–16 nuclear loci in the four species and analyzed their nucleotide variation. Three complementary approaches were used to understand the impact of demographic processes on population divergence, namely parametric models, such as the isolation-with-migration (IM) models, semiparametric methods, such as that implemented in the program Structure (Pritchard et al. 2000) and recent intensive simulation-based approaches. IM models allow the estimation of key evolutionary parameters, such as current and ancestral effective population sizes and divergence times, and the separation of ancestral shared polymorphisms from gene flow and introgression occurring during the divergence process. Although IM models allow us to distinguish parapatric from allopatric speciation (Hey 2006), they are, in most cases, still limited to a pair of species, a situation that might be inappropriate in some cases, and they might not give reliable estimates if, for example, the divergence took place too recently, leading to a dearth of fixed nucleotide sites between species. These methods are also based on simple demographic models and on a set of demanding assumptions, the violation of which can lead to poor estimates (Becquet and Przeworski 2007, 2009; Nielsen and Beaumont 2009). Semiparametric methods, such as that implemented in the program Structure (Pritchard et al. 2000), are powerful exploratory tools when trying to understand population genetic structure, in particular, admixture and hybridization (e.g., Lepais et al. 2009). However, in contrast to the parametric methods, they do not lead to estimates of demographic parameters. Finally, at the other end of the modeling spectrum, recent intensive simulation-based approaches are able to implement complex demographic processes that have often been neglected in population genetics studies, thereby allowing new insights into the forces shaping genetic variation. Excoffier and co-workers (e.g., Excoffier and Ray 2008; Currat et al. 2008) have, for instance, highlighted the importance of demographic processes occurring during range expansions in shaping the distribution of genetic variation within and between species. In particular, they have shown that simple demographic processes are often sufficient to explain the fact that introgression is generally observed from the local species to the invading one (Currat et al. 2008). Directional introgression is common in plants and has often been explained by selection. Currat et al. (2008) suggest that this is not a parsimonious interpretation. Similarly to the semiparametric methods, these exploratory approaches do not lead to the estimation of demographic parameters. The three groups of methods are therefore complementary.

We used these different approaches to address three main questions. First, is the nucleotide variation in the spruce species of the QTP higher than that of boreal conifer

species such as *P. abies* or *Pinus sylvestris*? The latter have enormous distribution ranges, but their diversity seems to have been curtailed by ancient bottlenecks (Heuertz et al. 2006; Pyhäjärvi et al. 2007). In contrast, species from the QTP may have been less affected by repeated glacial cycles. Second, when did species from the QTP diverge and did they also show evidence of range shifts and ancient changes in population size in response to the geological and climatic history of the QTP? Finally, is *P. purpurea* the result of recent introgression between *P. likangensis* and *P. wilsonii*, as suggested by morphological traits and chloroplast and mitochondrial data (Farjón 1990; Fu et al. 1999; Du et al. 2009) or does it have a more complex origin?

Methods

Plant Material

Seeds were harvested from 4, 6, 6, and 15 populations of *P. schrenkiana*, *P. wilsonii*, *P. purpurea*, and *P. likiangensis*, respectively (fig. 1 and supplementary table S1, Supplementary Material online). We sampled 1–16 individuals per population, resulting in 23–80 individuals per species. The seeds were stored at 4°C and then soaked in water at 4°C overnight before the haploid DNA from megagametophytes was isolated. DNA extraction was carried out with either QIAGEN DNeasy Plant Mini Kits or the CTAB method (Doyle and Doyle 1990).

Sequencing

In total, 12 loci for *P. schrenkiana*, 14 for *P. wilsonii* and *P. purpurea*, and 16 for *P. likiangensis* (2009, 4CL, COL2, EBS, FT3, GI, MOO2, M007D1, PCH, Sb16, Sb29, Sb62, se1364, se1390, xy1420, ZTL) were used for sequence analysis (supplementary table S2, Supplementary Material online). The loci were chosen to be single or low copy genes. Sequence reactions were carried out either on an ABI 3730XL DNA Analyzer using ABI Prism Bigdye 3.1 or on a MEGABACE 1000 DNA Analysis System using Terminator Cycle Sequencing Ready Reaction Kit. Sequence data were base called, assembled and edited with the Phred v. 0.020425.c, Phrap v. 0.990319, and Consed v. 19.0 software suite keeping only high-quality data (phred score $>$ 25) (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998) at Uppsala University or simply edited with Mega v. 3.1 (Kumar et al. 2004) at Lanzhou University. All putative polymorphic sites were finally validated by visual inspection of chromatograms. All sequence data have been deposited with the EMBL/GenBankData Libraries under accession numbers GU261725–GU262987. In a few cases, different loci were sequenced from different megagametophytes originating from the same mother tree. In situations where this occurred, they were treated as one single haplotype.

Nucleotide diversity

For each species, S, the number of segregating sites, N_H, the number of haplotypes, Watterson's parameter, θ_W (Watterson 1975), π, the nucleotide diversity (Tajima 1983), and the haplotype diversity, H_e (Nei 1987) were calculated on the pooled data set using compute (Thornton 2003).

Linkage disequilibrium and recombination

As in Heuertz et al. (2006), the square of the correlation coefficient between each single nucleotide polymorphism (SNP) pair, r^2, was calculated to estimate linkage disequilibrium (LD) using DnaSP v. 4.50 (Rozas et al. 2003). The significance level of the statistical association between alleles at different sites was measured using Fisher's exact test, and Bonferroni correction was used to correct for false positives. Multilocus estimates of the population recombination rate, ρ, and the population mutation rate, θ, were obtained with the program Seqlib v. 1.0 (http://sourceforge.net/projects/seqlib/, Mita et al. 2007), which was also used for demographic inferences (see below).

Population structure

The genetic structure of the four spruce species was assessed using two methods. First, an analog of Wright's fixation index (Wright 1951; Weir and Cockerham 1984), which also takes into account the distance among haplotypes, ϕ_{ST} (Excoffier et al. 1992), was used to assess population differentiation within and among species. ϕ_{ST} values were estimated with an analysis of molecular variance (AMOVA) approach implemented by Arlequin v. 3.1.1 (Excoffier et al. 2005). Results are reported as an average over loci weighted by the total variance in allele frequency in the main text as well as locus by locus in the Supplementary Material online. Significance of ϕ_{ST} values was tested by permuting haplotypes among populations. Second, Structure v. 2.3 (Hubisz et al. 2009) was used to assess the correspondence between species grouping and genotypic clustering. Compared with Structure v. 2.2 (Pritchard et al. 2000; Falush et al. 2003, 2007), Structure v. 2.3 makes use of information about sampling locations. In the present case, individuals were simply grouped by species. Hubisz et al. (2009) showed that Structure v. 2.3 produces more accurate ancestry estimates than Structure v. 2.2 when information is low but gives similar results when the data set is informative. Following Hubisz et al. 2009 and to insure that the resulting clustering was not due to too much weight being given to outcomes correlated with sampling location, we also used Structure v. 2.2. Because the results with Structure v. 2.3 were biologically easier to interpret, but otherwise did not reveal any substantial difference with results obtained with Structure v. 2.2, we will only report the former. As in Heuertz et al. (2006), sites that showed significant statistical association after Bonferroni correction were removed before Structure analysis. To infer the structure of the sampled populations, the likelihood of each number of clusters, K, where $1 \leqslant K \leqslant 6$, was assessed and allowance made for the correlation of allele frequencies between clusters. Two sets of runs were performed, both under an admixture model. First, we performed 10 runs with a burn-in of 50,000 and 500,000 iterations. Then to get a better estimate of the variation among runs, 50 runs were

performed with a burn-in of 20,000 and 500,000 iterations. The latter are reported here. Graphics were drawn using the program Distruct v.1.1 (Rosenberg 2004). The most likely number of clusters was estimated with the original method from Pritchard et al. (2000) and with the ΔK statistics given in Evanno et al. (2005).

Departure from the standard neutral model

Tajima's D (Tajima 1989) and Fu and Li's D^* and F^* statistics (Fu and Li 1993) were calculated using compute (Thornton 2003). Fay and Wu's H (Fay and Wu 2000; Zeng et al. 2006) was calculated using DnaSP v. 4.50 (Rozas et al. 2003). The polarization into ancestral and derived states needed for Fay and Wu's H was inferred using different outgroups (see supplementary table S3, Supplementary Material online, for a list of the outgroups used for different genes). Due to the fragmented distributions of some of the sampled species, each species could not realistically be assumed to be panmictic. This is particulary true for *P. likiangensis* that has been morphologically divided into a number of subspecies, and it was important to consider how our sampling scheme may have affected our conclusions. A recent study by Städler et al. (2009) showed that pooling samples from different populations in a species with population structure can artificially inflate values of Tajima's D under an equilibrium stepping-stone model and thereby potentially bias any conclusions that were reached. Städler et al. (2009) considered three sampling schemes: local samples contain n sequences from a single randomly chosen population; pooled samples contain n sequences originating from several populations; and scattered samples include individual sequences from n different populations (i.e., a single sequence is randomly picked from each population). Only the scattered and pooled sampling schemes were considered in this study as the number of sequences per population was limited. For the scattered sampling scheme, one individual was randomly selected from each subpopulation, whereas in the pooled sampling scheme, the same number of individuals was randomly sampled from all subpopulations. The summary statistics S, η_1 (number of singletons), θ_W, π, and Tajima's D were calculated for the resampled individuals and averaged over all loci. The pooled sampling scheme implemented in this study differs slightly from the scheme implemented by Städler et al. (2009) as they sampled evenly from a subset of demes, whereas we sampled randomly from all possible subpopulations. The resampling was repeated 100 times for each sampling scheme and for each of the four studied species.

The Hudson-Kreitman-Aguadé (HKA) test (Hudson et al. 1987) was performed with the program HKA (http://lifesci.rutgers.edu/~heylab) in the three species from the QTP using *P. schrenkiana* as an outgroup. It should be noted that the presence of shared polymorphisms among the species from the QTP and *P. schrenkiana* may affect the result of the HKA test and therefore, these results should be interpreted carefully.

We used an approximate Bayesian computation (ABC) approach (Marjoram and Tavaré 2006 and references therein) implemented in the program Seqlib v. 1.0 (Mita et al. 2007) and available at http://sourceforge.net/projects/seqlib/ to fit three simple demographic models to the data. The algorithm in Seqlib compares observed values of the number of segregating sites, S, the nucleotide diversity, π, and the number of haplotypes to the same summary statistics obtained through coalescent simulations. The program implements the regression ABC approach first introduced by Beaumont et al. (2002) and which is also clearly described in Thornton (2009). Its main advantage over a simple sample rejection method is its speed. In short, 1 million simulations were performed, sampling from wide uniform priors for all model parameters. From these simulations, the 5% closest to the observed data was kept and used in a local linear regression step to obtain estimates of model parameters. Three models were evaluated 1) the standard neutral model (SNM) that includes two parameters, the population mutation rate, θ, and the population recombination rate, ρ; 2) an exponential growth model (PEM) with three parameters θ, ρ, and an exponential growth factor, α; and, finally, 3) a bottleneck model (BNM) with five estimated parameters θ, ρ, the time since the end of the bottleneck, T, the duration of the bottleneck, d, and the size of the population during the bottleneck, f. Under all models, a uniform prior over a range from 0 to 0.1 was used for both θ and ρ. For the PEM, the growth factor (α) prior was uniformly distributed over 0 to 100 and in the BNM, uniform priors covering 0–10, 0–10, and 0–1 were used for the time since the end of the bottleneck (T), the duration of bottleneck (d), and the population size during the bottleneck (f), respectively. All parameters are scaled by the present-day θ which, in the case of a bottleneck, means that the size of the population during the bottleneck was constrained to be smaller or equal to the present-day size (0–1). Prior ranges were chosen based on previous results from conifers (Heuertz et al. 2006; Pyhäjärvi et al. 2007). In order to obtain a posterior probability for each of the models, a new ABC analysis was performed, but using the discretized (16 categories for each parameter) posterior from the first round as a prior. We also restricted the number of simulations to 10,000 and kept 1,000 for regression and calculation of model probability. For details regarding the procedure and methods implemented in this analysis, see the Seqlib v. 1.0 manual. We evaluated the evidence of model 1 against model 2 (where 1 stands for SNM, PEM, and BNM and 2 stands for SNM) using an approximation of the Bayes factor (Kass and Raftery 1995). We computed the Bayes factors as the ratios of the posterior probabilities of models 1 and 2.

Isolation-with-migration

Parameters of an IM model were estimated with the program Mimar (Becquet and Przeworski 2007). The IM model used in Mimar is defined by six 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_{ei} 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 migration rates between the two descendent populations, $m_{12}=4N_{e1}m_{12}$ and $m_{21}=4N_{e1}m_{21}$, where m is the migration rate per generation. Mimar implements a Markov chain Monte Carlo method to estimate IM parameters and is based on a slightly modified version of the 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 the samples 1 and 2, respectively, S_s is the number of sites with shared alleles between the two samples, and S_f is the number of sites where fixed alleles are found in one sample and no polymorphisms are found in the other sample. Mimar differs from previous implementations of the IM model by taking recombination into account, but otherwise, it relies on the same set of major assumptions. Namely, loci are assumed to be neutral, there is no population structure within each species, and the pair of species considered is assumed to be more closely related than any of them is to a third species. Finally, as pointed out by its authors, Mimar does not provide precise estimates unless data sets have both shared and fixed alleles between the two samples. We therefore limited our analyses to pairs of species involving *P. schrenkiana*, as the other pairs of species did not have any fixed sites. Importantly, this constitutes an indirect comparison of the three species of the southeast QTP, and in particular, it assesses the relationship of *P. purpurea* with its two putative parental species. 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 species i, f_i (see supplementary table S3, Supplementary Material online, for a list of the outgroups used). 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 $f_i = 0, f_j = 1$, and $i \neq j$, the allele is fixed in sample j; and if $f_i = 0$, $0 < f_i < 1$, and $i \neq j$, the allele is specific to sample j. The average multilocus estimate of ρ was used to estimate the population recombination rate. All pairwise comparisons were run for 20 million steps (with a 2 million step burn-in), with the number of ancestral recombination graphs per locus generated each step set to 100. The variance for each parameter of the normal kernel distributions used to propose each new parameter value was optimized in a series of preliminary runs by monitoring the acceptance rate and the degree of mixing of the chain. Due to problems to obtain good estimates for some of the parameters in the *P. wilsonii* /*P. schrenkiana* comparison, additional runs were performed to assess whether this ambiguity could also bias other parameters.

To further test the validity of the Mimar runs, a goodness-of-fit test was performed using the package MimarGOF, with simulations being carried out using the posterior distributions of the parameters estimated in Mimar. F_{ST} between the divergent species, π and Tajima's D for each descendent species, and the Wakeley and Hey (1997) S statistics were then calculated for this simulated data and compared with observed values to assess the fit of the IM model.

Results

Nucleotide diversity

Sequence diversity was detected at all 16 loci. An average number of 79 megagametophytes per locus were sequenced across species. Of 10,827 bp sequenced, around 60% was in coding regions. Insertions/deletions (indels) covered 160 bp and were excluded from the analyses unless specified. An average of 170 segregating sites were found per species, ranging from 49 in *P. schrenkiana* to 256 in *P. likiangensis*. Singletons constituted 44% of the total number of segregating sites. *Picea purpurea* has the highest silent nucleotide diversity, π_s (0.00996). *Picea likiangensis* (0.00930) and *P. wilsonii* (0.00874) have similar levels, whereas *P. schrenkiana* is nearly four times less variable (0.00258). A similar pattern is observed for the total nucleotide diversity (π_T) (fig. 2 and supplementary table S4, Supplementary Material online).

LD and recombination

Among the 1,814 pairwise comparisons between SNPs, 132 were significant after Bonferroni correction. Due to low levels of nucleotide variation, LD could not be estimated with confidence in *P. schrenkiana*. However, in the three remaining species, the value of r^2 averaged over all sequenced regions investigated 0.291. LD, as measured by r^2 decayed to below 0.2 within 250–500 bp (data not shown). The population recombination rate, ρ, was slightly higher in *P. likiangensis* and *P. wilsonii* than in *P. purpurea* and *P. schrenkiana*, but in all cases, the estimate was much wider than estimates of θ (table 1).

Population structure

ϕ_{ST} values across all loci within and among species are given in table 2, and more extensive results of the AMOVA within each species are given in supplementary tables S5–S7, Supplementary Material online. *Picea schrenkiana* is the most divergent species, the ϕ_{ST} values with each of the three other species being around 0.5. Among the latter, *P. wilsonii* showed the highest divergence. Overall, ϕ_{ST} values among populations within species were generally high (*P. schrenkiana*: 0.222, *P. likiangensis*: 0.158, and *P. wilsonii*: 0.169), except in *P. purpurea*, where covariation was mostly found within populations (−0.004).

We also used a Bayesian clustering algorithm, Structure v. 2.3, on the entire data set. The most likely number of clusters was $K = 3$ (average LnPD $= -1475$ when two outlier runs were removed and LnPD $= -1519$ otherwise) when the original method from Pritchard et al. (2000) was used. The difference with $K = 4$ and $K = 5$ was, however, minimal (LnPD $= -1515$ and LnPD $= -1516$, respectively) (fig. 3). The most likely number of clusters was $K = 2$ when we instead used the ΔK statistics given in Evanno et al. (2005) (data not shown). The latter primarily reflects the fact that a single cluster is much more unlikely than $K \geqslant 2$ but could also suggest a hierarchical structure of the data. Figure 4 gives the Structure v. 2.3 clustering results for $K = 2$–5. For $K = 2$, the first cluster is made of *P. schrenkiana* and *P. likiangensis* individuals. Both species also

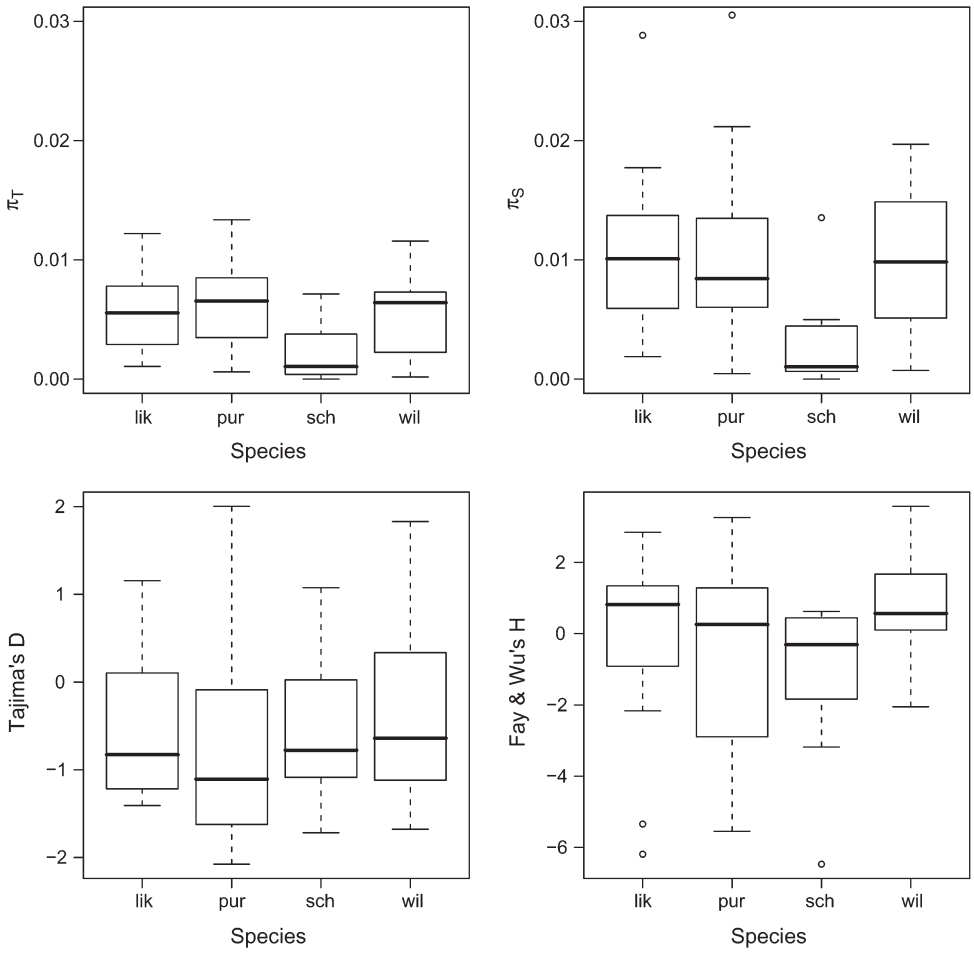

FIG. 2. Box plots of the summary statistics for *Picea likiangensis* (lik), *P. purpurea* (pur), *P. schrenkiana* (sch), and *P. wilsonii* (wil). The top row shows nucleotide diversity in the four species, with total diversity (π_T) and synonymous nucleotide diversity (π_s) in the left and right boxes, respectively. The bottom row shows Tajima's *D* and Fay and Wu's *H* for each of the species. Bars represent the median, the bottom and top of the boxes represent the 25% and 75% percentiles, respectively, and whiskers extend out to 1.5 times the interquartile range. Dots are outliers.

have a contribution from the second cluster, which is dominated by *P. wilsonii*. This contribution is generally very small in *P. likiangensis* and could reflect rare and recent introgression events. All *P. purpurea* individuals are admixed with a major contribution from the *P. schrenkiana*–*P. likiangensis* cluster and a minor one from the *P. wilsonii* cluster. For $K = 3$, *P. schrenkiana* and *P. wilsonii* constitute two "pure" clusters, whereas *P. likiangensis* individuals' ancestry traces back to a third cluster as well as, though to a much smaller extent, to the *P. schrenkiana* and *P. wilsonii* clusters. The genome of *P. purpurea* individuals traces back to the three clusters with major and almost equal contributions of the *P. schrenkiana* and *P. likiangensis* clusters and a smaller contribution from the *P. wilsonii* cluster. This pattern, together with the clustering obtained for $K = 4$ and $K = 5$, supports a complex origin of this species.

Table 1. Distribution of the Population Mutation Rate, θ, and the Population Recombination Rate, ρ, in *Picea likiangensis*, *P. purpurea*, *P. schrenkiana*, and *P. wilsonii* Estimated with Seqlib.

Species	θ/ρ	Average	MAP	2.5%	97.5%
P. lik.	θ	0.00649	0.0066	0.00514	0.0078
	ρ	0.00578	0.0066	0.00250	0.0111
P. pur.	θ	0.02445	0.0266	0.01377	0.0376
	ρ	0.00050	0.0000	0.00003	0.0139
P. sch.	θ	0.00272	0.0000	0.01593	0.0045
	ρ	0.00088	0.0000	0.00012	0.0025
P. wil.	θ	0.00990	0.0066	0.05502	0.0155
	ρ	0.00203	0.0000	0.00010	0.0055

NOTES.—In each case, estimates were obtained under the most likely demographic model (see text). MAP is the maximum a posteriori estimation and is the mode of the posterior distribution. The MAP is computed by seqlib as the midpoint of the most likely category in the discretized posterior distribution. There were 16 discrete classes. 2.5% and 97.5% are the 2.5% and 97.5% quantiles. *P. lik.*, *P. likiangensis*; *P. pur.*, *P. purpurea*; *P. sch.*, *P. schrenkiana*; and *P. wil.*, *P. wilsonii*.

Table 2. Φ_{ST} Values over All Loci among Populations within Each Species (Diagonal) and among Species (Lower Part).

	P. lik.	*P. pur.*	*P. sch.*	*P. wil.*
P. lik.	0.158***	—	—	—
P. pur.	0.069***	−0.004	—	—
P. sch.	0.529***	0.526***	0.222***	—
P. wil.	0.165***	0.105***	0.513***	0.169*

NOTES.—*P. lik.*, *Picea likiangensis*; *P. pur.*, *P. purpurea*; *P. sch.*, *P. schrenkiana*; *and P. wil.*, *P. wilsonii*.
$^{*}P < 0.05$, $^{***}P < 0.001$.

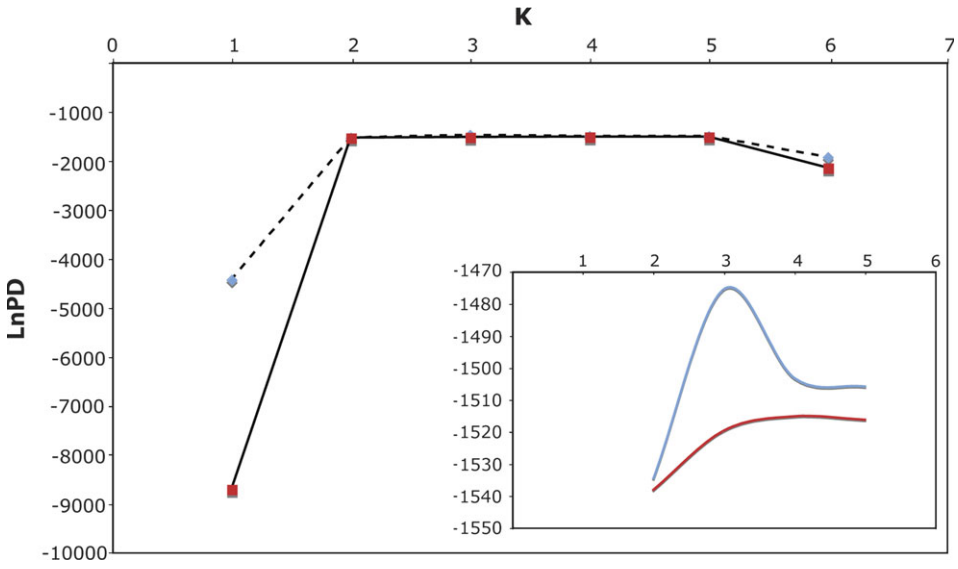

FIG. 3. Estimated number of clusters (K) obtained with Structure. The mean of LnPD is plotted over K (1–6) when 50 runs are considered (dotted line) and when the two most extreme values are removed (black line). The inset is a magnification of the area K =2–5 showing that $K = 3$ is more likely when the most two extreme values are removed.

Tests of the SNM and demographic inferences

An HKA test (Hudson et al. 1987) was performed to test departure from neutrality at individual loci. *Picea likiangensis*, *P. purpurea*, and *P. wilsonii* were each tested in turn with *P. schrenkiana* as an outgroup to ascertain whether the observed polymorphisms within species and the divergence between species deviated significantly from what would be expected under the SNM. In each case, the HKA test suggests no deviation from the SNM at any of the individual loci (data not shown).

The mean Tajima's D was negative for all four species with values ranging from −0.38 in *P. wilsonii* to −0.80 in *P. purpurea*, whereas the mean Fay and Wu's H was negative in *P. schrenkiana* (−1.19) and *P. purpurea* (−0.68), close to zero in *P. likiangensis* (−0.07), and positive in *P. wilsonii* (0.80) (fig. 2 and supplementary table S8, Supplementary Material online). Few individual loci departed significantly from the SNM, but most had negative Tajima's D values. Interestingly though, locus 2009 had positive values of both D and H in all species except *P. schrenkiana*, and the SNM could be rejected for Gigantea (GI) in one species (*P. purpurea*). Negative average values of both Tajima's D and Fay and Wu's H reveal the presence of skews toward both low-frequency variants (negative D) and high

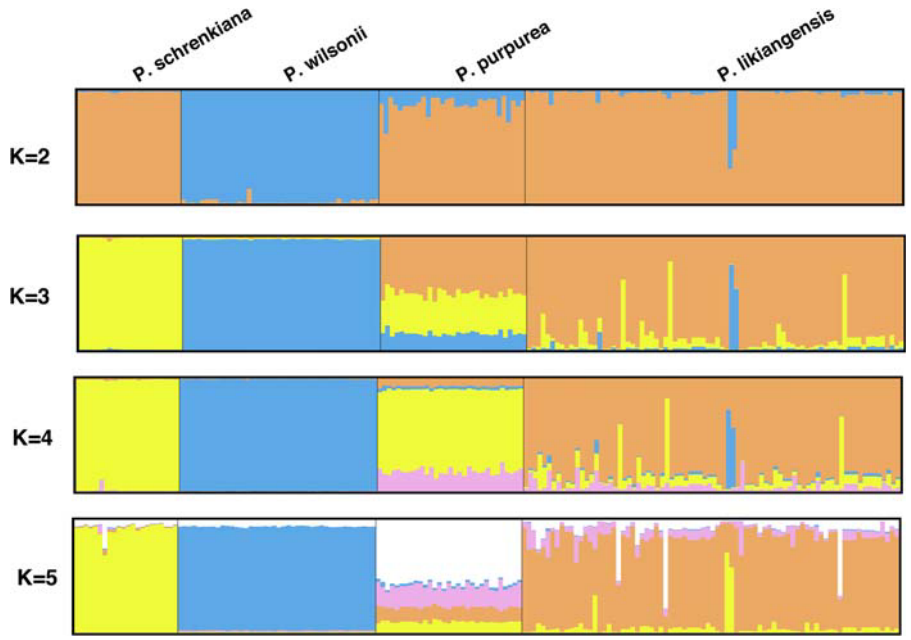

FIG. 4. Structure analysis of the four species when K =2–5 clusters are assumed. For each K value, results of the run with the highest value of LnPD were used. Variation among runs was limited.

Table 3. Bayes Factors of the Three Demographic Models.

	P. lik.	*P. pur.*	*P. sch.*	*P. wil.*
SNM	1	1	1	1
PEM	0.994	1.213	1.177	1.072
BNM	0.917	1.207	1.352	0.920

NOTES.—In each case, the demographic model was compared with the SNM. . *P. lik.*, *Picea likiangensis*; *P. pur.*, *P. purpurea*; *P. sch.*, *P. schrenkiana*; and *P. wil.*, *P. wilsonii*.

frequency–derived variants (negative H). This has generally been shown to reflect the presence of a relatively ancient bottleneck (Haddrill et al. 2005; Heuertz et al. 2006; Pyhäjärvi et al. 2007; Ingvarsson 2008). This could apply to *P. purpurea* and *P. schrenkiana*. On the other hand, *P. wilsonii* might have experienced an even more ancient bottleneck because for older bottlenecks, Fay and Wu's H would become more positive as new mutations accumulate. Finally, *P. likiangensis* did not show any skew toward high frequency–derived variants ($H \approx 0$) but exhibited a rather strong skew toward low-frequency variants, a pattern that could simply reflect recent population growth.

Supplementary table S9, Supplementary Material online, shows a comparison of different sampling schemes for a number of summary statistics. Values of Tajima's D for *P. likiangensis*, *P. purpurea*, and *P. wilsonii* under the pooled sampling scheme were more negative than under the scattered scheme, whereas for *P. schrenkiana*, the opposite trend was observed. For all species, however, the difference in Tajima's D between each of the sampling schemes was small, ranging from 0.044 to 0.075. It should be noted that, due to there being a smaller number of populations in some of the studied species, the sampling size was often small, and this may have limited the effect of the different sampling schemes.

In order to determine which simple demographic model best characterizes each of the species, we used an approximate Bayesian computation approach. We performed a two-step analysis, and the acceptance rate under the second round of analysis, where values were sampled from the posterior distribution of the first round, was used as an estimate of the posterior probability of the different models. In general, the posterior probabilities of the different models for a given species were quite similar leading to Bayes factors close to 1 for the different models. The SNM had the highest Bayes factor in *P. likiangensis*, but the difference with the growth model was small. Similarly, in *P. purpurea*, the Bayes factor of the population growth model was only marginally higher than that of the BNM. In *P. wilsonii*, the Bayes factor of the SNM was slightly lower than that of the Bayes factor of the growth model, but the estimate of the growth rate was close to zero and the simplest SNM may thus be preferred. Finally, in *P. schrenkiana*, the BNM had the highest Bayes factor (table 3). So in summary, the two major species *P. likiangensis* and *P. wilsonii* do not depart significantly from the neutral model; *P. schrenkiana* shows evidence of a bottleneck and *P. purpurea* of population growth. Under the selected model for each species, most parameters had obvious

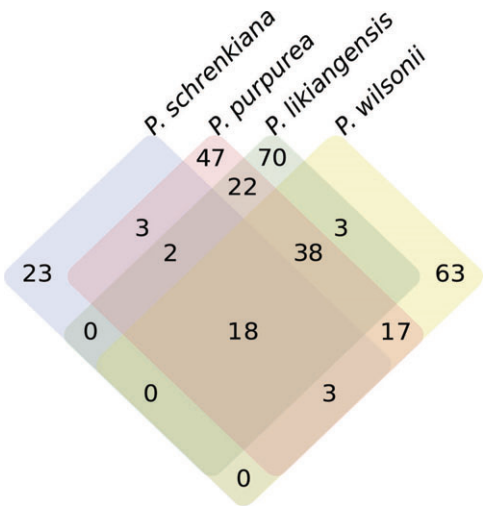

FIG. 5. Venn diagram representation of polymorphisms that are either shared among different species or private to a single species.

modes, suggesting that the data are informative for the parameters of interest (supplementary file 1, Supplementary Material online).

IM models

Shared polymorphism among species is extensive, whereas fixed sites are rare (fig. 5 and supplementary table S10, Supplementary Material online). Because of the absence of fixed sites among the three species of the QTP, IM analyses were restricted to pairs including *P. schrenkiana*. Estimates of divergence times, effective population sizes, and migration rates are given in table 3 (the posterior distributions are given in supplementary table S11 and supplementary figs. S1–S3, Supplementary Material online). These estimates are based on a mutation rate of $\mu = 1.0 \times 10^{-8}$ per site per generation and an average generation time of 50 years (Bousquet and Bouillé 2005; Chen et al. 2010). The mutation rate value used here is, of course, an approximate estimate but is close to the best estimates reported so far for conifer species (Willyard et al. 2007). In any case, we are here more interested in the relative values of the estimates of divergence times and effective population sizes, so possibly the main assumption is that the same mutation rate applies to all four species. The highest divergence time was between *P. likiangensis* and *P. schrenkiana* (10.9, 90% confidence interval [CI]: 5.0–15.4) and the lowest between *P. purpurea* and *P. schrenkiana* (3.2, 90% CI: 2.3–8.7) (fig. 6). Interestingly, the divergence time of *P. purpurea* was not intermediate between the respective divergence times of *P. likiangensis* and *P. wilsonii* (5.1, 90% CI: 3.9–23.61) from *P. schrenkiana* as one might have expected if *P. purpurea* was a simple hybrid between these two species. Although the CIs are broad, the distributions of the divergence times showed a clear mode (supplementary figs. S1–S3, Supplementary Material online). The inferred migration rates were symmetrical between *P. purpurea* and *P. schrenkiana*, whereas migration rates were asymmetrical in the two other pairs. In both cases, the migration rates from *P. schrenkiana*

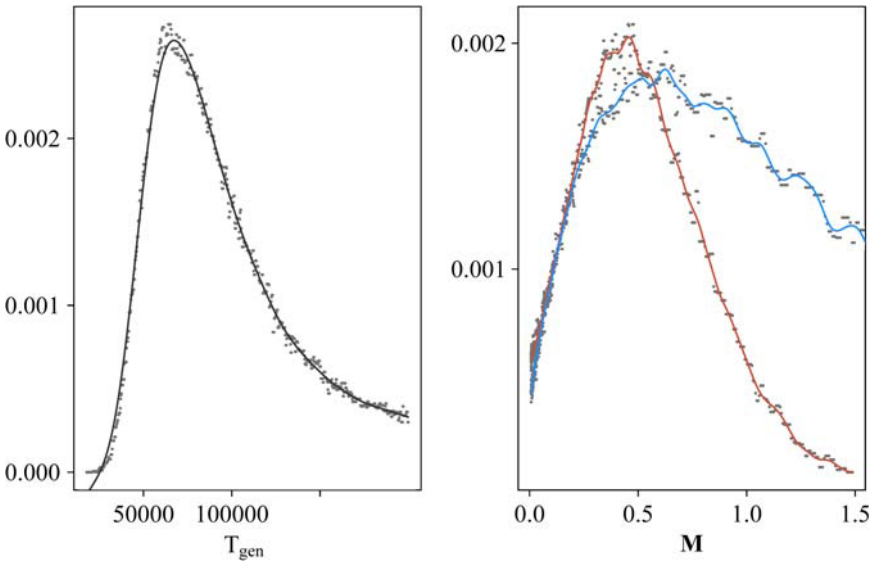

FIG. 6. Posterior distributions for the comparison between *Picea purpurea* and *P. schrenkiana* of the divergence time in generations (T_{gen}) and the migration rate (M) given by the IM model as implemented in Mimar. The red curve shows the migration rate from *P. purpurea* to *P. schrenkiana* (M_{12}), whereas the blue curve shows the migration rate from *P. schrenkiana* to *P. purpurea* (M_{21}).

were much higher than the migration rate to *P. schrenkiana* (table 4). *Picea schrenkiana* thus appears as either a source or closely related to a source population of both *P. wilsonii* and *P. likiangensis*.

The goodness of fit of the results was tested using the program MimarGOF. For all three comparisons, the distributions of the Wakeley and Hey (1997) S statistics, F_{ST} between species, and π for the two descendent species provided a good fit for the observed statistics. For Tajima's D, the distributions simulated with MimarGOF were less negative than the observed values in all comparisons (supplementary figs. S4–S6, Supplementary Material online), suggesting that departures from the SNM have not been fully accounted for.

Discussion

The present study confirms and extends the general picture of conifer evolution that emerged from recent studies: Spruces (Chen et al. 2010), as well as pines (Willyard et al. 2009), are characterized by large effective population sizes, incomplete lineage sorting, hybridization, and unexpectedly short divergence times. The large number of shared polymorphisms and the low number of fixed differences are perhaps not surprising because, under a simple allopatric model of speciation with no selection, it will take approximately $9 - 12N_e$ generations for the genealogies of $> 95\%$ of the loci to be reciprocally monophyletic (Hudson and Coyne 2002). Because many forest tree species have large effective population sizes and long generations, long separation times in calendar years will translate into rather short times on an evolutionary timescale and shared polymorphisms are thereby expected, even in the absence of gene flow among species. Yet, ecological and phylogeographical studies (e.g., Palme et al. 2004) have often tended to interpret allele sharing almost solely in terms of introgression rather than in terms of incomplete lineage sorting without actually attempting to tell these two processes apart. Although, in some cases, the current nonrandom geographical distribution of shared alleles may indeed suggest that introgression is a more parsimonious hypothesis than incomplete lineage sorting, our current results, as well as those obtained in pines (Willyard et al. 2009), indicate that one needs to be cautious and that assuming introgression is the sole cause of allele sharing may lead to overestimates of its importance. At least in many tree species, incomplete lineage sorting seems to have played an important role and may well be the major source, though likely not the only one, of allele sharing across species. Below, we will first put the results of the present study into perspective and then discuss their implications for the species history.

Table 4. Estimates of the IM Model Parameters.

	N_1	N_2	N_A	T_{My}	M_{12}	M_{21}
P. lik./P. sch.	**155**	**33**	**10.0**	**10.9**	**0.33**	**0.94**
P. pur./P. sch.	**206**	**32**	**151**	**3.2**	**0.48**	**0.50**
P. wil./P. sch.	**140**	**20**	**178**	**5.1**	**0.06**	**2.12**

NOTES.—N_i is the effective population size of the i^{th} species in thousands of individuals, T_{My} is the time since divergence in million of years, and M_{12} is the migration rate from species 1 to species 2 forward in time and M_{21} is the migration rate from species 2 to species 1. The generation time was assumed to be 50 years and the mutation rate was assumed to be 1×10^{-8} per site per generation. Posterior distributions are given in supplementary figs. S2–S4, Supplementary Material online. *P. lik.*, *Picea likiangensis*, *P. pur.*, *P. purpurea*, *P. wil.*, *P. wilsonii*; and *P. sch.*, *P. schrenkiana*.

Nucleotide diversity, LD, and species demography

Silent nucleotide diversity for *P. schrenkiana* (π_s = 0.00258) was much lower than in the other three species (*P. wilsonii*: π_s = 0.00874, *P. purpurea*: π_s = 0.00996, and *P. likiangensis*: π_s = 0.00930) but similar to that observed in *P. breweriana* (π_s = 0.00200), another spruce species with a

small distribution range (Chen et al. 2010). In contrast, the level of nucleotide diversity observed in the three species from the southern edge of the QTP is high compared with that obtained in other conifers (Savolainen and Pyhäjärvi 2007). For example, their silent nucleotide diversity is a little higher than estimates in *P. taeda* (π_s = 0.0064, Brown et al. 2004; π_s = 0.0079, Gonzalez-Martinez et al. 2006) and higher than in *P. abies* (π_s = 0.0039, Heuertz et al. 2006) and *P. mariana* (π_s = 0.00461), but similar to that observed in *P. glauca* (π_s = 0.00905, Chen et al. 2010), three boreal spruce species with continent-wide distributions. Levels of LD in the three species, where LD could be estimated, were somewhat higher than that observed in *P. abies* ($r^2 = 0.115$ and decaying to 0.2 within $\sim$ 100 bp) (Heuertz et al. 2006), but it should be noted that data were pooled over populations and so estimates of LD could be inflated. Despite this possible upward bias of estimates of LD, high estimates of the population recombination rate ρ in all species further indicate the importance of recombination in shaping diversity in conifers.

The high silent nucleotide diversity in the three QTP species seems in apparent conflict with their relatively restricted distributions (fig. 1) but agrees well with demographic inferences. First, in contrast to boreal species that have often gone through severe ancient bottlenecks, we did not detect any departure from the SNM in *P. likiangensis* and *P. wilsonii*, and in the case of *P. purpurea*, there was evidence of population growth. Their genetic stability suggests that the glacial climatic oscillations had little effect on their species range and rarely reduced their total diversity (Peng et al. 2007). Second, the high diversity observed in these species could be due to past gene exchange among species, which is reflected by the detection of ancient gene flow between each of the three species and *P. schrenkiana*. *Populus tremula*, which also shows a high level of polymorphism ($\pi_s = 0.0160$, Ingvarsson 2005; $\pi_s = 0.0120$, Ingvarsson 2008), might be another case where high polymorphism is due to hybridization with other species (Lexer et al. 2005; Ingvarsson 2008). In addition, we cannot also exclude the possibility of a genetic contribution from other local species that were not included in the present study, which might for instance be reflected in the unexpectedly large amount of unique polymorphisms present in *P. purpurea*, although this could also be caused by population expansion. Third, the relatively pronounced population genetic structure observed in *P. wilsonii* and *P. likiangensis* will also tend to increase their effective population size because, assuming a Wright's island model, the effective population size is $\frac{n_d N}{1-F_{ST}}$, where n_d is the number of demes and N the size of each deme (Wright 1943; Rousset 2003). The assumption of a Wright's island model, rather than a metapopulation model, may not be unreasonable here if the populations have indeed been fairly stable through time.

Genetic differentiation

Spruce species are wind pollinated with outcrossing rates being generally high (Bousquet and Bouillé 2005) and this should result in a low level of population differentiation. This is generally observed in conifers, although certain species with more fragmented and restricted distributions tend to show a higher level of population differentiation (e.g., Boyle and Morgenstern 1987; Lagercrantz and Ryman 1990; Furnier et al. 1991; Isabel et al. 1995; Ledig et al. 1997, 2000; Ledig 2000; Wang et al. 2005; Aizawa et al. 2007; Peng et al. 2007, and references therein). With the exception of *P. purpurea*, the overall ϕ_{ST} within species was high (*P. schrenkiana*, ϕ_{ST} = 0.222; *P. wilsonii*, ϕ_{ST} = 0.169; and *P. likiangensis*, ϕ_{ST} = 0.158). This high level of differentiation could, at least partly, be attributed to their restricted and fragmented distribution. The high level of differentiation in *P. schrenkiana* is similar to the value previously reported by Goncharenko et al. (1992) for isozyme loci. The high level of population differentiation for *P. wilsonii* and *P. likiangensis*, both of which grow mainly in mountain ranges along the southeastern edge of the QTP, is also consistent with previous studies on conifers from the same region, such as *P. asperata* Mast. (Wang et al. 2005), *Pinus densata* (Ma et al. 2006), and *P. likiangensis* (Peng et al. 2007). Presumably, the high population differentiation results from the complex topography of their habitats, high mountains, and deep valleys preventing gene flow among populations (Wang et al. 2005; Peng et al. 2007). Furthermore, the complex topography of the area might also have provided multiple isolated refuges during glacial stages (Tang and Shen 1996; Peng et al. 2007), thereby reinforcing their population genetics differentiation (Petit et al. 2003; Peng et al. 2007). Finally, the relatively lower level of differentiation observed in *P. purpurea* could simply reflect the recent origin and rapid growth of that species.

Species histories

The abundance and distribution of plant species on the Tibetan Plateau and neighboring areas have been closely associated with changes in temperature and humidity, which in turn were, at least partly, a reflection of rapid shifts in elevation. Palynological studies suggest that conifers were present in the region some 50 Ma but that spruce species only started to be an important fraction of the pollen record on the QTP around 38 Ma (Wang et al. 1990; Dupont-Nivet et al. 2008). Spruce species remained common until c. 20 Ma at which time their contribution to the pollen record started to decline and stayed low until 17 Ma. The ensuing years are unfortunately poorly documented, but the continuing elevation and climatic changes through which the QTP went during this period certainly led to further fluctuations in abundance. In particular, the significant uplift of southern Tibet some 7 Ma inferred by some authors (e.g., Wang et al. 2006, 2008) would certainly have affected spruce distribution.

The combined information from morphology, nuclear DNA, and cytoplasmic markers of the group of spruce species investigated in the present study indeed suggests a complex history over that period of time. As we noted earlier on, *P. schrenkiana* and *P. wilsonii*, on the one hand, and *P. purpurea* and *P. likiangensis*, on the other hand, are assigned to two different groups based on morphological

criteria (Farjón 1990; Fu et al. 1999). This may, at first glance, seem to conflict with the clustering based on cytoplasmic and nuclear markers: *P. purpurea* and *P. wilsonii* show extensive haplotype sharing for both cpDNA and mtDNA, whereas *P. likiangensis* and *P. schrenkiana* belong to a separate clade (cpDNA; Ran et al. 2006) (Liu J., unpublished results). Nuclear markers indicate yet another grouping: independently from the number of clusters assumed in the Structure analysis, *P. wilsonii* is clearly separated from a *P. schrenkiana–P likiangensis* group, whereas *P. purpurea* appears as a hybrid between these two main groups. Haplotype sharing for cytoplasmic markers could be explained by the following model: individuals from the *P. schrenkiana–P. likiangensis* group moved northward and met *P. wilsonii* in what is today the natural range of *P. purpurea*. Introgression of genes from *P. wilsonii*, the local species, to individuals from the *P. schrenkiana–P. likiangensis* group, the incoming one, ensued until the introgressed individuals became more frequent than the *P. wilsonii* individuals (Currat et al. 2008). Eventually, the resulting introgressed individuals carried the same chloroplast and mitochondrial markers as *P. wilsonii*. Under the Currat et al. (2008) model, the introgression of organelle genomes is expected to be more important than that of nuclear genes, simply because the population differentiation at organellar genes within each species is higher than the differentiation at nuclear genes. This relationship between population differentiation and levels of introgression is indeed observed in *Picea* species of the QTP (Du et al. 2009).

Under the scenario outlined above the nuclear genome of the introgressed individuals will be admixed, which is certainly the case, though not exactly in the way expected under the simple introgression model between *P. likiangensis* and *P. wilsonii* initially envisioned. The distribution of nuclear DNA variation and the IM analyses indeed suggest a somewhat different scenario. In the Structure analysis, individuals from *P. purpurea* assign with different probabilities to the three major clusters when $K = 3$, suggesting that the *P. purpurea* genome is a combination of the genomes of *P. likiangensis*, *P. wilsonii*, and *P. schrenkiana* or more likely an ancestor or a close relative of both *P. likiangensis* and *P. schrenkiana*. However, the contributions of *P. wilsonii* and *P. likiangensis* to the *P. purpurea* genome are already low when $K = 4$ and becomes extremely limited for $K = 5$. In contrast, the contribution of *P. schrenkiana* becomes dominant. Hence, at the nuclear level, *P. purpurea* does not look like a simple hybrid between *P. likiangensis* and *P. wilsonii*. To understand better this result, and in particular the timescale on which these events unfolded, we turn to the IM model analyses.

Contrary to our expectations, with an estimated divergence time of 3.2 million years (My), *P. purpurea* diverged from *P. schrenkiana* later than *P. wilsonii* (5.1 My) and considerably later than *P. likiangensis* (10.9 My). These divergence times together with the results of the Structure analysis suggest that *P. purpurea* primarily stemmed from *P. schrenkiana* or perhaps more likely from a common ancestor and acquired its current nuclear and cytoplasmic genome composition during secondary contacts with *P. wilsonii* and *P. likiangensis*. Interestingly, cpDNA markers show that *P. purpurea* belongs to the same clade as other high altitude species scattered along the Himalayas: *P. spinulosa*, *P. smithiana*, and *P. farreri* (Ran et al. 2006). The divergence time estimates of *P. likiangensis* and *P. purpurea* from *P. schrenkiana* suggests a scenario where these species that are today scattered along the mountain ridges of the QTP diverged sequentially with the progressive uplifts of these areas. In this respect, it will be crucial to include *P. spinulosa*, *P. smithiana*, and *P. farreri* in further studies in order to test this hypothesis.

The divergence time estimates rely, of course, on unwarranted assumptions about generation time and mutation rate but, unless these parameters vary strongly among species, the assumptions should not affect the sequence of events too strongly. Perhaps of more concern are departures from the assumptions of the IM model implemented in Mimar. First, the pair of species under analysis is assumed to be more closely related than any of them is to a third species. In other words, one assumes that there has not been gene flow with a third species. This is obviously a problem when carrying out Mimar analyses on a group of species as done in this study and as was also done in the first application of the program (Becquet and Przeworski 2007). We cannot also rule out the possibility of gene flow from species not considered in the present study, such as *P. spinulosa* and *P. smithiana*. The effect of gene flow from a third species will likely increase the divergence time among the pair under study if the species is a distant relative, but it could also have the opposite effect if the species is intermediate between the two species under consideration. In the present case, gene flow into *P. schrenkiana* from species situated north of its range would likely increase the divergence time, whereas gene flow between species of the QTP would give the opposite effect. Second, Mimar assumes a lack of population structure in each species and in the ancestral one. This assumption is not fulfilled in most of the species studied here as they show significant population structure, and the same was likely true for the ancestral population size. Population structure in the ancestral species will lead to an overestimate of the ancestral effective population size, whereas estimates of divergence time will be biased downward or upward depending on parameters such as the length of the period under which the ancestral population was structured (Becquet and Przeworski 2009). The effect of population structure in the descendent populations is similarly difficult to predict. Population structure will make genealogies within species both longer on average and more variable, but the ensuing effect on divergence time is less straightforward to predict. In any case, the timescale suggested by the divergence times obtained here corresponds roughly to the time of the last uplift of the southern part of the QTP (Wang et al. 2006, 2008) and indicates a very recent origin for the species. More than anything else, this recent origin, combined with gene flow among species, could be the main cause of the lack of clear delineation among species and thereby the seemingly complicated speciation process.

Conclusions

The present study sheds new light on the demographic processes that accompanied the speciation of the conifer species of the QTP (see also Ma et al. 2006). It illustrates the importance of explicitly considering ancestral shared polymorphisms when analyzing the joint history of tree species (see also Willyard et al. 2009). The conclusions on the speciation process itself that could be drawn at this stage remain, by necessity, limited as only a handful of loci were analyzed and major aspects of the speciation processes could therefore not be addressed. In particular, we have left aside questions about the nature and magnitude of the part played by natural selection in the establishment of barriers to gene flow. Under ecological speciation, reproductive isolation between populations occurs by adaptation to different environments or ecological niches. In contrast, under mutation-order speciation, reproductive isolation evolves by the chance occurrence and fixation of different alleles between populations adapting to similar selection pressures (Schluter 2009). Which mechanism prevailed under the fast changing environment of the QTP over the past 20 Ma will be an exciting question to address, but more extensive genomics and ecological data will be required.

The present study also illustrates how combining different markers and modern analysis methods can help understand complex past demographic processes. Although the resulting picture remains incomplete, the study nonetheless makes it clear that reliance on a single type of markers or a single analytical approach would have likely led to erroneous conclusions, especially when, as in the present case, the fossil record is scarce. In this respect, the suggested stability of the species range on the QTP obtained here is reminiscent of the recent paradigm shift in our understanding of the impact of the last glacial maximum (LGM), brought about by coalescent-based analysis of nuclear DNA and a reanalysis of the pollen record (Birks and Willis 2008; Lascoux et al. 2008). The "southern refugia paradigm," which stipulates that most species survived the LGM in small southern refugia, was primarily a consequence of the almost exclusive use of descriptive studies of cytoplasmic markers. Nuclear DNA studies, together with the discovery of LGM wood fossils in Central Europe (Willis et al. 2000), showed that this paradigm actually did not always apply to cold-tolerant plant species and a more nuanced picture emerged. Thanks to recent technical and conceptual advances we are no longer confined to a handful of approaches in our quest to understand the long- and short-term history of QTP species. This is most fortunate as the history of the conifer species of the QTP looks more like a tangled web of relationships than a nicely unfolding tree-like process.

Acknowledgments

The research was funded by grants from the Carl Tryggers Foundation, the Philip-Sörensen Foundation, the EVOLTREE network of excellence, and the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) to M.L. and the National Natural Science Foundation of China and Key Innovation Project of Ministry of Education of China to J.L. (grant numbers 30930072 and 30725004). T.K. acknowledges financial support from the Nilsson-Ehle Foundation, and M.L. thanks the Chinese Academy of Sciences for granting him a visiting professorship and Li Haipeng for his hospitality in Shanghai. Li Yuan was supported by the Chinese Scholarship Council. We thank Céline Becquet for kindly answering questions about Mimar, and Rémy Petit for comments on an earlier version of the manuscript. Part of this work was carried out by using the resources of the Computational Biology Service Unit from Cornell University, which is partially funded by Microsoft Corporation. We thank the associate editor and three anonymous reviewers for helpful comments.

Supplementary Material

Supplementary tables S1–S11 and figures S1–S6 and Supplementary file are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

References

Aizawa M, Yoshimaru H, Saito H, Katsuki T, Kawahara T, Kitamura K, Shi F, Kaji M. 2007. Phylogeography of a northeast Asian spruce, *Picea jezoensis*, inferred from genetic variation observed in organelle DNA markers. *Mol Ecol.* 16:3393–3405.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

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

Becquet C, Przeworski M. 2009. Learning about modes of speciation by computational approaches. *Evolution* 63:2547–2562.

Birks JHB, Willis KJ. 2008. Alpines, trees, and refugia in Europe. *Plant Ecol Divers*. 1:147–160.

Bousquet J, Bouillé M. 2005. Trans-species shared polymorphism 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.

Boyle T, Morgenstern E. 1987. Some aspects of the population-structure of Black spruce in central New-Brunswick. *Silvae Genet*. 36:53–60.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB. 2004. Nucleotide diversity and linkage disequilibrium in Loblolly pine. *Proc Natl Acad Sci USA*. 101:15255–15260.

Chen J, Källman T, Gyllenstrand N, Lascoux M. 2010. New insights on the speciation history and nucleotide diversity of three boreal spruce species and a Tertiary relict. *Heredity* 104:3–14.

Currat KM, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. *Evolution* 62: 1908–1920.

Doyle J, Doyle J. 1990. Isolation of plant DNA from plant tissue. *Focus* 12:13–15.

Du FK, Petit RJ, Liu JQ. 2009. More introgression with less gene flow: chloroplast vs. mitochondrial dna in the *Picea asperata* complex in China, and comparison with other Conifers. *Mol Ecol.* 18: 1396–1407.

Dupont-Nivet G, Hoorn C, Konert M. 2008. Tibetan uplift prior to the eocene-oligocene climate transition: evidence from pollen analysis of the Xining basine transition. *Geology* 6:987–990.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol.* 14:2611–2620.

Ewing B, Green P. 1998. Basecalling of automated sequencer traces using Phred. II error probabilities. *Genome Res*. 8:186–194.

Ewing B, Hillier L, Wendl M, Green P. 1998. Basecalling of automated sequencer traces using Phred. I accuracy assessment. *Genome Res*. 8:175–185.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online* 1:47–50.

Excoffier L, Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends Ecol Evol (Amst)*. 23:347–351.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479–491.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587.

Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* 7:574–578.

Farjón A. 1990. Pinaceae: drawings and descriptions of the genera *Abies, Cedrus, Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix* and *Picea*. Koenigstein (Germany): Koeltz Scientific Books.

Fay JC, Wu CI. 2000. Hitchhiking under positive darwinian selection. *Genetics* 155:1405–1413.

Fu L, Li N, Elias T. 1999. Flora of China. Vol. 4. Beijing (China): Beijing Science Press. Available from: http://www.efloras.org/.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709.

Furnier G, Stine M, Mohn C, Clyde M. 1991. Geographic patterns of variation in allozymes and height growth in White spruce. *Can J Forest Res*. 21:707–712.

Goncharenko GG, Potenko VV, Abdyganyev N. 1992. Variation and differentiation in natural populations of Tien-Shan spruce (*Picea schrenkiana* Fisch. et Mey). Genetika 28:83–96, translated into English as Soviet Genetics 1992.

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

Gordon D, Desmarais C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res*. 8:195–202.

Haddrill PR, 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, Paoli ED, Källman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N. 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce (*Picea abies* (l.) Karst). *Genetics* 174: 2095–2105.

Hey J. 2006. Recent advances in assessing gene flow between diverging populations and species. *Curr. Opin. Genet. Dev*. 16: 592–596.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour*. 9:1322–1332.

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

Hudson RR, Kreitman ME, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.

Ingvarsson PK. 2005. Molecular population genetics of herbivore-induced protease inhibitor genes in european aspen (*Populus tremula L*., salicaceae). *Mol Biol Evol*. 22:1802–1812.

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

Isabel N, Beaulieu J, Bousquet J. 1995. Complete congruence between gene diversity estimates derived from genotypic data at enzyme and random amplified polymorphic DNA loci in black spruce. *Proc Natl Acad Sci USA*. 92:6369–6373.

Kass RE, Raftery AE. 1995. Bayes factors. *J Am Stat Assoc*. 90:773–795.

Kumar S, Tamura K, Nei M. 2004. Mega3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform*. 5:150–163.

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

Lascoux M, Palmé AE, Cheddadi R, Latta RG. 2004. Impact of ice ages on the genetic structure of trees and shrubs. *Philos Trans R Soc Lond, B, Biol Sci*. 359:197–207.

Lascoux M, Pyhäjärvi T, Källman T, Savolainen O. 2008. Past demography in forest trees: what can we learn from nuclear DNA sequences that we do not already know? *Plant Ecol Divers*. 1:209–215.

Ledig F, Bermejo-Velazquez B, Hodgskiss P, Johnson D, Flores-Lopez C, Jacob-Cervantes V. 2000. The mating system and genic diversity in Martinez spruce, an extremely rare endemic of Mexico's sierra madre oriental: an example of facultative selfing and survival in interglacial refugia. *Can J Forest Res*. 30:1156–1164.

Ledig FT. 2000. Founder effects and the genetic structure of Coulter pine. *J Hered*. 91:307–315.

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

Lepais O, Petit RJ, Guichoux E, Lavabre JE, Alberto F, Kremer A, Gerber S. 2009. Species relative abundance and direction of introgression in oaksoaks. *Mol Ecol*. 18:2228–2242.

Lexer C, Fay MF, Joseph JA, Nica MS, Heinze B. 2005. Barrier to gene flow between two ecologically divergent Populus species, *P. alba* (white poplar) and *P. tremula* (European aspen): the role of ecology and life history in gene introgression. *Mol Ecol*. 14: 1045–1057.

Luo J, Wang Y, Korpelainen H, Li C. 2005. Allozyme variation in natural populations of *Picea asperata*. *Silva Fenn*. 39:167–176.

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. *Mol Biol Evol*. 23:807–816.

Marjoram P, Tavaré S. 2006. Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet*. 7: 759–770.

Meng L, Yang R, Abbott RJ, Miehe G, Hu T, Liu J. 2007. Mitochondrial and chloroplast phylogeography of *Picea crassifolia* kom. (pinaceae) in the Qinghai-Tibetan plateau and adjacent highlands. *Mol Ecol*. 16:4128–4137.

Mita SD, Ronfort J, McKhann HI, Poncet C, Malki RE, Bataillon T. 2007. Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signaling in *Medicago truncatula*. *Genetics* 177:2123–2133.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. *Mol Ecol*. 18:1034–1047.

Palme A, Su Q, Palsson S, Lascoux M. 2004. Extensive sharing of chloroplast haplotypes among European birches indicates hybridization among *Betula pendula, B. pubescens* and *B. nana*. *Mol Ecol*. 13: 167–178.

Peng XL, Zhao CM, Wu GL, Liu JQ. 2007. Genetic variation and phylogeographic history of *Picea likiangensis* revealed by RAPD markers. *Trees-Struct Funct*. 21:457–464.

Petit RJ, Aguinagalde I, de Beaulieu JL, et al. (17 co-authors). 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* 300:1563–1565.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

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.

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

Rosenberg N. 2004. DISTRUCT: a program for the graphical display of population structure. *Mol Ecol Notes* 4:137–138.

Rousset F. 2003. Effective size in simple metapopulation models. *Heredity* 91:107–111.

Royden LH, Burchfiel BC, van der Hilst RD. 2008. The geological evolution of the Tibetan plateau. *Science* 321:1054–1058.

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

Savolainen O, Pyhäjärvi T. 2007. Genomic diversity in forest trees. *Curr Opin Plant Biol.* 10:162–167.

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

Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P. 2009. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* 182:205–216.

Syring J, Farrell K, Businský R, Cronn R, Liston A. 2007. Widespread genealogical nonmonophyly in species of Pinus subgenus strobus. *Syst Biol.* 56:163–181.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

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

Tang LY, Shen CM. 1996. Late Cenozoic vegetation history and climatic characteristics of Qinghai-Xizang plateau. *Acta Micropaleont Sin*. 13:321–337.

Thornton KR. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.

Thornton KR. 2009. Automating approximate Bayesian computation by local linear regression. *BMC Genet.* 10:35.

Wakeley J. 2003. The evolution of population biology, chapter inferences about the structure and history of populations: coalescents and intraspecific phylogeography. New York: Cambridge University Press.

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

Wang C, Zhao X, Liu Z, Lippert PC, Graham SA, Coe RS, Yi H, Zhu L, Liu S, Li Y. 2008. Constraints on the early uplift history of the Tibetan plateau. *Proc Natl Acad Sci USA*. 105:4987–4992.

Wang DN, Sun XY, Zhao YN. 1990. Late Cretaceous to Tertiary palynofloras in Xinjiang and Qinghai, China. *Rev Palaeobot Palynol.* 65:95–104.

Wang Y, Deng T, Biasatti D. 2006. Ancient diets indicate significant uplift of southern Tibet after ca. 7 ma. *Geology* 34:309–312.

Wang Y, Luo J, Xue X, Korpelainen H, Li C. 2005. Diversity of microsatellite markers in the populations of *Picea asperata* originating from the mountains of China. *Plant Sci*. 168:707–714.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7:256–276.

Weir B, Cockerham C. 1984. Estimating f-statistics for the analysis of population-structure. *Evolution* 38:1358–1370.

Willis KJ, Rudner E, Sümegi P. 2000. The full-glacial forests of central and southeastern Europe. *Quat Res*. 53:203–213.

Willyard A, Ann W, Syring J, Gernandt DS, Liston A, Cronn R. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Mol Biol Evol.* 24:90–101.

Willyard A, Cronn R, Liston A. 2009. Reticulate evolution and incomplete lineage sorting among the Ponderosa pines. *Mol. Phylogenet. Evol.* 52:498–511.

Wright S. 1943. Isolation by distance. *Genetics* 28:114–138.

Wright S. 1951. The genetical structure of populations. *Ann Eugen*. 15:323–354.

Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174: 1431–1439.