MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 1212–1226 doi: 10.1111/j.1365-294X.2010.04546.x

Genomic diversity, population structure, and migration following rapid range expansion in the Balsam Poplar, *Populus balsamifera*

STEPHEN R. KELLER,* MATTHEW S. OLSON,† SALIM SILIM,‡ WILLIAM SCHROEDER‡ and PETER TIFFIN*

**Department of Plant Biology, University of Minnesota, St. Paul, MN 55108, USA, †Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, AK 99775, USA, ‡Agroforestry Division, Agriculture and Agri-Food Canada, PO Box 940, No. 2 Government Road, Indian Head, Saskatchewan S0G 2K0, Canada*

Abstract

Rapid range expansions can cause pervasive changes in the genetic diversity and structure of populations. The postglacial history of the Balsam Poplar, *Populus balsamifera*, involved the colonization of most of northern North America, an area largely covered by continental ice sheets during the last glacial maximum. To characterize how this expansion shaped genomic diversity within and among populations, we developed 412 SNP markers that we assayed for a range-wide sample of 474 individuals sampled from 34 populations. We complemented the SNP data set with DNA sequence data from 11 nuclear loci from 94 individuals, and used coalescent analyses to estimate historical population size, demographic growth, and patterns of migration. Bayesian clustering identified three geographically separated demes found in the Northern, Central, and Eastern portions of the species' range. These demes varied significantly in nucleotide diversity, the abundance of private polymorphisms, and population substructure. Most measures supported the Central deme as descended from the primary refuge of diversity. Both SNPs and sequence data suggested recent population growth, and coalescent analyses of historical migration suggested a massive expansion from the Centre to the North and East. Collectively, these data demonstrate the strong influence that range expansions exert on genomic diversity, both within local populations and across the range. Our results suggest that an in-depth knowledge of nucleotide diversity following expansion requires sampling within multiple populations, and highlight the utility of combining insights from different data types in population genomic studies.

Keywords: migration, phylogeography, population structure, *Populus*, range expansion

Received 25 September 2009; revision accepted 21 December 2009; accepted 22 December 2009

Introduction

Historical changes in range expanse have had pronounced effects on species at high latitudes, where episodic movements of continental ice sheets during Pleistocene glacial cycles have forced massive restructuring of populations. Demographic events during range expansions and contractions can drive pervasive changes to genome-wide patterns of population genetic diversity (Taberlet *et al.* 1998; Hewitt 2000; Petit *et al.* 2003; Excoffier *et al.* 2009). Founder effects and genetic drift are expected to be especially strong during range expansion, causing gradients or even abrupt transitions in allele frequencies, genetic diversity, and population structure (Ramachandran *et al.* 2005; Klopfstein *et al.* 2006; Excoffier & Ray 2008; Francois *et al.* 2008; Hofer *et al.* 2009). Further, expansions can affect the geographic

Correspondence: Peter Tiffin, Fax: +1 612 625 1738;
E-mail: ptiffin@umn.edu

distribution of both neutral and selectively important genetic variation and thereby complicate inferences drawn from clines in alleles or phenotypic traits (Vasemagi 2006; Keller *et al.* 2009), genome scans for selection (Coop *et al.* 2009), and genetic association mapping (Pritchard *et al.* 2000; Price *et al.* 2006). Therefore, a thorough understanding of how expansion has shaped diversity and population structure not only provides insight into the demographic history of a species, but also is prerequisite for studying the molecular basis and evolution of locally adaptive traits in natural populations (Nielsen 2005; Wright & Gaut 2005).

Analyses of genetic diversity following the worldwide expansion of humans have revealed a complex demographic history, including bottlenecks and rapid demographic growth (Rogers 1995; Marth *et al.* 2003), serial founder events and allele frequency clines (Cavalli-Sforza *et al.* 1993; Ramachandran *et al.* 2005), population subdivision (Rosenberg *et al.* 2002; Li *et al.* 2008), and genetic drift along the expanding wavefront (Hofer *et al.* 2009). Many plant species have also experienced large historical changes in population structure and range expanse. Although there have been several surveys of nucleotide polymorphism in plants (Nordborg *et al.* 2005; Wright *et al.* 2005; Caicedo *et al.* 2007; Ingvarsson 2008), most studies have relied on a 'species-wide' sample of relatively few individuals collected across a broad geographic area. Comparatively few studies have sampled within multiple populations to show how diversity in the genome has been shaped locally by demographic events (Arunyawat *et al.* 2007; Moeller *et al.* 2007; Ross-Ibarra *et al.* 2008).

Following the last glacial maximum (*c.* 18 thousand years ago), the forest tree *Populus balsamifera* colonized the portion of north-temperate and boreal North America that was widely covered by continental ice sheets (Webb & Bartlein 1992; Williams *et al.* 2004). In this paper, we describe patterns of diversity and population structure at 412 SNPs surveyed across the *P. balsamifera* genome, and estimate historical population demography and migration using coalescent models applied to DNA sequences from 11 nuclear genomic regions. We find that despite the evidence of both historical and ongoing migration, populations are far from equilibrium and exhibit large variation in genetic diversity across the range, attributable to a recent and large-scale range expansion.

Methods

Study species and sample collection

Populus balsamifera L. (family Salicaceae) is a dioecious, and thus obligately outcrossing, deciduous tree common in floodplain and upland sites throughout boreal North America. It is widespread across the northern USA and Canada, attaining the northernmost distribution of any tree species in North America. It is fast growing and generally short-lived, with specimens reaching first reproduction in *c.* 8–10 years and rarely living longer than 200 years (Burns & Honkala 1990). Seeds are small and born on silky hairs (pappus) that facilitate long-distance dispersal by wind. Pollen also is windborne and dispersed prior to spring leaf out, allowing for potentially very long-distance gene flow.

Samples used for this study were obtained from the Agriculture Canada Balsam poplar (AgCanBap) collection. Plant tissue originated from natural populations as stem cuttings from individual trees, collected so as to minimize the possibility of sampling clonal genotypes. Stems were rooted and transplanted into common gardens in Indian Head, Saskatchewan and leaf or bud tissues were dried in silica gel prior to DNA extraction using QIAGEN DNeasy Plant Maxi kits.

SNP discovery, assay development, and genotyping

We identified potential SNPs based on data from a previous genome-wide survey of 590 nuclear genomic regions resequenced in a discovery panel of 15 individuals chosen to broadly sample the geographic diversity of the species (Olson *et al.* 2010). These 590 loci were chosen from randomly selected transcripts in the *P. trichocarpa* genome sequence (Tuskan *et al.* 2006), and primers were designed to amplify ~600 bp regions while excluding gene paralogs (details in Olson *et al.* 2010). Genomic sequences were provided to Cogenics (Morrisville, North Carolina), where a single polymorphic site per region was arbitrarily selected for development of multiplex SNP assays using Sequenom's MALDI-TOF mass spectrometry. From the 590 sequences, 512 SNPs were initially selected and 474 SNPs passed assay design with Cogenics and went into production. Of these, 423 assays produced genotype clustering patterns that allowed stringent calling of homo- vs. heterozygous genotypes.

SNP genotyping at 423 loci was conducted on total genomic DNA isolated from 474 individuals from 34 populations collected across *P. balsamifera*'s geographic range (Table S1). 408 of these loci had call rates (genotypes with no missing values) >90%, while the remaining loci all had call rates >80%. Nine loci were found to be monomorphic. We attribute this to error during sequencing or genotype misclassification, because the individuals from the discovery panel were included in this sampling and hence SNPs should be polymorphic (9/423 loci = 2.1% error rate). A similar error rate estimate (2.2%) was obtained by comparing nucleotide

data for 95 loci from the resequenced discovery panel to their respective genotypes from the SNP assays.

We tested each locus for Hardy–Weinberg equilibrium and observed two loci that significantly deviated from the observed empirical distribution of observed vs. expected heterozygosity (Fig. S1). One of these loci (*P. trichocarpa* transcript I.D. 557231) had no observed heterozygous genotypes, while the other locus (757632) had a large excess of observed heterozygotes. These two loci as well as the nine monomorphic loci were not included in subsequent analyses. The remaining 412 loci were annotated for their functional classification (synonymous, nonsynonymous, intron, UTR) using megaBLAST searches of the *P. trichocarpa* genome (build 1.1) on GenBank. All SNP data are available for download from the Poplar population genomics website (http://www.popgen.uaf.edu/).

Population-level resequencing

To complement the SNP genotypes, we haphazardly selected 11 of the 590 reference loci. We PCR amplified and sequenced these loci from 11 to 15 individuals from each of seven geographically dispersed populations (FBK, FRE, GPR, KUU, LAB, POR, and STL; Table S1 and Fig. 1). Sequences were trimmed for low quality flanking sequence, aligned, and edited manually in Aligner v3.0.0 (Codon Code Corporation, Dedham, MA, USA). Heterozygous sites were called automatically based on secondary peak height using Aligner's algorithm ('mutation detection preferences' set to 33% overlap). All polymorphic sites and regions of low sequence quality were visually examined for quality control of base calls. Insertion–deletion polymorphisms (indels) were fairly common in these regions, including some

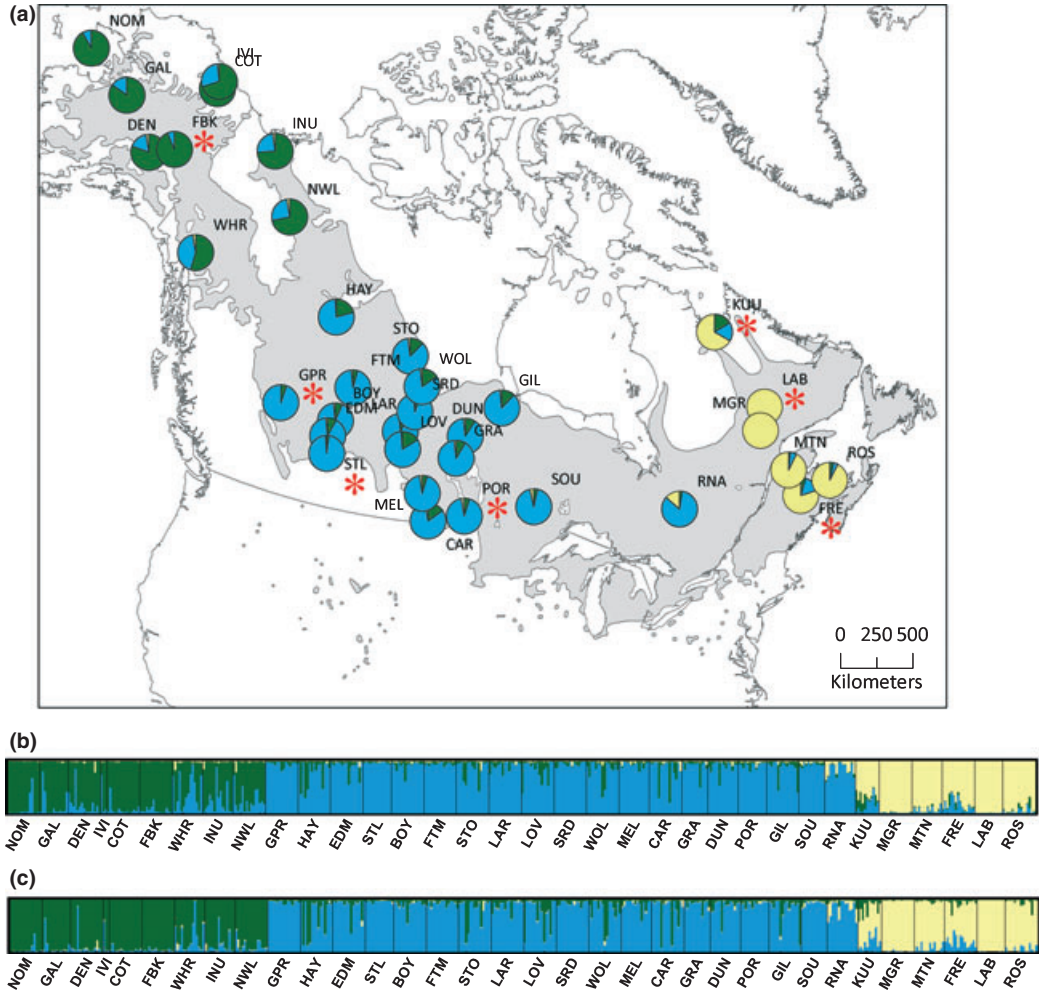

Fig. 1 Bayesian clustering of 412 SNP loci across 474 individuals and 34 populations. (a) Population frequencies of deme assignments from the INSTRUCT K = 3 model. Gray shading denotes the geographic range of P. balsamifera. Individual assignments to K = 3 demes from (b) INSTRUCT (c) STRUCTURE. Asterisks denote populations included in the 11 sequence loci dataset.

individuals that were heterozygous for indel polymorphisms. Indels were reconstructed by inference from information from forward and reverse sequences and the locations at which overlapping chromatograms began. Because evolution of indel polymorphism is complex and difficult to ascribe to a particular mutation model for incorporation into coalescent-based analysis of sequence evolution, we chose to remove indels from the alignments prior to further analysis. All sequences were deposited in GenBank under Accession nos GU380356–GU381373.

Data analyses

We tested for population structure using two Bayesian model-based clustering methods. First, we used the correlated allele frequency model (*F*-model) with admixture implemented in STRUCTURE v2.2 (Pritchard *et al.* 2000; Falush *et al.* 2003). The *F*-model corresponds to a demographic scenario of simultaneous divergence of subpopulations from an ancestral population, with each subpopulation undergoing genetic drift in allele frequencies at a unique rate inversely proportional to its effective size (Falush *et al.* 2003). We conducted 10 replicate runs for each value of K (the number of demes) from 1 to 20, with no prior placed on the population of origin. Each run consisted of a burn-in of 50 000 iterations, followed by data collection over 200 000 iterations. We evaluated inference of K using multiple methods: the ad-hoc method of Pritchard *et al.* (2000), which favours the model with the highest $\ln \Pr(X \mid K)$, and the ΔK method (Evanno *et al.* 2005) which favours the model with the greatest second-order rate of change in $\ln \Pr(X \mid K)$. We also tested for population structure using the INSTRUCT algorithm (Gao *et al.* 2007), which relaxes STRUCTURE's assumptions of Hardy–Weinberg equilibrium within clusters and computes the Deviance Information Criterion (DIC) to provide a more formal means of model selection. Following established guidelines, we considered a difference in DIC (ΔDIC) between the best and next best model >>2 as indicative of substantial support for the best model (Spiegelhalter *et al.* 2002). Settings for INSTRUCT runs were identical to those described for STRUCTURE, except that we allowed INSTRUCT to estimate the population-level inbreeding coefficient (F_{IS}). Two independent runs were performed for each value of K from 1 to 5. For both STRUCTURE and INSTRUCT runs, we conducted model averaging of individual ancestry coefficients across replicates and calculated the average pairwise similarity of individual assignments across runs (H') using CLUMPP (Jakobsson & Rosenberg 2007) and plotted using DISTRUCT (Rosenberg 2004). For comparison to the Bayesian clustering analyses, we also conducted a principal components analysis (PCA) on the population allele frequencies using the R PACKAGE of P. Legendre.

To assess how diversity within and among populations varied across the range, we calculated several summary statistics based on SNP allele frequencies. First, we quantified the degree of genetic divergence among populations (F_{ST}) using FSTAT v2.9.4 (Goudet 1995). Significance was determined from 2000 random permutations. We also used FSTAT to determine the magnitude of diversity contained within each population, estimated as the expected heterozygosity (H_e). To assess how SNP diversity varied among different historically isolated regions and to look for putative refugia harbouring high diversity, we grouped populations according to their majority deme membership based on results from Bayesian clustering and calculated the average H_e within these demes. Among-deme differences in H_e and F_{ST} were tested by 2000 random permutations of populations among demes. A hierarchical partitioning of genetic variance among demes, populations, and individuals was performed using analysis of molecular variance (AMOVA) in Arlequin v3.1 (Excoffier *et al.* 2005). The observed number of private SNPs per deme and the expected number after sample size correction based on rarefaction were calculated using HP-RARE (Kalinowski 2005). Finally, we interpolated range-wide gradients in H_e and polymorphic loci using kriging in ARCGIS v9.2 (ESRI, Redlands, CA, USA).

Insights into a species' demographic history can be described by the distribution of allele frequencies across loci – the Site Frequency Spectrum (SFS) (Hein *et al.* 2005). However, ascertained SNP data must be corrected for bias if the data are to provide an accurate representation of the true SFS (Nielsen *et al.* 2004). The maximum likelihood approach of Nielsen *et al.* (2004) reconstitutes the true (unobserved) frequency distribution based on the probability of an allele being polymorphic in a discovery sample of size d haploid sets of chromosomes, given its observed frequency in a final sample of n haploid sets of chromosomes (where $d < n$). Importantly, the method makes no distributional assumptions about the data or the demographic history of the population from which the SNPs were sampled (Nielsen *et al.* 2004), and is robust to population structure, provided the discovery sample is not geographically restricted relative to the final sample (Rosenblum & Novembre 2007). We calculated the SFS based on the frequency of the minor allele at each locus, and corrected for ascertainment bias using the equations of Nielsen *et al.* (2004) adapted for the folded SFS, i.e., estimates of the minor allele when the derived allelic state is unknown (Rosenblum & Novembre 2007). We set $d = 30$, given our discovery sample of 15 diploid individuals. We chose to use a folded SFS because

P. trichocarpa is the only outgroup taxon for which we have widespread coverage at our SNP loci, but *P. trichocarpa* and *P. balsamifera* are recently diverged. Models of isolation with migration between these two species support shared ancestral polymorphism due to incomplete lineage sorting and ongoing introgression (N. Levsen, personal communication), preventing reliable inference of the ancestral allelic state. For comparison to a neutral equilibrium model of no growth, we also calculated the folded SFS expected in a neutral equilibrium Wright-Fisher population (eqn 14 of Nielsen *et al.* 2004). Because the method does not explicitly account for the effects of selection, we report the SFS for the full data set and for just the silent SNPs (synonymous and noncoding). A MATLAB script implementing the ascertainment bias correction is available from http://www.popgen.uaf.edu/.

We used the DNA sequence loci to gain additional insights into the molecular diversity and evolutionary history of *P. balsamifera* that was not possible from the SNP loci. Our diploid sequence haplotypes were first phased using PHASE v2.1 with default parameters (Stephens *et al.* 2001; Stephens & Scheet 2005), resulting in 97% of haplotypes pairs with >90% posterior probability. We then estimated the number of segregating sites, Watterson's θ, and nucleotide diversity (π) at replacement and silent sites (synonymous and noncoding) using DnaSP v.5 (Librado & Rozas 2009). We also used DnaSP to test for departure from demographic equilibrium in the SFS by calculating Tajima's *D* for each locus and population and comparing the observed values to the distribution from 5000 simulations of the neutral coalescent without recombination. We recognize that such an approach is not robust as a test of selection, but our intent here is to test general departures from an equilibrium population. Historical population size and migration rates were estimated from the phased haplotypes using Bayesian Markov Chain Monte Carlo (MCMC) coalescent modelling implemented in LAMARC v2.1.3 (Kuhner 2006, 2009). Bayesian MCMC methods provide parameter estimates based on full likelihood estimation at the expense of the flexibility and decreased computation time of approximate likelihood approaches such as ABC (Beaumont *et al.* 2002). For each of the three major demes identified by our clustering analysis, we estimated the scaled population mutation rate ($\theta = 4Ne\mu$), population growth (g, where $\theta_t = \theta_{\text{present}} \exp(-gt)$) and pairwise migration ($M = m/\mu$), as well as the recombination parameter ($r = \rho/\mu$, where ρ is the estimated frequency of recombination per site per generation). We used the Felsenstein 84 model of evolution, and set the transition:transversion ratio to 2. Uniform priors were placed on θ[0, 0.01], g[−500, 10 000], and M[0, 10 000]. Three independent MCMC runs of varying length and burn-in were conducted and produced similar results; here, we present results from the longest run, which consisted of an initial chain of 2×10^4 iterations (burn-in of 1×10^3) followed by a long chain of 10^6 iterations (burn-in of 5×10^3, sampling every 20 iterations). Using TRACER v1.5, (http://beast.bio.ed.ac.uk/Tracer), we observed convergence of the likelihood in MCMC chains for all loci following the burn-in.

To examine more recent migration among demes and for comparison to the historical migration estimated under the coalescent, we conducted assignment tests on the multilocus SNP genotypes using STRUCTURE. Assignment tests were implemented in a separate STRUCTURE run with an informative prior placed on deme membership based on the value of *K* resulting from model selection. We allowed for detection of migrants up to two generations before present (option GENSBACK = 2). Run times and burn-in were as described previously.

Results

Population structure

Based on the 412 SNPs from 474 individuals, STRUCTURE revealed increasing model likelihoods as the number of demes (*K*) increased from 2 through 8, with lower likelihoods and higher variance among runs for $K > 8$ (Table 1). Model selection based on the ΔK criterion supported $K = 2$, while Pritchard's ad-hoc method supported $K = 8$ (with $P = 0.997$). Clustering of SNP genotypes using the INSTRUCT algorithm showed an increase in model likelihood with increasing *K*, but the deviance information criterion (DIC) clearly supported $K = 3$, with the next closest model ($K = 2$) receiving substantially less support (ΔDIC = 714; Table 1). Individual ancestry coefficients were highly consistent across replicate runs within STRUCTURE and INSTRUCT, as well as highly congruent between the two programs (Fig. 1). Average pairwise similarity (H') = 0.877 for values of *K* from 2 to 8, while H' was slightly higher ($H' = 0.993$) for $K \leq 3$. Based on the pattern and consistency of individual assignments, and the more formal model selection criteria provided by DIC, we consider $K = 3$ to capture most of the biologically relevant information in the data.

The three demes identified by INSTRUCT showed a striking biogeographic pattern, with a Northern deme consisting of 9 populations from Alaska and the Yukon territory, a large Central deme containing 19 populations from across western and central Canada, and an Eastern deme containing 6 populations from Ontario eastward (Quebec, Labrador, New Brunswick; Fig. 1).

Table 1 Bayesian model-based clustering likelihoods and model selection for the number of demes (*K*) present in the SNP data set. Vales of $K > 10$ are not reported due to their low likelihood

	Structure				Instruct	
K	Average ln Pr($X \mid K$)*	SD ln Pr($X \mid K$)†	Pr (*K*)‡	Δ*K*§	LL¶	DIC**
1	−129024	0	0.000		−102603	205588
2	−125562	10	0.000	**282.69**	−98397	202739
3	−124901	38	0.000	4.75	−97129	**202025**
4	−124422	333	0.000	0.37	−95261	204488
5	−123822	22	0.000	6.61	−93693	205503
6	−123079	204	0.000	2.83		
7	−122915	166	0.003	0.96		
8	−122909	593	**0.997**	1.87		
9	−124012	1513	0.000	1.05		
10	−123527	1399	0.000	1.02		

*Averaged over 10 replicate runs.
†Standard deviation over 10 replicate runs.
‡Ad-hoc model selection method of Pritchard *et al.* (2000).
§Model selection method of Evanno *et al.* (2005).
¶Log likelihood of the data averaged over two runs.
**Model selection method according to the Deviance Information Criterion (DIC) of Gao *et al.* (2007). The best model within each class of selection criteria is indicated in bold.

These groupings were also identified by STRUCTURE for $K = 3$. When $K = 2$ (identified as the best model using the *post hoc* Δ*K* method), both STRUCTURE and INSTRUCT collapsed the Central and Northern demes into one, while the Eastern deme remained distinct. However, we note that three demes were also apparent in the clustering of populations in PCA space, including a clear separation of the Northern and Central demes (Fig. S3).

The three demes differed in the magnitude of divergence from the reconstructed ancestral population under a STRUCTURE model of simultaneous splitting followed by drift (Falush *et al.* 2003). Compared to ancestral allele frequencies, the Central deme differed minimally ($F_{ST} = 0.008$), whereas the Northern deme was slightly more diverged ($F_{ST} = 0.031$) and the Eastern deme had diverged substantially ($F_{ST} = 0.196$). The large divergence was accompanied by evidence from INSTRUCT for a significant inbreeding coefficient within the Eastern deme ($F_{IS} = 0.072$, 95% credible interval 0.043–0.095), probably the result of population substructure within this region (e.g. a Wahlund effect). In contrast, there was no evidence that F_{IS} differed from zero within the Northern and Central demes (95% credible intervals overlapped zero in each case).

Divergence in SNP frequencies among present day populations was low, but significant (mean $F_{ST} = 0.053$; 95% confidence interval 0.048–0.058), and showed a multimodal distribution suggesting the presence of higher-order genetic structure, consistent with results from Bayesian clustering (Fig. 2). Hierarchical partitioning of SNP diversity with AMOVA showed 4.4% of the genetic variance was distributed among demes, while 2.4% was among populations within demes (Table S3). Pairwise population divergence (F_{ST}) was generally low within the Central (0.009) and Northern (0.032) demes, while Eastern populations were more highly diverged from one another (0.090), in agreement with the substructure identified by INSTRUCT. This resulted in significant heterogeneity among demes in the magnitude of F_{ST} (permutation test: $P = 0.034$). Between demes, the largest divergences involved comparisons with eastern populations (especially North vs. East), while divergences between Central and Northern populations were much lower (Fig. 2; F_{ST} between each pair of populations is reported in supplementary Table S2).

Genomic diversity

SNPs. Of the 412 SNP loci, 112 (27%) were nonsynonymous, 135 (33%) were synonymous, and the remaining 165 loci were located in either introns (126), 5′ or 3′ UTRs (11), or apparent pseudogenes (28 loci). Within-population polymorphism and expected heterozygosity (H_e) varied geographically; central populations in southwestern Canada had notably high H_e, while all Eastern and several Northern populations located outside the main contiguous portion of the range had noticeably lower H_e (Fig. 3A). SNP diversity varied significantly among demes (permutation test: $P < 0.001$), with diversity greatest within Central and Northern populations (mean $H_e = 0.207$ and 0.199, respectively) and less in Eastern populations (mean = 0.175). The number of

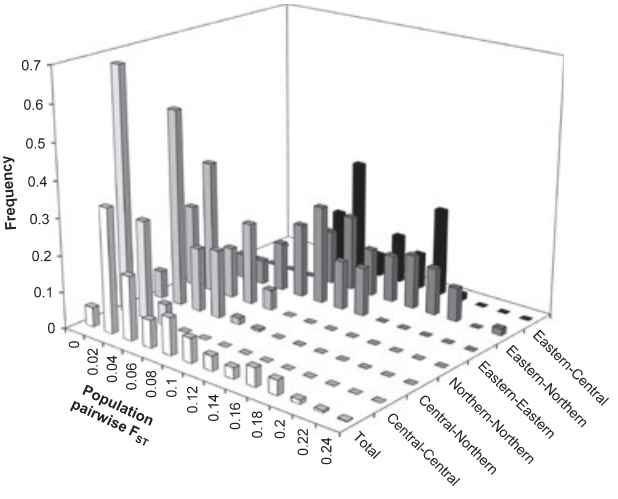

Fig. 2 Pairwise genetic divergence among populations within and among demes.

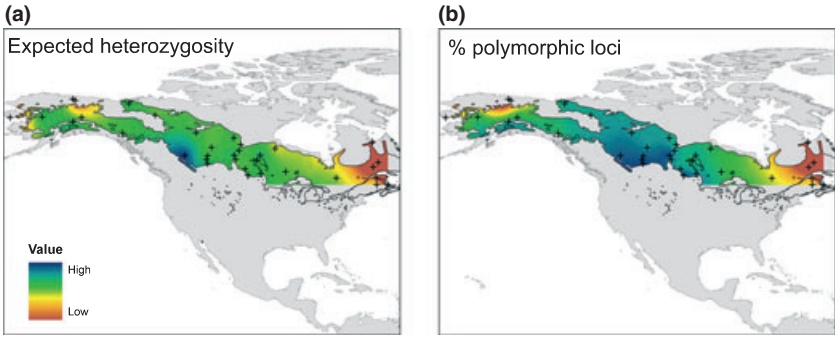

Fig. 3 Geographic gradients in diversity at SNP loci within populations, as summarized by (a) expected heterozygosity and (b) the percentage of polymorphic loci. Crosses mark population locations.

private SNPs (before and after rarefaction) and the percentage of polymorphic loci were greater in Central populations than in Northern or Eastern populations (Table 2). These patterns created range-wide gradients in allelic polymorphism, with diversity highest in the Centre of the range and declining towards the Northern and Eastern peripheries (Fig. 3B).

Among the global sample, the observed SNP Site Frequency Spectrum (SFS) showed a deficit of low frequency alleles and an excess of mid to high frequency alleles, consistent with ascertainment bias during SNP discovery (Fig. 4A). After correcting for ascertainment bias, the range-wide SFS exhibited an excess of low frequency alleles compared to neutral equilibrium expectations (Fig. 4A). Within demes, the bias-corrected SFS showed a large excess of low frequency variants in the Central and Eastern demes, consistent with a very recent population expansion, although it is possible that population substructure within the Eastern deme may be partially contributing to the abundance of rare alleles (Fig. 4B). The Northern deme also showed an excess of rare alleles, although of lesser magnitude than the other two demes, as well as a slight excess of mid-frequency alleles relative to equilibrium expectations.

DNA sequences. The sequence data from the 11 loci revealed 112 segregating sites across ~6 kb (Table 3). Nucleotide diversity estimates based on segregating sites (θ_W = 0.0033) and pairwise differences (π_{total} = 0.0020) were of similar magnitude, whereas diversity at replacement sites (π_{rep} = 0.0008) was lower than at silent sites (π_{silent} = 0.0030), consistent with a history of purifying selection.

Table 2 Private alleles within demes

	Central	Eastern	Northern
# Fixed SNPs*	12	73	50
# Polymorphic SNPs	400	339	362
# Private SNPs	22	10	2
# Private SNPs, rarefaction†	16.4	15.5	6.7
# Private haplotypes‡	44	14	5

*Based on 412 SNP loci polymorphic in the global sample.
†Based on a standardized sample of six populations per deme and 30 haploid genotypes per population.
‡Summed values over 11 sequence loci.

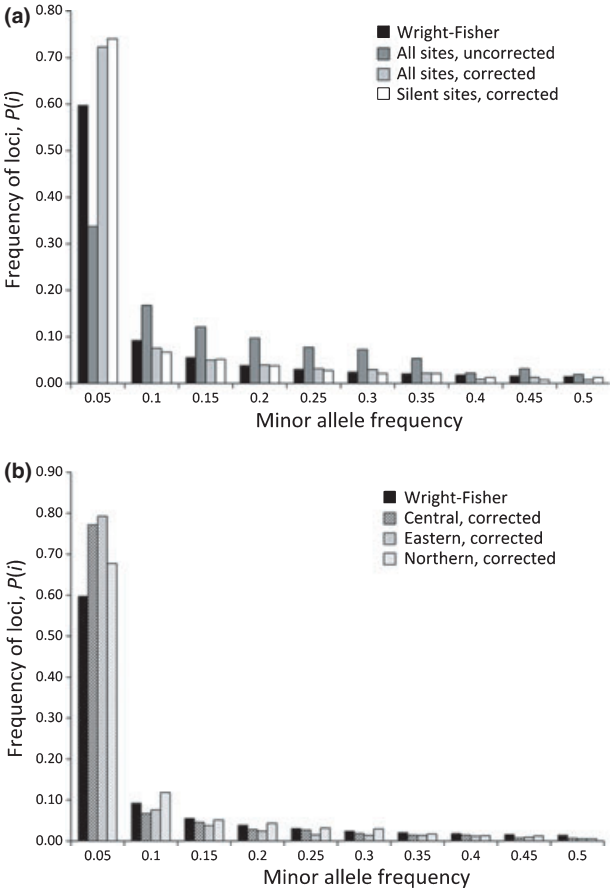

Fig. 4 Folded site frequency spectrum (SFS) across SNP loci, before and after ascertainment correction. Wright-Fisher refers to SFS expectations under neutral equilibrium. (a) Global SFS, (b) SFS by deme.

We obtained coalescent estimates of nucleotide diversity that estimated the input of population demography, migration, and recombination using LAMARC. The mutation-scaled recombination rate (ρ/μ) was r = 0.22 (95% credible interval 0.0001–0.53), indicating the effect of

Table 3 Molecular diversity at 11 nuclear sequence loci

Locus	N*	Length (bp)	Silent sites	Nonsyn sites	S†	θ_W	π_{total}	π_{silent}	π_{rep}
174078	186	475	364.0	82.5	9	0.0033	0.0011	0.0014	0.0000
195487	186	623	326.7	293.3	8	0.0022	0.0006	0.0011	0.0001
230673	178	590	540.5	49.5	14	0.0041	0.0036	0.0039	0.0000
554813	188	526	122.7	402.3	8	0.0026	0.0014	0.0035	0.0008
554898	186	634	558.7	71.3	14	0.0038	0.0034	0.0037	0.0006
558978	188	505	325.3	178.7	10	0.0034	0.0012	0.0004	0.0026
560714	188	631	383.0	246.0	4	0.0011	0.0001	0.0001	0.0001
563785	184	498	379.0	118.3	9	0.0031	0.0035	0.0045	0.0006
566362	186	516	114.4	398.7	7	0.0023	0.0020	0.0002	0.0026
567670	178	494	118.5	370.5	10	0.0035	0.0014	0.0049	0.0002
588180	188	507	188.6	316.4	19	0.0065	0.0042	0.0090	0.0013
Average						0.0033	0.0020	0.0030	0.0008

*Number of phased haploid sequences.
†Number of segregating sites.

recombination on haplotype diversity was low relative to mutation. Multilocus estimates of nucleotide diversity exhibited large variation among demes (Fig. 5A). The Bayesian most probable estimates showed that population diversity was highest in the Central deme (θ = 0.0028, 95% credible interval: 0.0020–0.0049), and was significantly greater than diversity found in either the Northern (θ = 0.0008, 0.0005–0.0021) or Eastern demes (θ = 0.0004, 0.0003–0.0007). Assuming the neutral mutation rate does not vary among demes, this translates into effective population sizes that are 3.5 to 7 times larger in the Centre than in the North or East, respectively.

The frequency distribution of polymorphisms from the sequence loci was skewed towards rare variants, as indicated by negative values of Tajima's D (Tables 4 and S4), in agreement with the excess of low frequency alleles observed at the SNP loci. However, D also varied among the demes, suggesting potentially different demographic histories; D was most negative in the Centre but near zero in both the East and North (Table 4). When integrated across loci using the demographic model in LAMARC, all three demes showed small but positive values of g, consistent with recent exponential population growth. The variances of the posterior distributions of these estimates were large, however, and credible intervals included zero in each deme (Fig. 5B).

Historical and contemporary migration

Coalescent-based estimates from LAMARC indicated that the input of polymorphism due to historical migration among demes was very high relative to the mutation rate ($M = m/\mu$). The most probable estimates of M ranged from 1216 to 9598, with the highest migration observed out of the Centre and into the North and East (M = 9572–9578). Both pathways of migration from the Centre exhibited broad posterior distributions that clearly excluded values of $M < 2000$, but were otherwise consistent with a range of very high migration values, including modal values that were near the upper bound of the prior ($M \sim 10\,000$, Fig. 5C). In contrast, estimates of migration into the Centre from the North and East were much lower (M = 1216–1684) and showed narrower posterior distributions. Migration between the North and East was also low but significantly greater than zero (Fig. 5C). The direction of migration was asymmetrical, with the Eastern deme experiencing a greater input of polymorphism due to migration (M = 2732) than the North (M = 1684).

The high rates of historical migration inferred from the sequence genealogies were largely consistent with recent gene flow inferred by assignment tests on the SNP genotypes. Within a given demographic population, most individuals had ancestry that assigned strongly to the same deme (Fig. 1). However, we detected 25 individuals whose multilocus genotypes had >50% probability of ancestry in a deme different from their resident population—likely due to recent immigration (Table S5). Most of these migration events involved exchange with the Eastern deme (23 of 25), including 16 migrants from Central or Northern demes found within an Eastern population and 7 migrants originating in the Eastern deme but sampled elsewhere. Migrants were commonly observed within populations that were in close proximity to populations from different demes (e.g., between the 'RNA' population and Eastern deme populations). In other cases, observed migrants were quite distant from the nearest sampled population belonging to the migrant's deme, including eight individuals identified as migrants between the North and East, echoing the inference of trans-continental migration identified by LAMARC.

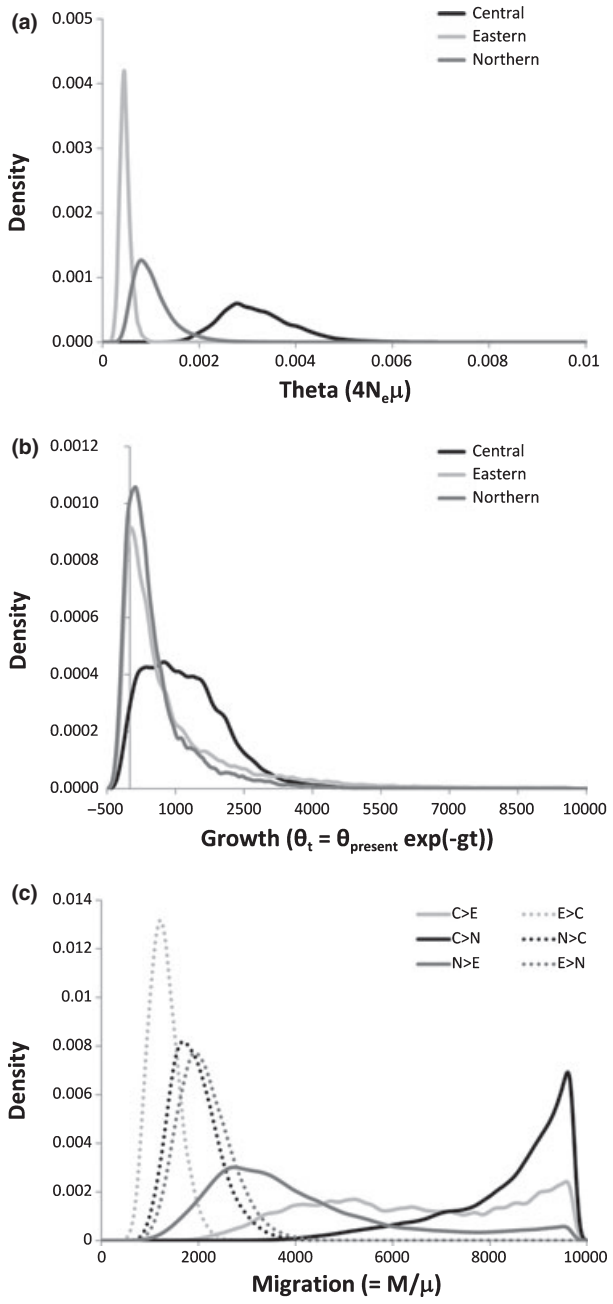

Fig. 5 Posterior density distributions from LAMARC for (a) population size, (b) growth, and (c) migration among demes. The legend for (c) gives migration from deme I into deme J (I>J).

Table 4 Tajima's *D* estimated by deme and pooled across the entire sample. Statistical tests were made against a standard neutral model using 5000 coalescent simulations

Locus	Central	Eastern	Northern	Overall
174078	−2.020**	0.218	−1.416†	−1.368†
195487	−1.415†	−0.206	−1.022	−1.619*
230673	−0.372	0.027	1.301	−0.328
554813	−1.238	−1.417†	0.800	−1.038
554898	0.299	−1.165	0.813	−0.304
558978	−1.317†	−1.491*	0.983	−1.533*
560714	−1.639*	−0.907	n/a‡	−1.656**
563785	−0.052	1.203	1.257	−0.308
566362	−0.131	0.465	0.638	−0.299
567670	−1.131	−0.321	−1.450†	−1.451*
588180	−1.542*	2.033*	0.048	−0.951
Average	−0.960	−0.142	0.195	−0.931

† $P < 0.1$, *$P < 0.05$, **$P < 0.01$.
‡Locus 560714 was fixed in the Northern deme.

Discussion

Population structure and diversity

Populus balsamifera is an ecologically important tree species with a widespread geographic range throughout much of northern North America, portions of which were under continental ice sheets during the last glacial maximum. Therefore, most present day populations result from a large-scale range expansion that has occurred since the last glacial maximum *c.* 18 000 years ago. In order to characterize this expansion and its effects on genetic diversity and population structure, we assayed each of 474 individuals sampled from throughout the species' range at 412 SNP loci, and sequenced 11 nuclear loci from 94 individuals. Based on the pattern of covariance among SNPs, Bayesian clustering identified three major genetically and spatially separated demes in the Northern, Central, and Eastern parts of the range. The distinctiveness of these demes was evident from the distribution of genetic diversity: demes differed in expected heterozygosity, levels of polymorphism and private alleles, frequency distribution of polymorphic sites, effective population size, and relative contribution to the global pool of migrants.

Bayesian clustering analyses of the SNP data revealed evidence for population structure, but the most probable number of demes depended upon the method used to select *K*. Model selection is a critical component of Bayesian clustering analyses, yet is still largely implemented as ad-hoc summaries of the likelihood (Pritchard *et al.* 2000) or its rate of change with *K* (Evanno *et al.* 2005). One advantage of the INSTRUCT algorithm is its reporting of the Deviance Information Criterion (DIC) which provides a more formal assessment of the information content in the posterior distribution (i.e., model likelihood) as estimated by MCMC, while penalizing model complexity. The ad-hoc method of Pritchard supported the model with the highest ln Pr($X|K$) at $K = 8$ (Table 1); however the ΔK method favoured clustering

the data into $K = 2$ demes (Eastern vs. others), while the DIC from INSTRUCT supported separating the Northern from the Central deme at $K = 3$. Given that changes in DIC >> 2 are generally interpreted as strong support for one model over another (Spiegelhalter *et al.* 2002), the change in DIC we observed from $K = 2$ to 3 was substantial (>700; Table 1). This suggests a large gain in information content was achieved by separating out the Northern and Central demes. This separation seems justified biologically by the geographic isolation of the two demes and their very different patterns of polymorphism. Moreover, the individual ancestry coefficients suggested little gain in meaningful structure was apparent at higher values of K (Fig. S3). We also note that a principal components analysis clearly separated all three groups, including North vs. Central (Fig. S3). Two points can be made here. First, model selection is a critical aspect of Bayesian clustering, and difficult problems (e.g. low divergence) may benefit from more defined statistical approaches such as DIC. Second, the matrices of individual ancestry coefficients from STRUCTURE and INSTRUCT were highly congruent, suggesting that while issues of model selection are crucial, the inference of ancestry for a given level of K is robust to clustering method, at least for these data.

Several aspects of the SNP data suggest that the ancestral population from which the three current demes were derived was most similar to the Central deme. The Central deme exhibited the least divergence from the ancestral allele frequencies reconstructed by STRUCTURE ($F_{ST} < 0.01$), lacked population substructure (mean $F_{ST} = 0.009$), and showed the highest levels of expected heterozygosity, percentage of polymorphic SNPs, and private SNPs. As a result, measures of within-population diversity exhibited geographical gradients away from the Centre, extending outwards towards the margins of the range (Fig. 3). Coalescent analyses on DNA sequences showed that the Central deme had an effective population size that was 3–7 times larger than in the North or East, while also producing the greatest outflow of migration into neighbouring demes (Fig. 5).

Admixture is known to generate regions of high genetic diversity along pathways of expansion where lineages from different refugial populations meet and exchange alleles (Widmer & Lexer 2001; Petit *et al.* 2003). However, admixture is unlikely to explain the high diversity we found in the Central deme for two reasons. First, the frequency of private alleles is clearly elevated in the Central deme (Table 2), an outcome inconsistent with admixture blending diversity (Widmer & Lexer 2001). Second, admixture between demes should result in genotypes with mixed ancestry coefficients in the Bayesian clustering analysis (e.g. Whitfield *et al.* 2006). In contrast, most genotypes from the Centre of the range assigned strongly to the Central deme, with no indication that genotypes resulted from introgression between the North and East (Fig. 1). The abundance of private alleles and absence of admixture among demes, as well as the lack of population substructure within the Central deme and high estimates of nucleotide polymorphism, indicate that the central core of the species' range consists of a large set of genetically diverse populations with little evidence of admixture, drift, or founder effects, supporting the Central deme as descended from the main demographic refuge of *P. balsamifera* during Pleistocene range restrictions. The long-term persistence of a refugial population south of the ice sheets also is consistent with biogeographical evidence from pollen records (Williams *et al.* 2004) and the phylogeography of other plant species in western North American (Soltis *et al.* 1997). However, it should also be cautioned that *P. balsamifera* is known to hybridize with congeners, and it is possible that interspecific introgression may partially contribute to the higher diversity in the Central deme. We do note that there were no highly divergent haplotypes in the 11 sequence loci, as would point to introgression (Neiman *et al.* 2009).

While present day Central populations likely descended from a refugial deme that resided south of the ice sheets and contained the primary store of diversity, present day Northern and Eastern populations exhibited lower diversity and polymorphism, smaller N_e, and received a high frequency of migrants from the Central deme (Figs 3 and 5). Divergence among local populations was also higher in the North and especially the East, as expected from founder effects during colonization (Whitlock & McCauley 1999). These results are in agreement with population genetic predictions for range expansions, wherein founder effects and the establishment of marginal populations cause gradients in genetic diversity, migration, and population structure away from the Centre towards to the edges of the range (reviewed in Excoffier *et al.* 2009).

While our data suggest that genotypes in the Eastern and Northern demes were derived from the Central deme during expansion, it does not appear that the Eastern and Northern demes have had identical histories. Eastern populations harbour a large percentage of fixed loci and exhibit dramatically lower SNP diversity, higher population substructure, and greater allele frequency divergence with populations from other demes (range in F_{ST}: 0.074–0.231; Fig. 2). These patterns are consistent with a hypothesis of allele surfing during range expansion, in which strong genetic drift in recently founded populations causes initially rare alleles to increase rapidly to high frequency (Edmonds *et al.*

2003; Klopfstein *et al.* 2006; Excoffier & Ray 2008; Hofer *et al.* 2009). If this hypothesis is correct, we would expect some mutations that were rare in the ancestral population to have drifted to high frequency during expansion into the East, resulting in an excess of globally rare (though not necessarily private) SNPs in Eastern populations (Klopfstein *et al.* 2006). To test this, we examined the per-deme frequency of the 50 loci with the lowest frequency in the global sample. In agreement with the prediction, Eastern populations had significantly greater numbers of rare SNPs (mean = 8.8) compared to Central (mean = 5.1) and Northern populations (mean = 2.1) (Kruskal–Wallis test = 19.31, d.f. = 2, $P < 0.0001$). Further, we might expect that some alleles would have drifted to high frequency while others remained rare (Hofer *et al.* 2009), generating higher variance in the frequency of minor alleles within Eastern populations compared to Central or Northern populations. To test this second prediction, we determined the Minor Allele Frequency (MAF) at each SNP within the global data set and calculated the among-locus variance in MAF separately for each population. Eastern populations on average had higher variance in MAF (mean $\sigma^2 = 0.040$) compared to Central (0.020) and Northern populations (0.023); a highly significant difference (Kruskal–Wallis test = 18.13, d.f. = 2, $P < 0.0001$). Thus, while most loci retained low minor allele frequencies in the Central and Northern demes and hence had low variance, a significantly larger number of SNPs increased in frequency in the East, causing higher variance among loci. We interpret the low overall diversity, the strong population substructure, the divergence in allele frequencies, the abundance of rare polymorphisms, and the high variance in MAF as collectively supporting a scenario of strong genetic drift during a recent expansion. This most likely occurred into the East after the Central region had been colonized following glacial retreat.

Like the Eastern deme, the Northern deme harbours lower diversity than the Central deme and migration appears to be largely from the Centre to the North. Unlike the East, however, the North shows no evidence of strong population substructure, abundance of rare SNPs, or elevated variance in minor allele frequency. It is unclear why the North and East should show such different patterns of diversity, but we offer three nonexhaustive scenarios. One possibility is that migration into the North occurred earlier than migration into the East, providing more time for subsequent gene flow to erase evidence of founder effects. This could have occurred along the ice-free corridor that opened between the Laurentide and Cordilleran ice sheets during the early stages of glacial retreat (*c.* 12 000 years ago), permitting plant migration into the northwest while the northeast remained largely under ice (Williams *et al.* 2004). Another possibility is that the form of dispersal differed during expansion into the North and East. Long-distance dispersal that results in population establishment ahead of the main migration front can cause strong founder effects and genetic subdivision relative to shorter dispersal distances and more contiguous expansion (Ibrahim *et al.* 1996; Petit *et al.* 2004). Thus, differences between the Northern and Eastern demes in within-population diversity and between-population divergence could have arisen if migration into the north occurred as a steady expansion while colonization of the eastern part of the range was characterized by a greater frequency of long-distance dispersal events. Lastly, much of Alaska remained ice-free during the Pleistocene and constituted an important refugium for some plant species (Hultén 1937; Anderson *et al.* 2006; Provan & Bennett 2008), possibly including one or more species of *Populus* (Williams *et al.* 2004). If there was a small refugial population in the North during the last glaciation, then gene flow may have kept Northern deme populations similar to those in the Centre, while a small population size may explain the lack of private alleles, as typically expected for a refugial population (Provan & Bennett 2008).

Historical population size and migration

Levels of nucleotide diversity indicate that *P. balsamifera* has a historically small effective population size and has experienced a recent population expansion. The 11 nuclear loci we sequenced harboured low diversity at neutral sites (mean $\pi_{silent} = 0.0030$), similar to other recent estimates for *P. balsamifera* where π at silent or synonymous sites ranged from 0.0033 to 0.0045 (Breen *et al.* 2009; Olson *et al.* 2010). Diversity was similarly low in the closely related *P. trichocarpa* where π ranged from 0.0029 to 0.0035 (Gilchrist *et al.* 2006; Tuskan *et al.* 2006). These are among the lowest values reported for other forest trees with widespread distributions (Savolainen & Pyhäjärvi 2007). By contrast, our estimates of diversity are ~20% of that found in the congener *P. tremula* ($\pi_{syn} = 0.0120$–0.0220) (Ingvarsson 2005, 2008), reflecting either a much smaller historical population size or a shift in the mutation rate since the split between the common ancestors of *P. balsamifera* and *P. tremula*. If we take our estimate of neutral diversity ($\pi_{silent} = 0.0030$) and assume a stable population size, a neutral mutation rate of ~2.5 substitutions per site per billion years (Tuskan *et al.* 2006) and a 15 year generation time (Tuskan *et al.* 2006; Ingvarsson 2008), we can estimate the historical N_e ($= \theta / 4\mu$) for *P. balsamifera* as only ~20 000 individuals. We caution that many of these assumptions are likely to be inaccurate and thus

this estimate should be interpreted as only a very rough approximation. Nevertheless, it is considerably smaller than the N_e estimated for *P. tremula* (118 000) under identical assumptions (Ingvarsson 2008).

Despite the indication of an historically small N_e in *P. balsamifera*, the abundance of low frequency alleles in both the SNP (Fig. 4) and sequence data (Table 4), as well as shallow genealogies recovered in the coalescent analysis with LAMARC (Fig. 5B), indicate a recent expansion in population size. The combination of low overall polymorphism and the shallow genealogies suggests that population growth may have been very recent, as expected given that the majority of the species' range was under ice until *c.* 10 000 years ago, and growth has not yet resulted in the introduction of many new mutations. While few of our individual estimates of Tajima's *D* were significantly different from neutral equilibrium expectations, they were often negative, especially in the large Central deme (Table 4). A recent survey of nucleotide diversity at 590 genomic-regions in *P. balsamifera* also recovered consistently negative values of Tajima's *D* at synonymous sites (mean $D_{syn} = -0.116$; Wilcoxon sign rank test: $P < 0.011$; Olson et al. 2010). Thus, there is consistent evidence for the signature of population expansion beginning to accumulate in the genome, although the magnitude of effect is small, likely due to the expansion being very recent. The weak signature of expansion highlights the need for large multi-locus data sets when assessing population history, especially when demographic events have occurred recently relative to the mutation rate.

Our analysis of historical migration among demes is also consistent with recent range expansion. LAMARC's estimates of the mutation-scaled migration rates were very high, meaning that most genetic variation is entering demes via migration rather than mutation. In fact, the posterior distributions for the highest values of *M* were truncated by the maximum value allowable for the prior, indicating that the actual values of *M* between some demes may be even higher than we estimated (Fig. 5C). The pattern of migration was also highly asymmetrical, with ~7-fold greater migration from the Central deme into the peripheral demes. Migration histories affected by very recent expansions are not well modelled by LAMARC (Kuhner 2006), especially when one or more populations are recently derived from another in the analysis (LAMARC documentation). Thus, the very high migration estimates we observed may in part reflect the sampling of ancestral (Central) diversity during colonization of the peripheries of the range, in addition to post-expansion gene flow between established populations. In contrast, migration between the Northern and Eastern demes was much lower but still highly significant (Fig. 5C). These migration estimates probably reflect rare long-distance gene flow across the northern tier of the continent. Poplar seeds and pollen are specialized for wind dispersal, and thus migration may reflect movement of seeds or pollen in high altitude wind currents. Similar cases of extreme long-distance dispersal have been reported for high-latitude plant species in Europe (Alsos *et al.* 2007).

Inferences of contemporary migration from assignment tests also suggested that recent migration has affected the geographic distribution of diversity. The observation of migrant individuals between Eastern and Northern populations is especially remarkable, and corroborates the historical migration estimated from LAMARC. For example, five individuals with large ancestry coefficients in the Northern deme were observed within the KUU population in Quebec, located at the northernmost extent of *P. balsamifera*'s distribution in eastern Canada. The contemporary migration we observed may be underestimates, given that the low divergences among demes probably limited the power of assignment tests to detect migrants (Manel *et al.* 2005). Thus, we interpret these results as strong evidence for recent and ongoing migration in *P. balsamifera*, and the potential for seed and/or pollen dispersal over very long distances.

Implications for association mapping and identifying evidence for selection

This extensive characterization of genome-wide diversity within and among populations of *P. balsamifera* revealed low but significant regional-scale population structure, likely originating during rapid expansion from an effectively small ancestral population. The level of population structure we observed in *P. balsamifera* is similar to that of other outcrossing forest trees, with most genetic variation occurring within populations. Nevertheless, even weak population structure causes gametic disequilibrium among physically unlinked alleles, thereby complicating the discovery of functionally important variation using association genetics. The three demes discovered in this study also are geographically structured by latitude and other environmental gradients; thus demic stratification in *P. balsamifera* will be especially important to control for in association analyses on phenotypic traits exhibiting clinal variation. Further, there is evidence from Eastern populations that drift during expansion acted on a subset of alleles to raise them to high frequencies. These processes can produce genomic regions of locally reduced polymorphism as well as clines in gene frequency—both canonical signatures of selection (Nielsen *et al.* 2007). Controlling for these demographic influences on diversity presents future challenges for the mapping of functionally important variation in natural populations.

Acknowledgements

We thank Jennifer Reese, Molly Peterson, and Iswariya Jayachandran for DNA sequencing, Robert Neville for assistance with MATLAB, Amanda Robertson for aid in developing Perl scripts for SNP identification and the Life Sciences Informatics group at UAF for aid in SNP discovery and website development. LAMARC runs were completed with the help of the University of Alaska Life Sciences Informatics Portal. STRUCTURE and INSTRUCT runs were completed with the help of the Computational Biology Service Unit from Cornell University, which is partially funded by Microsoft Corporation. This work was funded by NSF Plant Genome award DBI-0701911 to MSO and PT.

References

Alsos IG, Eidesen PB, Ehrich D, et al. (2007) Frequent long-distance plant colonization in the changing Arctic. *Science*, **316**, 1606–1609.

Anderson LL, Hu FS, Nelson DM, Petit RJ, Paige KN (2006) Ice-age endurance: DNA evidence of a white spruce refugium in Alaska. *Proceedings of the National Academy of Sciences, USA*, **103**, 12447–12450.

Arunyawat U, Stephan W, Stadler T (2007) Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular Biology and Evolution*, **24**, 2310–2322.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in Population Genetics. *Genetics*, **162**, 2025–2035.

Breen AL, Glenn E, Yeager A, Olson MS (2009) Nucleotide diversity among natural populations of a North American poplar (*Populus balsamifera*, Salicaceae). *New Phytologist*, **182**, 763–773.

Burns RM, Honkala BH (1990) *Silvics of North America: Volume 2. Hardwoods*. United States Department of Agriculture (USDA), Forest Service, Washington, D.C.

Caicedo AL, Williamson SH, Hernandez RD, et al. (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. *Plos Genetics*, **3**, 1745–1756.

Cavalli-Sforza LL, Menozzi P, Piazza A (1993) Demic expansions and human evolution. *Science*, **259**, 639–646.

Coop G, Pickrell JK, Novembre J, et al. (2009) The role of geography in human adaptation. *Plos Genetics*, **5**, e1000500.

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2003) Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences, USA*, **101**, 975–979.

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, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**, 347–351.

Excoffier L, Foll M, Petit RJ (2009) Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 481–501.

Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**.

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

Francois O, Blum MGB, Jakobsson M, Rosenberg NA (2008) Demographic history of European populations of Arabidopsis thaliana. *Plos Genetics*, **4**, e1000075, doi:1000010.1001371/journal.pgen.1000075.

Gao H, Williamson S, Bustamante CD (2007) A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*, **176**, 1635–1651.

Gilchrist EJ, Haughn GW, Ying CC, et al. (2006) Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. *Molecular Ecology*, **15**, 1367–1378.

Goudet J (1995) FSTAT (Version 1.2): A computer program to calculate F-statistics. *Journal of Heredity*, **86**, 485–486.

Hein J, Schierup MH, Wiuf C (2005) *Gene Genealogies, Variation, and Evolution*. Oxford University Press, Oxford.

Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Annals of Human Genetics*, **73**, 95–108.

Hultén E (1937) *Outline of the History of Arctic and Boreal Biota During the Quaternary Period; Their Evolution During and After the Glacial Period as Indicated by the Equiformal Progressive Areas of Present Plant Species*. Bokförlags aktiebolaget Thule, Stockholm.

Ibrahim KM, Nichols RA, Hewitt GM (1996) Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity*, **77**, 282–291.

Ingvarsson PK (2005) Nucleotide polymorphism and linkage disequilbrium within and among natural populations of European Aspen (*Populus tremula* L., Salicaceae). *Genetics*, **169**, 945–953.

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

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Kalinowski ST (2005) HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. *Molecular Ecology Notes*, **5**, 187–189.

Keller SR, Sowell DR, Neiman M, Wolfe LM, Taylor DR (2009) Adaptation and colonization history affect the evolution of clines in two introduced species. *New Phytologist*, **183**, 678–690.

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482–490.

Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, **22**, 768–770.

Kuhner MK (2009) Coalescent genealogy samplers: windows into population history. *Trends in Ecology & Evolution*, **24**, 86–93.

Li JZ, Absher DM, Tang H, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–1104.

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

Manel S, Gaggiotti O, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology and Evolution*, **20**, 136–142.

Marth G, Schuler G, Yeh R, et al. (2003) Sequence variations in the public human genome data reflect a bottlenecked population history. *Proceedings of the National Academy of Sciences, USA*, **100**, 376–381.

Moeller DA, Tenaillon MI, Tiffin P (2007) Population structure and its effects on patterns of nucleotide polymorphism in teosinte (*Zea mays* ssp. parviglumis). *Genetics*, **176**, 1799–1809.

Neiman M, Olson M, Tiffin P (2009) Selective histories of protease inhibitors: elevated polymorphism, purifying selection, and positive selection driving divergence of recent duplicates. *New Phytologist*, **183**, 740–750.

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

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, **8**, 857–868.

Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics*, **168**, 2373–2382.

Nordborg M, Hu TT, Ishino Y, et al. (2005) The Pattern of Polymorphism in *Arabidopsis thaliana*. *PLoS Biology*, **3**, e196.

Olson MS, Robsertson AL, Takebayashi N, et al. (2010) Nucleotide diversity and linkage disequilibrium in Balsam Poplar (*Populus balsamifera*). *New Phytologist*, in press. doi: 10.1111/j.1469-8137.2009.03174.x.

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

Petit RJ, Bialozyt R, Garnier-Gere P, Hampe A (2004) Ecology and genetics of tree invasions: from recent introductions to Quaternary migrations. *Forest Ecology and Management*, **197**, 117–137.

Price AL, Patterson NJ, Plenge RM, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

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

Provan J, Bennett KD (2008) Phylogeographic insights into cryptic glacial refugia. *Trends in Ecology & Evolution*, **23**, 564–571.

Ramachandran S, Deshpande O, Roseman CC, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences, USA*, **102**, 15942–15947.

Rogers AR (1995) Genetic-evidence for a Pleistocene population explosion. *Evolution*, **49**, 608–615.

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.

Rosenberg NA, Pritchard JK, Weber JL, et al. (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. *Journal of Heredity*, **98**, 331–336.

Ross-Ibarra J, Wright SI, Foxe JP, et al. (2008) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS One*, **3**, 1–13.

Savolainen O, Pyhäjärvi T (2007) Genomic diversity in forest trees. *Current Opinion in Plant Biology*, **10**, 162–167.

Soltis DE, Gitzendanner MA, Strenge DD, Soltis PS (1997) Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. *Plant Systematics and Evolution*, **206**, 353–373.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, **76**, 449–462.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, **7**, 453–464.

Tuskan GA, DiFazio S, Jansson S, et al. (2006) The genome of the black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Vasemagi A (2006) The adaptive hypothesis of clinal variation revisited: single-locus clines as a result of spatially restricted gene flow. *Genetics*, **173**, 2411–2414.

Webb T, Bartlein PJ (1992) Global changes during the last 3 million years – climatic controls and biotic responses. *Annual Review of Ecology and Systematics*, **23**, 141–173.

Whitfield CW, Behura SK, Berlocher SH, et al. (2006) Thrice out of Africa: Ancient and recent expansions of the honey bee, *Apis mellifera*. *Science*, **314**, 642–645.

Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: F_{ST} does not equal $1/(4Nm + 1)$. *Heredity*, **82**, 117–125.

Widmer A, Lexer C (2001) Glacial refugia: sanctuaries for allelic richness, but not for gene diversity. *Trends in Ecology & Evolution*, **16**, 267–269.

Williams JW, Shuman BN, Webb T, Bartlein PJ, Leduc PL (2004) Late-quaternary vegetation dynamics in North America: scaling from taxa to biomes. *Ecological Monographs*, **74**, 309–334.

Wright SI, Bi IV, Schroeder SG, et al. (2005) The effects of artificial selection of the maize genome. *Science*, **308**, 1310–1314.

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution*, **22**, 506–519.

S.K. studies the ecological, genetic, and evolutionary changes that accompany major transitions in geographic range size and extent, especially during biological invasions and climate change. M.O. studies the genetics of local adaptation and inference of historical demography and selection using nucleotide

diversity data. S.S. and W.S. are interested in the domestication and conservation of *Populus balsamifera* to develop cultivars for wood-based commodities, energy feedstock, and environmental services. P.T.'s research uses both population genetics and manipulative field experiments to understand how biological interactions and environmental change affect evolution.

Supporting Information

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

Table S1 Sampling localities of *Populus balsamifera*.

Table S2 Pairwise F_{ST} among populations.

Table S3 Hierarchical analysis of molecular variance (AMOVA) for 412 SNP loci genotyped for 474 individuals of balsam poplar. Samples were grouped into 34 populations, clustered into 3 regional demes identified by Bayesian clustering. Significance was assessed with 10,000 permutations of the data.

Table S4 Tajima's *D* estimated by locus and population. Populations fixed for a single haplotypes are undefined for Tajima's D and are reported as ''n/a''.

Table S5 Inference of recent migrants from STRUCTURE assignments.

Fig. S1 Tests of Hardy-Weinberg equilibrium for the 423 SNP loci. Points in red are loci that were removed from further analyses due to extreme deviation from the empirical distribution.

Fig. S2 Individual assignment results from STRUCTURE clustering for increasing models of K.

Fig. S3 Principal components analysis (PCA) of population allele frequencies at 412 SNPs. PCA was performed using the R package of P. Legendre.

Please note: Wiley-Blackwell are 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.