Appendix S1. Supplementary methods and results.

Additional sampling methods.—We obtained samples of leaves or stem cuttings from 96 *P. tremuloides* trees from 33 local sites across the species native range in western North America for sequencing (Fig. 1). Samples preserved for genetic work included 1–5 individuals per site (mean = 5 individuals per site) from a Canadian site in Alberta as well as US samples from the following states: Washington (WA), Oregon (OR), California (CA), Idaho (ID), Utah (UT), Montana (MT), and Colorado (CO). Additional information on the locations of the sampling sites and tissue sources of individual trees is provided in Data S1 of the Supporting Information.

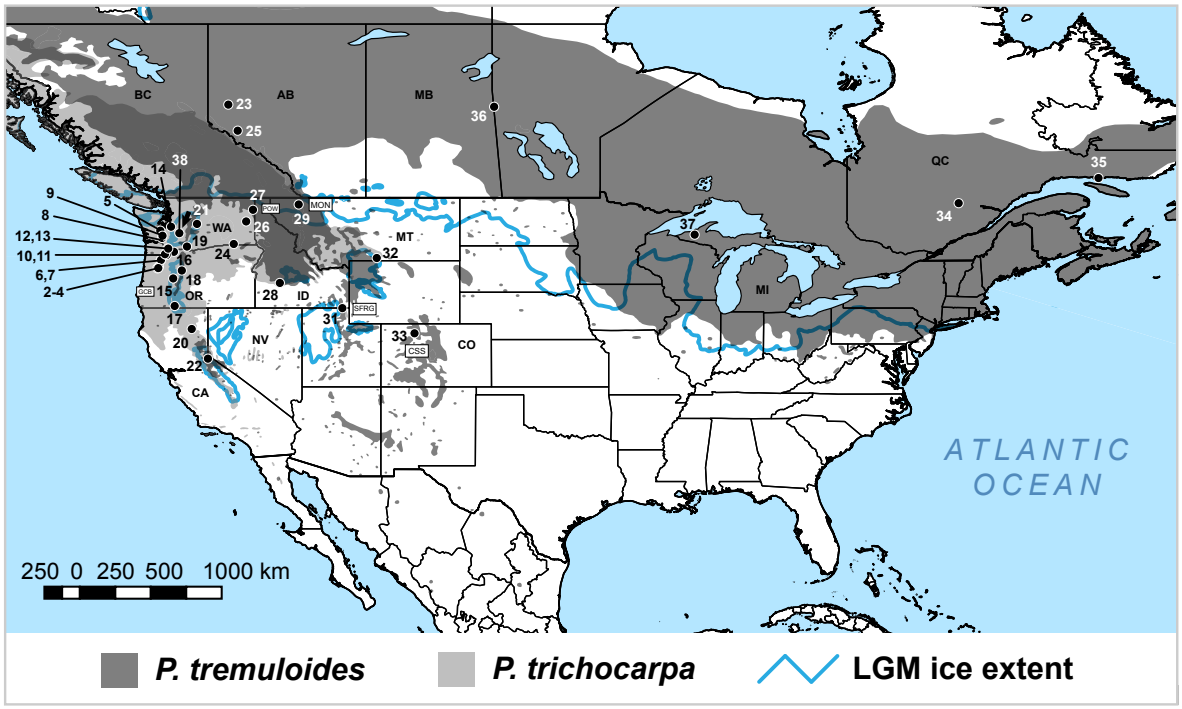

Figure A1. Map of geographical sampling sites for sequenced *P. tremuloides* in the final dataset. Map is similar to Fig. 1, except sampling sites (black dots) are numbered to match Data S1 of the Supporting Information, and the extent of continental ice sheets during the Last Glacial Maximum (LGM) is shown with a solid light blue line. State abbreviations follow the main text and Fig. 1.

We obtained a final dataset comprised of 34,796 SNPs representing 183 *Populus tremuloides* Michx. from 36 natural populations from throughout the western and central–northern portions of the species native distribution (Little 1971). However, due to space limitations, our presentation of sampling localities jointly with population structure results in the main text (Fig. 1) precluded full presentation of mapped sampling locality occurrence points and numbers, which would allow the reader to cross-reference between the main text and other files including geographical sampling details in supplementary file Data S1. *Above*, in appendix **Figure A1**, we provide a map of numbered local sites represented in the final *P.*

tremuloides dataset with site numbers corresponding to Data S1. The map uses the same coordinate system (WGS 1984) as that shown in Fig. 1, and was generated in QGIS v2.14 'Essen'.

Additional bioinformatics and sequencing methods and results.—Illumina sequencing on our *Ape*KI GBS library yielded a total of 382 million reads after initial base calling but prior to quality control or further analyses, with an average of 3.9 million reads per sample. The total number of 'good' barcoded reads with clear assignment to samples was 321 million (85%), and the total number of unique tags retained was 17.5 million (5.5%). Out of 96 samples, 17 samples failed quality checks, based on having <10% of mean reads per sample (Supporting Information Table 2/S2). Combining our GBS dataset with 313 million raw reads from Schilling *et al.* (2014) (see Methods) increased the total number of barcoded reads to 634 million, with an average of 3.1 million reads per sample.

In our final set of filtered SNPs, single nucleotide patterns were biased towards A/G and C/T SNPs [A/C: 3704 (10.6%); A/G: 9847 (28.3%); A/T: 4758 (13.7%); C/G: 2597 (7.46%); C/T: 10,125 (29.1%); G/T: 3765 (10.8%)]. Also, in this dataset, transitions (Ts: 19,972) outnumbered transversions (Tv: 14,824) by a ratio of 1.35.

As noted in the main text, we re-ran the TASSEL-GBSv2 pipeline (Glaubitz *et al.* 2014) on a second version of our final dataset (our GBS data plus raw GBS data from Schilling *et al.* 2014) while removing the 45 technical replicates from Schilling *et al.* (2014). This was accomplished by re-running the pipeline while removing the technical replicate barcodes and IDs from the key file. We then used vcf-compare to calculate the numbers of SNPs within, and shared between, VCF (variant call format) files resulting from the original TASSEL-GBSv2 run and the no-technical-replicates run. Additionally, we determined the IDs of all SNPs in each file, as well as the intersection of SNP IDs shared between the two VCF files from these independent runs using vcftools and regex. To visualize these results, we generated a Venn diagram by inputting SNP IDs from each file into the jvenn web server (http://jvenn.toulouse.inra.fr/app/example.html; Bardou *et al.* 2014). Results are shown in **Figure A2** *below* (*next page*) and demonstrate 99.4% similarity of the no-technical-replicate run SNPs and the original SNPs, suggesting that inclusion of technical replicates did not have a large influence on our SNP discovery results.

Additional ecological niche modeling methods and results.—We modeled the existing fundamental 'Grinnellian' niche of *P. tremuloides* to infer the geographical distributions of climatically suitable conditions for the species both contemporaneously and in the past (see Peterson *et al.* 2011 for a review of niche theory and ENM terminology). We used ENMeval (Muscarella *et al.* 2014) to optimize regularization multiplier (RM) and feature class (FC) parameters of the MaxEnt model (Phillips *et al.* 2006), which are known to affect model complexity, overfitting, and predictions. The RM parameter is important because it penalizes overly complex models, whereas the FCs are functions of the raw environmental data (Phillips *et al.* 2006, 2017; Phillips & Dudík 2008; Elith *et al.* 2011). Also in ENMeval, we implemented a geographic partitioning scheme. Specifically, because we transferred the present-day ENMs backwards into late Pleistocene environments by projecting the models onto paleoclimatic scenarios listed in Table 1, we used the 'block' partitioning scheme, which is recommended for modeling applications that requires model transferences across time periods (Muscarella *et al.* 2014).

Additionally, our ENM analyses relied on the 19 bioclimatic-environmental variables in the WorldClim 1 dataset (Hijmans *et al.* 2005). These data include monthly averages of variables derived from precipitation and temperature values initially recorded by weather stations (worldwide)

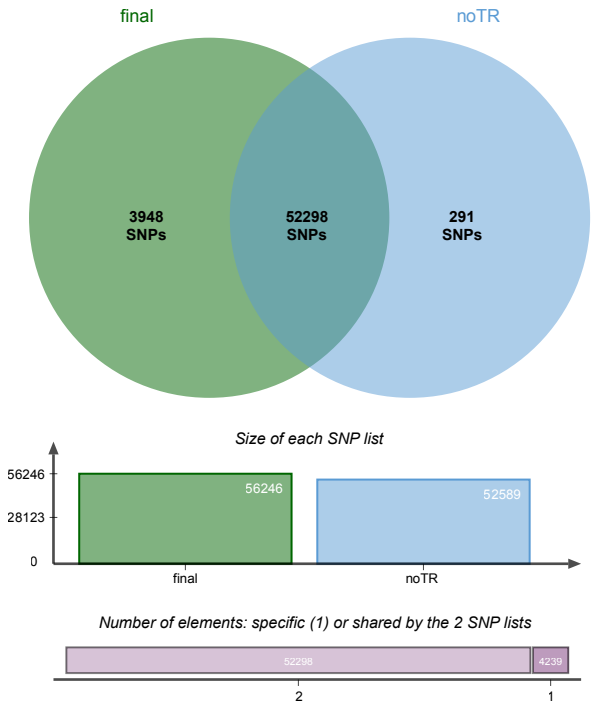

Figure A2. Venn diagram describing the patterns and intersection of SNPs resulting from two independent runs of the TASSEL-GBSv2 pipeline (Glaubitz *et al.* 2014) on our final dataset: the original run ('final') and a run excluding technical replicates ('noTR'). Results were generated from SNP lists in VCF files from each run.

from 1960 to 1990 and then subsequently interpolated across weather stations. We used bioclimatic variables at a resolution of 2.5 decimal degrees, as indicated in Table 1. Any other layers that were at a higher resolution were downsampled to 2.5 decimal degrees prior to analyses; this only applied to the dataset for the LIG scenario (from Otto-Bliesner *et al.* 2006).

We conducted ENM analyses in MaxEnt at the whole-species level, as well as the level of intraspecific genetic clusters within *Populus tremuloides* inferred herein and shown in Fig. 1. Species occurrence datasets often contain biases in geographic space reflective of unequal sampling effort, for example with more intense sampling near cities, research institutions, or more easily accessible areas and field sites (e.g. Reddy & Dávalos 2003). By thinning occurrences, we expected to decrease biases in geographical or environmental space reflecting unequal sampling effort, an issue that is known to adversely affect ENM analyses (see Reddy & Dávalos 2003; Boria *et al.* 2014). The final set of filtered occurrence records for *P. tremuloides* and each genetic cluster is provided in the Mendeley Data accession listed under Data Accessibility in the main text.

When conducting ENM analysis under a hierarchical design, as in this study (species, clusters as hierarchical levels), it is important to define separate and appropriate calibration areas for each level of the analysis. The area used for model calibration at the species level was determined using methods including a minimum convex polygon (MCP) around all of the points in the final set of filtered occurrence records, using raster (Hijmans 2017), and this calibration area is available in vector shapefile format (alongside that for all other calibration areas) in the 'Calibration_Areas' folder of our Mendeley Data accession (see Data Accessibility, main text). However, to generate

calibration areas for each cluster, a different approach was required, and it was important to remove two kinds of areas from each cluster's MCP. First, we removed areas included within the MCPs for the other clusters, because a given genetic cluster could be absent from the areas of closely related clusters/lineages due to competition or other unknown biotic interactions (see Anderson & Raza 2010). Second, we removed areas with *P. tremuloides* occurrence records that could pertain to unidentified lineages (due to lack of samples and genetic data from those areas). To expand, records outside the cluster MCPs cannot be reliably assigned to any of the clusters without introducing additional assumptions or biases. Thus, occurrence records within the cluster MCPs represent a geographically (and possibly environmentally) biased sample across the area potentially occupied by each lineage. This artificial bias is similar to having a barrier that prevents a species/lineage from reaching climatically suitable areas (Anderson & Raza 2010) and thus must be removed from the calibration areas (Radosavljevic & Anderson 2014; Peterson *et al.* 2011, pp. 161–162).

As a result, we delimited the whole species distribution by creating a minimum concave polygon (MCcP) around all occurrence records in the filtered set. Then, for each *P. tremuloides* genetic cluster, we excluded areas within the species MCcP but outside of the corresponding cluster MCP. The excluded areas are included as 'MCPea' calibration areas in our Mendeley Data accession. Under this procedure, the resulting calibration area (or 'occurrences polygon') for each lineage was the combination of its own MCP with areas of the species-level MCP but excluding the species-level MCcP. In appendix **Figure A3** *below*, we provide a map showing areas covered by the MCPs for the whole species ('species buffered MCP') and for clusters 1–3, as well as the MCcP for the species ('species MCcP').

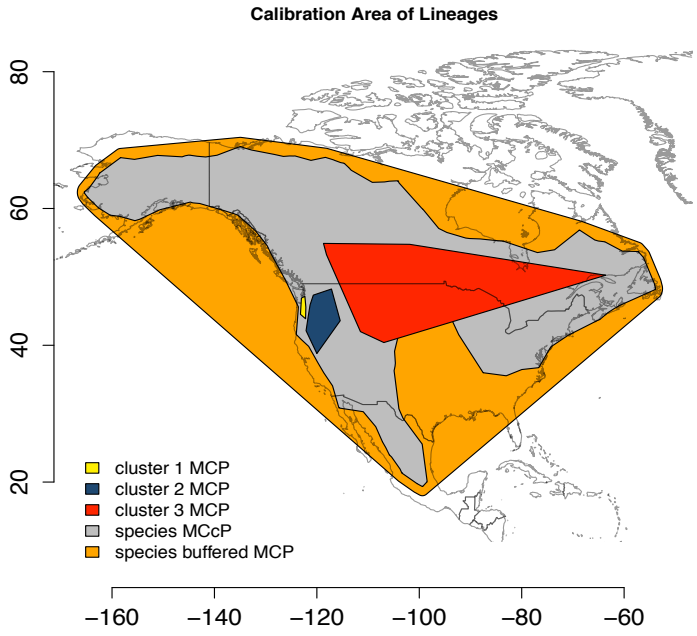

Figure A3. Map of calibration areas for lineages analyzed using ecological niche modeling in this study. Different polygons were created as described in the main text and this appendix for *P. tremuloides* and each of the ADMIXTURE-inferred intraspecific genetic clusters (clusters 1–3) shown in Fig. 1.

Crowdfunding.—As noted in the Acknowledgments section, this research was funded in part through a crowdfunding effort hosted by Experiment.com. The crowdfunding project was entitled, "The lost aspens of the Willamette Valley: Did catastrophic floods carry them from the Rockies?", and was led by Co-PIs Collin Peterson, Steve Strauss, Bill Ripple, and Rich Cronn. To acknowledge the contributions of the 68 crowdfunding donors in full, we here provide a full list of their names, as follows:

> Timothy S Leatherman, David B. Wagner, Logan Norris, David Altman, Bruce P Dancik And Brenda L Laishley, Haven Baker, Bruce Chassy, Denny Luan, Terri Lomax, Jenny Kao, Ellen Watrous, David Oates, Bob Latham, Jeff Clark, Stefan Rauschen, Tina Wasem, Nicholas Wheeler, James Rinehart, Gabriela Ritokova, Sophie Duckett, Chris Wozniak, Dave Moses, Tom Adams, John Vendeland, Malory K. Peterson, Drew L. Kershen, Haiwei Lu, Kendrick Moholt, David Dalton, Al Goldner, Jim Kuhlman, Cathy L. Peterson, Linda M. Hardie, Tina Loop-Duckett, John E. Carlson, Deian Moore, Lecia Schall, Jason Holliday, Jeff Peterson, Norman Ellstrand, Jonathan Gressel, Joel Corcoran, Naomi Weidner, Laurie Simmons, Gleb Bazilevsky, Valerie Boggs, Chuck Cannon, Nick Houtman, Susan Bexton, J. Keith Gilless, Andrew Groover, Jim Border, Aaron R Leichty, David Showalter, Nancy Allen, Peggy Lemaux, Peg Silloway, Austin Strauss, Liz Swan, Oscar Jasklowski, John DeFrancisco, Mary Garrard, Tanying, Jennifer Preston Brennan, Taylor Helfand, Christina Tran, Cindy Wu, and Ryan Lower.

We gratefully acknowledge each of these individuals for their support of our work, without which this paper would not have been possible.

References

Anderson, R.P. & Raza, A. (2010) The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography*, *37*, 1378–1393.

Bardou, P., Mariette, J., Escudié, F., Djemiel, C., & Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, *15*(1), 293.

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One*, *9*(2), e90346

Hijmans, R. J. (2017) *Package 'raster': Geographic data analysis and modeling.* R package version 2.6-7. Available at: <https://CRAN.R-project.org/package=raster>

Hijmans, R. J., Cameron, S.E., Parra, J.L., Jones, P.G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*, 1965–1978.

Little, E. L. Jr. (Ed.) (1971). Vols. Miscellaneous Publication 1146. Digitized 1999 by US Geological Survey. US Department of Agriculture.

Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014). ENMeval: an R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models. *Methods in Ecology and Evolution*, 5, 1198–1205.

Peterson, A., Soberón, J., Pearson, R., Anderson, R., Martínez-Meyer, E., Nakamura, M., & Araújo, M. (2011) *Ecological niches and geographic distributions*. Princeton University Press, Princeton, N.J.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259.

Phillips, S. J., & Dudík, M. (2008). Modelling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, *31*, 161–175.

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: an open-source release of Maxent. *Ecography*, *40*(7), 887–893.

R Core Team (2018) *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna.

Radosavljevic, A. & Anderson, R.P. (2014) Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, *41*, 629–643.

Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, *30*(11), 1719–1727.

Schilling, M. P., Wolf, P. G., Duffy, A. M., Rai, H. S., Rowe, C. A., Richardson, B. A., & Mock, K. E. (2014). Genotyping-by-sequencing for *Populus* population genomics: an assessment of genome sampling patterns and filtering approaches. *PLoS One*, *9*(4), e95292.