MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 4014–4028 doi: 10.1111/mec.12347

Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*

BRYAN C. CARSTENS,* REID S. BRENNAN,† VIVIEN CHUA,†‡ CAROLINE V. DUFFIE,†‡ MICHAEL G. HARVEY,†‡ RACHEL A. KOCH,† CALEB D. MCMAHAN,†‡ BRADLEY J. NELSON,† CATHERINE E. NEWMAN,†‡ JORDAN D. SATLER,* GLENN SEEHOLZER,†‡ KARINE POSBIC,† DAVID C. TANK§¶ and JACK SULLIVAN¶**

**Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH 43210, USA, †Department of Biological Sciences, 202 Life Sciences Building, Louisiana State University, Baton Rouge, LA 70803, USA, ‡Museum of Natural Science, 119 Foster Hall, Louisiana State University, Baton Rouge, LA 70803, USA, §College of Natural Resources, University of Idaho, Room 204D, Natural Resources Building, PO Box 441133, Moscow, ID 83844-1133, USA, ¶Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Room 441, Life Sciences South, PO Box 443051, Moscow, ID 83844-3051, USA, **Department of Biological Sciences, University of Idaho, Room 274, Life Sciences South, PO Box 443051, Moscow, ID 83844-3051, USA*

Abstract

Phylogeographic inference has typically relied on analyses of data from one or a few genes to provide estimates of demography and population histories. While much has been learned from these studies, all phylogeographic analysis is conditioned on the data, and thus, inferences derived from data that represent a small sample of the genome are unavoidably tenuous. Here, we demonstrate one approach for moving beyond classic phylogeographic research. We use sequence capture probes and Illumina sequencing to generate data from >400 loci in order to infer the phylogeographic history of *Salix melanopsis*, a riparian willow with a disjunct distribution in coastal and the inland Pacific Northwest. We evaluate a priori phylogeographic hypotheses using coalescent models for parameter estimation, and the results support earlier findings that identified post-Pleistocene dispersal as the cause of the disjunction in *S. melanopsis*. We also conduct a series of model selection exercises using IMa2, Migrate-n and ∂a∂i. The resulting ranking of models indicates that refugial dynamics were complex, with multiple regions in the inland regions serving as the source for postglacial colonization. Our results demonstrate that new sources of data and new approaches to data analysis can rejuvenate phylogeographic research by allowing for the identification of complex models that enable researchers to both identify and estimate the most relevant parameters for a given system.

Keywords: genetic structure, Illumina sequencing, phylogeography, Pleistocene refugia, solution-based capture probes

Received 6 November 2012; revision received 9 April 2013; accepted 9 April 2013

Introduction

One of the central objectives of phylogeographic research has been to investigate how climatic shifts that occurred during the Pleistocene have influenced the distribution and genetic structure of organisms (e.g. Hewitt 1996; Avise *et al.* 1998). Of particular interest has been the identification of glacial refugia, using genetic (Soltis *et al.* 1997; Crespi *et al.* 2003; Tribsch & Schonswetter 2003) and environmental (Hugall *et al.* 2002; Carstens & Richards 2007; Waltari *et al.* 2007) data, largely because understanding refugial dynamics is critical both to conservation/management and to understanding patterns and processes of speciation. Phylogeographers have commonly located putative refugia by identifying regions of high genetic diversity (e.g. Brunsfeld & Sullivan 2005), but some studies have

Correspondence: Bryan Carstens, Fax: +1 (614) 292-2030; E-mail: carstens.12@osu.edu

conducted statistical tests of refugial locations by employing either parametric simulation (e.g. Knowles *et al.* 2007), Bayesian (Morgan *et al.* 2011) or information theoretic (Provan & Maggs 2011) approaches. Investigations such as these exemplify statistical phylogeography, an increasingly quantitative discipline that relies on analytical methods that implement complex and parameter-rich models.

The willow *Salix melanopsis* is one of ~150 plant and animal complexes with disjunct populations in both the inland rainforests of the northern Rocky Mountains and the Cascades and coastal ranges (Nielson *et al.* 2001). Because of the diversity of taxa that display this disjunction, biogeographers have been motivated to explain the environmental processes that produced this pattern (Brunsfeld *et al.* 2001; Shafer *et al.* 2010), with the disjunction attributed to either pre-Pleistocene vicariance or post-Pleistocene dispersal. Carstens *et al.* (2005) presented evidence that dispersal led to the disjunction in *S. melanopsis*, but the posterior probability of this model was lower than that exhibited by other species. This result inspired additional work in *S. melanopsis*.

Brunsfeld *et al.* (2007) investigated northern populations of *S. melanopsis* using ~1800 bp of cpDNA from the *mat*K gene; however, demographic history was difficult to infer from the ML estimate of genealogy. Haplotypes from most *S. melanopsis* grouped into two clades that were considered products of separate refugia in the Salmon and Clearwater River drainages. In addition, two other clades of *S. melanopsis* haplotypes were present, but corresponded to phenotypes that were considered separate races (lowland, subalpine) by Brunsfeld *et al.* (1992). Carstens & Richards (2007) used parametric simulation to conduct tests of phylogeographic models. The tests rejected a model where *S. melanopsis* was partitioned into multiple refugia, but could not reject a model of expansion from a single refuge at the end of the Pleistocene. These results have two potential shortcomings: they are based on fixed parameter estimates from a single genetic locus that are not likely to be accurate (e.g. Felsenstein 2006), and they only reflect the history of the chloroplast genome, which may not match that of the nuclear genome. Consequently, Tsai & Carstens (2012) expanded this work by collecting data from four additional nuclear loci using Sanger sequencing. They compared demographic models using approximate Bayesian computation (ABC) and thus did not rely on fixed parameter estimates. The model with the highest posterior probability was a single refuge model with population expansion ($P = 0.684$), but a model of compartmentalized refugia (in the Salmon and Clearwater drainages) contained a nontrivial portion of the posterior density ($P = 0.264$). In summary, *S. melanopsis* probably dispersed to the coast from inland populations at the close of the Pleistocene, and the available data suggest but do not demonstrate conclusively that this expansion was from a single refuge.

Our approach to data analysis proceeds by considering the results of existing work as tentative findings that should be evaluated. As such, we analyse our data using several approaches. First, we use Lamarc 2.0 (Kuhner 2006), a coalescent-based method that estimates the extrinsic rate of population expansion from a single population, as suggested by Tsai & Carstens (2012). Second, we analyse our data using IMa2 (Hey & Nielsen 2007), which incorporates a model that was previously used to test regional biogeographic hypotheses in a range of organisms (Carstens *et al.* 2005). Third, we utilize two methods designed to help us explore Pleistocene refugial structure. Our approach will evaluate the findings of previous research using a larger data set and devote particular attention to the question of the refugial dynamics, which has proven difficult to address. Where possible, we evaluate multiple demographic models per analysis and quantify the relative fit of these models to the data. This model selection approach allows for parameter estimation, but also the identification of the demographic parameters that are required to understand the history of *S. melanopsis* in the Pacific Northwest.

Although recent phylogeographic work has expanded beyond single-locus studies, genomic scale data sets remain rare. Here, we utilize two emergent technologies, sequence capture probes (Good 2011; McCormack *et al.* 2012) and Illumina sequencing, to sequence targeted regions of the *Salix melanopsis* genome in samples from 12 river drainages in the temperate rainforests of the Pacific Northwest of North America (Fig. 1). The resulting data (408 loci, 335 SNPs) are used to explore demographic models and estimate relevant genetic parameters in a manner that accounts for uncertainty in the underlying models. Our data are expected to have several advantages over those collected in previous studies. First, it is likely that these data are more representative of the actual genetic variation across the *S. melanopsis* genome because we do not screen for polymorphism prior to data collection. Second, the amount of data should enable a more detailed parameterization of demographic models and allow us to develop a more clear understanding of how *S. melanopsis* responded to deglaciation.

Materials and methods

Design of capture probes and library preparation

Because *Salix melanopsis* is a riparian specialist, samples were chosen from the major river drainages within its range in Washington and Idaho (Fig. 1), including 13 from the Clearwater and six from the Salmon drainages

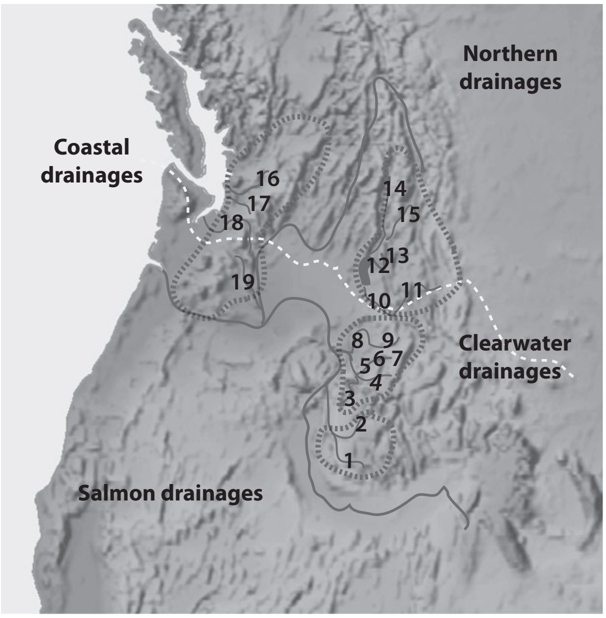


Fig. 1 Map of sample collection sites. Collection localities are marked with numbers that correspond to those of Table S1. Also shown are the four biogeographic regions (named and outline in dashed dark lines) as well as the maximum extent of glacial advance during the last glacial maximum (marked with a dashed white line).

(sites of putative refugia), as well as 12 from the northern drainages (recently deglaciated) and eight from the western coastal and Cascades drainages. In addition, sequence from a single *Salix alba* was collected to polarize SNPs and for potential use as an outgroup. Leaf tissue was sampled from herbarium specimens at the University of Idaho Stillinger Herbarium (Table S1, Supporting information), and whole-genomic DNA was extracted using a Qiagen DNeasy Plant Extraction kit. For each sample, 200 ng of whole-genomic extraction was sheared to lengths of 200–300 bp using a Bioruptor in the Louisiana State University Genomics Facility. After shearing, samples were quantified using a Qubit (Life Technologies), end-repaired and A-tailed using enzymes purchased from NEB. Illumina sequencing adapters (each containing a 12 bp barcode; Hamady *et al.* 2008) were ligated to the product, and the libraries were again quantified using a Qubit. Rather than sequencing RADTags (e.g. Emerson *et al.* 2010) or some other reduced representation library (e.g. Barbazuk *et al.* 2005), we used RNA capture probes (e.g. Good 2011) to target specific regions of the *S. melanopsis* genome. Capture probes were designed from sequences acquired by randomly sequencing DNA fragments from a single *S. melanopsis* individual on a partial Roche 454 sequencing plate during a project described by Zellmer *et al.* (2012). RNA baits were designed by J-M Rouillard (Mycroarray, Inc), and a custom set of biotinylated 120-*mer* baits was synthesized by Mycroarray. Nearly 800 loci, each 120–300 bp, were targeted by the probe set, and baits were tiled at 60-bp intervals for loci that exceeded 120 bp. The library was hybridized to the baits following manufacturer instructions, washed, rehybridized and washed again. To generate a sufficient sample for sequencing, emulsion PCR (Williams *et al.* 2006) was used to amplify the library following the wash steps. Sequencing was conducted by multiplexing eight individuals in each of five channels of an Illumina GA*ii*X located at the LSU Genomics facility using 108-bp reads. An average of 3.02×10^7 sequence reads per channel were obtained, or 3.76×10^6 per sample (Table S2, Supporting information). In addition, we included sequence data from four autosomal loci collected by Tsai & Carstens (2012) and from a single chloroplast locus collected by Brunsfeld *et al.* (2007).

Bioinformatics processing

Initial data processing occurred using the FASTX-Toolkit (Gordon & Hannon 2012; http://hannonlab.cshl.edu/fastx_toolkit/index.html). A Perl script was written to process sequence offloads from each channel as follows. First, quality scores were converted to Sanger format, and reads were filtered to retain only those reads with >90 bases of >q20 (Phred scale), corresponding to a >99% chance of the base being correct. Sequences were demultiplexed (allowing a single error in the barcode sequence) and trimmed to remove the barcode sequence. Sequences were then mapped to a reference sequence, which was a combination of sequences from two assemblies. For the first assembly, we mapped the sequences from *S. melanopsis* that were used to design the initial probe set to the *Populus trichocarpa* (also Salicaceae) genome using tools provided by Phytozome (http://www.phytozome.net/poplar). Specifically, we used a BLASTN search (Altschul *et al.* 1990), with the E threshold set to −1, and the BLOSUM6 comparison matrix. We then excised a 2-kb region, centred on the mapped sequence and used these sequences as references. Because not all of the sequences for the capture probes mapped to *Populus*, we also generated *de novo* assembly of reads from the *S. alba* individual using Velvet (Zerbino & Birney 2008) and Velvet Optimiser 2.1.7 (S Gladman; http://bioinformatics.net.au/software.shtml). In cases where *de novo* alignments also mapped to *Populus*, we chose the former for use as our reference. Once these reference sequences were generated, we mapped sequence reads from each individual to these sequences using BWA (Li & Durbin 2009). SAMtools (Li *et al.* 2009) was then used to generate a pile-up of all samples and to call SNPs; we generated a pile-up consisting of all individuals with the MPileup command and required >100X coverage per SNP with a minor allele frequency of >0.2. Note that the high coverage requirement did not impact the coverage across samples in the various alignments, as we had nearly the same number of individuals in the alignments when we reduced the requisite coverage to 10X. To infer haplotypes, data were phased using Phase version 2.1 (Stephens *et al.* 2001; Stephens & Donnelly 2003). For each alignment, the Phase input was generated using a Perl script, and sites were inferred with a certainty threshold of 95%. In the Phase output, all sites that could not be inferred within the threshold were replaced with missing data for downstream analyses. After phasing, input files for the analyses described below were assembled with custom Python scripts. Scripts used for the processing or analysis of the data are available from Dryad.

Characterizing patterns of genetic variation

Summary statistics were estimated using Compute, a component of the Libsequence package (Thornton 2003). To test for population structure among river drainages and among regions (coastal, northern drainages, Salmon, Clearwater; see Fig. 1), we conducted an analysis of molecular variance (AMOVA) in GeneticStudio version 131 (Dyer 2009) with 1000 permutations. In addition, because isolation by distance is often a driver of intraspecific genetic divergence, we performed a Mantel test to test for isolation by distance. To estimate population structure, we used the software Structure (Pritchard *et al.* 2000), which implements a clustering algorithm that groups individuals based on multilocus genotypes. Invariant sites from variable loci were excluded and linkage was assumed if more than one SNP was present at a particular locus. We then used these genotypic data as the input for Structure. The number of populations (*k*) was tested from 1 to 13, corresponding to the number of sampled river drainages (Table S1, Supporting information), using sample locality as a prior. This represents a conservative approach, as the use of a locality prior does not generally bias the results in cases where there is no signal for structure in the data (Hubisz *et al.* 2009). Fifteen independent Markov Chain Monte Carlo (MCMC) replicates of 500 000 steps were run, with the first 100 000 discarded as burn-in. The most likely set of cluster membership coefficients for each *k* was identified using the Greedy algorithm in CLUMPP (Jakobsson & Rosenberg 2007), and data were visualized in DISTRUCT (Rosenberg 2004).

Estimating the rate of population expansion using Lamarc

Previous work has identified expansion from a single refuge as a demographic model with a good fit to *S. melanopsis*. Accordingly, we analysed our data using Lamarc 2.0 (Kuhner 2006) to estimate $\theta = 4N_e\mu$ and population growth (γ). The analysis was conducted using a Bayesian search with priors for $\theta = 0.00001$, 100; and $\gamma = -500$, 1000. Four adaptively heated Markov chains (with temperatures of 1.0, 1.1, 1.2 and 3.0) were utilized, with 100 000 recorded trees (with the first 10 000 discarded as burn-in).

Exploring regional biogeographic patterns

Several biogeographic hypotheses have been proposed to explain the observed disjunct distribution of *S. melanopsis* as having resulted either from ancient, pre-Pleistocene vicariance, or from post-Pleistocene dispersal along a southern or northern route (Brunsfeld *et al.* 2001). Because divergence times and migration rates between eastern and western populations are the key parameters that differentiate biogeographic models for each of these hypotheses, we followed Carstens *et al.* (2005) and estimated these parameters using an isolation-with-migration model. The vicariance model predicts deep population divergence between the inland and coastal forests with low rates of migration, while models that include recent dispersal predict shallow population divergence with high rates of migration.

An isolation-with-migration model, as implemented in IMa2 (Hey & Nielsen 2007; Hey 2010), was used to test these hypotheses by dividing the samples into coastal and inland populations. This approach estimates key demographic parameters: $\theta = 4N_e\mu$, the timing of population divergence from an ancestral population (τ) and gene flow between the populations (m). We analysed two data sets; the first consisted of all polymorphic loci that included samples from both the inland and Cascades regions. The second consisted of 20 loci with sequence available across all individuals. For each, we experimented with prior bounds before defining the upper bounds of the priors ($\theta = 15$; $\tau = 5$; $m = 20$). Runs included a burn-in period of 10 000 steps and a geometric heating scheme that varied between $-g1 = 0.9$ and $-g2 = 0.999$, and consisted of 100 coupled Markov chains. Convergence was explored by examining the effective sample sizes and plots of the posterior densities of parameters. We computed the model probabilities of nested submodels following Carstens *et al.* (2009).

Inferring the structure of Pleistocene refugia

The Clearwater River drainage has been considered a likely refugium for inland rainforest species because it was the northernmost drainage to the south of the maximum advance of the last Pleistocene glacier and because there are numerous inland rainforest endemics whose distributions are restricted to this drainage (Brunsfeld *et al.* 2001). Immediately to the south of the Clearwater drainage is the Salmon River drainage, which has also been suggested as a putative site for a Pleistocene refugium (Nielson *et al.* 2001). Because *Salix melanopsis* is found in both drainages, we explored both single refugium (Clearwater) and dual refugia (Clearwater and Salmon) biogeographic models. Because refugial populations are expected to contribute to the newly founded populations, estimating gene flow is critical to inferring postglacial migration dynamics. We estimated gene flow under a coalescent framework using the program Migrate-n (version 3.2.16; Beerli & Felsenstein 2001; Beerli 2006). Parameters were first estimated under a model that allowed gene flow to occur among all regions. For the full analysis, we used all variable loci that included at least one sample in each of four biogeographic regions (i.e. the coastal, northern, Clearwater and Salmon drainages). Initial values were calculated using F_{ST}, and we used model averaging to estimate migration rates and θ values. Migrate-n analyses were conducted using a static heating strategy with four short chains (with temperature values of 1.0, 1.5, 3.0, and 1.0×10^6) and a single long chain. 200 000 steps were recorded every 100 generations with 10 000 steps discarded as burn-in. Stationarity of the Markov chain was assessed by examining ESSs for each parameter.

In addition to estimating gene flow using the full model in Migrate-n, we also designed several reduced models that were intended to represent various hypothesized refugia (Fig. 2). In addition to estimating parameters under these reduced models, we calculated the marginal likelihood of each and then compared these migration models following Beerli & Palczewski (2010). Thermodynamic integration with Bézier approximation (implemented in Migrate-n) was used to estimate marginal likelihood and calculate model probability. This approach allows us to calculate the relative posterior probability of the models so that we identify the models with high probability given the data, and subsequently make inferences regarding the putative refugia. Note that the data for this analysis consisted of the 20 polymorphic loci with the sequence available across all individuals (as in the IMa2 analysis, above) because patterns of missing data may introduce spurious results in the model comparison (P. Beerli, personal cummunication). Search strategies were as above.

One potential shortcoming of the Migrate-n analysis is that it does not explicitly model among-population divergence or population expansion. Therefore, we also evaluated a set of models that parameterize population divergence along with other processes (including migration and population expansion) using ∂a∂i 1.6.2 (Gutenkunst *et al.* 2009). Whereas Migrate-n uses sequence data to evaluate population histories using the coalescent, ∂a∂i uses SNP frequency data to evaluate histories. ∂a∂i summarizes SNP data in the form of allele frequency spectra (AFS) generated from all biallelic SNPs. Evolutionary processes such as population divergence, expansion and migration influence the AFS. ∂a∂i uses a diffusion equation to simulate changes in the AFS density distribution under each model and then compares the resulting AFS with an AFS generated from empirical data. Parameter estimates were optimized using the nonlinear BFGS optimization routine implemented in *∂a∂i*.

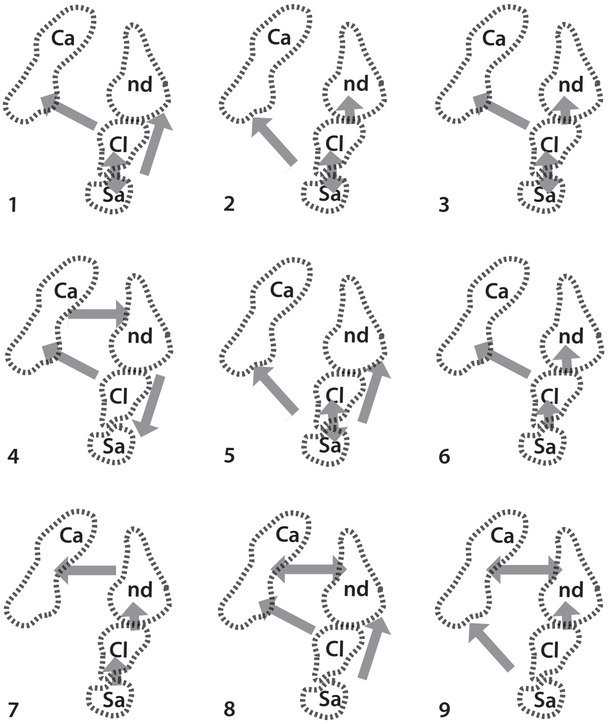
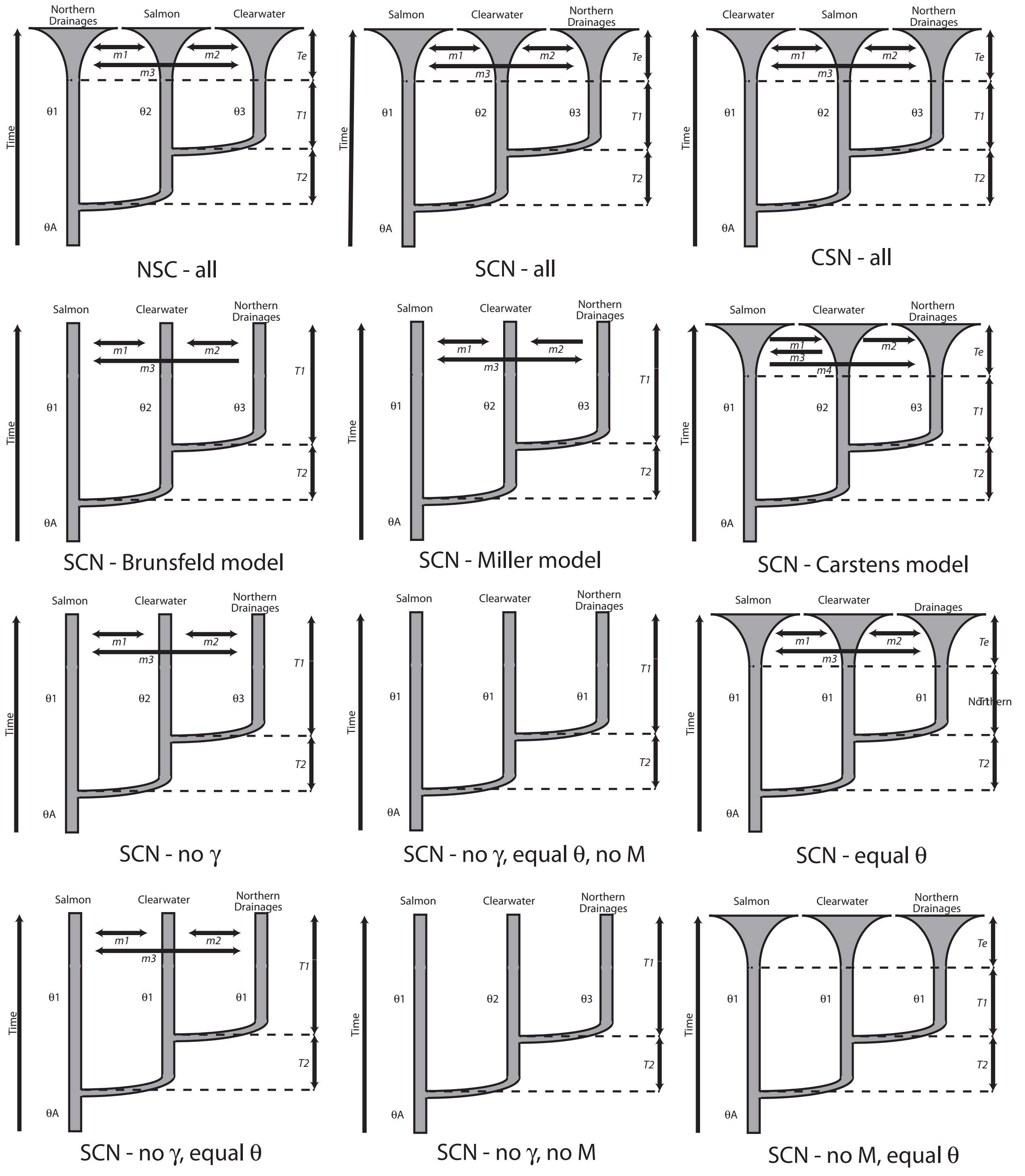


Fig. 2 Schematic of migration models. Nine biogeographic models (numerically labelled at lower left) used for model testing in Migrate-n are shown. Biogeographic regions are abbreviated as follows: Salmon (Sa), Clearwater (Cl), northern drainages (nd) and Cascades and coastal drainages (Ca). For each, arrows depict the directionality of gene flow among populations. For example, model 1 (upper left) posits that the Clearwater and Salmon exchange alleles, while the former is the source for the Cascades and coastal drainages, while the latter is the source for the northern drainages.

We used three sets of models for analysis of the history of the three inland populations (Clearwater, Salmon, and northern drainages): the full model parameterized with each of the three possible patterns of divergence (i.e. topologies), a set of three models described by the authors of an earlier study (Brunsfeld



NSC - all

SCN - all

CSN - all

SCN - Brunsfeld model

SCN - Miller model

SCN - Carstens model

SCN - no γ

SCN - no γ, equal θ, no M

SCN - equal θ

SCN - no γ, equal θ

SCN - no γ, no M

SCN - no M, equal θ

Fig. 3 Demographic details of each of the 12 three-dimensional models included in the *dadi* analysis are shown. Population sizes (θ_i), migration rates (m_i), timing of population divergence (T_i) and expansion (T_e) are estimated from the data using allele frequency spectra.

et al. 2007) and a set of models with simplified parameters (Fig. 3). In addition, we evaluated two sets of two-dimensional models; the first collapsed two populations in the inland region, and the second included samples from the coastal ranges and was intended as comparable to the IMa2 analysis (Fig. S1, Supporting information). We compared the three- and two-population models separately because the AFS used for each class was of a different dimensionality (due to the different number of populations), and thus, the likelihoods calculated by ∂a∂i are not equivalent. The composite likelihoods of the models given the data are calculated by ∂a∂i under the assumption the SNPs are not linked, and we are violating this assumption with approximately half of our data (i.e. in those loci that contain more than one SNP). This violation would be problematic if we were to apply a standard statistical cut-offs in a test for statistical significance (e.g. using a χ^2 distribution in a likelihood ratio test). However, because the pattern of linkage is consistent across model comparisons, there is no reason to expect that the composite likelihoods are biased as a function of a particular model, and thus, we apply an information theoretic approach to model selection by calculating the Akaike information criterion score (Akaike 1974) and other metrics as described in the study by Burnham & Anderson (2002).

Results

Bioinformatics processing

Our raw data consisted of approximately 75 million sequence reads, with ~99% matched to one of the eight Golay barcodes. 59 million reads passed the quality filter and were aligned to the reference sequences using BWA, and ~53% of these reads were assembled into contigs that were used to call SNPs in downstream analyses (Table S2, Supporting information). The remaining sequences appear to represent either fragments of genomic DNA that did not hybridize to the capture probes and were not removed during the purification step, or genomic DNA that hybridized to one of the capture probes that did not assemble *de novo* or align to the *Populus* genome. A total of 408 loci were assembled to reference sequences (~½ mapped to *Populus*); approximately three-fourths of these were invariant, and the polymorphic loci contained a total of 335 SNPs (Fig. S2; Table S3, Supporting information).

Descriptive analyses

Summary statistics were calculated for each polymorphic locus (Table 1). AMOVA results (Table 2) did not find significant genetic structure among the four biogeographic regions (P = 0.945), but genetic structure was present among the individual river drainages (P < 0.001). Additionally, the Mantel test did not show a significant signal for isolation by distance among the samples (P = 0.483), suggesting that factors other than simple geographic distances are influencing genetic structure in *Salix melanopsis*. This result is in contrast with the isolation-by-distance results in the study by Tsai & Carstens (2012), suggesting that this analysis is sensitive to differences in the number of samples, loci or polymorphism. The clustering algorithm implemented in Structure revealed genetic structure at several levels of K (Fig. S3, Supporting information), but this genetic structure did not appear to be concordant with geographic locations (Figs. S4 and S5, Supporting information). As our analyses were conducted

Table 1 Average, minimum and maximum values for assorted summary statistics for variable loci. Summary statistics are shown for three different classes of polymorphic loci, along with the number in each class (in parentheses). Shown are the number of *Salix melanopsis* samples sequenced (number), the length of each locus (length), the number of segregating sites (S), the number of haplotypes (nhap), the haplotype diversity (hap div), Watterson's Theta (θ_W), nucleotide diversity (π) and Tajima's D (D)

Loci	Number	Length (bp)	S	nhap	hap div	θ_W	π	D
Identified via Sanger sequencing* (5)								
Ave	24.70	568.6	15.0	10.40	0.81	0.00805	0.00794	−0.29
Min	24	141	8	8	0.71	0.00338	0.00175	−1.54
Max	38	1159	23	12	0.87	0.01261	0.01611	0.75
Assembled by aligning to *Populus trichocarpa* (51)								
Ave	13.17	1144.1	2.75	1.78	0.22	0.01071	0.01106	0.65
Min	4	95	1	2	0.12	0.00079	0.01055	−1.14
Max	37	1673.1	5	4	0.67	0.13716	0.08658	2.26
Assembled *de novo* using Velvet (49)								
Ave	28.5	311.2	2.98	6.23	0.36	0.00699	0.00446	−0.96
Min	8	76	1	2	0.09	0.00037	0.00395	−2.00
Max	38	1327	22	19	0.93	0.02820	0.02711	1.73

*Includes four nuclear and one chloroplast locus (Tsai & Carstens 2012).

Table 2 AMOVA results. Samples are divided into four regions (Cascades, Salmon, Clearwater and northern Drainages) and within each region into river drainages. Significance was assessed using 1000 bootstrap replicates. Shown for each level are the degrees of freedom (d.f.), sum of squares (SS) and mean of squares (MS)

Source	d.f.	SS	MS
Among region	3	56.4647	18.8216
Among population	9	157.9982	17.5554
Error	25	391.5988	15.6640
Total	37	606.0617	

Parameter	Value	*P*
ϕ_{RT}	0.0050	0.9450
ϕ_{PR}	0.0432	0.5480
ϕ_{ST}	0.0480	0.0010

with the location prior, we consider this strong evidence for a lack of genetic structure in these data.

Estimating the rate of population expansion using Lamarc

Lamarc was used to estimate $\theta_{\text{per site}} = 0.0105$ (0.0099, 0.0113) and $\gamma = 146.2$ (89.5, 268.2). As in the study by Tsai & Carstens (2012), the extrinsic rate of population expansion is positive and does not include 0 in the 95% credible interval, indicating that the population is likely larger in size in the present than it was in the past (Lamarc documentation). However, these results are smaller than those reported previously (Tsai & Carstens 2012), possibly due to the overall lower levels of polymorphism in the data.

Exploring regional biogeographic hypotheses

Regional biogeographic hypotheses were tested using an isolation-with-migration model (implemented in IMa2). Because not all loci were sequenced in each individual, we initially analysed a data set containing all variable loci in which at least one individual from both the coastal and inland regions was sequenced (i.e. for the IMa2 analysis, there were 79 variable loci with samples sequenced in both regions). When these loci were included in an analysis, the Markov chain had difficulty reaching a stable region of high probability, and after 200+ million generations, the ESS values for τ were <5 (for our data, τ consistently represented the lowest ESS value). However, the Markov chains of an analysis of a 20-loci data set (consisting of five loci generated with Sanger sequencing and 15 loci for which there were no missing data) were able to reach large ESS for the τ parameter of 1058 in 97 h (864 865 steps for the 100 coupled chains), and we report these results here (Table 3). Estimates of divergence time ($\tau = 0.0095$) and migration rates to the coastal drainages ($m = 2.22$) are consistent with the recent dispersal hypothesis posited by Brunsfeld *et al.* (2001).

To further explore these models, we also conducted model selection following Carstens *et al.* (2009). In this approach, the model probabilities of the nested models contained within the full model (i.e. implemented in IMa2) are calculated and compared. Gene flow is included in all of the most probable of the nested IMa2 models, and it is symmetrical in the submodel with the highest model probability (Table 4). Because models that do not include gene flow contain none of the total model probability, we can strongly reject ancient vicariance as a contributing factor to the disjunct distribution in *S. melanopsis*. Assuming a generation time of 5 years (as assumed by Brunsfeld *et al.* 2007) and a mutation rate of 7.0×10^{-9} (estimated for *Arabidopsis thaliana*; Ossowski *et al.* 2010), inland and coastal populations are estimated to have been isolated for at least 450 generations (2700 generations if we assume a mutation rate inferred for *Populus*; Tuskan *et al.* 2006). Effective population sizes (assuming the *Arabidopsis* mutation rate) for the inland populations and coastal populations were ~25 000 and 2000 individuals, with an intermediate ancestral population size (~15 000). These sizes would be larger by a factor of ~6 assuming the rate inferred for *Populus*. In addition, the effective migration rates (i.e. 2 Nm) into the coastal and inland populations were 0.19 and 0.18, respectively. These values support previous assertions that the coastal populations were the result of post-Pleistocene expansion from the inland populations (Carstens *et al.* 2005).

Table 3 Parameter estimates under the full isolation-with-migration model. Shown are estimates made using the program IMa2 (Hey & Nielsen 2007). The point estimates and credible range for population divergence (τ), $\theta = 4\ N_e\mu$ and gene flow backwards in time (*M*) are shown

IMa2 (20 loci)	τ	θ_A	θ_I	θ_C	m_{IC}	m_{CI}
HiPt	0.0095	1.265	2.09	0.1650	2.220	0.1800
95% HPD interval	0.003, 0.048	0.94, 1.71	1.71, 10.63	0.06, 0.72	1.86, 10.82	0.30, 3.93
Model averaged	0.082	1.421	1.945	0.678	10.318	1.293

Table 4 Shown are models considered by IMa2, the number of parameters for each model, its AIC score, AIC differences (Δ_i), model likelihoods and model probabilities (w_j). All values were calculated following Burnham & Anderson (2002)

Description	Model	k	log(P)	AIC	Δ_i	Model likelihood	w_i
$\theta_A\ \theta_{/}\theta_2$ M$_1$ M$_1$	ABCDD	4	−1.588	11.175	0.000	1.000	0.386
$\theta_A\ \theta_{/}\theta_2$ M$_1$ M$_2$	FULL	5	−1.047	12.094	0.919	0.632	0.244
$\theta_A\ \theta_{/}\theta_A$ M$_1$ M$_1$	ABADD	3	−3.330	12.660	1.484	0.476	0.184
$\theta_A\ \theta_{/}\theta_A$ M$_1$ M$_2$	ABADE	4	−3.196	14.391	3.216	0.200	0.077
$\theta_A\ \theta_{/}\theta_1$ M$_1$ M$_1$	ABBDD	3	−4.653	15.307	4.131	0.127	0.049
$\theta_A\ \theta_{/}\theta_1$ M$_1$ M$_2$	ABBDE	4	−4.504	17.008	5.833	0.054	0.021
$\theta_A\ \theta_A\ \theta_2$ M$_1$ M$_2$	AACDE	4	−4.527	17.054	5.878	0.053	0.020
$\theta_A\ \theta_A\ \theta_A$ M$_1$ M$_2$	AAADE	3	−5.654	17.308	6.132	0.047	0.018
$\theta_A\ \theta_A\ \theta_A$ M$_1$ M$_1$	AAADD	2	−30.354	64.708	53.532	0.000	0.000
$\theta_A\ \theta_A\ \theta_2$ M$_1$ M$_2$	AACDD	3	−30.212	66.424	55.248	0.000	0.000
$\theta_A\ \theta_1\ \theta_2$ M$_2$	ABC0D	4	−55.391	118.782	107.607	0.000	0.000
$\theta_A\ \theta_A\ \theta_A$	AAA00	1	−460.517	923.034	911.859	0.000	0.000
$\theta_A\ \theta_A\ \theta_2$	AAC00	2	−460.517	925.034	913.859	0.000	0.000
$\theta_A\ \theta_1\ \theta_A$	ABA00	2	−460.517	925.034	913.859	0.000	0.000
$\theta_A\ \theta_1\ \theta_1$	ABB00	2	−460.517	925.034	913.859	0.000	0.000
$\theta_A\ \theta_1\ \theta_2$	ABC00	3	−460.517	927.034	915.859	0.000	0.000
$\theta_A\ \theta_1\ \theta_2$ M$_1$	ABCD0	4	−460.517	929.034	917.859	0.000	0.000
					Σ Model likelihood = 2.588		

Inferring the structure of Pleistocene refugia

In addition to conducting tests of regional biogeographic models, we wanted to further explore refugial dynamics by exploring rates of gene flow among the four main biogeographic regions. Estimates of migration were generally high into the coastal and Cascades drainages, as well as the Clearwater and northern drainages (Table 5). While rates into the Salmon drainages were lower, the credible intervals of all estimates were broadly overlapping and, based on the full model, we can conclude that gene flow is prevalent in *S. melanopsis*.

We also used Migrate-n following Beerli & Palczewski (2010) and Provan & Maggs (2011) to calculate the posterior probability of nine models that restrict migration to patterns that characterized different putative refugia (Fig. 2). We found strong support for a model first proposed by Brunsfeld *et al.* (2007). This model posits that the Clearwater drainage was the source of the western populations and that the Salmon drainage served as the source for the northern populations. Given the high posterior probability of this model ($P > 0.99$; Table 6), results from the Migrate-n analyses suggest that (i) there were multiple refugia in the inland temperate rainforest located to the south of the glacier and (ii) that these refugia served as the source for different regions at the close of the Pleistocene. However, because Migrate-n does not parameterize the historical divergence of populations, we also conducted model selection using a method that estimates this parameter.

Demographic models were explored using *∂a∂i*. Similar to the approach used with Migrate-n, we calculated the probability of model$_i$ | data for multiple models, ranked these probabilities using information theory and conducted model averaging of the parameter estimates. The model with the highest probability given the data describes the Clearwater drainage as the source for the northern drainages (rather than the Salmon), but does include gene flow among all drainages. This model does not include population expansion, but does incorporate differing θ values among drainages and posits that ancestral populations were substantially smaller than the modern populations (consistent with both the Lamarc and IMa2 results). As in the Migrate-n results, this model is supported at a very high probability ($w_i = 0.95$; Table 7), indicating that we are capable of discriminating among complex models with our data. We also explored two-dimensional models and found that a model that treated the Salmon and Clearwater as a single refuge that was the source for the northern drainages had a substantially higher likelihood given the data than models that treated the Salmon and Clearwater as separate.

Discussion

Phylogeographic inferences in Salix melanopsis

The goal of this investigation is to explore how the riparian specialist *Salix melanopsis* responded to the environmental changes that accompanied the retreat of the last glacier some 20 000 years before present. The current distribution of *S. melanopsis* is disjunct in the Pacific Northwest, with populations in the coastal and Cascades ranges as well as in the inland temperate rainforest of the northern Rocky Mountains. Analyses presented here support earlier findings that attribute the formation of this disjunct distribution to post-Pleistocene expansion from an inland refuge. First, direct estimate of population divergence indicates that coastal populations have been separate for hundreds of generations, but given the estimate of 5 years per generation provided by Brunsfeld *et al.* (2007) and a mutation rate between that of *Arabidopsis* and *Populus*, this divergence is likely Holocene in origin. Second, estimated migration rates from the inland to coastal populations are high in Migrate-n (2 Nm = ~2.8), IMa2 (~ 2.2) and ∂a∂i (~2.8). Taken together, these results confirm that recent dispersal can account for the disjunct distribution in *S. melanopsis* and are consistent with prior work in the system (Carstens *et al.* 2005). A more intriguing question involves the refugial dynamics within the inland regions.

Table 5 Parameter estimation using Migrate-n. Parameters estimated under the full model are shown, with gene flow shown using values that are not scaled by the mutation rate. These values are comparable to those estimated using IMa2 and *dadi*. Populations are abbreviated as follows: Cascades (1); Clearwater (2); northern drainages (3); and Salmon (4)

Parameter	2.50%	Mode	97.50%
θ_1	0.002	0.00397	0.00574
θ_2	0.00247	0.00417	0.0058
θ_3	0.0022	0.00404	0.0058
θ_4	0.00067	0.00237	0.00407
$M_{2->1}$	1.18	2.92	4.95
$M_{3->1}$	1.03	2.56	4.48
$M_{4->1}$	1.22	2.96	5.19
$M_{1->2}$	1.27	2.64	4.56
$M_{3->2}$	1.49	3.06	5.21
$M_{4->2}$	1.37	2.89	4.92
$M_{1->3}$	1.33	2.85	5.22
$M_{2->3}$	1.53	3.45	5.80
$M_{4->3}$	1.49	3.47	5.77
$M_{1->4}$	0.34	1.45	3.18
$M_{2->4}$	0.37	1.58	3.41
$M_{3->4}$	0.31	1.41	3.03

We used two model selection approaches to explore refugial population dynamics (Migrate-n and ∂a∂i), and the optimal model was strongly supported in each (Fig. 4). However, these models are not entirely comparable. In the optimal model from the Migrate-n analysis, the northern drainages are colonized from the Salmon refuge, whereas the Clearwater served as the source for the northern drainages in the ∂a∂i analysis. While migration rates are much higher in the Migrate-n analysis than they are in the ∂a∂i analysis, estimates of $\theta = 4 N_{e}\mu$ are similar. This suggests that these approaches are interpreting the shared polymorphism in a different manner. In Migrate-n, shared polymorphism is largely attributed to gene exchange, while ∂a∂i also incorporates temporal population divergence. The later should be an important parameter to consider because it allows us to differentiate between a single and dual refuge model. However, gene flow between the Salmon and Clearwater drainages is present in all of the models with any support given the data (i.e. both Migrate-n and ∂a∂i). Rather than expecting to find a single refuge located in either of these drainages, it might be better to envision either multiple refugia in individual river canyons in both drainages or a refuge that encompassed regions of both drainages. It is also worth

Table 6 Model selection with Migrate-n. Shown are model number (see Fig. 2), verbal descriptions of the models, the pattern of migration, harmonic estimates of the marginal likelihood (lmL) and the posterior model probability (*P*). Bolded text in the third column indicates which of the populations are refugia in a particular model. Populations are abbreviated as follows: Cascades (Ca); Clearwater (CL); Salmon (Sa); and northern drainages (N)

No.	Description of model	Migration pattern	Harmonic lmL	*P*
1	Two refugia, with Clearwater as source of Cascades and Salmon as source for northern	Ca<-**Cl**<->**Sa**->No	−43872.35	0.99976
2	Two refugia, with Salmon as source of Cascades and Clearwater as source for northern	Ca<-**Sa**<->**Cl**->No	−43880.67	0.00024
3	Single refuge (Clearwater)	Sa<->**Cl**->N; Cl->Ca	−43950.06	0.00000
4	Stepping stone (Clearwater)	**Cl**->Ca->N->Sa	−43966.94	0.00000
5	Single refuge (Salmon)	Cl<->**Sa**->N; Sa->Ca	−43997.26	0.00000
6	Stepping stone (Salmon), with Clearwater source for Cascades	**Sa**->Cl->Ca; Cl->N	−44039.82	0.00000
7	Stepping stone (Salmon)	**Sa**->Cl->N->Ca	−44122.03	0.00000
8	Clearwater and Salmon are sources, sending migrants to Cascades and northern, respectively, and those exchange migrants	**Cl**->Ca<->N < -**Sa**	−44150.8	0.00000
9	Clearwater and Salmon are sources, sending migrants to northern and Cascades, respectively, and those exchange migrants	**Cl**->N < ->Ca<-**Sa**	−44259.9	0.00000

Table 7 Model selection and parameter estimates using *∂a∂i*. For each model, the log likelihood (−lnL), number of parameters (*k*), Akaike information criterion score (AIC), AIC differences (Δ_i), model likelihoods and probabilities (w_i) and parameter estimates (Fig. S2) are shown. Parameters include $\theta_{\#} = 4\ N_e\mu$, timing of population divergence ($T_{\#}$) and expansion (T_e; if applicable) as well as rates of gene flow into various lineages ($m_{\#}$). Models are grouped into three groups, corresponding to the dimensionality of the allele frequency spectrum (AFS). The first analyses samples from the Salmon, Clearwater and northern drainages; the second collapses two of these three groups into a single population; the third compares the inland to the coastal samples

Inland temperate rainforest—three-dimensional AFS						Optimized parameter estimates										
Model name	−lnL	k	AIC	Δ_i	w_i	θ_A	T_e	T_1	T_2	m_1	m_2	m_3	m_4	θ_1	θ_2	θ_3
SCN—no γ	331.664	10	683.33	0.00	0.950	0.15		4.08	0.33	0.01	1.69	1.49	—	1.30	4.42	3.04
SCN—equal θ	337.202	8	690.40	7.08	0.028	0.04	2.69	4.52x	0.18	0.01	1.20	1.43	—	3.75	—	—
SCN—no γ, no M	338.687	7	691.37	8.05	0.017	0.06	—	0.01	0.00	—	—	—	—	0.07	0.24	0.46
SCN—all	336.939	10	693.88	10.55	0.005	0.61	1.27	3.36	0.59	0.43	3.00	1.67	—	0.57	3.51	1.92
CSN—all	342.744	10	705.49	22.16	0.000	0.31	0.78	3.85	0.41	0.74	3.00	1.35	—	0.55	4.38	1.70
SCN—Brunsfeld model	345.453	9	708.91	25.58	0.000	0.08	—	5.53	2.68	0.23	1.76	4.68	—	0.44	3.50	6.13
SCN—no γ, equal θ	349.778	7	713.56	30.23	0.000	0.01	—	0.01	0.01	—	—	—	—	0.21	—	—
SCN—no γ, no M, equal θ	354.255	5	718.51	35.18	0.000	0.14	—	0.01	0.01	—	—	—	—	0.30	—	—
SCN—Carstens model	371.995	11	765.99	82.66	0.000	0.26	6.34	9.99	0.64	0.37	3.05	0.96	4.10	0.66	1.13	3.23
NSC—all	409.803	10	839.61	156.28	0.000	0.13	2.00	5.64	0.35	1.10	1.00	0.54	—	0.78	2.67	3.99
SCN—Miller model	492.43	9	1002.86	319.53	0.000	0.13	—	1.16	6.06	0.39	0.48	2.14	—	0.72	1.48	1.37

Inland temperate rainforest—two-dimensional AFS						Optimized parameter estimates						
Model name	−lnL	k	AIC	Δ_i	w_i	θ_A	T_e	T_1	m_1	m_2	θ_1	θ_2
S+C, N—all	102.69	7	219.38	0.00	1.000	0.31	2.99	7.02	3.01	5.00	4.10	1.11
S+N, C—all	118.169	7	250.34	30.96	0.000	0.07	2.53	10.00	1.61	1.76	2.30	2.82
C+N, S—all	124.587	7	263.17	43.79	0.000	4.46	1.46	0.35	1.68	5.00	0.45	1.14

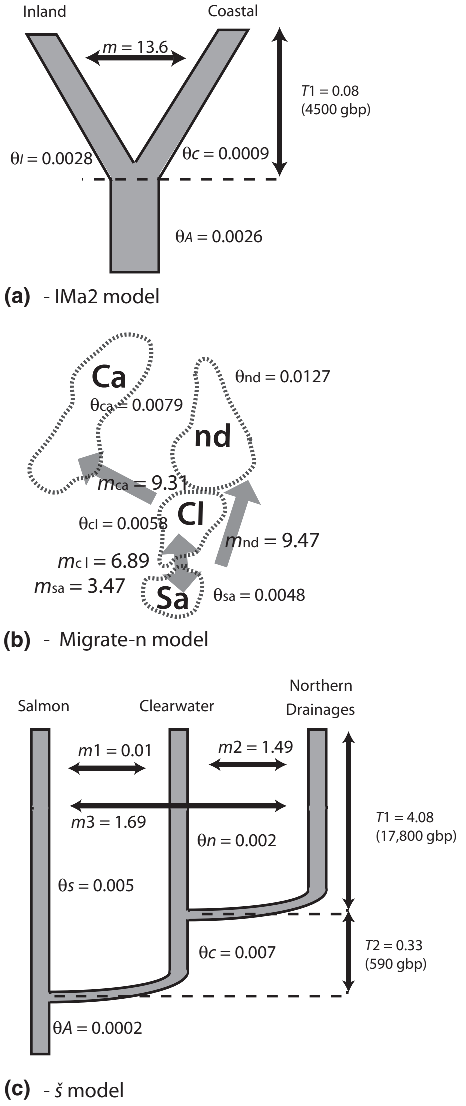
Biogeographic hypotheses—two-dimensional AFS						Optimized parameter estimates					
Model name	−lnL	k	AIC	Δ_i	w_i	θ_A	T_1	m_1	m_2	θ_1	θ_2
Isolation only	107.656	4	223.31	0.00	0.742	0.08	0.001	—	—	0.12	0.03
Isolation-with-migration	106.711	6	225.42	2.11	0.258	0.06	10.00	2.41	2.82	3.27	2.75
Dispersal from inland	112.648	5	235.30	11.98	0.002	0.06	5.92	1.43	—	5.47	0.66
Dispersal from Cascades	118.721	5	247.44	24.13	0.000	0.09	3.44	2.80	—	0.23	6.24

noting that in contrast to previous findings (Tsai & Carstens 2012), the ∂a∂i analysis did not find support for models that included population expansion. This finding could result from the smaller number of sampled individuals in this study, from the conservative SNP calling that may eliminate true SNPs that are present at low frequency, or be indicative that previous studies were errant.

Pull-down methods

Recent advances in sequencing technology have dramatically expanded the data available to phylogeographic investigations. While this increase has clear benefits, such as the expansion in the analytical methods applicable to the discipline (e.g. AFS approaches such as ∂a∂i), the data collected using next-generation sequencing are also likely to be improved samples (in the statistical sense) of the actual genomic variation of the focal species. In this regard, it is useful to compare the data collected here to previous efforts in *S. melanopsis*. Tsai & Carstens (2012) collected sequence data from four autosomal loci identified by designing primers derived from the *Populus* genome and screening with three individuals suggested by Brunsfeld *et al.* (1992). Four loci that contained multiple variable sites were selected for further sequencing. The loci collected by Tsai & Carstens (2012) are more variable, on average, than those collected here (Fig. S2, Supporting information). Researchers who use similar screening protocols should view this as a cautionary tale. While it is unclear how broadly such sampling strategies can bias phylogeographic analyses, it is clear that ascertainment bias can lead to an overestimation of heterozygosity (Rogers &

Jorde 1996), which may in turn influence our ability to detect putative refugia. However, because sequencing capture probes allow us to collect both polymorphic and monomorphic data, the question of whether to analyse all the loci or only the polymorphic loci is newly relevant. We have chosen to analyse only polymorphic loci, but it is worth exploring the extent to which this might influence our analyses. Analyses such as the AMOVA, Structure and ∂a∂i utilize only the SNPs in the data set and thus are not influenced by our decision. However, coalescent-based methods such as Lamarc, IMa2 and Migrate-n might be influenced, because overall estimates of $\theta = 4\,N_e\mu$ may vary depending on whether monomorphic loci are included. To explore this subject, we re-ran some analyses using additional invariant loci and found that estimates of some parameters were slightly different (Tables S4 and S5, Supporting information). However, because all analyses are conditioned on the data, they might be expected to change slightly as different data are analysed.



(a) - IMa2 model

(b) - Migrate-n model

(c) - š model

Fig. 4 The optimal models from each of the model selection analyses are shown. (a) depicts the model from the Migrate-n analysis, (b) depicts the model from the IMa2 analysis, and (c) depicts the model from the *∂a∂i* analysis. The parameters estimated under each model are shown (model averaged values are shown for the IMa2 model) and have been converted to comparable units of *θ* per site, gene flow (measured by the rate that alleles come into the population per generation) and population divergence in generations before present (gbp).

How should we analyse phylogeographic data?

Traditionally, phylogeographic data have been analysed by summarizing genetic variation with statistics (e.g. F_{ST}, θ_w) or estimating parameters using a particular model (e.g. Nm with Wright's island model, genealogies under a phylogenetic model). Although such summaries and estimates are formally generated, researchers often interpret these values in a qualitative manner, resulting in overinterpretation (Knowles & Maddison 2002) and leaving the inferences vulnerable to confirmation bias (Nickerson 1998). While model-based methods have become common in phylogeographic investigations (Knowles 2009), the role that these models play in the inference process remains varied. Most commonly, the models are used to estimate parameters, and the parameter estimates form the basis of the phylogeographic inferences.

Various analytical methods implement specified demographic models, including some of the approaches used to analyse our data (i.e. IMa2 and Migrate-n). These methods often include advanced machinery for parameter estimation, but at the cost of flexibility in the particular details of the demographic model. Our goal for this work was to understand the recent history of *Salix melanopsis* and, in particular, how postglacial habitat was colonized. Parameter estimates contain clues regarding this history but may be difficult to interpret. For example, consider estimates made under the Migrate-n model: rates of migration are uniformly high, the credible intervals broadly are overlapping, and θ for the coastal population is much lower than the inland populations (Table 5). We might infer from this pattern that coastal populations have been recently founded from inland refugia, but if this is the case, why are the levels of gene flow from the inland to the coast so nearly matched by those in the opposite direction? Faced with difficulties in inference such as these, researchers have turned to custom tests of phylogeographic hypotheses. For example, Knowles (2001) designed explicit models of refugial structure that differed in the timing and pattern of diversification.

The probability of data given the assumption that the model is true was calculated using coalescent simulations to build null distributions of a summary statistic and rejecting or failing to reject based on some cut off. The inference that diversification was in part due to isolation in multiple glacial refugia from Knowles (2001) was largely informed by the rejection of certain explicit models.

Phylogeographic hypothesis tests have two substantial shortcomings: they rely on simulations conducted under a fixed set of parameter values, and they can only reject or fail to reject a model. These shortcomings result from the statistical framework; the null distributions and critical values are tools developed for experimental science. Phylogeography is a historical science lacking experimental controls and replication, and phylogeographic hypothesis testing is an attempt to adopt conventional statistical tools to the discipline. We did not employ hypothesis testing here because we lack a clear null hypothesis and thus wish to evaluate multiple models. This is a difficult task for hypothesis testing, due to both the reduced power brought about by multiple comparisons and the requisite interpretation of and model that cannot be rejected as equivocal. Rather, we have proceeded in several analyses by calculating the probability of multiple models given the data and then by ranking these models using information theoretic approaches (e.g. Anderson 2008). In doing so, the models serve as exploratory tools that guide inference through the identification of models that are probable given the data, rather than a hypothesis to be tested.

Demographic model selection should be a tool for phylogeographic inference because it enables a quantitative evaluation of which parameters (and thus which biological processes) are important to consider. For example, our results indicate that both temporal divergence and gene flow among populations have contributed to the current patterning of genetic diversity within *S. melanopsis*. This is evident both across the disjunction (i.e. the inland and coastal populations) and in the postglacial dynamics of the inland populations. However, if demographic model selection is to play an important role in the next generation of phylogeographic analysis, several challenges must be met. One of these is to increase the flexibility of the methods. For example, IMa2 can calculate the relative probabilities of submodels for a pair of diverging populations, but not more. A second challenge is to fairly define the set of models under consideration. Although programs such as ∂a∂i allow users to specify a wide range models, in most systems (as in *S. melanopsis*) the analyses will not exhaust the possible set of models, and in these cases, it is possible that some very realistic models were not considered. Finally, the question of sample partitioning remains. In most cases, our samples were divided in part using prior information and geography, but also as dictated by the particulars of the analysis (i.e. into two groups for the IMa2 analysis, three for the ∂a∂i and four for the Migrate-n analyses), and these groupings of samples might not be biologically meaningful. Consequently, it can be difficult to compare the results from the various approaches (e.g. the optimal models of the Migrate-n and ∂a∂i analyses). Regardless of these challenges, demographic model selection and multimodel inference have allowed us to reduce our dependence on the results of parameter estimation and to base our inferences on the identification of important biological processes. The implementation of model selection in approaches such as ∂a∂i, Migrate-n, IMa2 as well as ABC methods (reviewed by Csilléry *et al.* 2010) represents an important advance for a field that is no longer data limited.

Acknowledgements

We thank S Hird, J McVay, T Pelletier, N Reid, Y-H Tsai and A Zellmer for discussions related to this work. We thank J Carstens for extracting the DNA from *Salix melanopsis* samples. We thank R Gutenkunst and P Beerli, who provided valuable assistance and advice related to the ∂a∂i and Migrate-n analyses, respectively. We thank S Herke for assistance with the Illumina GA*ii*X. We thank R Abbott, D Weisrock and three anonymous reviewers who provided many thoughtful comments that improved this manuscript. Funding for this work was provided by NSF-DEB 0918212 and NSF-DEB 0956069 to BCC.

References

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC*, **19**, 716–723.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Anderson DR (2008) *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer Science, New York.

Avise JC, Walker D, Johns GC (1998) Species durations and Pleistocene Effects on vertebrate phylogeography. *Proceedings Royal Society of London B*, **265**, 1707–1712.

Barbazuk WB, Bedell JA, Rabinowicz PD (2005) Reduced representation sequencing: a success in maize and a promise for other plant genomes. *BioEssays*, **27**, 839–848.

Beerli P (2006) Comparison of Bayesian and maximum likelihood inference of population genetic parameters. *Bioinformatics*, **22**, 341–345.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations using a coalescent approach. *Proceedings of the National Academy of Sciences USA*, **98**, 4563–4568.

Beerli P, Palczewski M (2010) Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics*, **185**, 313–326.

Brunsfeld SJ, Sullivan J (2005) A multi-compartmented glacial refugium in the northern Rocky Mountains: evidence from the phylogeography of *Cardamine constancei* (Brassicaceae). *Conservation Genetics*, **6**, 895–904.

Brunsfeld SJ, Soltis DE, Soltis PS (1992) Evolutionary patterns and processes in *Salix* sect. Longifoliae: evidence from chloroplast DNA. *Systematic Botany*, **17**, 239–256.

Brunsfeld SJ, Sullivan J, Soltis DE, Soltis PS (2001) Comparative phylogeography of northwestern North America: a synthesis. In: *Integrating Ecological and Evolutionary Processes in a Spatial Context* (eds Silvertown J & Antonovics J), pp. 319–339. Blackwell Science, Oxford.

Brunsfeld SJ, Miller TR, Carstens BC (2007) Insights into the biogeography of the Pacific Northwest of North America: evidence from the phylogeography of *Salix melanopsis*. *Systematic Botany*, **32**, 129–139.

Burnham KP, Anderson DR (2002) *Model Selection and Multimodal Inference: A Practical Information Theoretic Approach*, 2nd edn. Springer-Verlag, New York.

Carstens BC, Richards CL (2007) Integrating coalescent and ecological niche modeling in comparative phylogeography. *Evolution*, **61**, 1439–1454.

Carstens BC, Brunsfeld SJ, Demboski JR, Good JM, Sullivan J (2005) Investigating the evolutionary history of the pacific northwest mesic forest ecosystem: hypothesis testing within a comparative phylogeographic framework. *Evolution*, **59**, 1639–1652.

Carstens BC, Stoute HN, Reid NM (2009) An information theoretical approach to phylogeography. *Molecular Ecology*, **18**, 4270–4282.

Crespi EJ, Rissler LJ, Brown RA (2003) Testing Pleistocene refugia theory: phylogeographical analysis of *Desmognathus wrighti*, a high-elevation salamander in the southern Appalachians. *Molecular Ecology*, **12**, 969–984.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, **25**, 410–418.

Dyer R (2009) GeneticStudio: a suite of programs for spatial analysis of genetic-marker data. *Molecular Ecology Resources*, **9**, 110–113.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences USA*, **107**, 16196–16200.

Felsenstein J (2006) Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution*, **23**, 691–700.

Good JM (2011) Reduced representation methods for subgenomic enrichment and next-generation sequencing. In: *Molecular Methods for Evolutionary Genetics* (eds Orgogozo V, Rockman MV), pp. 85–103. Humana Press, New York, NY.

Gordon A, Hannon GJ (2012) "FASTX-Toolkit", FASTQ/A short-reads pre-processing tools (unpublished) http://hannonlab.cshl.edu/fastx_toolkit/

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamente CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**, 1–11.

Hamady M, Walker J, Harris J, Gold N, Knight R (2008) Error-correcting barcoded primers allow hundreds of samples to be pyrosequenced in multiplex. *Nature Methods*, **5**, 235–237.

Hewitt GM (1996) Some genetic consequences of ices ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, **58**, 247–276.

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

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the USA*, **104**, 2785–2790.

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.

Hugall A, Moritz C, Moussalii A, Stanisic J (2002) Reconciling paleodistribution models and comparative phylogeography in the Wet Tropics rainforest land snail *Gnarosophia bellendenkerensis* (Brazier 1975). *Proceedings of the National Academy of Sciences USA*, **99**, 6112–6117.

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

Knowles LL (2001) Did the Pleistocene glaciation promote divergence? Tests of explicit refugial models in montane grasshoppers. *Molecular Ecology*, **10**, 691–701.

Knowles LL (2009) Statistical phylogeography. *Annual Reviews of Ecology, Evolution and Systematics*, **40**, 593–612.

Knowles LL, Maddison DR (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623–2635.

Knowles LL, Carstens BC, Keat ML (2007) Coupling genetic and ecological-niche models to examine how past population distributions contribute to divergence. *Current Biology*, **17**, 1–7.

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

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2012) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

Morgan K, O'Loughlin SM, Chen BIN *et al.* (2011) Comparative phylogeography reveals a shared impact of pleistocene environmental change in shaping genetic diversity within nine *Anopheles* mosquito species across the Indo-Burma biodiversity hotspot. *Molecular Ecology*, **20**, 4533–4549.

Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, **2**, 175–220.

Nielson M, Lohman K, Sullivan J (2001) Phylogeography of the Tailed frog (As*caphus truei*): implications for the biogeography of the Pacific Northwest. *Evolution*, **55**, 147–160.

Ossowski S, Schneeberger K, Lucas-Lledó JI *et al.* (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92–94.

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

Provan J, Maggs CA (2011) Unique genetic variation at a species' rear edge is under threat from global climate change. *Proceedings Royal Society of London B*, **279**, 39–47.

Rogers AR, Jorde LB (1996) Ascertainment bias in estimates of average heterozygosity. *American Journal of Human Genetics*, **58**, 1033–1041.

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

Shafer ABA, Cullingham CI, CÔTÉ SD & Coltman DW (2010) Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, **19**, 4589–4621.

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.

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction. *American Journal of Human Genetics*, **73**, 1162–1169.

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.

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

Tribsch A, Schonswetter P (2003) Patterns of endemism and comparative phylogeography confirms paleo-environmental evidence for Pleistocene refugia in the Eastern Alps. *Taxon*, **52**, 477–497.

Tsai Y-H, Carstens BC (2012) Assessing model fit in phylogeographic investigations: an example from the North American sandbar willow *Salix melanopsis*. *Journal of Biogeography*, **40**, 131–141.

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

Waltari E, Hijmans RJ, Peterson AT, Nyari AS, Perkins SL, Guralnick RP (2007) Locating Pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS ONE*, **720**, e563.

Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD (2006) Amplification of complex gene libraries by emulsion PCR. *Nature Methods*, **3**, 545–550.

Zellmer AJ, Hanes MM, Hird SM, Carstens BC (2012) Deep phylogeographic structure and environmental differentiation in the carnivorous plant S*arracenia alata*. *Systematic Biology*, **61**, 763–777.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

This work represents the product of a graduate course taught by B.C. at Louisiana State University during the spring of 2012. With two exceptions (J.S. and D.T.), the coauthors were students in this course. Most of the data processing and analyses were conducted by students as a class project. All of the authors are interested in the application of computational approaches to phylogeography and curious about the advantages and disadvantages of using next-generation sequence data in these investigations.

Data accessibility

Data are deposited in GenBank (KC863649–KC863939; JY474042–JY475417; JX187155–JX187361; DQ060270–DQ060277, DQ875023–DQ875049, DQ922722, DQ922769–DQ922773). Input files used in some analyses and scripts used in the processing of these data are available in Dryad (doi:10.5061/dryad.t7r78).

Supporting information

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

Table S1 Sample information. Shown for each *Salix melanopsis* sample are the collection ID, river and locality number, drainage, and latitude and longitude.

Table S2 Number of sequence reads by channel and barcode. Only sequence reads that passed the quality filters are reported.

Table S3 Sequencing data for all polymorphic loci.

Table S4 Results from two IMa2 runs are shown.

Table S5 Results from Migrate-n and Lamarc runs are shown.

Fig. S1 Demographic details of each of the 7 two dimensional models included in the *dadi* analyis are shown.

Fig. S2 Level of SNP variability in *S. melanopsis.*

Fig. S3 Results from the Structure analysis: the change in likelihood as a function of the number of clusters (K).

Fig. S4 Structure diagram.

Fig. S5 Analysis of data using *Beast.