MOLECULAR ECOLOGY

Molecular Ecology (2012) doi: 10.1111/j.1365-294X.2012.05712.x

Demography and speciation history of the homoploid hybrid pine *Pinus densata* on the Tibetan Plateau

JIE GAO,*†[1] BAOSHENG WANG,‡[1] JIAN-FENG MAO,* PÄR INGVARSSON,‡ QING-YIN ZENG* and XIAO-RU WANG*‡

*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China, †Graduate School, Chinese Academy of Sciences, Beijing 100049, China, ‡Department of Ecology and Environmental Science, UPSC, Umeå University, SE-901 87 Umeå, Sweden

Abstract

***Pinus densata* is an ecologically successful homoploid hybrid that inhabits vast areas of heterogeneous terrain on the south-eastern Tibetan Plateau as a result of multiple waves of colonization. Its region of origin, route of colonization onto the plateau and the directions of introgression with its parental species have previously been defined, but little is known about the isolation and divergence history of its populations. In this study, we surveyed nucleotide polymorphism over eight nuclear loci in 19 representative populations of *P. densata* and its parental species. Using this information and coalescence simulations, we assessed the historical changes in its population size, gene flow and divergence in time and space. The results indicate a late Miocene origin for *P. densata* associated with the recent uplift of south-eastern Tibet. The subsequent differentiation between geographical regions of this species began in the late Pliocene and was induced by regional topographical changes and Pleistocene glaciations. The ancestral *P. densata* population had a large effective population size but the central and western populations were established by limited founders, suggesting that there were severe bottlenecks during the westward migration out of the ancestral hybrid zone. After separating from their ancestral populations, population expansion occurred in all geographical regions especially in the western range. Gene flow in *P. densata* was restricted to geographically neighbouring populations, resulting in significant differentiation between regional groups. The new information on the divergence and demographic history of *P. densata* reported herein enhances our understanding of its speciation process on the Tibetan Plateau.**

Keywords: coalescent simulation, effective population size, gene flow, hybrid speciation, isolation history, nucleotide diversity

Received 19 March 2012; revision received 4 June 2012; accepted 6 June 2012

Introduction

Understanding the process of divergence (and ultimately, speciation) requires quantitative characterization of pivotal evolutionary parameters such as current and historical effective population sizes, the timing of divergence events and the rate of gene flow between populations or sister species. This information is particularly important for understanding the process of homoploid hybrid speciation (HHS), a form of hybrid speciation that generates derivative hybrid species with chromosome numbers unchanged from the parents (Grant 1981; Arnold 1997). Unlike hybrid speciation accompanied by ploidy changes, which can cause near-instantaneous reproductive isolation of the incipient hybrids from their progenitors, reproductive isolation in HHS is often facilitated by ecological selection, which promotes the adaptation of the hybrid to a novel habitat (Lewontin & Birch 1966; Buerkle *et al.* 2000; Gross & Rieseberg 2005; Abbott *et al.* 2010; Mao & Wang 2011).

Correspondence: Dr Xiao-Ru Wang, Fax: +46 90 7866705; E-mail: xiao-ru.wang@emg.umu.se

[1]These authors contributed equally to this work.

Such extrinsic isolating barriers are often incomplete, especially for outcrossing plants, and may evolve during episodes of environmental change. When dealing with ancient HHS events in species with complex extant population structures, the speciation process can be better understood by dating the divergence among derived populations and estimating demographic scenarios. However, it can be difficult to make inferences about past evolutionary events in cases where the investigated species evolved under different levels of connectivity with the parental populations. Gene flow between populations, including local introgressions between the hybrid and parental species, and secondary contacts between diverged populations can quickly erase the historical genetic signatures of initial speciation events in haploid cytoplasmic genomes and create complex patterns of genetic diversity in hybrid populations (Arnold 1993; Gross *et al.* 2003).

Reliably estimating population demographic parameters is typically considered challenging because the observed patterns of genetic diversity and differentiation are generally at least approximately consistent with a wide range of divergence times and gene flows (Slatkin & Maddison 1989), and even in the absence of gene flow, estimates of divergence time can be confounded by changes in effective population sizes (Edwards & Beerli 2000). Isolation-with-migration (IM) models (Nielsen & Wakeley 2001; Hey & Nielsen 2004; Hey 2010a) have been developed for conducting quantitative assessments of demographic parameters to study the process of species divergence. With multilocus data, these methods can be used to estimate multiple parameters simultaneously in a way that is informed by genealogies from two (or more) closely related populations and can provide insights into historical gene flow processes involving ancestral populations (Hey 2010a; Dixon *et al.* 2011). However, the likelihood-based IM methods are limited by the assumption that the sizes of the ancestral and descendant populations remain constant. As a result, they cannot be used to determine whether population sizes have changed or how descendant populations were founded (Hey 2005). An alternative approach that can potentially be used to fit very complex models, including changes in population size (Beaumont *et al.* 2002), is approximate Bayesian computation (ABC). ABC involves simulating large numbers of data sets using parameters drawn from a prior distribution. Each simulated data set is then compared to the observed data using summary statistics. Parameter samples that yield simulated data whose summary statistics are in good agreement with those for the experimental observations are accepted, and the accepted parameters jointly approximate the posterior. Because ABC methods are based on comparing summary statistics for simulated and observed data, they are less accurate than methods using the likelihood of the complete data (Beaumont *et al.* 2002). In situations involving multiple populations where many parameters need to be considered, it is not obvious how many summary statistics will be required to accurately capture complex demographic processes such as divergence, gene flow and population size changes (Beaumont *et al.* 2002). In addition, ABC methods are very computation-intensive and suffer from the curse of dimensionality as model complexity increases (Blum & François 2010). ABC methods are thus more useful for testing demographic hypotheses in a single gene pool whose population demography can be summarized by a relative small number of summary statistics, especially when the available sequence information is not extensive. In the work reported herein, we used likehood-based IM methods to investigate the history of isolation and migration between populations of the hybrid pine *Pinus densata* and subsequently used an ABC method to infer the demographic characteristics of each descendant population.

Pinus densata is the dominant pine in south-eastern (SE) Tibet. Genetic analyses suggest that *P. densata* is a homoploid hybrid between two other Asian pine species, *Pinus tabuliformis* and *Pinus yunnanensis* (Wang & Szmidt 1990, 1994; Wang *et al.* 1990, 2001; Liu *et al.* 2003; Song *et al.* 2003). The distribution of the three pine species forms a geographical succession, with *P. tabuliformis*, *P. densata* and *P. yunnanensis* generally being found in northerly, intermediate and southerly latitudes, respectively (Mao & Wang 2011). *Pinus tabuliformis* is widely distributed across northern and central China, *P. yunnanensis* has a relatively limited range in south-western China and *P. densata* forms extensive pure forests that regenerate well on the SE Tibetan Plateau at elevations ranging from 2700 to 4200 m above sea level (Wu 1956; Mao *et al.* 2009). Ecological niche modelling suggests that *P. densata* could potentially inhabit a vast landscape and that it may be the most successful homoploid hybrid plant species in terms of its known geographical scale of establishment (Mao & Wang 2011). Our recent analysis based on maternally inherited mitochondrial (mt) and paternally inherited chloroplast (cp) DNA sequence variation revealed that the extensive distribution of *P. densata* on the plateau arose from multiple waves of colonization out of an ancient hybrid zone located in the north-east periphery of its current distribution range (Wang *et al.* 2011). These waves of westward colonization created a strong geographical structure in the distribution of mtDNA sequences and established three clusters of populations along the route of colonization, each of which has a unique and nearly fixed mitotype composition. In addition, analysis of cpDNA variation revealed that the

direction and intensity of introgressions from the two parental species varied among these geographical regions: the primary pollen inflow to *P. densata* trees in the north-east of the species' range comes from *P. tabuliformis*, whereas trees in the central region are more strongly affected by *P. yunnanensis* and those at the western edge of its range have been isolated from both parent species for a long time (Wang *et al.* 2011).

The strong geographical structuring of the genetic diversity of *P. densata* implies distinct divergence between populations established by the different colonization events over the large and heterogeneous landscapes of SE Tibet. Reconstructing the species' divergence history will shed light on the ecology and genetics of HHS in *P. densata*. It is expected that coalescence analyses of nucleotide polymorphisms in nuclear genes will provide further insights into the evolutionary forces that have shaped the genetic structure in *P. densata* and will make it possible to examine the species' historical speciation and divergence processes. In a previous study, we analysed nucleotide variation at seven loci in five populations of *P. densata* (Ma *et al.* 2006) to assess their allelic histories and general pattern of nucleotide diversity. However, our current understanding of the colonization history of the species indicates that the previous study did not consider the western range of the species at all and that the central region was inadequately represented. Our knowledge of the species' divergence history across its range is thus incomplete. In the study reported herein, we sampled populations of *P. densata* that are representative of different colonization events and surveyed nucleotide polymorphism over eight nuclear loci (that were not analysed in our previous work) in 19 populations of *P. densata* and its parental species. We used the IM and ABC methods to understand the divergence process of *P. densata* during its origin and subsequent colonization of vast territories. Our specific objectives were (i) to establish time frames for the origin and subsequent divergence of *P. densata* lineages, (ii) to determine how the size of the *P. densata* population varied throughout its speciation history, and (iii) to characterize the patterns of introgression from parental species and the gene flow between gene pools of *P. densata*. Improving our understanding of these issues will provide new insights into the mechanisms of HHS and especially those that are unique to the case of *P. densata*.

Materials and methods

Population sampling and DNA extraction

Initially, a total of 54 populations were sampled throughout the distribution ranges of the three pine species. For each population, cones and needles were collected from 10 to 20 randomly selected mature trees that were at least 100 m apart. The needles were used for genotyping of the mt and cp genomes (Wang *et al.* 2011). The mt genotyping results revealed an ancestral hybrid zone in the north-eastern periphery of the current range of *Pinus densata*. Three successive waves of westward colonization from this zone into the plateau established three distinct clusters of populations in the central and western ranges of the species' distribution (Wang *et al.* 2011). In this study, we selected 10 *P. densata* populations to represent these hybridization and colonization events (Fig. 1a): three from the ancestral hybrid zone (Nos. 1–3, referred to as group I—see the Results section for a more extensive discussion on population grouping), five from the two clusters of the central distribution (Nos. 4–8, referred to as group II) and two from the western range (Nos. 9–10, referred to as group III). For *Pinus tabuliformis* and *Pinus yunnanensis*, five and four populations were selected, respectively, to represent the major mitotypes over the distribution ranges of each species. The mitotype and chlorotype composition of each population are shown in Fig. 1b, c. One seed from each of 10–20 trees from each population was germinated on Petri dishes. Genomic DNA was extracted from the haploid megagametophyte of each seed using a Plant Genomic DNA Kit (Tiangen Biotech, Beijing, China) according to the manufacturer's instructions. The geographical distributions of the 19 sampled populations are shown in Fig. 1a. The name, location and sample size of each population are listed in Table S1 (Supporting information).

Locus selection and sequencing

The search for suitable loci began with a set of 132 single or low copy genes identified in *Arabidopsis thaliana*. We then used the *Pinus taeda* EST database of the NCBI (http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantESTBLAST.shtml?3352) to search for homologous copies of these genes in *Pinus*; the sequences of these homologues were used to design primers for amplifying the corresponding genes in the studied pines. A stringent set of criteria were enforced during locus selection. First, all loci that yielded multiple bands during PCR amplification were discarded. Loci with single PCR bands were cloned using a PGEM T-easy vector (Promega Inc.), and 7–8 clones were sequenced for each locus to determine whether they represented a single sequence; those that did not were excluded, leaving only 37 for further analysis. Each of these 37 loci was amplified and sequenced in four *P. tabuliformis* and four *P. yunnanensis* individuals. We then examined the level of polymorphism at these loci in the two species;

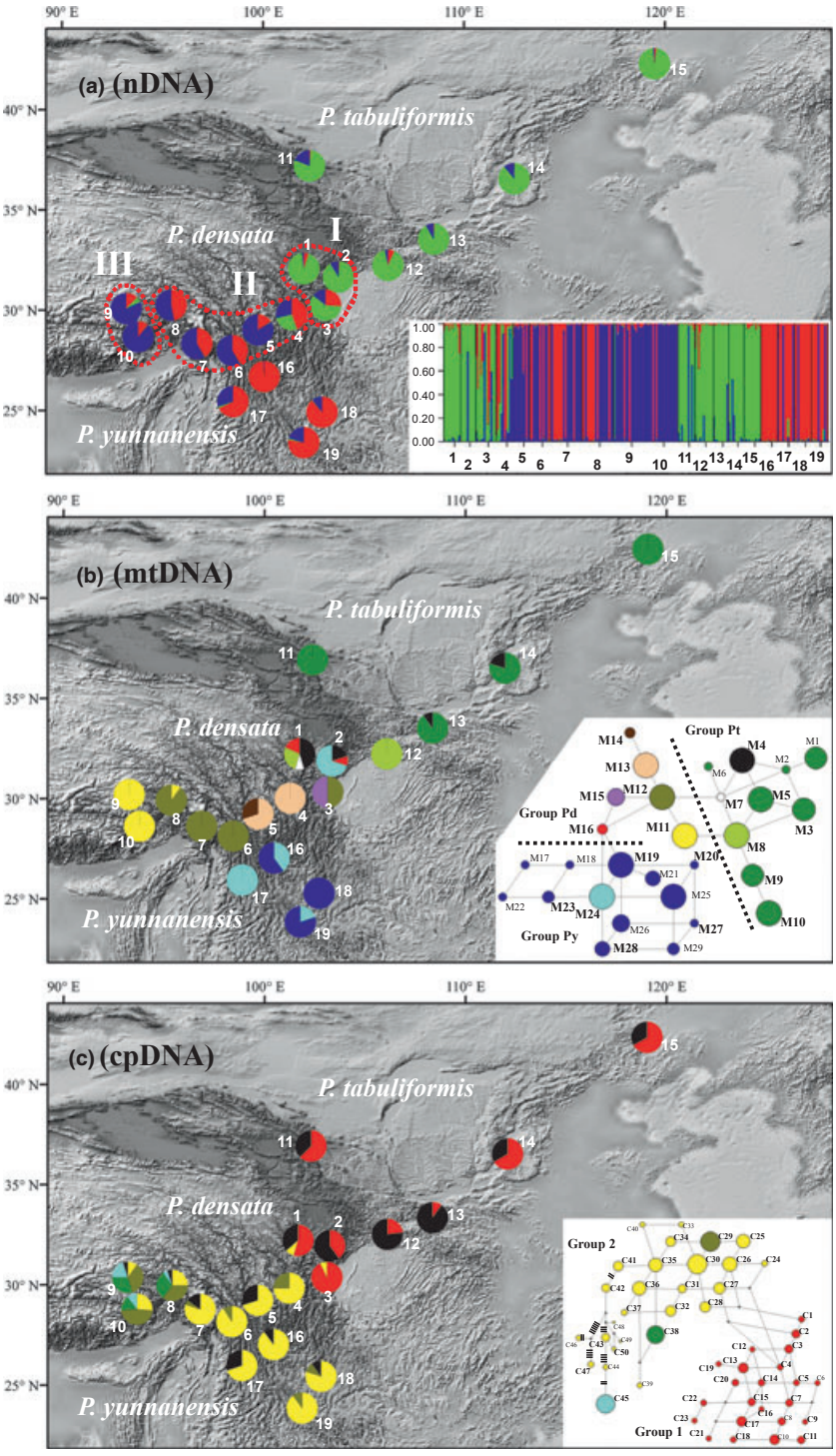

Fig. 1 (a) Assignment of the population frequency of demes by Structure at $K = 3$, based on eight nuclear loci across all 19 populations of the three pine species. The three *P. densata* groups indentified by their nuclear genetic composition were designated I, II and III. (b) Mitotype composition of the 19 analysed populations of the three pine species. Pie charts show the proportions of mitotypes in each population. The network was retrieved from our previous study based on range-wide population sampling (Wang *et al.* 2011). Nineteen mitotypes were detected in the populations of this study and they are indicated in bold. (c) Chlorotype composition of the 19 populations. The network was retrieved from our previous study based on range-wide population sampling (Wang *et al.* 2011). Forty common chlorotypes (each occurred more than twice) were detected in the 19 populations of this study; they are indicated in bold. Pie charts show the proportions of chlorotypes in each population. Rare chlorotypes that occurred only once or twice over all population are in black.

23 exhibited monomorphism or extremely low polymorphism and were therefore excluded from further analyses. Of the remaining 14 loci, six had alignment patterns suggestive of paralogy, that is, each locus harboured distinct classes of sequences, which is indicative of gene duplication or nonspecific priming. For the sake of rigour, we also excluded these six loci, leaving eight for population sequencing. The selected genes consist of a dehydrin gene (*DEH*), a gene belonging to the protein kinase C superfamily (*PKc*); an early light-induced protein gene (*ELIP*); an alpha-trehalose-phosphate synthase gene (*TPS*); a glutathione-S-transferase gene (*GSTG*) and three genes from the *RAV* transcription factor family (*RAV*1, *RAV*2 and *RAV*3). For six of the eight loci

(*DEH*, *RAV*1, *RAV*2, *RAV*3, *ELIP* and *PKc*), we sequenced the complete coding regions, while for the other two (*TPS* and *GSTG*) only a fraction of the gene was sequenced (Table S2, Supporting information).

The primers used to amplify and sequence each locus and the annealing temperature used in each amplification are listed in Table S2 (Supporting information). The PCR products were subjected to electrophoresis using a 1.0% agarose gel, after which the desired band was cut from the gel and purified using a GFX PCR DNA and Gel Band Purification Kit (Amersham Pharmacia Biotech, Buckinghamshire, UK). The purified PCR products were sequenced directly on an ABI 3730 automated sequencer (Applied Biosystems, Foster City, CA, USA). Unique haplotype sequences for each locus have been deposited in the GenBank under accession nos JQ070979–JQ071373.

The genus *Pinus* is divided into two subgenera, *Pinus* and *Strobus*. The split between the two subgenera is the best dated node in pine phylogeny, being supported by both molecular and fossil evidence, and is often used to calibrate the divergence of other clades in the genus (Willyard *et al.* 2007). In this study, we chose *Pinus armandii* as a representative of subgenus *Strobus*. Megagametophyte DNA from this species was sequenced for the eight loci included in this study, and the resulting sequences were used as outgroups in neutrality tests. The timing of split between *Pinus* and *Strobus* and the nucleotide divergence between *P. armandii* and the species analysed in this study (all of which belong to the *Pinus* subgenus) were used to estimate mutation rates in our pines for IM analysis (for details, see the section on *Isolation-with-migration analysis*).

Nucleotide diversity and neutrality tests

Sequences were aligned using ClustalX (Thompson *et al.* 1997) and further manually adjusted using BioEdit v. 5.0.9.1 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Indels were excluded from all analyses in this study, and coding regions and open reading frames were identified by comparison to ESTs for *P. taeda*.

For each locus, the number of segregating sites (S); the extent of nucleotide polymorphism in terms of θ_w (Watterson 1975) and the nucleotide diversity in terms of π (Nei & Li 1979) at total sites (π_t), silent sites (π_s) and nonsynonymous sites (π_a) were estimated for each locus and the combined multilocus sequences of each population, group of populations, and species using DnaSP v. 5.10.01 (Librado & Rozas 2009).

Each locus was tested for departures from neutral expectation using Tajima's *D* (Tajima 1989), Fu and Li's *D** and *F** (Fu & Li 1993) and the standardized Fay and Wu's *H* (Fay & Wu 2000; Zeng *et al.* 2006) statistics, as well as the McDonald and Kreitman (MK) test (McDonald & Kreitman 1991) using DnaSP (Librado & Rozas 2009). In addition, a multilocus HKA (Hudson *et al.* 1987) test was conducted to examine whether the level of polymorphism within species correlated with the degree of divergence between species across loci. The HKA test was performed for three pairwise comparisons between *P. densata*, *P. tabuliformis* and *P. yunnanensis* using the multilocus HKA program (http://genfaculty.rutgers.edu/hey/software). The orthologous sequence from *P. armandii* was used as an outgroup for the *H*, *D**, *F** and MK tests. The significance of each test was determined using 1000 coalescence simulations.

Population structure

Population differentiation (F_{ST}) was estimated at each locus and for the combined sequence set for each group of populations and species using AMOVA in Arlequin v. 3.1 (Excoffier *et al.* 2005), with significance testing based on 1000 permutations. The genetic structure of *P. densata* was also investigated using the model-based clustering algorithm implemented in Structure v. 2.3 (Hubisz *et al.* 2009). This program employs a Bayesian algorithm to infer the true number of clusters (*K*) in a sample of individuals. The most likely number of clusters (*K*) was evaluated using the ΔK method (Evanno *et al.* 2005), in which ΔK is an ad hoc statistic based on the rate of change in the log probability of data between successive *K* values; the chosen value of *K* was that which gave the highest value of ΔK. A total of twenty replicate runs were conducted for every value of *K* between 2 and 10, with a burn-in of 50 000 iterations and a run length of 500 000 iterations. An admixture model was used without prior population information. We used DnaSP to estimate the level of linkage disequilibrium for each pair of informative sites within genes; only sites between which Fisher's exact test with Bonferroni correction indicated no significant correlation were used for Structure analyses. Structure analyses were performed on all 19 populations of the three species, resulting in the identification of three *P. densata* population groups (see Results). This grouping was used in the subsequent IM and approximate Bayesian computation (ABC) analyses.

Isolation-with-migration analysis

Divergence and demographic parameters within *P. densata* and between its progenitor taxa were estimated using two IM analyses (doi:10.5061/dryad.566n8). The basic IM model includes population–size parameters for one ancestral and two descendant populations, a parameter for the timing of the split and two gene exchange parameters. A genealogical tree

reflecting ancestral–descendant relationships was used in multipopulation IM analyses for the three identified *P. densata* groups. Relationships among the groups were inferred from the suggested colonization events based on mtDNA data (see Results, Wang *et al.* 2011). An additional pairwise IM analysis was conducted for *P. tabuliformis* and *P. yunnanensis* to estimate a lower bound for the timing of the speciation event of *P. densata*. The extant population sizes of the two parental species were compared to the ancestral population size of *P. densata*. All IM analyses were conducted using the IMa2 program (Hey 2010b). Because the IM model assumes no intralocus recombination, we extracted the largest non-recombining block for each gene using IMgc program (Woerner *et al.* 2007) and performed the IM analysis with these reduced data sets. Gene flow is expressed as the population migration rate ($2Nm$, forward in time), which is the product of the effective number of gene copies in a population and the rate at which its genes are replaced by incoming migrating genes (Hey 2010b).

To estimate the IM parameters, the generation time was set to 50 years, and the mutation rate (μ, per locus per year) at all sites for each locus was estimated as $\mu = \pi \times L/2T$, where L is the length of the locus, π is the nucleotide divergence at all sites between the subgenera *Pinus* and *Strobus* for that locus and T is the divergence time of these two subgenera. In this study, the value of π for the subgenera *Pinus* and *Strobus* was calculated using *P. armandii* as a representative of *Strobus*. The divergence time for the two subgenera was assumed to be 85 Ma, as proposed by Willyard *et al.* (2007) based on molecular and fossil evidence.

After a few test runs to optimize the prior boundary for demographic parameters, we conducted each IM simulation for 15 million steps with a burn-in of 20 million steps and ran 10 Metropolis Coupling of Markov chains (MCMC) under the infinite-sites (IS) mutation model (Kimura 1969) of sequence evolution. The mixing properties of MCMC were assessed by the number of independent points sampled for each parameter (ESS), the swapping rates between successive chains of MCMC and by monitoring the trend-line plots of the parameters. Long, well-mixing runs were repeated with different random seed numbers until similar posterior distributions were generated at least three times. When this happened, the analysis was considered to have converged on a stationary distribution, and one of the three runs was chosen as the final result.

Demographic inference for each Pinus densata *group*

We used ABC simulations to infer the demographic history of each of the three *P. densata* groups. A range of demographic scenarios were fitted to the observed sequence data following the procedure described by Ingvarsson (2008). Briefly, a large number of replicate simulations were performed for each demographic model, where the parameters of the model were drawn from prior distributions that were chosen to be uninformative about the true value of the parameters unless otherwise specified. Simulated data were summarized using θ_w (Watterson 1975), Tajima's D (Tajima 1989), the standardized Fay and Wu's H (Fay & Wu 2000; Zeng *et al.* 2006) and Kelly's Z_{nS} (Kelly 1997) statistics, and the same set of summary statistics were calculated for the observed data. The simulated samples were accepted only when they were sufficiently close to the observed data. The accepted data points were then used to estimate the posterior distribution for the parameters of the model (Beaumont *et al.* 2002). Model selection was conducted as described by Beaumont (2008) using the VGAM package in R (Thomas 2012).

We tested three demographic models which comprised (i) an instantaneous size change model, (ii) an exponential growth model and (iii) a bottleneck model (Fig. S1, Supporting information). The instantaneous size change model assumes an ancestral population of size N_A that instantaneously expands to the current population size (N_0) at time T_0; the exponential growth model assumes an exponential increase in population size to N_0, starting at time T_0 with a constant exponent ($\alpha = \log(N_0/N_A)/T_0$). For the bottleneck model, the ancestral population N_A was assumed to have instantaneously increased to population size N_0 at time T_0 and then shrunk because of a bottleneck, with a subsequent exponential expansion. The bottleneck was characterized by three parameters: the time since the end of the bottleneck (T_1), the duration of the bottleneck (T_b) and the reduction in population size during the bottleneck (N_1). The growth rate (α) after the bottleneck was given by the function $\alpha = \log(N_0/N_1)/T_1$. The prior distributions of N_A and T_1 were based on IMa2 results; the prior ranges of T_0, T_b and N_1 were chosen to cover a broad range of possible demographic scenarios (Table S3, Supporting information). In all simulations, locus-specific θ and ρ values were used, which were derived by multiplying the length of each gene (L) by the *per site* values of θ and ρ, respectively. The *per site* values of θ and ρ were drawn from uniform priors covering ranges of 10^{-5} to 0.02 and 10^{-5} to 0.1, respectively (Table S3, Supporting information).

For model selection, 3×10^5 simulations were generated for each of the three demographic models, and the 900 points closest ($P_\delta = 0.001$) to the obtained data were used. An additional 7×10^5 samples were subsequently simulated for the bottleneck model. In total, 10^6 samples were generated for the bottleneck model, and 1000

closest data points ($P_\delta = 0.001$) were used to estimate the posterior distributions of the model parameters. We tested different values of P_δ (0.01–0.0005) but obtained similar posterior modes for the estimated parameters (data not shown), confirming that the ABC estimates were insensitive to P_δ (Beaumont *et al.* 2002). Finally, we used posterior predictive simulations (Gelman *et al.* 2004) to assess the fit of the parameters estimated from the posterior distributions. One hundred thousand new data sets were generated using parameters sampled from the posterior distributions. These simulated data sets were summarized using θ_w, nucleotide diversity π, Tajima's *D*, the standardized Fay and Wu's *H* and Kelly's Z_{nS} and then compared to the corresponding observed data. All simulations were performed and analysed using the program ms (Hudson 2002). The ABC analyses were performed using R scripts provided by Beaumont (http://www.rubic.rdg.ac.uk/~mab/stuff/).

Results

Nucleotide variation and neutrality tests

A total of eight unlinked nuclear loci were sequenced for 242 individuals representing 19 populations of three pine species. The sequenced fragments ranged from 447 to 1078 bp with a total concatenated length of 5763 bp after excluding gaps and missing data (Table S2, Supporting information). The nucleotide diversity at each locus and over the eight loci for each population and species are presented in Fig. 2 and Table S1 (Supporting information). In general, the levels of nucleotide polymorphism (θ_w and π_t) over the eight loci were higher in both *Pinus densata* (0.0098 and 0.0065) and *Pinus tabuliformis* (0.0092 and 0.0070) than in *Pinus yunnanensis* (0.0060 and 0.0046). The overall nucleotide polymorphism observed in this study was similar to that previously reported in these three species (Ma *et al.* 2006). The levels of polymorphism differed between loci; *PKc* was the most polymorphic locus ($\theta_w = 0.0178$, 0.0186 and 0.0110 in *P. densata*, *P. tabuliformis* and *P. yunnanensis*, respectively), and *RAV*3 the least ($\theta_w = 0.0053$, 0.0044 and 0.0030 in *P. densata*, *P. tabuliformis* and *P. yunnanensis*, respectively) in all three pine species. For all loci considered, *P. tabuliformis* and *P. densata* exhibited a greater degree of diversity than *P. yunnanensis*. The exceptions were *DEH* and early light-induced protein gene (*ELIP*), which had slightly higher nucleotide diversity (π_t) in *P. yunnanensis* (0.0129 and 0.0056) than in *P. tabuliformis* (0.0126 and 0.0035) and *P. densata* (0.0127 and 0.0053). The diversity at silent sites (π_s) was approximately six times greater than at nonsynonymous sites (π_a), with each gene exhibiting a π_a/π_s ratio of <1 in all three species (Table S1, Supporting information).

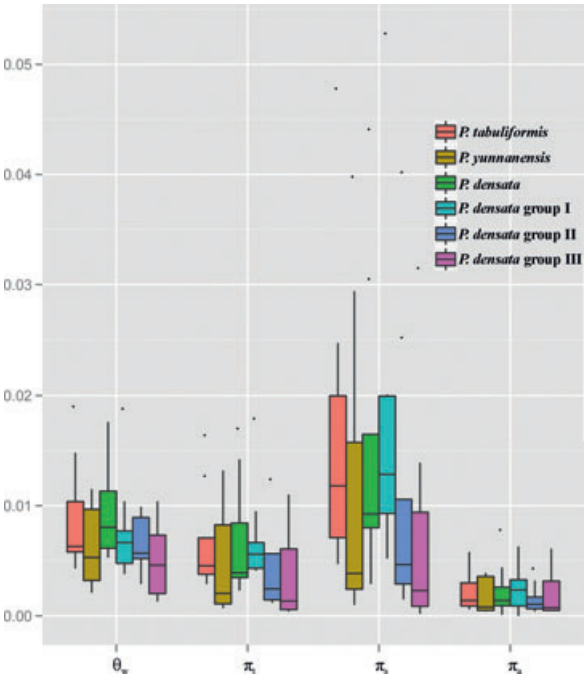

Fig. 2 The distribution of nucleotide diversity across different categories of site for each species and *P. densata* group. θ_w, nucleotide polymorphism measured at all variable sites; π_t, nucleotide diversity at all sites; π_s, nucleotide diversity at silent sites; π_a, nucleotide diversity at nonsynonymous sites.

Using 85 Ma as the divergence time between the two *Pinus* subgenera, the estimated mutation rate (per site per year) at all sites for each gene ranged from 1.56–8.69×10^{-10} for *P. tabuliformis* to 1.46–8.61×10^{-10} for *P. yunnanensis* and 1.55–8.75×10^{-10} for *P. densata*. The average mutation rates over all loci were 3.81, 3.79 and 3.77×10^{-10} for *P. yunnanensis*, *P. tabuliformis* and *P. densata*, respectively. These values are similar to the estimates by Willyard *et al.* (2007) for the genus *Pinus* and Li *et al.* (2010) for spruce.

With regard to the single locus neutrality tests, all three species yielded negative but not significant values for Tajima's *D*, Fu and Li's *D** and *F**, and Fay and Wu's *H*, with only a few exceptions (Table S4, Supporting information). MK tests yielded significant results only for the *DEH* locus in *P. tabuliformis* and for *ELIP*, *RAV*2 and *RAV*3 in *P. densata*. For the multilocus HKA test, none of the pairwise species comparisons departed significantly from the neutral equilibrium model (data not shown). All together, we did not find strong evidence for selection at the analysed loci in any of the three species.

Genetic differentiation and population structure

Significant population differentiation (F_{ST}) was detected for all loci in *P. densata*. The multilocus F_{ST} value for *P. densata* (0.282) was much greater than those for *P. tabuliformis* (0.034) or *P. yunnanensis* (0.036; Table 1). Such a marked difference in F_{ST} between *P. densata* and the two parental species are observed in all previous investigations (Wang *et al.* 2001; Ma *et al.* 2006). Further investigation of the population structure in *P. densata* and the two parental species using Structure revealed the highest likelihood for $K = 3$ clusters (average log probability of data Ln*P*(D) = −7891). For $K = 3$, populations of *P. tabuliformis* and *P. yunnanensis* were homogenous within species but distinctly differentiated between species. The ancestry of all *P. tabuliformis* individuals was tracked to cluster 1 (Fig. 1a, green) and that of *P. yunnanensis* to cluster 2 (Fig. 1a, red). The genetic composition of *P. densata* varied across geographical regions; its populations were approximately divided into three groups. The ancestry for the northeastern populations (Nos. 1–3) was dominated by admixture from cluster 1, so these populations were collectively labelled as group I. For the western populations (Nos. 9–10), the dominant proportion of ancestry was from a *P. densata*-specific cluster (cluster 3, blue); these populations were collectively referred to as group III. The central populations (Nos. 4–8) had fairly admixed ancestry from all three genetic clusters to different degrees, and we named them group II.

Based on the mtDNA data, the group I region was recognized as an ancestral hybrid zone that harbours distinct mitotypes found in all three species (Fig. 1b, Wang *et al.* 2011). This region is geographically adjacent to the range of *P. tabuliformis*, and the cpDNA sequences of the group I *P. densata* population were only weakly differentiated from those of *P. tabuliformis* (Fig. 1c, Wang *et al.* 2011). This indicates that pollen-mediated introgression is dominant from this parental species. Strong introgression from *P. tabuliformis* into this region over a long period of time is believed to have caused the nuclear genome composition of group I to become similar to that of *P. tabuliformis* and distinct from the rest of *P. densata* range. The group II region contained two distinct mtDNA clusters, both of which experienced similar pollen flow from *P. yunnanensis* (Fig. 1b, c). At the nuclear loci, these populations showed varying degrees of admixture for all three species, suggesting a variable intensity of introgression from the parental species. Group III occupied the westernmost portion of the *P. densata* range and was established from group II during the most recent colonization event (Wang *et al.* 2011). Trees from this group had distinct mt and cpDNA compositions (Fig. 1b, c, Wang *et al.* 2011) and clustered as a separate unit in the Structure analysis of the nuclear loci. The discrepancy in population clustering among the three genomes is probably the result of the relatively limited ability of seeds to migrate between populations compared to pollen-mediated gene flow. Such asymmetric diversity is often observed in comparisons between maternally inherited mtDNA and paternally or bi-parentally inherited cpDNA and nuclear DNA (e.g. Petit *et al.* 2005). Collectively, the nuclear DNA-based grouping in the analysed samples reflected both the cytoplasmic composition of the populations and the intensity of introgression from parental species.

The three *P. densata* groups differed in terms of the level of nucleotide polymorphisms (θ_w and π_t). Groups I and II had higher nucleotide diversity ($\theta_w = 0.0078$ and $\pi_t = 0.0072$ for group I; $\theta_w = 0.0066$, $\pi_t = 0.0049$ for group II) than group III ($\theta_w = 0.0054$, $\pi_t = 0.0036$; Fig. 2 and Table S1, Supporting information). For each group, the diversity at silent sites (π_s) was approximately twofold greater than at total variable sites and sixfold greater than at nonsynonymous sites (π_a; Fig. 2 and Table S1, Supporting information).

Table 1 Population differentiation (F_{ST}) in the three pine species

Species	No. of populations	Locus								
		DEH	*RAV*1	*PKc*	*ELIP*	*TPS*	*GSTG*	*RAV*2	*RAV*3	Multilocus
Pinus tabuliformis	5	0.083*	0.192**	−0.011	−0.018	0.004	0.037	0.041	0.045	0.034*
Pinus yunnanensis	4	0.095*	0.042	−0.032	0.008	0.068	0.111*	0.069*	0.022	0.036
Pinus densata	10	0.225**	0.390**	0.103**	0.140**	0.496**	0.524**	0.457**	0.342**	0.282**
Within *P. densata*										
Group I	3	0.043	−0.033	0.061*	0.032	0.040	0.077	0.028	0.019	0.061*
Group II	5	0.027	0.095**	0.036	0.077**	0.053*	0.065	0.365**	0.264**	0.073**
Group III	2	0.154**	0.017	−0.028	0.029	0.021	0	0.012	−0.025**	0.042

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

Isolation-with-migration history

We performed multiple IM simulation runs for the three *P. densata* population groups using IMa2 to recover the marginal posterior distributions of the probabilities of the 15 demographic parameters in a three-population system. Mutation rates at all sites for each gene locus appropriate for a calibration of 85 Ma were used in this analysis. The maximum-likelihood estimates (MLEs) and 95% highest posterior density (HPD) intervals for each demographic parameter are summarized in Table 2 and illustrated in Fig. 3; their marginal distributions are shown in Fig. 4. The first divergence occurred between group I and the ancestral population of groups II and III at *c.* 6.64 Ma, and the divergence between groups II and III at about 0.30 Ma. The effective population sizes (N_e) for group I (2.11×10^5) and group II (2.62×10^5) were much greater than that of group III (0.80×10^5). The estimated N_e of the ancestral population of all three groups (0.41×10^5) and the ancestor of groups II and III (<1000 individuals) were all smaller than their descendent populations.

Asymmetric migration was observed for all pairs of groups. The estimated long-term gene flow between groups I and III was very low, with $2Nm = 0.39$ and 0.02 for migration from group I into III and from III into I, respectively. There was substantial migration from group I to group II (6.93) but little in the opposite direction (0.15). Similarly, for groups II and III, there was a very large amount of migration from group III to II (23.49), but much less in the opposite direction (1.45). There was also considerable migration from the ancestral population of groups II and III to group I (24.00) but much less in the opposite direction (0.27).

We also performed pairwise IM analysis for *P. tabuliformis* and *P. yunnanensis* (Fig. S2 and Table S5, Supporting information). The estimated effective population sizes were 1.07×10^5, 0.37×10^5 and 118.29×10^5 for *P. tabuliformis*, *P. yunnanensis* and their ancestral population, respectively. The marginal posterior probability distribution of the divergence time between *P. tabuliformis* and *P. yunnanensis* showed a major peak at *c.* 20 Ma (Fig. S2 and Table S5, Supporting information).

The estimated migration from *P. tabuliformis* to *P. yunnanensis* was effectively zero (0.07), whereas that in the opposite direction was 3.13. The contemporary distribution of *P. tabuliformis* and *P. yunnanensis* is allopatric, and each species is characterized by distinct mitotypes and chlorotypes (Wang *et al.* 2011). Thus, the observed asymmetric gene flow between *P. tabuliformis* and *P. yunnanensis* probably represents a historical pattern of gene exchange in which movement from the south (*P. yunnanensis*) to the north (*P. tabuliformis*) was predominant. This direction of gene flow mirrors the advance in the timing of phenological events on moving from the south to the north, which would result in asymmetric gene flow between then-parapatric populations of the two species.

Bottlenecks and population expansion in Pinus densata *groups*

Gene-based summary statistics (Tajima's *D* and Fay and Wu's *H*) revealed a pronounced difference between the three groups (Fig. S3, Supporting information). Groups II and III displayed much more negative values ($D = -1.14$ and $H = -1.99$ in group II and $D = -1.44$ and $H = -2.39$ in group III) than group I ($D = -0.14$ and $H = -0.47$), which suggests that these regions have distinct evolutionary histories. We thus examined various demographic models for each of the groups.

The approximate Bayesian computation (ABC) model selection approach indicated that all three *P. densata* groups had gone through a bottleneck, with posterior probabilities of 0.956, 0.932 and 0.988 for groups I, II and III, respectively (Table 3). In the bottleneck models for each group, most parameters had distinct modes in

Table 2 Estimated demographic parameters for the three *Pinus densata* groups using IMa2

	N_0	N_1	N_2	N_3	N_4	T_0	T_1	$2N_0m0 > 1$	$2N_1m1 > 0$	$2N_0m0 > 2$	$2N_2m2 > 0$	$2N_1m1 > 2$	$2N_2m2 > 1$	$2N_0m0 > 3$	$2N_3m3 > 0$
MLEs	2.11	2.62	0.80	713	0.41	0.30	6.64	0.15	6.93	0.02	0.39	23.49	1.45	24.00	0.27
HPD95Lo	1.12	0.87	0.25	248	0.02	0.21	3.10	0	0	0	0	0	0	9.66	0.15
HPD95Hi	4.00	10.69	2.46	1819	2.02	0.43	8.73	7.63	34.99	7.11	2.92	285.4	25.56	55.53	0.56

MLEs, maximum-likelihood estimates; HPD95Lo, the lower bound of the estimated 95% highest posterior density (HPD) interval; HPD95Hi, the upper bound of the estimated 95% HPD interval; N_0, effective population size of group I; N_1, effective population size of group II; N_2, effective population size of group III; N_3, effective population size of the ancestor of groups II and III; N_4, effective population size of the ancestor of all three groups; T_0, split time between groups II and III; T_1, split time between group I and the ancestor of groups II and III; N_0, N_1, N_2 and N_4 are scaled by 10^5 individuals; T_0 and T_1 are scaled by Myr. $2N_imi > j$, population migration rate from population j to population *i* forwards in time, for example, $2N_0m0 > 1$ represents population migration rate from group II to group I.

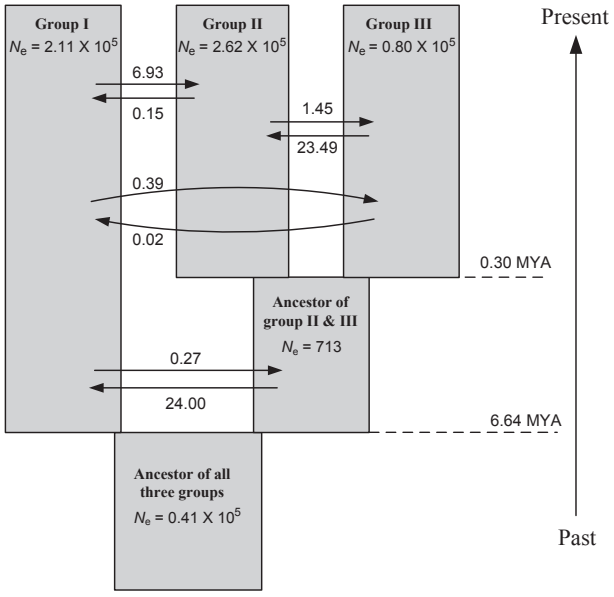

Fig. 3 Summary of the isolation-with-migration model for the three *P. densata* groups. Fifteen demographic parameters estimated by IMa2 are shown. Each block represents a current or ancestral population with their estimated effective population size (N_e). Arrows denote the direction of gene flow with the estimated migration rate labelled above or below the arrow. The timings of the two splitting events are indicated in Ma.

the posterior distributions, suggesting that the data contain enough information to estimate these parameters (Table S6, Supporting information). However, the nature of the best-fitted bottleneck scenario differed for each group (Table S6, Supporting information). For group I, the posterior mode of the population size during bottleneck (N_1, in units of N_0) was 0.032 (95% credible interval, 0.002–0.173), the duration of the bottleneck (T_b, in units of $4N_0$ generations) was 0.001 and the time since the end of the bottleneck (T_1, in units of $4N_0$ generations) was 0.003. For groups II and III, the N_1 values were 0.080 and 0.070, the T_b values were 0.024 and 0.022 and the T_1 values were 0.002 and 0.004, respectively. The estimated current population sizes were 2.11×10^5, 2.62×10^5 and 0.80×10^5 for groups I, II and III, respectively (Table 2). If we assume a generation time of 50 years for *P. densata*, the estimated T_b and T_1 values would correspond to 0.05 Myr and 0.12 Ma, 1.27 Myr and 0.12 Ma and 0.35 Myr and 0.06 Ma for groups I, II and III, respectively (Table S6, Supporting information).

Posterior predictive simulations (Fig. S3, Supporting information) showed a generally good agreement between the observed and simulated data sets, except that the mean values of *H* were substantially higher for the simulated data in groups II and III. The failure to fit the simulated *H* values in groups II and III may be due to the population structure in these groups, as population structure can produce very negative *H* values (Przeworski 2002). The simulations assumed no population structure, which would generate much higher (and in most cases, positive) *H* values. More complex simulations that incorporate this factor could potentially provide more appropriate models for groups II and III. However, parameter-rich models require more sequence information to resolve and so could not be drawn up using our current data set.

Discussion

Late Miocene origin of Pinus densata *and its subsequent divergence*

The estimated divergence time between the two parental species was about 20 Ma (95% HPD: 15–32 Ma) under 85 Ma calibration. This is consistent with the suggestion of a recent origin (14–67 Ma) for most extant pine taxa, especially those belonging to the subgenus *Pinus* (Willyard *et al.* 2007). In *Pinus densata*, the first split between the ancient hybrid zone (group I) and the ancestor of the central and western groups (II and III) occurred *c.* 6.6 Ma, and the divergence between the central (group II) and western range (group III) occurred around 0.3 Ma. The splitting time between the parental species can be regarded as a lower bound for the speciation time of *P. densata*, with the first splitting time within *P. densata* serving as the upper bound. We thus estimate the origin of *P. densata* to have occurred around 6.6–20 Ma. This estimate is coincident with a number of major recent tectonic events in the SE Tibetan Plateau. The uplift of the Tibetan Plateau dates back at least 20 Ma (Harrison *et al.* 1992; Ruddiman 1998), and significant increases in the altitude of the eastern part of the Tibetan Plateau are thought to have occurred about 8–10 Ma (Harrison *et al.* 1992; Zhisheng *et al.* 2001). The drastic geographical changes brought about by the uplift and climatic changes during the Plio-Pleistocene could have altered the regional flora and separated overlapping or parapatric species (Florin 1963; Frenzel 1968). The contemporary distributions of *Pinus tabuliformis* and *Pinus yunnanensis* are allopatric. The discovery of the ancestral hybrid zone in the north-eastern periphery of the *P. densata* range suggests that hybridization could have occurred in the previously overlapping zone between *P. yunnanensis* and *P. tabuliformis*. The uplift of the eastern Tibetan Plateau and associated climate changes could then have gradually pushed the northern edge of the *P. yunnanensis* range southward to its present distribution, and populations in the ancestral hybrid zone would have become isolated from *P. yunnanensis* (Wang *et al.* 2011). The creation of new territories on the Tibetan

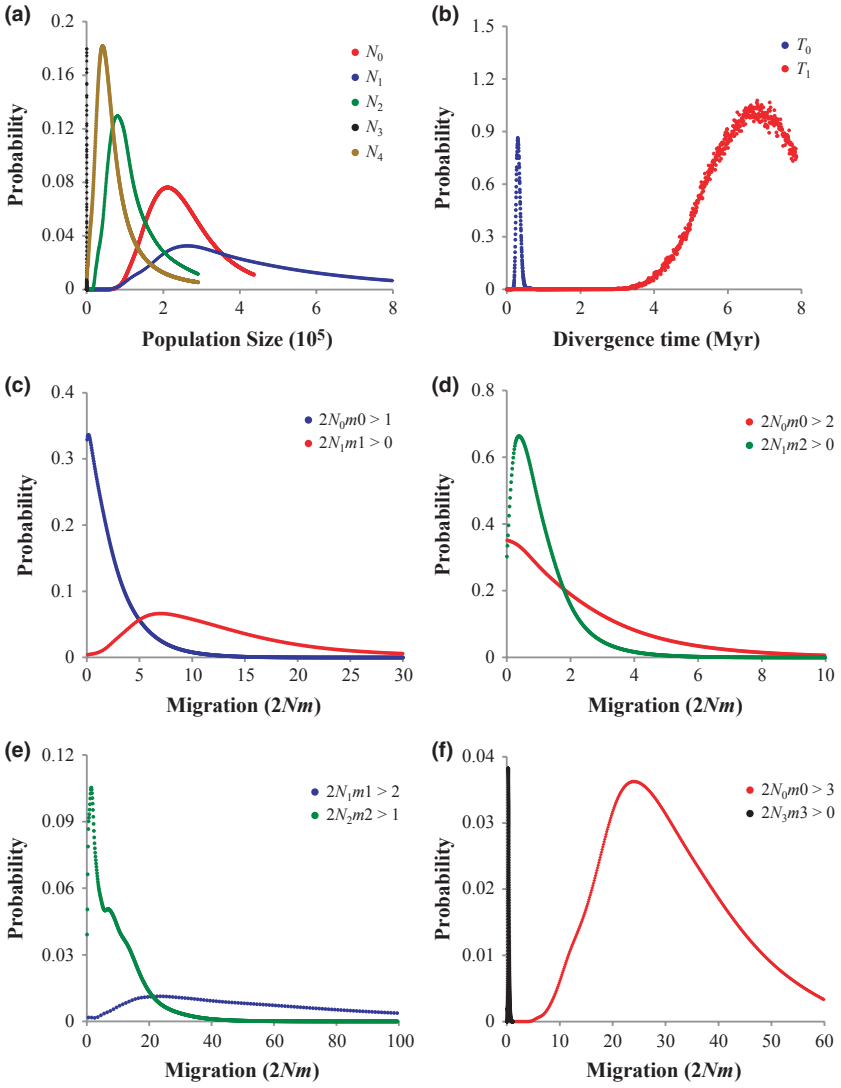

Fig. 4 Marginal distribution of the posterior probability of 15 demographic parameters estimated by IMa2 for the three groups of *P. densata* populations. (a) Effective population size of group I (N_0, red), group II (N_1, blue), group III (N_2, green), the ancestor of groups II and III (N_3, black) and the ancestor of all three groups (N_4, brown). (b) Splitting time between groups II and III (T_0, blue) and between group I and the ancestor of groups II and III (T_1, red). (c) Migration rate from group I to group II ($2N_1m1 > 0$, red) and from group II to group I ($2N_1m1 > 0$, blue). (d) Migration rate from group I to group III ($2N_2m2 > 0$, green) and from group III to group I ($2N_0m0 > 2$, red). (e) Migration rate from group II to group III ($2N_2m2 > 1$, green) and from group III to group II ($2N_1m1 > 2$, blue). (f) Migration rate from group I to the ancestor of groups II and III ($2N_3m3 > 0$, black) and from the ancestor of groups II and III to group I ($2N_0m0 > 3$, red). Values of the probability of N_3, T_0 and $2N_3m3 > 0$ are scaled down by factors of 100, 25 and 100, respectively, to allow the presentation of all parameters in the same panel in a, b and f, respectively.

Plateau would have provided opportunities for the hybrid lineages to colonize and establish extensive forests. Ecological niche modelling suggests that the habitat occupied by *P. densata* is distinctly shifted from that of the parental species (Mao & Wang 2011). The fact that *P. densata* grows and regenerates well in the high plateau habitat demonstrates its unique ecological adaptation (Mao *et al.* 2009). Rapid speciation events triggered by the uplift of the plateau have also been proposed for other plant species in this region (Liu *et al.* 2006; Yue *et al.* 2009; Wang *et al.* 2010; Xu *et al.* 2010). Collectively, our results suggest a late Miocene origin for *P. densata* that was associated with major tectonic events in the region.

The isolation and genetic differentiation within *P. densata* across geographical regions should be affected by both large-scale climate fluctuations and topographical changes. The first splitting event, between group I and the ancestor of groups II and III, was dated to 6.6 Ma with a 95% credible interval of 3.1–8.73 Ma. This event overlaps with the latest episodes of the uplift of eastern Tibetan Plateau, when major river drainage systems in SE Tibet were rearranged and reinforced in the Pliocene (Clark *et al.* 2004). High mountain chains and deep valley systems created by these tectonic events could have functioned as effective barriers to gene flow. The Dadu River system, which separates groups I and II, was redirected and further incised by 800 m during this process of paleo-drainage rearrangement (Clark *et al.* 2004). Currently, the Dadu River flows in inner gorges more than 2000 m deep. The topography of SE Tibet is characterized by a number of similarly massive valley systems, such as those of the Yangtze, Mekong, Salween and the Yarlung Zangbo Rivers (the upper stream of the Brahmaputra River). These have been identified as strong

Table 3 Posterior probabilities of three demographic models for *Pinus densata* groups I, II and III, obtained from ABC model selection

Model	Posterior probability		
	Group 1	Group II	Group III
Instantaneous size change	0.016	0.068	0.007
Exponential growth	0.029	<0.001	0.005
Bottleneck	0.956	0.932	0.988

ABC, approximate Bayesian computation.

geographical barriers to dispersal and defined the phylogeographic history of regional flora, such as the 'Mekong-Salween Divide' (Gao *et al.* 2007; Li *et al.* 2011).

The divergence between groups II and III (0.3 Ma) occurred much later, after the plateau had reached its present elevation; it coincides with the middle Pleistocene glaciations (Zheng & Rutter 1998; Zhou *et al.* 2006). Thus, the differentiation between the central and western groups should be mainly affected by regional climate fluctuations. This view is supported by geological studies, which suggested that Pleistocene glaciations in Tibet were geographically limited to isolated mountain groups or smaller plateaus in higher mountains, and there is no evidence for extensive whole plateau glaciations (Shi 2002; Zhou *et al.* 2006). This would have made it possible for *P. densata* populations to persist during the glaciations in isolated refugia and accumulate genetic divergence. Similar isolation histories during recent glaciations have been proposed for other Tibetan conifers (Opgenoorth *et al.* 2010), shrubs (Shimono *et al.* 2010) and herbs (Yang *et al.* 2008; Gao *et al.* 2009; Wang *et al.* 2009). Our results thus provide more information on the pre-Last Glacial Maximum history of Tibetan vegetation.

Demographic history of Pinus densata

The demographic history of *P. densata* is complex, involving strong bottlenecks followed by population expansions during the westward colonization. The effective population size (N_e) estimated for the ancestor of *P. densata* (ancestor of all three groups, 0.41×10^5) was comparable to that of *P. yunnanensis* (0.37×10^5) but smaller than that of *P. tabuliformis* (1.07×10^5). The large N_e of the hybrid species reflects its broad genetic base. The estimated N_e of *Helianthus paradoxus*, a homoploid hybrid of single origin, is less than one-fifth as large as that of its parental species (Strasburg *et al.* 2011). The large N_e of the ancestor of *P. densata* indicates that large fractions of the gene pool of the two parental species were involved in the hybrid speciation during either initial hybridization or subsequent backcrossing. In contrast, the estimated N_e for the ancestor of groups II and III was much smaller, comprising fewer than 1000 individuals. This marked difference in N_e between the ancestor of all groups and that of the central and western groups suggests that a severe bottleneck occurred during the first split between group I and the ancestor of groups II and III. The very low migration (0.27) detected from group I to the ancestor of groups II and III also supports this bottleneck event and indicates limited seed and pollen migration. These findings are consistent with the significant decline in genetic diversity at both the organelle (Wang *et al.* 2011) and nuclear DNA levels along the route of westward colonization.

Approximate Bayesian computation simulations suggested that all three *P. densata* groups experienced a bottleneck followed by exponential growth. The effect of a bottleneck is determined by the magnitude of population size reduction, the duration and the time that has elapsed since the end of the bottleneck. Among the three groups, the bottleneck detected in groups I and II (0.12 Ma) was more ancient than that in group III (0.06 Ma), and the duration and the severity of the bottleneck was shorter and milder in group I (0.05 Myr and $0.03N_0$, respectively) as compared with the other two groups. A weak and ancient bottleneck, as revealed in group I, may not be always detectable. Our organelle genome-based mismatch distribution test did not reveal any bottleneck affecting group I (Wang *et al.* 2011); this was probably due to differences in inheritance and N_e between nuclear and organelle genomes. The biparentally inherited nuclear genome has a greater N_e than the uniparentally inherited organelle genomes and can thus store more genetic information, facilitating the tracking of more ancient events (Birky *et al.* 1989; Petit *et al.* 2005).

Compared to their ancestor population, the N_e values for groups II and III have increased by factors of 367 and 112, respectively. One factor that may have contributed to this expansion in the *P. densata* population size is introgression from one or both of the parental species. Groups I and II, which have large N_e values (2.11×10^5 and 2.62×10^5), showed predominant introgression from *P. tabuliformis* and *P. yunnanensis*, respectively (Fig. 1, Wang *et al.* 2011). Conversely, group III, which has the smallest N_e (0.80×10^5) of the three groups, was suggested to have been isolated from its parent species in terms of seed and pollen flow for a long time (Wang *et al.* 2011). Potential violations of the assumption that unsampled populations make no genetic contribution are an unavoidable problem when using IM models to study closely related species separated by a weak genetic barrier (Ikeda *et al.* 2009; Nadachowska & Babik 2009; Li

et al. 2010). A simulation study conducted to evaluate the performance of IM in cases where such violations had occurred showed that gene flow between the first species and a third unsampled species inflates the N_e of the first species and ancestor but causes underestimation of the N_e for the second species (Strasburg & Rieseberg 2010). This suggests that our estimated N_e values for groups I and II could have been biased upwards by introgression from *P. tabuliformis* and *P. yunnanensis*, respectively, and the N_e for the ancestor of all three groups would have been increased by introgression from both parental species. For the ancestor of groups II and III, the N_e may have been biased downwards by introgression from *P. tabuliformis* to group I (the ancestor of groups II and III was used as the second species for group I), but biased upwards by introgression from *P. yunnanensis* to group II in a three-population IM model. It is not possible to state quantitatively which of these opposite effects was more dominant in shaping the N_e of the ancestor population of groups II and III. Furthermore, it has been suggested that the presence of genetic structure in the ancestral population can result in overestimation of the ancestral N_e (Becquet & Przeworski 2009). Thus, considering the multiple hybrid origin and significant geographical structure in *P. densata*, the IM estimate of N_e for the ancestral population of all three groups may have an upward bias.

By using all silent polymorphic sites for $N_e = \theta_{ws}/4\mu$, we directly calculated N_e values for the three *P. densata* groups ($N_e = 1.6 \times 10^5$, 1.1×10^5 and 0.9×10^5 for groups I, II and III, respectively). These values all fall within the 95% HPD of the IMa2 estimates (Table 2). Thus, while N_e is affected by several factors in different ways, our estimates for the sampled populations from the multi-population IM model can be considered reliable. In summary, our results suggest that the ancestral population of *P. densata* was large and that its descendent populations west of the ancient hybrid zone were established by limited founders. After separation from the ancestral populations, population expansions occurred in all *P. densata* groups at different times in history.

Variable gene flow between geographical regions across the range of Pinus densata

Different *P. densata* groups experienced different patterns of gene flow. Relatively high gene flow ($2Nm > 1.0$) was detected between the geographically neighbouring groups (groups I and II, and II and III), with considerably less ($2Nm < 0.5$) between geographically distant groups (groups I and III). This suggests that gene flow in *P. densata* is restricted to adjacent regions. Gene flow between groups I and III is probably prevented by the long distance separating them and the complex topography of SE Tibet, a pattern that is also found in other conifers on the Tibetan Plateau (Gao *et al.* 2007; Cun & Wang 2010; Li *et al.* 2010).

The high levels of gene flow detected by IM could represent continuous gene flow from the initial divergence through to the present, a historic gene flow after divergence that lasted for a short period or a secondary contact after a period of allopatric divergence. Groups I and II had different and distinct mitotype and chlorotype compositions (Wang *et al.* 2011), which suggests there was no strong current seed and pollen flow between the two groups because the haploid mt and cp genomes are often replaced by introgression. Thus, the gene flow detected in the nuclear genome between groups I and II is more likely to be a signature of ancient common ancestry. Additionally, this residual signature is only preserved in the neighbouring regions between the two groups because gene flow estimates significantly decreased in both directions ($2Nm < 0.3$) after removing two adjacent marginal populations (Nos. 3 and 4, data not shown). The gene flow detected between groups II and III could reflect a secondary contact after the most recent glaciations, because the most western populations (No. 8) in group II, adjacent to the group III region, had chloroplast genome components that were similar to those of group III (Wang *et al.* 2011; Fig. 1). However, we cannot absolutely rule out the possibility that gene flow occurred during the process of population divergence. When we removed population No. 8 from the IMa2 simulation, a considerable gene flow still existed ($2Nm = 13.5$ from group III to group II and 10.4 in the opposite direction). This means that gene exchange has affected a large proportion of groups II and III since the secondary contact between them or alternatively (and perhaps more likely) that a signature of historic gene flow has been maintained owing to the short isolation and divergence time separating the two regions.

Interestingly, the gene flows from groups I and III into group II were much higher than those in the opposite directions. Asymmetric gene flows between populations or species can be caused by various intrinsic or extrinsic factors, such as asymmetry in the compatibility, survival and fertility of the two kinds of hybrids (Wachowiak *et al.* 2011) and phenology or wind direction during flowering (Bai *et al.* 2010). Additionally, the gene flow estimated with the IM model could be biased by genetic exchange between the first focal species and a third, unsampled species, resulting in an increase in estimated gene flow from the second focal species into the first species but a decrease in the opposite direction (Strasburg & Rieseberg 2010). Thus, introgression from *P. yunnanensis* into group II populations may result in upward-biased gene flow for transfers from groups I and III into II and downward-biased transfers in the

opposite direction. Similarly, introgression from *P. tabuliformis* into group I could also lead to overestimation of the gene flow from the ancestor of groups II and III into group I and underestimation of the flow in the opposite direction. The intensity of these biases caused by introgression from parental species remains to be defined quantitatively using simulations that incorporate different levels of gene flow from unsampled species.

As one of the most ecologically successful homoploid hybrid plant species, *P. densata* is remarkable for its successful colonization of vast areas of heterogeneous terrain on the Tibetan Plateau. Across its distribution, its distinct population histories and genetic compositions illustrate the complexity of its speciation process (Wang & Szmidt 1994; Wang *et al.* 2001, 2011; Song *et al.* 2003; Ma *et al.* 2006). The present study based on nuclear coalescence analyses further resolved details of the pine's speciation history and patterns of gene flow, yielding four key findings. First, *P. densata* originated in the late Miocene during the recent uplift of SE Tibetan Plateau. The divergence between *P. densata* populations in different geographical regions began in the late Pliocene and was induced by regional topographical changes and subsequent Pleistocene glaciations. Second, the ancestral *P. densata* population was large, but the central and western populations were established by limited founders, indicating that there were severe bottlenecks during the westward migration out of the ancestral hybrid zone. Third, recent population expansions have occurred in all geographical regions, especially in the western populations, but at different times in the past. Finally, gene flow between *P. densata* populations is limited and is observed only between adjacent geographical regions. Such genetic isolation could be due to geophysical, ecological and intrinsic genetic factors. Further research into these areas would advance our understanding of the origin and maintenance of biological diversity on the Tibetan Plateau, especially in terms of identifying the population genetic processes that serve as the engines of diversification and speciation (Lexer & Stölting 2011).

Acknowledgements

This manuscript was improved by the comments of three anonymous reviewers and subject editor Dr M. Heuertz. This study was supported by grants from the Natural Science Foundation of China (NSFC 30830010 and 31100158), the National Basic Research Program of China (2009CB119104) and Vetenskapsrådet, Sweden.

Reference

Abbott RJ, Hegarty MJ, Hiscock SJ, Brennan AC (2010) Homoploid hybrid speciation in action. *Taxon*, **59**, 1375–1386.

Arnold ML (1993) *Iris nelsonii* (Iridaceae): origin and genetic composition of a homoploid hybrid species. *American Journal of Botany*, **80**, 577–583.

Arnold ML (1997) *Natural Hybridization and Evolution*. Oxford University Press, New York, New York.

Bai WN, Liao WJ, Zhang DY (2010) Nuclear and chloroplast DNA phylogeography reveal two refuge areas with asymmetrical gene flow in a temperate walnut tree from East Asia. *New Phytologist*, **188**, 892–901.

Beaumont MA (2008) Joint determination of topology, divergence time and immigration in population trees. In: *Simulations, Genetics and Human Prehistory* (eds Matsumura S, Forster P and Renfrew C), pp. 135–154. McDonald Institute for Archaeological Research, Cambridge.

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

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

Birky CW, Fuerst P, Maruyama T (1989) Organelle gene diversity under migration, mutation, and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. *Genetics*, **121**, 613–627.

Blum MGB, François O (2010) Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, **20**, 63–73.

Buerkle CA, Morris RJ, Asmussen MA, Rieseberg LH (2000) The likelihood of homoploid hybrid speciation. *Heredity*, **84**, 441–451.

Clark MK, Schoenbohm LM, Royden LH *et al.* (2004) Surface uplift, tectonics, and erosion of eastern Tibet from large-scale drainage patterns. *Tectonics*, **23**, TC1006.

Cun YZ, Wang XQ (2010) Plant recolonization in the Himalaya from the southeastern Qinghai-Tibetan Plateau: geographical isolation contributed to high population differentiation. *Molecular Phylogenetics and Evolution*, **56**, 972–982.

Dixon CJ, Kapralov MV, Filatov DA (2011) Gene flow and species cohesion following the spread of *Schiedea globosa* (Caryophyllaceae) across the Hawaiian Islands. *Journal of Evolutionary Biology*, **24**, 1–11.

Edwards SV, Beerli P (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, **54**, 1839–1854.

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

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

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.

Florin R (1963) The distribution of conifer and taxad genera in time and space. *Acta Horticulturae Bergiani*, **20**, 122–312.

Frenzel B (1968) The Pleistocene vegetation of northern Euras: recent vegetation of northern eurasia resulted from a relentless contest between steppe and forest. *Science*, **161**, 637–649.

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

Gao LM, Moeller M, Zhang XM *et al.* (2007) High variation and strong phylogeographic pattern among cpDNA haplotypes in *Taxus wallichiana* (Taxaceae) in China and North Vietnam. *Molecular Ecology*, **16**, 4684–4698.

Gao QB, Zhang DJ, Chen SY *et al.* (2009) Chloroplast DNA phylogeography of *Rhodiola alsia* (Crassulaceae) in the Qinghai-Tibet Plateau. *Botany*, **87**, 1077–1088.

Gelman A, Carlin JB, Stern HS, Rubinand DB (2004) *Bayesian Data Analysis*, 2nd edn. Chapman& Hall/CRC Press, Boca Raton, Florida.

Grant V (1981) *Plant Speciation*. Columbia University Press, New York, New York.

Gross BL, Rieseberg LH (2005) The ecological genetics of homoploid hybrid speciation. *Journal of Heredity*, **96**, 241–252.

Gross BL, Schwarzbach AE, Rieseberg LH (2003) Origin(s) of the diploid hybrid species *Helianthus deserticola* (Asteraceae). *American Journal of Botany*, **90**, 1708–1719.

Harrison TM, Copeland P, Kidd WSF, Yin A (1992) Raising Tibet. *Science*, **255**, 1663–1670.

Hey J (2005) On the number of New World founders: a population genetic portrait of the peopling of the Americas. *Plos Biology*, **3**, 965–975.

Hey J (2010a) The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Molecular Biology and Evolution*, **27**, 921–933.

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

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

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

Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

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

Ikeda H, Fujii N, Setoguchi H (2009) Application of the isolation with migration model demonstrates the Pleistocene origin of geographic differentiation in *Cardamine nipponica* (Brassicaceae), an endemic Japanese alpine plant. *Molecular Biology and Evolution*, **26**, 2207–2216.

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

Kelly JK (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.

Kimura M (1969) Number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. *Genetics*, **61**, 893–903.

Lewontin RC, Birch LC (1966) Hybridization as a source of variation for adaptation to new environments. *Evolution*, **20**, 315–336.

Lexer C, Stölting KN (2011) Tracing the recombination and colonization history of hybrid species in space and time. *Molecular Ecology*, **20**, 3701–3704.

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

Li Y, Zhai S-N, Qiu Y-X *et al.* (2011) Glacial survival east and west of the 'Mekong–Salween Divide' in the Himalaya–Hengduan Mountains region as revealed by AFLPs and cpDNA sequence variation in *Sinopodophyllum hexandrum* (Berberidaceae). *Molecular Phylogenetics and Evolution*, **38**, 31–43.

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

Liu ZL, Zhang D, Hong DY, Wang XR (2003) Chromosomal localization of 5S and 18S-5.8S-25S ribosomal DNA sites in five Asian pines using fluorescence in situ hybridization. *Theoretical and Applied Genetics*, **106**, 198–204.

Liu JQ, Wang YJ, Wang AL, Hideaki O, Abbott RJ (2006) Radiation and diversification within the *Ligularia-Cremanthodium-Parasenecio* complex (Asteraceae) triggered by uplift of the Qinghai-Tibetan Plateau. *Molecular Phylogenetics and Evolution*, **38**, 31–49.

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

Mao JF, Wang XR (2011) Distinct niche divergence characterizes the homoploid hybrid speciation of *Pinus densata* on the Tibetan Plateau. *American Naturalist*, **177**, 424–439.

Mao JF, Li Y, Wang XR (2009) Empirical assessment of the reproductive fitness components of the hybrid pine *Pinus densata* on the Tibetan Plateau. *Evolutionary Ecology*, **23**, 447–462.

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

Nadachowska K, Babik W (2009) Divergence in the face of gene flow: the case of two newts (Amphibia: Salamandridae). *Molecular Biology and Evolution*, **26**, 829–841.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences, USA*, **76**, 5269–5273.

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

Opgenoorth L, Vendramin GG, Mao KS *et al.* (2010) Tree endurance on the Tibetan Plateau marks the world's highest known tree line of the Last Glacial Maximum. *New Phytologist*, **185**, 332–342.

Petit RJ, Duminil J, Fineschi S *et al.* (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology*, **14**, 689–701.

Przeworski M (2002) The signature of positive selection at randomly chosen loci. *Genetics*, **160**, 1179–1189.

Ruddiman W (1998) Geology: early uplift in Tibet? *Nature*, **394**, 723–725.

Shi Y (2002) Characteristics of late Quaternary monsoonal glaciation on the Tibetan Plateau and in East Asia. *Quaternary International*, **97–8**, 79–91.

Shimono A, Ueno S, Gu S *et al.* (2010) Range shifts of *Potentilla fruticosa* on the Qinghai-Tibetan Plateau during glacial and

interglacial periods revealed by chloroplast DNA sequence variation. *Heredity*, **104**, 534–542.
Slatkin M, Maddison WP (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, **123**, 603–613.
Song BH, Wang XQ, Wang XR, Ding KY, Hong DY (2003) Cytoplasmic composition in *Pinus densata* and population establishment of the diploid hybrid pine. *Molecular Ecology*, **12**, 2995–3001.
Strasburg JL, Rieseberg LH (2010) How robust are ''isolation with migration'' analyses to violations of the IM model? A simulation study *Molecular Biology and Evolution*, **27**, 297–310.
Strasburg JL, Kane NC, Raduski AR *et al.* (2011) Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Molecular Biology and Evolution*, **28**, 1569–1580.
Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
Thomas WY (2012) VGAM: vector generalized linear and additive models. R package version 0.8-6. URL http://CRAN.R-project.org/package=VGAM.
Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, **25**, 4876–4882.
Wachowiak W, Palme AE, Savolainen O (2011) Speciation history of three closely related pines *Pinus mugo* (T.), *P. uliginosa* (N.) and *P. sylvestris* (L.). *Molecular Ecology*, **20**, 1729–1743.
Wang XR, Szmidt AE (1990) Evolutionary analysis of *Pinus densata* (Masters), a putative tertiary hybrid. 2. A study using species specific chloroplast DNA markers. *Theoretical and Applied Genetics*, **80**, 641–647.
Wang XR, Szmidt AE (1994) Hybridization and chloroplast DNA variation in a *Pinus* species complex from Asia. *Evolution*, **48**, 1020–1031.
Wang XR, Szmidt AE, Lewandowski A, Wang ZR (1990) Evolutionary analysis of *Pinus densata* (Masters), a putative tertiary hybrid. 1. Allozyme variation. *Theoretical and Applied Genetics*, **80**, 635–640.
Wang XR, Szmidt AE, Savolainen O (2001) Genetic composition and diploid hybrid speciation of a high mountain pine, *Pinus densata*, native to the Tibetan Plateau. *Genetics*, **159**, 337–346.
Wang LY, Abbott RJ, Zheng W *et al.* (2009) History and evolution of alpine plants endemic to the Qinghai-Tibetan Plateau: *Aconitum gymnandrum* (Ranunculaceae). *Molecular Ecology*, **18**, 709–721.
Wang H, Qiong L, Sun K *et al.* (2010) Phylogeographic structure of *Hippophae tibetana* (Elaeagnaceae) highlights the highest microrefugia and the rapid uplift of the Qinghai-Tibetan Plateau. *Molecular Ecology*, **19**, 2964–2979.
Wang B, Mao JF, Gao J, Zhao W, Wang XR (2011) Colonization of the Tibetan Plateau by the homoploid hybrid pine *Pinus densata*. *Molecular Ecology*, **20**, 3796–3811.
Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.
Willyard A, Syring J, Gernandt DS, Liston A, Cronn R (2007) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Molecular Biology and Evolution*, **24**, 90–101.
Woerner AE, Cox MP, Hammer MF (2007) Recombination-filtered genomic datasets by information maximization. *Bioinformatics*, **23**, 1851–1853.
Wu CL (1956) The taxonomic revision and phytogeographical study of Chinese pines. *Acta Phytotaxonomica Sinica*, **5**, 131–163 (in Chinese).
Xu TT, Abbott RJ, Milne RI *et al.* (2010) Phylogeography and allopatric divergence of cypress species (*Cupressus* L.) in the Qinghai-Tibetan Plateau and adjacent regions. *BMC Evolutionary Biology*, **10**, 194.
Yang FS, Li YF, Ding X, Wang XQ (2008) Extensive population expansion of *Pedicularis longiflora* (Orobanchaceae) on the Qinghai-Tibetan Plateau and its correlation with the Quaternary climate change. *Molecular Ecology*, **17**, 5135–5145.
Yue JP, Sun H, Baum DA *et al.* (2009) Molecular phylogeny of *Solms-laubachia* (Brassicaceae) s.l., based on multiple nuclear and plastid DNA sequences, and its biogeographic implications. *Journal of Systematics and Evolution*, **47**, 402–415.
Zeng K, Fu YX, Shi SH, Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, **174**, 1431–1439.
Zheng BX, Rutter N (1998) On the problem of quaternary glaciations, and the extent and patterns of Pleistocene ice cover in the Qinghai-Xizang (Tibet) Plateau. *Quaternary International*, **45–6**, 109–122.
Zhisheng A, Kutzbach JE, Prell WL, Porter SC (2001) Evolution of Asian monsoons and phased uplift of the Himalayan Tibetan Plateau since Late Miocene times. *Nature*, **411**, 62–66.
Zhou SZ, Wang XL, Wang J, Xu LB (2006) A preliminary study on timing of the oldest Pleistocene glaciation in Qinghai-Tibetan Plateau. *Quaternary International*, **154**, 44–51.

J.G. and B.W. performed the sequencing and data analyses and drafted the manuscript. J.-F.M. collected the samples and participated in manuscript writing. P.I. participated in data analyses and manuscript writing. Q.-Y.Z. instructed the sequencing experiment and participated in manuscript writing. X-R.W. designed the research, analysed data and wrote the manuscript.

Data accessibility

DNA sequence: GenBank accessions nos JQ070979–JQ071373.

IMa2 input and reference files: DRYAD entry doi:10.5061/dryad.566n8.

Supporting information

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

Table S1 Geographical location, sample sizes (N), the number of segregating sites (S), nucleotide polymorphism (θ_w), nucleotide diversity (π_t, total sites; π_s, silent sites; π_a, nonsynonymous), number of haplotypes (n_h) and cytoplasmic diversity (H_e) of the 19 populations of the three pine species.

Table S2 Descriptions of the eight investigated loci.

Table S3 Prior distributions of the demographic parameters for instantaneous size change (S), exponential growth (G) and bottleneck model (B).

Table S4 Neutrality test at each locus as measured by Tajima's *D*, Fu and Li's *D** and *F**, Fay and Wu's *H*, and the MK test.

Table S5 Estimated demographic parameters for *Pinus yunnanensis* and *Pinus tabuliformis* by IMa2.

Table S6 Posterior distributions for the demographic parameters of the bottleneck model estimated by ABC.

Fig. S1 Schematic presentation of the three models we examined by ABC.

Fig. S2 Marginal distribution of the posterior probability of six demographic parameters estimated by IMa2 for *Pinus yunnanensis* and *Pinus tabuliformis*.

Fig. S3 Mean and variance of five summary statistics calculated from 10^5 posterior predictive simulations.

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