Genotyping-by-sequencing and ecological niche modeling illuminate phylogeography, admixture, and Pleistocene range dynamics in quaking aspen (Populus tremuloides)

Justin C. Bagley1,2,*, Neander M. Heming2, Eliécer E. Gutiérrez2,3, Upendra K. Devisetty4, Karen E. Mock5, Andrew J. Eckert1, and Steven H. Strauss6

1Plant Evolutionary Genetics Laboratory, Department of Biology, Virginia Commonwealth University, 1000 W Cary St., TRANI 126, Richmond, VA 23284, USA
2Departamento de Zoologia, Instituto de Ciências Biológicas, Universidade de Brasília, 70910-900, Brasília, DF, Brazil

*Corresponding author. E-mail: jcbagley@vcu.edu.

JCB: jcbagley@vcu.edu; NMH: neandermh@gmail.com; EEG: ee.gutierrez.bio@gmail.com; UD: upendra@cyverse.org; KM: Karen.E.Mock@usu.edu; AE: Andrew.J.Eckert@vcu.edu; SHS: Steven.H.Strauss@oregonstate.edu

3Programa de Pos-Graduação em Biodiversidade Animal, Centro de Ciências Naturais e Exatas, Universidade Federal de Santa Maria, Santa Maria, RS 97105-900, Brazil
4CyVerse, Bio5, University of Arizona, Keating Bioresearch Building, 0400D12, 1657 E. Helen St., Tucson, AZ 85721, USA
5Department of Wildland Resources and Ecology Center, Utah State University, 5230 Old Main Hill, Logan, UT 84322, USA
6Department of Forest Ecosystems and Society, Oregon State University, 321 Richardson Hall, Corvallis, OR 97331, USA

Author E-mail Addresses: JCB: jcbagley@vcu.edu; NMH: neandermh@gmail.com; EEG: ee.gutierrez.bio@gmail.com; UD: upendra@cyverse.org; KM: Karen.E.Mock@usu.edu; AE: Andrew.J.Eckert@vcu.edu; SHS: Steven.H.Strauss@oregonstate.edu

https://doi.org/10.7287/peerj.preprints.27162v1 | CC BY 4.0 Open Access | rec: 29 Aug 2018, publ: 29 Aug 2018
Abstract (249 words)

*Populus tremuloides* is the widest-ranging tree species in North America, and an ecologically important component of mesic forest ecosystems displaced by the Pleistocene glaciations. Using phylogeographic analyses of genome-wide SNPs (34,796 SNPs, 183 individuals) and ecological niche modeling, we inferred population structure, admixture, and Pleistocene range dynamics of *P. tremuloides*, and tested several historical biogeographical hypotheses. We found three genetic lineages located in coastal (cluster 1), Cascadian/Northern Rocky Mountains (cluster 2), and Southern Rocky Mountains to northern regions (cluster 3) of the *P. tremuloides* range, with phylogenomic relationships of the form ((cluster 1, cluster 2), cluster 3). The main vector of admixture was from cluster 3 into cluster 2, with the admixture zone trending northwest through the Rocky Mountains along a recognized phenotypic cline (Utah to Idaho). Clusters 1 and 2 provided mixed support for the ‘stable-edge hypothesis’ that unglaciated southwestern populations persisted *in situ* since the last glaciation. By contrast, cluster 3 exhibited ‘trailing-edge’ dynamics, e.g. clinal
genetic variation and niche suitability predictions signifying complete northward post-glacial expansion. Results were also consistent with the ‘inland dispersal hypothesis’ predicting post-glacial assembly of Pacific Northwestern forest ecosystems, but rejected the hypothesis that Pacific-coastal populations were colonized during outburst flooding from glacial Lake Missoula. Overall, congruent patterns between our phylogeographic and ecological niche modeling results and fossil pollen data demonstrate complex mixtures of stable-versus trailing-edge dynamics and refugial locations within *P. tremuloides*. These findings confirm and refine previous genetic studies, while strongly supporting a distinct Pacific-coastal genetic lineage of quaking aspen.

**Introduction**

Many North American forest tree species are broadly distributed across large areas of the continent (Little 1971; Prasad *et al.* 2007). Closely related populations and species also frequently occur across major geographical barriers, including western-Pacific mountain ranges (e.g. Cascades, Sierra Nevada), ranges of the greater continental divide (Rocky Mountains) and eastern divide (Appalachian Mountains), or xeric habitats of the Great Basin and major deserts. Explaining such distributions requires historical biogeographical processes of range fragmentation in a wide-ranging ancestral lineage spanning both sides of a barrier, and/or dispersal into areas on either side (e.g. Rosen 1978). Pleistocene glaciations represent another major factor influencing the distributions and genetic diversity of North American tree species (Hewitt 1996, 2001; Soltis *et al.* 2006; Jaramillo-Correa *et al.* 2009). Fossil pollen and plant macrofossil data show that the advance and retreat of massive continental ice sheets during the mid-late Pleistocene
cyclically reduced population sizes and forced extirpations of forest trees from northern
glaciated areas, followed by predominantly northward population/range expansion during
post-glacial recolonization (Jackson et al. 2000; Williams et al. 2004; but see Provan &
Bennett 2008). Teasing apart these historical and evolutionary processes in a geographical
context to explain the diversity, demography, and assembly of North American tree
communities has been a key goal of phylogeography for more than 20 years (Sewell et
al. 1996; Soltis et al. 1997; Mitton et al. 2000; Cheddadi et al. 2006; Fazekas & Yeh
2006; Godbout et al. 2008; O’Connell et al. 2008; Eckert et al. 2010; Gugger et al. 2010;
Keller et al. 2010; Breen et al. 2012).

Quaking aspen, *Populus tremuloides* Michx., is the most widely distributed North
American tree species, ranging from northern Canada southward into pockets of central
Mexico (Fig. 1; Little 1971). While common above 500 m elevation, *P. tremuloides*
occur in distinct microhabitats throughout their range, with montane western populations
occurring in riparian corridor, treeline, and krümmholz stands (Shepperd et al. 2006) and
eastern populations more likely encountered along rivers and riparian zones (Barnes &
Wagner 2002). In addition to striking morphological variation (Barnes 1967, 1975), *P.
tremuloides* possess among the highest genetic diversity reported for *Populus* species to
date (e.g. Jelinski & Cheliak 1992; Cole 2005; Callahan et al. 2013; Wang et al. 2016),
and genetic resources include linkage maps and a draft annotated genome (Pakull et
al. 2009; Sjödin et al. 2009; Sundell et al. 2015). These characteristics make quaking
aspen an ideal system for studying the genomic and ecological contexts of speciation and
local adaptation in forest trees. Historically unglaciated portions of the species range
correspond to major glacial-stage refugia in the Cascades/Northern Rocky Mountains
and the mid-southern Rocky Mountain Front, based on phylogeographic data from many
Plant and animal taxa (reviewed in Brunsfeld et al. 2001; Jaramillo-Correa et al. 2009). Therefore, P. tremuloides also presents outstanding opportunities for testing historical biogeographical hypotheses on the locations of forest tree refugia and the formation of mesic, temperate forest ecosystems of the Pacific Northwest and Rocky Mountains.

Several important gaps in our knowledge of P. tremuloides evolutionary history remain open to investigation, including whether additional data from high-throughput sequencing might confirm or refine previously described intraspecific genetic clusters (Callahan et al. 2013), and the nature of Pleistocene range dynamics [e.g. refugia since the Last Interglacial (LIG); but see Ding et al. (2017)] and their influence on intraspecific genetic variation. To address these knowledge gaps, we investigate the genetic structure and population history of P. tremuloides by integrating phylogeographical analyses of genotyping-by-sequencing (GBS) data (Elshire et al. 2011) with ecological niche modeling (ENM; Peterson et al. 2011) analyses predicting the species Pleistocene to recent geographical range-dynamics. Using a high ratio of genome-wide SNPs to individuals is an optimal sampling design (e.g. Felsenstein 2006) that has been shown to yield high-resolution inferences of population history, even with small–moderate sample sizes (e.g. Gutenkunst et al. 2009; Willing et al. 2012; Robinson et al. 2014; Boehm et al. 2015).

Thus, we sought a balance between numerical and genomic sampling by offsetting losses of information from sampling on average ~ 5 (median: 3; range: 1–22) individuals per local subpopulation (=site) with thousands of unlinked single nucleotide polymorphism (SNP) loci from throughout the genome. Our specific goals were four-fold: 1) to infer broad-scale patterns of population structuring within P. tremuloides using genomic SNP data; 2) to test the presence and directionality of admixture between intraspecific gene pools; 3) to infer Pleistocene range dynamics of the species and its genetic lineages using
ENMs and assess whether genetic differentiation is explained by connectivity or isolation of predicted suitable habitats over the last glacial cycle; and 4) to test several a priori historical biogeographical hypotheses for North American forest trees described below.

Materials and Methods

Historical biogeographical hypotheses.—Our study design permitted testing two pairs of competing a priori historical biogeographical hypotheses, and one standalone hypothesis. First, we expected (H_1) the ‘stable-edge hypothesis’ of long-term persistence of _P. tremuloides_ populations in the intermountain west to be strongly supported if predicted suitable areas were fully or partly stable from the LGM to present, had higher genetic or phylogeographic diversity (number of lineages), and exhibited significant isolation by geographical distance (Callahan _et al._ 2013; Hampe & Petit 2005). The opposite (H_2) ‘trailing-edge hypothesis’ would be supported if predicted suitable areas were completely latitudinally displaced during the LGM and populations exhibited clinal genetic patterns indicating large-scale post-glacial population expansions (e.g. Hewitt 1996, 2001; Hampe & Petit 2005; Excoffier _et al._ 2009). Based on previous studies, we expected stable-edge dynamics in ‘southwestern cluster’ populations and trailing-edge dynamics in ‘northern cluster’ populations of _P. tremuloides_ (Callahan _et al._ 2013). _Populus tremuloides_ also exhibits the ‘mesic forest disjunct pattern’ of Brunsfeld _et al._ (2001), with Pacific-coastal populations separated from interior Rocky Mountain populations by intervening arid habitats of the Columbia Plateau in east–central Washington. Thus, we used our genetic results to test two non-mutually exclusive regional biogeographical hypotheses proposed to explain this distributional pattern. We tested (H_3) the well-known ‘ancient vicariance hypothesis’ positing that Cascade/Cosat range forests
became isolated from Northern Rocky Mountain forests during the Pliocene formation of the Cascades Range (Daubenmire 1975) against (H₄) the ‘inland dispersal hypothesis’ that mesic forests recolonized the Northern Rockies since the LGM without substantial genetic divergence from coastal populations (Brunsfeld et al. 2001; Brunsfeld & Sullivan 2005; Carstens et al. 2005). During post-glacial retreat of continental glaciers, large lakes formed and eventually overflowed, flooding to coastlines and sparking extirpations or dispersals of northern populations of plants and animals from the Pacific Northwest (Pielou 1991). Among the largest, glacial Lake Missoula in western Montana repeatedly flooded the eastern Washington scablands and inundated the Willamette Valley in Oregon and the Columbia Basin to the Pacific coast, with the most massive events occurring since ~20 ka, and most recently ~14.7 ka (Balbas et al. 2017, refs. therein). Thus we tested the ‘Missoula floods hypothesis’ that Pacific-coastal *P. tremuloides* populations were colonized by transfer of propagules during these latest pronounced Missoula floods, which would predict that at least some populations in coastal Oregon and Idaho–Montana exhibit genetic similarity.

**Sampling and laboratory methods for genomic data.**—We obtained samples of leaves or stem cuttings from 96 *P. tremuloides* trees from 33 local sites (1–5 trees per site) across the species native range in western North America (Fig. 1), including samples from 11 sites in Callahan et al. (2013). Additional information on the locations of the sampling sites and tissue sources of individual trees is provided in Data S1 of the Supporting Information. We extracted genomic DNA from tissues using an in-house, phenol–chloroform DNA extraction protocol available in our Mendeley Data accession. Presence and quantity of extracted DNA was quantified using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific), and DNA quality was assessed by electrophoresis.
on 1% agarose gels. DNA concentrations were normalized before a GBS library was prepared containing 96 multiplexed samples according to the genotyping-by-sequencing (GBS) protocol developed for maize (Elshire et al. 2011) using a 5-bp single-cutter restriction enzyme, ApeKI. Single-end sequencing generated 100-bp reads on a single lane of an Illumina HiSeq 3000 flowcell at the Center for Genome Research and Biocomputing at Oregon State University. Base calling was performed in Casava v1.8 (Illumina, San Diego, CA).

**Dataset construction and bioinformatics.**—To increase genome coverage and read depth for SNP discovery, we generated a final genome-wide dataset for analysis by complementing our novel Illumina GBS dataset above with a published *P. tremuloides* dataset from Schilling et al. (2014), obtained directly from the authors. Combining datasets was justified because Schilling et al. (2014) employed the same ApeKI GBS protocol used herein, with samples prepped in two multiplexed libraries each containing 96 individuals and sequenced on separate Illumina HiSeq 2000 lanes. Schilling et al.’s (2014) dataset comprised 107 samples from individuals (including 45 technical replicates) representing six additional subpopulations, including U.S. sites WWA, KFO, and MI and Canadian sites FLFL, HSPQ, and SPQ (details in Data S1).

We conducted reference-based assembly, SNP discovery, and genotyping using the TASSEL-GBSv2 pipeline (Glaubitz et al. 2014) in the TASSEL v5.0 software distribution (Bradbury et al. 2007). In TASSEL-GBSv2, unfiltered Illumina data were demultiplexed by barcode (removing sequences that mismatched barcodes), trimmed to 64 bp, and stored in bit format to reduce computational time. Identical reads were collated as haplotypes, or ‘tags’ (Lu et al. 2013). To allow singletons for improving demographic/gene flow inferences, we set the minimum k-mer count across taxa and read counts per tag to
1. For the reference genome, we used the updated annotated genome for *P. tremuloides* (v1.1, ‘Potrs01b’, 30.1 Mb of sequence) available from The Populus Genome Integrative Explorer website (http://www.popgenie.org/; Sjödin *et al.* 2009; Sundell *et al.* 2015). Reads were mapped to the reference genome based on sequence similarity using the Burrows-Wheeler alignment tool, bwa v0.7.17 (Li & Durbin 2009). SNPs were called from tag alignments starting at the same physical position along the reference genome and filtered using default 10% minimum locus coverage (proportion of individuals) and minimum 0.01 minor allele frequency (MAF) settings. We filtered these SNPs to generate a ‘final’ set of SNPs for downstream analyses by removing indels and four problematic individuals with low barcode assignment success, keeping only biallelic loci, requiring at least two individuals per allele (MAF = 0.0025), and removing SNPs with minimum locus coverage <50%, in vcf tools v0.1.14 (Danecek *et al.* 2011). To evaluate potential effects of including technical replicates, we ran the SNP discovery pipeline a second time on the dataset while excluding technical replicates and compared the results to our original SNPs using Venn diagrams based on the output from vcf-compare (Danecek *et al.* 2011).

**Population genetic diversity, structure, and admixture.**—We used two methods to infer overall patterns of population genetic structure and individual ancestry from our final GBS SNPs. First, we used the model-based approach implemented in ADMIXTURE v1.3.0 to infer individual ancestries in a maximum-likelihood (ML) framework that outperforms STRUCTURE (Pritchard *et al.* 2000) in computational efficiency for large, genome-wide SNP datasets such as ours (Alexander *et al.* 2009). We estimated ancestry coefficients (*Q*) for each individual in 10 replicate ADMIXTURE runs for each of *K* = 1–10 current/ancestral gene pools or ‘clusters’. We also performed two replicate runs of 50-fold cross-validation for *K* = 1–10 to determine potential error in ancestry
estimation of each $K$. We then plotted the errors and identified the $K$-value having the lowest cross-validation error as the best $K$. Second, we estimated genetic clusters and membership probabilities using the model-free discriminant analysis of principal components (DAPC; Jombart et al. 2010) method available in the adegenet R package (Jombart & Ahmed 2011). Unlike ADMIXTURE, DAPC does not make assumptions of panmixia within genetic clusters, or of linkage equilibrium among loci. We objectively identified the appropriate number of PCs using the ‘xvalDapc’ cross-validation procedure, which evaluates prediction success of using DAPC on a training set of 90% of observations from each local subpopulation to group the remaining 10% of observations (Jombart et al. 2010). Results were compared for the number of PCs minimizing mean squared error, and the maximum number of PCs with >90% prediction success.

We evaluated general patterns of genetic diversity by calculating observed ($H_O$) versus expected ($H_E$) heterozygosity, gene diversity ($H_S$), $F_{IS}$, and counts of heterozygote and singleton alleles on per-individual and per-locus bases, and compared these among local subpopulations and ADMIXTURE clusters. Calculations were conducted using vcftools (Danecek et al. 2011), adegenet (Jombart & Ahmed 2011), and hierfstat (Goudet 2005). Given that $P. tremuloides$ likely experienced a drastic northward range expansion following the LGM (e.g. Jackson et al. 2000; Callahan et al. 2013), we tested for prevailing latitudinal or longitudinal clines in $H_O$ and $H_S$ (standardized for sampling effort by dividing population values through by sample size; hereafter, ‘$H_O^*$’ and ‘$H_S^*$’ using generalized linear models estimated in R for the species, and within each genetic cluster herein (see Results). We expected negative linear relationships in populations that experienced extensive post-glacial expansion consistent with the trailing-edge hypothesis (Hewitt 1996; Hampe & Petit 2005; Excoffier et al. 2009). We calculated
pairwise Nei’s (1972) $D$ and $F_{ST}$ (unbiased estimator of Weir & Cockerham 1984) genetic distances between clusters and subpopulations in StAMPP (Pembleton et al. 2013).

We assessed the differentiation of clusters further using a hierarchical analysis of genetic variance in hierfstat (‘varcomp.glob’ and ‘boot.vc’ functions), with trees (‘ind’) nested within local subpopulations (‘pop’), nested within genetic clusters (‘clust’), and significance of $F$-statistics estimated using 95% confidence intervals (CIs) from 100 bootstrap pseudoreplicates.

To evaluate spatial patterns of genetic variation and test the prediction of the stable-edge hypothesis that *P. tremuloides* populations from southwestern North America exhibit isolation by distance (IBD) due to long-term migration-drift equilibrium (Hampe & Petit 2005; Callahan et al. 2013), we tested for IBD using Mantel randomization tests and generalized linear modeling of linearized $F_{ST} = F_{ST}/(1 - F_{ST})$ versus log-geographic distance (after Rousset 1997). Tests used straight-line geographic distance between sampling locations estimated while accounting for the curvature of the earth’s surface (Vincenty inverse solution) in Imap (Wallace 2012). Mantel significance ($\alpha = 0.05$) was assessed using 10,000 randomizations of the data in ade4 (Dray & Dufour 2007). IBD tests were conducted globally, and within each ADMIXTURE cluster.

**Phylogenomic analyses with admixture.**—Prior to phylogenomic analyses, we removed three individuals with more than $\sim$50% missing data (CSS11, GCB5, and SFRG5-4-1, all from different populations; Data S1), yielding a dataset of 180 samples. We used TreeMix v1.13 (Pickrell & Pritchard 2012) to infer a maximum-likelihood tree topology of relationships among genetic clusters identified during our ADMIXTURE and DAPC analyses (see Results) while accounting for admixture among ancestral *Populus* populations. In the first set of analyses, where a *P. trichocarpa* outgroup sample was used to fix
the root, we accounted for linkage disequilibrium by using 500 bp blocks of SNPs, and a no-migration run was followed by a series of replicated runs sequentially adding migration events (−m flag). Migration edges were added until the proportion of variance in the SNP data explained by the model reached ≥ 99.8% (Pickrell & Pritchard 2012). We evaluated the consistency and significance of migration edges by conducting five replicate runs at the final migration level while estimating standard errors of the migration weights. We estimated nodal support for the final topology using 500 bootstrapping pseudoreplicates with $k = 500$ bp blocks of contiguous SNPs, and we determined $p$-values indicating whether supported migration edges significantly improved the fit to the data using the default jackknifing procedure (Pickrell & Pritchard 2012). Support for migration edges or novel admixture patterns were also assessed using plots of the residual fit of the tree graph. We evaluated potential effects on our TreeMix results of the substantial proportion of missing data in the outgroup sample by re-running the TreeMix procedure above while excluding the outgroup sample, which produced unrooted tree topologies. We also conducted a performance analysis assessing the impacts of choice of $k$ (block size) on our results, by reanalyzing the full dataset over the following range of $k$ values: 10, 100, 250, 500, 750, 1000, 2000, and 5000 bp.

**Ecological niche modeling.**—We used *P. tremuloides* occurrence data from throughout the species native distribution from Worrall et al. (2013), as well as from our genetic sampling sites. As the dataset contained $>100,000$ occurrence records and such a large dataset would likely carry elevated geographical or environmental space biases (Reddy & Dávalos 2003; Boria et al. 2014), we decreased the number of records by spatially filtering occurrences located $\leq 10$ km from other occurrences using the spThin R package (Aiello-Lammens et al. 2015). The filtered dataset comprised 14,146 occurrences and was
used in subsequent analyses. Our ENM analyses employed environmental data layers for
19 bioclimatic variables at a resolution of 2.5 decimal degrees, available in the WorldClim
1 dataset (Hijmans et al. 2005). The area used for calibration of the models was created
as a 1.5° buffer around the minimum convex polygon encompassing all occurrence sites
in the filtered dataset (sensu methods described in Anderson & Raza 2010; Barve et al.
2011).

To calibrate our models, we employed the maximum entropy method implemented
in MaxEnt v3.3.3k76 (Phillips et al. 2006, 2017; Phillips & Dudík 2008). Given the
importance of both evaluating model performance with spatially independent data and bal-
ancing model complexity and predictive power (Warren & Seifert 2011; Radosavljevic &
Anderson 2014), we fine-tuned MaxEnt using ENMeval (Muscarella et al. 2014) accessed
through the ENMwizard R package (Heming et al. 2018). We evaluated models using a
geographic partition scheme, and we optimized two important MaxEnt parameters that
impact model complexity and predictive power: the regularization multiplier (RM) and
feature classes (FCs) (Muscarella et al. 2014). To optimize these MaxEnt parameters,
we calibrated preliminary models employing a range of settings including the geographic
‘block’ partitioning scheme (Muscarella et al. 2014), but only varying the RM and FC
values. We used 8 values of RM from 0.5 to 4.0, incremented by 0.5. For each RM,
we conducted 15 preliminary analyses, one with each of the following feature classes or
combination thereof: L, P, Q, H, LP, LQ, LH, PQ, PH, QH, LPQ, LPH, LQH, PPH,
and LPQH, where ‘L’ is linear, ‘P’ is product, ‘Q’ is quadratic, and ‘H’ stands for hinge.
In total, 120 preliminary models were built to select the best settings, including RM and
FC combinations. Model selection was based on the small sample size-corrected Akaike
information criterion (AICc) as the main optimality criterion. Model omission rates cal-
culated with the 10th percentile and minimum training presence thresholds, and the Area Under the receiver operating characteristic Curve (AUC), were also used as secondary criteria for model selection (Peterson et al. 2011; Warren & Seifert 2011). Selected MaxEnt parameters (Supporting Information file Data S2) were used to calibrate a final model using occurrences in the final filtered dataset, but without a geographic partition scheme. The final model was projected onto multiple climatic and paleoclimatic scenarios from late Pleistocene to present (Table 1) within a geographical area covering an extent of 12.5°–75.9° N and 47.0°–172.2° W.

We conducted a second set of ENM analyses to infer potential past to present geographical distributions of genetic clusters within P. tremuloides (see Results), and to test for different range shifts predicted under the stable- versus trailing-edge hypotheses. Calibration areas should not include regions that species/lineages cannot disperse to due to geographical or ecological barriers (Anderson & Raza 2010). Assuming intraspecific genetic clusters have similar dispersal capabilities, we defined calibration areas for each cluster by building minimum convex polygons using the coordinates of genetic sampling sites exclusive to each cluster. We also excluded areas pertaining to other clusters (for example, possible focal cluster absence due to competitive exclusion) and areas with species occurrences pertaining to unidentified lineages, i.e. areas not covered by our genetic sampling. Procedures were performed in R using ENMwizard and raster (Hijmans 2017) (see details in Supporting Information Appendix S1).

Results

Dataset construction, SNP discovery, and SNP filtering.—Sequencing our ApeKI GBS library and combining the raw reads with raw data from Schilling et al. (2014)
yielded a total of 634 million reads (see additional details in Appendix S1 of the Supporting Information). Reference-based assembly and SNP calling in the TASSEL-GBSv2 pipeline yielded a total of 52,410 SNPs representing 202 Populus individuals, and results were 99.4% similar to the initial SNPs when technical replicates were excluded (Venn diagram, Appendix S1). After filtering in vcftools and quality controls, the final variant set contained 34,796 SNP loci for 182 P. tremuloides individuals and one P. trichocarpa individual from a total of 36 sampling sites. The final dataset also contained considerable variation, with global $F_{ST}$ and $F_{IS}$ over all loci by subpopulation (Weir & Cockerham 1984) of 0.148 and 0.186, respectively. Most genetic variation was present in the ingroup, and removing the P. trichocarpa outgroup sample yielded nearly identical global $F$-statistics. Mean depth of coverage per final SNP locus (hence individual) was 13.95, but per-locus values ranged from $\sim$1 to 1091. Genetic diversity at these loci was moderate, for example with mean observed heterozygosity ($H_O$) of 0.13 and mean overall gene diversity ($H_t$) of 0.18 (see details in Table S1 of the Supporting Information).

**Population genetic diversity, structure, and admixture.**—Our ADMIXTURE and DAPC clustering analyses each identified three genetic clusters (Fig. 1). Cross-validation error estimates for the ADMIXTURE models decreased rapidly from $K = 1$ to reach a low point at $K = 3$, the ‘best’ $K$, and then increased steadily to higher levels (Fig. S1). In the final ADMIXTURE model (Fig. 1A), cluster 1 was located along the Pacific Northwestern coastal plain and Coast Range/Cascades in Washington and Oregon. Consistent with predictions of the inland dispersal hypothesis, $H_4$, cluster 2 had a disjunct distribution between the eastern Cascades/Sierra Nevada mountain ranges and the Northern Rocky Mountains, but exhibited limited genetic divergence (Figs. 1B–D). Cluster 3 was present across the remainder of the study area including areas east of the Northern Rock-
ies continental divide in the U.S. and Canada (Fig. 1C). Individuals in clusters 2 and 3, especially *P. tremuloides* from sites CSS, GCB, SFRG, POW, and MON had admixture proportions consistent with putative genomic backgrounds involving introgression (*Q* = 0.15–0.5 for at least one cluster). Also, the *P. trichocarpa* sample had admixture proportions equally divided between clusters 2 (*Q* = 0.48) and 3 (*Q* = 0.52). There were geographical gradients in *Q*-values indicating admixture between neighboring clusters and little or no admixture between disjunct clusters 1 and 3, with the exception being the CSS population from east–central Colorado, which had individuals partially assigned to all three clusters. The *k*-means clustering step of DAPC identified three clusters as the best solution based on the BIC inflection point (Fig. S2). During DAPC cross-validation, prediction success peaked at 20 retained PCs and remained above 90% up to 100 PCs, indicating that lower values optimally minimized error but retaining 100 PCs minimized error while maximizing information content (Fig. S3). DAPC results were identical across 20–100 PCs, and driven by the same SNPs in loading plots of each allele (e.g. Fig. S4); thus only results based on 100 PCs are presented. The final DAPC used two discriminant functions and yielded three clusters that were clearly differentiated along the first axis, matched the geographical pattern of the ADMIXTURE clusters (Fig. 1D), and were 95% similar to those from ADMIXTURE (see additional details in Appendix S1). Thus, we took the ADMIXTURE results as our best estimate of distinct genetic clusters and used them as a priori groups in subsequent genetic analyses.

Patterns of genetic diversity were consistent with deviations from Hardy–Weinberg equilibrium due to pronounced population structure in our dataset, with heterozygosity being substantial and similar among ADMIXTURE clusters (Fig. 3A) but lower than expected under random mating (Fig. 3C). Cluster 1 exhibited substantial heterozygosity,
relatively fewer singletons, and lower inbreeding $F_{IS}$ consistent with limited admixture (Fig. 3B). By contrast, clusters 2 and 3 had relatively elevated inbreeding $F_{IS}$, possibly related to higher $N_e$ or admixture levels. Cluster 3 also exhibited the lowest per-individual counts of singleton SNPs, or private genetic variation (Fig. 3). Re-analyses excluding putatively admixed cluster 2 and 3 edge populations ($Q_{\text{max}} < 0.75$) yielded nearly identical results (Fig. S5), indicating that hybridization likely has not biased our genetic diversity estimates. Additionally, while triploid $P. tremuloides$ are common in intermountain west populations (Mock et al. 2012), we obtained higher cluster 2 and 3 $F_{IS}$ values than we would expect under high triploid frequencies (Fig. 3B), even after excluding admixed populations (Fig. S5), suggesting triploidy likely has not substantially biased our heterozygosity estimates.

Nei’s $D$ estimates, $F_{ST}$ estimates (Table S2), and heatmaps plotted against clustering trees of the distances (Figs. S6 and S7) suggested that cluster 2 was the most divergent from other clusters. Consistent with phylogenetic results below, $F_{ST}$ was also lowest between clusters 1 and 2. Hierarchical genetic differentiation was moderate and significant between clusters relative to the whole species ($F_{\text{clust/total}} = 0.092$, 95% CIs: 0.090–0.094) and between populations within clusters ($F_{\text{pop/clust}} = 0.089$, 95% CIs: 0.088–0.091) (Table S3), and these two levels made relatively equal contributions to genetic variance partitioning among subpopulations ($F_{\text{pop/total}} = 0.173$, 95% CIs: 0.170–0.176).

In line with predictions of the stable-edge hypothesis, standardized $H_{O}^*$ and $H_{S}^*$ relationships with latitude and longitude were non-significant, indicating lack of a genetic signal of directional spatial expansions in clusters 1 and 2 in west–central portions of the species range (Fig. 5). Consistent with genetic predictions of trailing-edge hypothesis and genetic signatures of northward and eastward post-glacial expansions, our generalized
linear modeling results showed that standardized estimates of population genetic diversity significantly declined with longitude overall ($H_O^*: r^2 = 0.13, t = -2.23, df = 32, p = 0.03; H_S^*: r^2 = 0.15, t = -2.19, df = 28, p = 0.037$) and within cluster 3 ($H_O^*: r^2 = 0.53, t = -2.62, df = 6, p = 0.039; H_S^*: r^2 = 0.71, t = -3.14, df = 4, p = 0.035$) (Fig. 5). Genetic diversity within cluster 3 also exhibited a significant negative clinal relationship with increasing latitude ($H_S^*: r^2 = 0.66, t = -2.81, df = 4, p = 0.048$).

Contrasting previous results of Callahan et al. (2013), who reported significant IBD in their southwestern $P. tremuloides$ cluster, we found no evidence for significant IBD overall, or within any of the ADMIXTURE clusters. Although linearized genetic distance declined with log-geographical distance in km between sites when analyzing the full dataset (Mantel’s $r = -0.25$), the pattern was diffuse ($p = 0.99; r^2 = 0.063$). Relationships were even more diffuse but slightly positive during separate analyses of clusters 1–3 ($p > 0.05$, Mantel’s $r$ range = 0.07–0.23; Fig. S8 and Table 2).

**Phylogenomic analyses with admixture.**—The proportion of variance among our SNP loci explained by the tree graph estimated without migration in TreeMix was 99.7% and surpassed the target value of 99.8% when one migration edge was added (‘m1’ model; proportion of variance explained = 100%). Subsequent to this, five replicate runs using the final m1 model consistently yielded the same tree topology and a migration edge from ADMIXTURE cluster 3 into cluster 2 was significant in all runs ($p < 0.001$). The final m1 run that maximized the log likelihood of the model yielded a tree topology placing clusters 1 and 2 more closely related to one another than to cluster 3, and that was strongly supported by bootstrap proportions (Fig. 2A). The corresponding residual plot (Fig. 2B) agreed with patterns of admixture inferred from ADMIXTURE and DAPC results but suggested moderate admixture between cluster 1 and $P. trichocarpa$, despite
current crossability barriers between these species. The ingroup-only analysis excluding
the outgroup sample (~48% missing data) yielded a similar migration event from the
ancestral cluster 3 population into cluster 2, with slightly different residuals (Fig. S9),
suggesting that inclusion of this sample had no adverse effects. Likewise, TreeMix per-
formance was similar over $k$ values varying by two orders of magnitude for the analysis
of the full dataset, indicating that results were insensitive to our choice of block size
parameter to account for LD (Fig. S10).

**Ecological Niche Modeling.**—Our ENMeval analyses identified RM = 0.5 and a com-
bination of product and hinge feature classes as the best-performing parameters for cali-
brating the final ENM. These parameters yielded a single ‘best’ candidate model with the
lowest AICc score (135032.19) and an AICc weight of ~1.0. This best-supported model
had mean omission rates of 0.24 and 0.06 for the 10th percentile and the lowest presence
training thresholds, respectively, as well as a mean test AUC of 0.733. Values of these
diagnostic metrics are provided for all candidate models in Supporting Information file
Data S2.

Projecting the final model onto geographic space revealed large areas holding suit-
able conditions for *P. tremuloides* during the present, ~6 ka in the Mid-Holocene (MH),
and ~125 ka during the LIG, whereas suitable climatic conditions for the species were
restricted to a markedly smaller area ~22 ka during the LGM (Figs. 4 and S11). Dur-
ing present-day, MH, and LIG periods, areas with suitable conditions for *P. tremuloides*
showed a similar pattern, nearly all being located in northern North America and mainly
covering modern-day Alaska, much of Canada, and small patches throughout the north-
western and western US (e.g. Fig. S11A, B, and D). By contrast, predicted suitable
conditions for *P. tremuloides* during the LGM only occurred in areas of the modern-day
United States—mostly in the central and eastern U.S., but also in small scattered disjunct areas mainly throughout the southern Cascade Range, northern Sierra Nevadas, Great Basin, and south–central Rockies (Fig. S11C). These general patterns were consistently found in results based on different general circulation models (Table 1) and indicated an overall pattern of southward range contraction to one or more refugia in south–central areas of North America (Fig. 4).

The minimum convex polygon approach to defining areas for calibrating models of *P. tremuloides* clusters yielded areas from which \( n = 11, 45, \) and 1504 filtered occurrence points were extracted for clusters 1, 2, and 3, respectively (Appendix S1). Projecting final models for each cluster onto geographic space indicated that the overall pattern of range dynamics within *P. tremuloides* resulted from distinct late Pleistocene range shifts of the clusters relative to one another and to the species as a whole. Broadly consistent with the stable-edge hypothesis, cluster 1 predictions exhibited relative stability over the last glacial cycle, with contiguous or disjunct areas of high-predicted habitat suitability signaling refugial areas along the Pacific Northwestern coast, Olympic Peninsula, and Aleutian Islands (Fig. 4). By contrast, predicted suitable habitat areas for cluster 2 extended along coastal mountain ranges and the Northern Rockies for much of the late Pleistocene, but expanded to cover much of northern North America during the LGM and progressively contracted to its modern geographical range through an intermediate MH stage. Predicted interglacial (LIG and present) suitable areas for cluster 3 were scattered across the Rockies and parts of the Great Basin to the west, with a more or less contiguous area across southern Canadian boreal forests from Saskatchewan east to the Atlantic coast (Fig. 4). However, closely matching trailing-edge hypothesis predictions, LGM suitable areas for cluster 3 completely shifted to mid-latitudes in the continental
interior (similar to species-level results), with two minor potential refugial areas located to the west and east. Consistent with the inference of gene flow between clusters 2 and 3 in our TreeMix results, our paleodistribution modeling results inferred overlapping areas of suitable habitat for these two lineages during interglacial periods of the late Pleistocene, including LIG, MH and present-day areas in the Northern Rockies, Middle Rockies, and scattered areas of the intermountain west (Fig. 4). Suitable geographical areas for the genetic clusters were again consistently inferred across MaxEnt analyses, suggesting they were robust to differences among circulation models.

**Discussion**

We inferred the phylogeographic history of *Populus tremuloides* by combining broad-scale inferences of population structure and admixture based on genome-wide SNP data (Elshire *et al.* 2011) with spatially explicit predictions of the past to present geographical distributions of the species and its intraspecific lineages using ENM hindcasting (Waltari *et al.* 2007; Peterson *et al.* 2011). Under this framework, we found strong evidence for significant patterns of population divergence and admixture among three intraspecific genetic clusters. Our genetic and geospatial results also lent mixed support to the stable-edge hypothesis positing that ‘rear edge’ *P. tremuloides* populations persisted long-term in southwestern portions of the range since the LGM. By contrast, trailing-edge dynamics and canonical genetic predictions of post-glacial range expansion (e.g. clinal genetic variation; Hewitt 1996, 2001; Excoffier *et al.* 2009) were strongly supported overall and within cluster 3 (Callahan *et al.* 2013; Hampe and Petit 2005). These results agree with previous genetic findings of Callahan *et al.* (2013), but present a more nuanced picture of *P. tremuloides* evolution and diversification refining the geographical positions of genetic subdivisions, past or on-going admixture, and putative Pleistocene refugia.
Quaking aspen phylogeography, admixture, and Pleistocene range shifts.—

Maximum-likelihood and model-free DAPC analyses strongly supported three intraspecific genetic clusters located in coastal (cluster 1), Cascades Range to Northern Rocky Mountains (cluster 2), and mid-southern Rocky Mountain to northern regions (cluster 3) of the *P. tremuloides* range (Figs. 1, 2, S1, S2 and S9). These clusters were significantly differentiated based on hierarchical $F$-statistics. The genetic and geographical distinctiveness of cluster 1 also supports a unique evolutionary history for the Pacific aspen populations, as suggested by taxonomists more than 100 years ago (Piper & Beattie 1915), and hence may warrant formal taxonomic recognition.

Following the divergence of *P. tremuloides* from a common ancestor with *P. trichocarpa*, our TreeMix species tree registered the deepest split between clusters 1 + 2 versus 3. This matches the principal genetic break reported in Callahan *et al.* (2013) based on microsatellite DNA markers and reveals our clusters 2 and 3 to be roughly analogous to their southwestern and northern clusters, respectively. However, Callahan *et al.*’s (2013) sampling in the zone between their main clusters was too sparse to clearly delineate the break beyond reference to the continental divide. We sequenced material from site POTR and two other sites near their break zone (BNF and their USF site, labeled SFRG herein) but also added samples from nearby stands in northeastern Washington, northwestern Montana, and Colorado. Probably due to our much broader sequencing of random nuclear loci from throughout the genome, piecemeal advances in geographical sampling coverage permitted us to redefine POTR as belonging to our southwestern cluster 2 rather than northern cluster 3, and to show that populations on a diagonal from northern Montana to central Colorado belong to cluster 3. Rather than strictly tracking the continental divide, the deepest genetic subdivision within *P. tremuloides* matches a
set of geological and elevational barriers in the Northern to Middle Rockies, but then de-
viates westward from the divide (which trends southward through central Colorado and
western New Mexico), with the presumed barrier then correlating to cold xeric desert
and shrubland habitats of the northern Great Basin and Snake River Plain.

A common Pleistocene biogeographical pattern in the Northern Hemisphere is
isolation in allopatric refugia followed by post-glacial dispersal and secondary contact of
independent evolutionary lineages (Avise 2000; Hewitt 2001; Swenson & Howard 2005).
Cold desert regions of the break zone above agree well with the split between the “south-
west-south” and “south-west-north” sub-clusters uncovered in Callahan et al. (2013),
which they hypothesized to form a secondary contact zone in the eastern Great Basin.
By contrast, we found individual admixture proportions within several populations (ad-
mixed individuals, left side of cluster 3 in Fig. 1A) indicating a zone of admixture or
secondary contact between the cluster 2 and 3 lineages, but restricted to habitats along
the Rocky Mountains, and our TreeMix analyses clarified the main direction of admixture
as being from cluster 3 into 2 (Figs. 2, S9 and S10). We hypothesize that this putative
area of hybridization spans from the Wasatch Range and Southern Rocky Mountains of
Utah and Colorado (including sites SFRG and CSS) through the Northern Rockies (Bit-
terroot, Lewis, and Absaroka–Beartooth Ranges; including MON site) and into northern
Idaho (POW site). Recent comparative analyses suggest that the Rocky Mountains rep-
resent a hotspot for phylogeographic breaks, hybrid zones, and contact zones across plant
and animal taxa (Swenson & Howard 2005). This area, from southern Utah through
northern Montana and Idaho, also corresponds to a known phenotypic cline in P. tremu-
loides leaf shape, size, and tooth number (Barnes 1975). Taken together, the intermediate
SNP genotypes and clinal morphology of P. tremuloides in this region match theoretical
and empirical expectations for tension zones formed by secondary contact along physical barriers, with clines maintained not by steep environmental gradients but by a balance between dispersal and local adaptation (e.g. Barton & Hewitt 1985). A biogeographical scenario in which this zone was formed by post-glacial secondary contact is supported by generally disjunct areas of niche suitability of these lineages during the LGM, followed by predicted MH and present-day distributional overlap within the Rocky Mountains (Fig. 4), which may have facilitated gene flow since ~6 ka. Several paleoecological records from the western US and Canada suggest that expansion and contact of cluster 2 and 3 lineages, if post-glacial, probably happened following ~13 ka. First, fossil pollen records from Washington (Whitlock 1992), Alberta (Lichti-Federovich 1970), Wyoming (Whitlock 1993), and southeast to Utah (Howard 2016) do not indicate substantial (stable) percentages of *Populus* pollen in this area until ≤12–10 ka, when aspen joined pioneer forest and shrub communities that overtook preceding tundra and grassland/steppe habitats. Second, the final retreat of the Laurentide Ice Sheet opened an ice-free Canadian corridor by ~13 ka, which was only recolonized northward by closed forests ~5 ka (Lichti-Federovich 1970; Pielou 1991).

**Stable- vs. trailing-edge dynamics and isolation by distance.**—Testing for stable-versus trailing-edge dynamics (cf. Hampe & Petit 2005) is challenging, and distributional or genetic approaches alone are insufficient for distinguishing between multiple historical scenarios or processes that could produce similar patterns (e.g. Knowles et al. 2007; Gugger et al. 2010; reviewed by Gavin et al. 2014). As expected from previous genetic results (Callahan et al. 2013) and theoretical predictions that populations near refugial locations should harbor greater genetic diversity (Hewitt 1996, 2001; Excoffier et al. 2009), we recovered clear genetic imprints of post-glacial range expansion for *P. tremuloides* as a
whole and for cluster 3 (e.g. Figs. 3 and 5), supporting the trailing-edge hypothesis (H$_2$). These results are indirectly supported by spatially explicit predictions of Pleistocene suitable areas for $P. $tremuloides$ and its intraspecific clusters from ENM hindcasting (Fig. 4). While standardized genetic diversity showed significant relationships with longitude, the best genetic evidence for northward and eastward spatial-demographic expansions came from significantly negative clinal patterns of cluster 3 $H_{O^*}$ and $H_{S^*}$ with latitude and longitude, respectively (Fig. 5). These are precisely the patterns of expansion suggested by our ENM hindcasting results for $P. $tremuloides$ cluster 3 over LGM to present. These integrated genetic and ENM results agree with other studies of Pacific Northwestern plants that found reductions of genetic diversity during northward post-glacial expansions in perennials and conifers (reviewed by Soltis et al. 1997; Jaramillo-Correa et al. 2009). The inferred contraction of cluster 3 to a mid-continental glacial refugium likewise matches well with previous LGM–recent ENM hindcasting results (Ding et al. 2017) and paleobotanical records of $Populus$ species shifting to cool mixed boreal and non-analog forests in the same area ∼18 ka during the LGM and subsequently expanding northward to higher temperate latitudes with other arboreal species (e.g. $Alnus$ and $Abies$ pollen types; Jackson et al. 2000; Williams et al. 2004; Breen et al. 2012).

We found evidence of stable-edge dynamics in southwestern areas of the $P. $tremuloides range, including a greater diversity of phylogeographic lineages (Fig. 1), fully or partly stable niche suitability areas for clusters 1 and 2 over LGM–present (Fig. 4), and higher population genetic diversity in the west, with heterozygosity declining with longitude across the species range (Fig. 5). This supports Callahan et al.’s (2013) hypothesis that mountainous terrain promoted elevational range shifts consistent with stable-edge dynamics in southwestern $P. $tremuloides$ populations during the late Quaternary. How-
ever, support for stable-edge dynamics in clusters 1 and 2 was mixed, with the chief departure from expectations being a lack of significant isolation by distance (sensu Wright 1943) during Mantel tests (Mantel 1967) and linear modeling of genetic and geographic distances (Fig. S8 and Table 2). This outcome is difficult to interpret, given it could indicate that assumptions of IBD tests such as migration–drift genetic equilibrium or spatial and environmental homogeneity (in ≥1–2 dimensions) are not met for a particular gene pool. We interpret non-significant IBD test results for clusters 1 and 2 as suggesting that a stepping-stone or hierarchical model does not fit the data (e.g. Slatkin 1993; Bohonak 2002), and that a scenario of genetic drift within genetically cohesive clusters is too simplistic. If clusters 1 and 2 actually experienced stable-edge dynamics suggested by our genetic and ENM results, then this could be explained by local subpopulations experiencing varying levels of isolation and migration. For example, this could reflect the Cascadian–Northern Rockies disjunction seen in cluster 2 subpopulations, or clusters having low $N_e$ or migration levels near range margins could have caused $F_{ST}$ to become under-predicted by an IBD model (see McRae 2006). Still, cluster 1 exhibited greater variation in linearized $F_{ST}$ than the other clusters, consistent with more ancient and genetically differentiated populations along stable edges (upper right, Fig. S8). We hypothesize that differences in the significance of IBD tests for the southwestern subpopulations between the two studies may be attributable to sampling artifacts caused by Callahan et al.’s (2013) inclusion of a highly disjunct Mexican subpopulation. Nevertheless, to obtain a better understanding of the effects of dispersal limitation in $P. tremuloides$, or potential for long-distance dispersal disrupting IBD patterns, we recommend additional studies extending our analyses of genome-wide SNPs to a greater density of southwestern subpopulations.
Biogeography of Pacific Northwest and Rocky Mountain mesic forests.—The question of how disjunct mesic forest species of the Pacific Northwest came to obtain their present distributions has long fascinated biogeographers (Daubenmire 1975; Brunsfeld et al. 2001; Brunsfeld & Sullivan 2005). While dominated by Douglas-fir and ‘cedar–hemlock’ forests interspersed with xerophytic pinelands (Brunsfeld et al. 2001), this ecosystem hosts three species of *Populus*—*P. angustifolia*, *P. tremuloides*, and *P. trichocarpa* (Little 1971; Eckenwalder 1996). *Populus tremuloides* cluster 2 exhibits a disjunct distribution between the Cascades/Sierra Nevada ranges and the Northern Rocky Mountains, which is a common forest pattern thought to be maintained by the action of arid shrubland/steppe habitats of the Columbia Plateau as a barrier to gene flow (Brunsfeld et al. 2001). Phylogeographic structuring and past Pleistocene niche suitability patterns inferred herein suggest that *P. tremuloides* cluster 2 entered this ecosystem and obtained a disjunct mesic forest distribution in the Pleistocene, after two major historical events impacted the region: 1) Pliocene uplift of the Cascades Range and 2) xerification of the Columbia Plateau ~2 million years ago (Ma) in the early Pleistocene (Daubenmire 1975; Brunsfeld et al. 2001). The chief reason for this is that, given Pleistocene glacial stages were much longer than relatively short ~10,000- to 20,000-year interglacials, the stability of *P. tremuloides* cluster 2 suitable areas in these mountain ranges over the last glacial cycle (Fig. 4) opens the possibility that the ancestral cluster 2 stock persisted in these areas during earlier glaciations over the last ~800,000 years (Pielou 1991; Lambeck et al. 2002). Nevertheless, we find limited genetic divergence between Cascade and Northern Rocky Mountain populations, as indicated by $F_{ST}$, Nei’s $D$, and clustering tree topologies and heatmaps of the distances (Figs. 2, 3, S6 and S7), suggesting that major Plio–Pleistocene vicariant events above never sundered a continu-
ous ancestral population. Overall, these results reject the ancient vicariance hypothesis (H₃) but more closely resemble patterns of genetic variation expected under the inland dispersal hypothesis, our H₄ (e.g. Brunsfeld et al. 2001). Genetic predictions of the inland dispersal hypothesis have similarly been supported by previous phylogeographic results for other forest tree species (Carstens et al. 2005; O’Connell et al. 2008).

Our phylogeographic results also clearly demonstrate that catastrophic post-glacial floods that swept across the Pacific Northwestern landscape repeatedly during the Pleistocene likely have not played a major role in shaping population structuring within *P. tremuloides*. In contrast to initial predictions for the Missoula floods hypothesis, *P. tremuloides* from the Pacific coast and Idaho–Montana were classified by ADMIXTURE and DAPC into separate genetic clusters (Fig. 1). This is despite our ENM results for the LGM, under multiple paleoclimatic scenarios, showing that suitable habitat likely existed for *P. tremuloides* across the Columbia Basin and nearby regions from LGM to present, particularly for cluster 2 (Fig. 4). Moreover, if Missoula outburst flooding had transferred *P. tremuloides* from Montana to the Pacific coastal zone, we might expect similar patterns of admixture in these areas; however, this expectation was neither met by patterns of admixture coefficients (Fig. 1A) nor migration edges supported by multiple TreeMix analyses (Figs. 2, S9, and S10). We hypothesize that the formation and draining of post-glacial lakes including Missoula did not majorly impact *P. tremuloides* population structure due to the wind-dispersed nature and peculiar life-history strategy of this species (e.g. clonal stands throughout many western areas of the species range). Given evidence of an important effect of Missoula floods on cold-tolerant freshwater fishes and mammals (e.g. Miller et al. 2006; Young et al. 2017, refs. therein), we suggest that aquatic plants or rodent-dispersed tree species may be more likely to exhibit the genomic
signatures of such events, and thus might be more fruitful targets for future studies of outburst flooding effects on regional plant communities.

**Comparative biogeography of North American *Populus***.—Our results yield a late Pleistocene biogeographical scenario for *P. tremuloides* starkly contrasting that proposed for some closely related *Populus* species, while resembling others. For example, balsam poplar, *P. balsamifera*, is thought to have persisted in a Beringian refugium whose genetic variation was subsequently swamped by post-glacial gene exchange with southern colonists (Breen *et al.* 2012), which possibly expanded northward from Central refugium source populations (Keller *et al.* 2010). By contrast, our ENM results indicate that *P. tremuloides* was probably never present in sheltered microhabitats or contiguous ice-free areas of the Beringian land bridge refuge during the last ice age (Fig. 4), as previously suggested for *P. balsamifera*. Results for *P. tremuloides* herein more closely resemble previous ENM hindcasting predictions for *P. trichocarpa*, which Levsen *et al.* (2012) found to suggest patterns of southern refugia and coastal stability, with northward post-glacial recolonization of Pacific Northwestern and Alaskan areas.

**Acknowledgements**

Research was crowdfunded through Experiment.com (https://experiment.com/projects/the-lost-aspens-of-the-willamette-valley-did-catastrophic-floods-carry-them-from-the-rockies), and we thank all 68 donors (Appendix S1) but recognize T. S. Leatherman for an especially generous donation. JCB received stipend support from U.S. NSF grant EF-1442486 to AJE, and NMH was supported by a postdoctoral fellowship from the Brazilian federal agency CAPES. K. House and N. Bruttell are thanked for articles in *The Oregonian* and *Corvallis Gazette-Times* popularizing the project. We thank E. R. Alverson (The Nature Conservancy), B. Ripple and K. Ault (Oregon State University), S. Gremel (National Park Service), K. Karoly (Lewis and Clark College),
and V. Hipkins and A. Chrisman (USDA Forest Service) for providing stand locations, DNA, or leaf material from aspen populations. We also recognize K. Vining (Oregon State University) and C. Zonick and A. Whitacre (Portland Metro) for conducting aspen collections, B. Greer for insights on aspen biogeography, and G. Meacham and J. F. Schmidt & Son Co. for support toward native aspen variety development using our genetic results.

References


Bohonak, A. J. (2002). IBD (isolation by distance): a program for analyses of isolation by


Gavin, D. G., Fitzpatrick, M. C., Gugger, P. F., Heath, K. D., Rodríguez-Sánchez, F., Do-


**Data Accessibility**

Raw sequence data are deposited in the NCBI Sequence Read Archive database (BioProject XXXXXXX). The DNA extraction protocol, 183-taxon genotype file with SNP calls in 012 format, species occurrence records, and R scripts used during the ENM analysis are available in our Mendeley Data accession (<http://dx.doi.org/10.17632/jhkhvdgyfy.1>).

**Supporting Information**
Appendix S1  Supplementary methods and results.

Data S1  Sample list, collection site details, and summary of individual assignments to genetic clusters in ADMIXTURE and DAPC analyses.

Data S2  ENM modeling results summary, including model selection results and diagnostic statistics (e.g., AUC) from ENMeval (Muscarella et al. 2014) and MaxEnt model metrics from ENMwizard (Heming et al. 2018).

Table S1  Per-locus genetic summary statistics for *P. tremuloides* and *P. trichocarpa* individuals in this study.

Table S2  Pairwise $F_{ST}$ estimates between *P. tremuloides* genetic clusters, and their 95% confidence intervals (CIs).

Table S3  Hierarchical analysis of genetic variance among *P. tremuloides* SNPs within genetic clusters ($F_{\text{clust/total}}$), within subpopulations by cluster ($F_{\text{pop/clust}}$), and within individual trees relative to populations ($F_{\text{ind/pop}}$) and the total variance ($F_{\text{ind/total}}$). Ranges given in parentheses are 95% confidence intervals.

Figure S1  Plot of ADMIXTURE cross-validation error versus $K$, showing that $K = 3$ is the best fit for the full *P. tremuloides* and *P. trichocarpa* dataset of 34,796 SNPs (sensu Alexander & Lange 2011).

Figure S2  Plot of Bayesian information criterion (BIC) scores for $k$-means clustering solutions over a range of $K$, from the first step of DAPC, with $K = 3$ being the best solution.

Figure S3  Results of DAPC cross-validation in R establishing that the appropriate number of principal components to retain ranges from 20–100 with similarly high (>90%) prediction success.

Figure S4  DAPC loading values plotted for all 34,796 SNPs, with SNP name labels beside the SNPs with the highest loadings.

Figure S5  Genetic patterns of heterozygote and singleton alleles within and among *P. tremuloides* genetic clusters, calculated while excluding putatively admixed edge populations ($Q_{\text{max}} < 0.75$). Results are analogous to corresponding panels of Fig. 3 (see Fig. 3 caption for details).

Figure S6  Heatmap of interindividual Nei’s $D$ estimates, reordered by row and column means, and flanked by dendrograms of the values. Color key and histogram at top left show the distribution of mean $D$ values.

Figure S7  Heatmap of unordered interpopulation $F_{ST}$ estimates flanked by clustering dendrograms from the distances. Color key and histogram at top left show the distribution of mean $F_{ST}$ values.

Figure S8  Results of isolation by distance tests based on generalized linear modeling analyses of linearized $F_{ST}$ versus log[geographic distance (km)] of *P. tremuloides* (top left) and its genetic clusters 1 (top right), 2 (bottom left), and 3 (bottom right).
**Figure S9**  Unrooted maximum-likelihood tree topology from the ingroup-only TreeMix analysis allowing a single migration event (A), and residual plot of the graph (B). Scale bars and legends same as in Fig. 2.

**Figure S10**  Results of TreeMix analyses on the full dataset conducted over varying levels of \( k \) block sizes (10 to 5000 bp) accounting for linkage disequilibrium. Scale bars and legends same as in Figs. 2 and S9.

**Figure S11**  Projection of the final present-day ecological niche model of *P. tremuloides* onto three late Pleistocene climatic scenarios. The ENM was built with MaxEnt and bioclimatic variables obtained from WorldClim 1, and the three Pleistocene time-slices (B–D) correspond to scenarios described in the Fig. 4 caption and Table 1. Model projections show continuous suitability values obtained using the cloglog format of MaxEnt after the application of a 10\(^{th}\)-percentile threshold. Extent of LGM ice sheets is indicated in white.
Tables and Figures

**Table 1** List of climatic scenarios onto which *P. tremuloides* ecological niche models were projected. Raster files were obtained from WorldClim (Hijmans *et al.* 2005). Time period is given either as date in the Gregorian calendar system, or as thousands of years ago (ka) in the case of paleoclimatic environmental reconstructions.

<table>
<thead>
<tr>
<th>Climatic scenario</th>
<th>Time period</th>
<th>GCM†</th>
<th>Resolution‡</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mid-Holocene (MH)</td>
<td>~6 ka</td>
<td>CCSM4</td>
<td>2.5°</td>
<td>Gent <em>et al.</em> (2011)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IPSL-CM5A-LR</td>
<td></td>
<td>Kageyama <em>et al.</em> (2013a,b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MPI-ESM-P</td>
<td></td>
<td>Stevens <em>et al.</em> (2013)</td>
</tr>
<tr>
<td>Last Glacial Maximum (LGM)</td>
<td>~22 ka</td>
<td>CCSM4</td>
<td>2.5°</td>
<td>Gent <em>et al.</em> (2011)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MPI-ESM-P</td>
<td></td>
<td>Stevens <em>et al.</em> (2013)</td>
</tr>
<tr>
<td>Last Interglacial (LIG)</td>
<td>~120–140 ka</td>
<td>NCAR-CCSM</td>
<td>0.5°</td>
<td>Otto-Bliesner <em>et al.</em> (2006)</td>
</tr>
</tbody>
</table>

†General circulation model names.

‡Resolutions of corresponding raster files in decimal degrees.
**Table 2** Results of Mantel tests for isolation by distance investigating relationships between $F_{ST}$ and geographical distances within *P. tremuloides*

<table>
<thead>
<tr>
<th>Group</th>
<th>Mantel test</th>
<th>$r$</th>
<th>$r^2$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. tremuloides</em></td>
<td>linearized $F_{ST} \times \log[\text{straight-line dist. (km)}]$</td>
<td>$-0.252$</td>
<td>$0.063$</td>
<td>$0.99$</td>
</tr>
<tr>
<td>cluster 1</td>
<td>linearized $F_{ST} \times \log[\text{straight-line dist. (km)}]$</td>
<td>$0.228$</td>
<td>$0.052$</td>
<td>$0.13$</td>
</tr>
<tr>
<td>cluster 2</td>
<td>linearized $F_{ST} \times \log[\text{straight-line dist. (km)}]$</td>
<td>$0.103$</td>
<td>$0.011$</td>
<td>$0.27$</td>
</tr>
<tr>
<td>cluster 3</td>
<td>linearized $F_{ST} \times \log[\text{straight-line dist. (km)}]$</td>
<td>$0.071$</td>
<td>$0.005$</td>
<td>$0.33$</td>
</tr>
</tbody>
</table>

Results are shown for simple Mantel (1967) tests performed for the species as a whole, and for each of the ADMIXTURE-inferred genetic clusters within *P. tremuloides* presented in Fig. 1. Mantel’s $r$ is the standardized test statistic and is equivalent to Pearson’s $r$, and $r^2$ is the coefficient of determination.
Figure 1 Map of sampling sites and patterns of population genetic structure of *Populus tremuloides* and outgroup (*P. trichocarpa*) individuals inferred from 34,796 SNP loci. Results are shown for ADMIXTURE (Alexander *et al.* 2009) assignment of individuals to each of $K = 3$ clusters (A) plotted onto topographic maps of the study area (B, C) with the present species ranges from Little (1971) shown in varying shades of gray. Pie charts show the per-site average ancestry assignment to each genetic cluster ($Q \geq 0.5$) per site and colors correspond to the ADMIXTURE barplot. Discriminant analysis of principal components (DAPC; Jombart *et al.* 2010) in adegenet (Jombart & Ahmed 2011) yielded a classification that was 95% similar to ADMIXTURE, with results plotted along the first and second discriminant functions and colored by cluster (D). Intermediate, encircled individuals in panel D were assigned to cluster 1 (solid ellipse) or cluster 3 (dotted ellipse) in ADMIXTURE. Positions of the outgroup sample are indicated by bold arrows throughout.
**Figure 2** Phylogenomic inference of relationships among *P. tremuloides* genetic clusters, and admixture from cluster 3 into cluster 2. The final maximum likelihood tree graph inferred by phylogenomic analysis in TreeMix v1.13 (Pickrell & Pritchard 2012) allowing a single migration event (A) is presented with the migration event arrow colored by migration weight and clusters indicated by colors corresponding to Fig. 1. Scale bar: 10 times the mean standard error in the sample covariance matrix. To show putative patterns of admixture not apparent from the tree and migration edges, the residual fit of the graph is plotted and colored by the standard error color palette (B).
**Figure 3** Genetic patterns of heterozygote and singleton alleles within and among *P. tremuloides* genetic clusters. Panels present ranges of observed numbers of heterozygote genotypes (A); ranges of inbreeding $F_{IS}$, a measure of heterozygote deviation from Hardy–Weinberg expectations (B); observed heterozygosity ($H_O$) plotted against expected heterozygosity ($H_E$), with deviations from 1:1 mainly below the dotted identity line (C); and ranges of observed numbers of singleton genotypes.
Figure 4 Predicted geographical distributions of *P. tremuloides* and its intraspecific genetic clusters from the late Pleistocene to present. MaxEnt models predicting present-day ranges of *P. tremuloides* and clusters 1–3 (top row) based on WorldClim 1 (Hijmans et al. 2005) were projected onto data layers for three late Pleistocene time-slices described in Table 1: Mid-Holocene (MH), Last Glacial Maximum (LGM), and Last Interglacial (LIG). Predictions are shown as color gradients from low (gray–dark yellow) to high (dark green) habitat suitability based on logistic output of the models (probability from 0 to 1), and the extent of land during each time period is shown in gray. The extent of glacial ice during the LGM is indicated in white.
**Figure 5** Latitudinal and longitudinal clines in SNP genetic diversity within *P. tremuloides* subpopulations. Linear relationships of standardized observed heterozygosity (\(H_O^*\)) and gene diversity (\(H_S^*\)) with latitude (top two rows) and longitude (bottom two rows) are presented for *P. tremuloides* as a whole, and for each ADMIXTURE-inferred genetic cluster. Each dot represents a population and is colored to match the colors of genetic clusters defined in Fig. 1, and regression lines are drawn from fitted values of generalized linear models. For significant relationships, *p*-values are given in the upper right of the plot.