

Ancient Bottleneck events, natural selection, and increased risk of maladaptation shape the past, present and future population structure of a widespread conifer

Sean Collins

Northern Arizona University

Anna Lincoln

Bureau of Land Management

Amanda R De La Torre

Amanda.de-la-torre@nau.edu

Northern Arizona University

Research Article

Keywords: exome capture, *Pinus ponderosa*, demographic history, genome x environment association, genomic offset

Posted Date: February 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8614253/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Demographic events and natural selection can have independent and confounding effects on population structure and adaptation capability. This means different species could have had very different paths to their current population structure. The influences of demography and selection along a species' evolutionary timeline can influence responses to current and future stressors. We used two genome-wide and range-wide datasets in Ponderosa pine (*Pinus ponderosa*) to better understand the influences of demography and selection on current population structure and future maladaptation risk.

Results

Samples were sequenced with 55,000 capture-seq probes, producing 1.2-1.5 million SNPs (400Mbp per sample) in each dataset. We examined current population structure, changes in effective population size, evidence for selection, genome-environment associations and future maladaptation risk using genomic offset. We identified four genetic clusters within the species; numerous bottleneck events that mirror glacial advances and past droughts; signatures of purifying and balancing selection; important environmental variables; and many geographic areas with high amounts of genomic offset.

Conclusions

Differentiation across varieties has mostly been driven by two processes (1) bottleneck events causing random losses of alleles and decreased connectivity, (2) spatially heterogeneous selection and purifying selection in response to varied moisture levels. Many areas are predicted to be at risk of future maladaptation, which coincide with areas showing increased mortality events and decreased regeneration in previous studies. Previous evolutionary events and rapidly changing climate have shaped future maladaptation risk. Drastic changes from current genomic structure are needed to meet a rapidly changing climate.

Background

Demographic events can have drastic impacts on population structure whether these events occurred recently or hundreds of thousands of years ago. Effective population size heavily influences standing genomic diversity (Wright 1931; Nei et al. 1975) and probability of fixation in new mutants (Wright 1931; Ohta 1992). Following large bottlenecks, there is a drastic decrease in diversity (Nei et al. 1975) and a greater chance of a new polymorphism to become abundant through drift, regardless of whether the polymorphism is advantageous, deleterious or neutral (Ohta 1992; Charlesworth 2002). In widely distributed species, unique bottleneck events can cause asymmetric losses of heterozygosity across locations within a species (Wakeley and Alickar 2001). Depending on gene flow corridors, bottlenecks can fragment continuous populations which can further lead to asymmetric heterozygosity levels within a species (Hewitt 1996). Differing levels of heterozygosity from bottleneck events can lead to varied allele

frequency across a species range creating unique phenotypes and adaptation capabilities, in part due to the random collection of alleles from surviving individuals (Nei et al. 1975; Hewitt 1996). Understanding a species' demographic history can aid in understanding whether a species is structured due to heterogeneous history that has led to isolated islands with unique assortments of alleles (Wright 1931; Wakeley 1998, 2000; Wakeley and Alicar 2001; Lynch 2007), or due to selection where the assortment of alleles across populations represents the optimum fitness for specific environments (Darwin 1859), or more likely a combination of the two. Furthermore, that understanding can provide inferences on a species adaptation capability, which is important in the face of a rapidly changing climate.

Natural selection and gene flow also play important roles in shaping species' population structure and adaptation capability. Often demographic events, type of selection, and levels of gene flow have confounding effects on differentiation and adaptation capability (Lenormand 2002; Charlesworth 2003). Mutations can be advantageous, deleterious or neutral and based on the infinite site model individual mutations can occur in any area of the genome and specifically in new areas (Kimura 1969; Hudson 1983). In fragmented populations, mutations can then occur in completely different areas of the genome, and these populations can become differentiated over time simply by random chance (Kimura 1969; Hudson 1983; Kimura 1983; Ohta 1992). Further complicating matters, purifying selection shrinks genetic diversity, but its efficacy is dependent upon population size and gene flow (Lanfear et al. 2014; Chen et al. 2017). Therefore, if fragmented populations have varied sizes and connectivity, purifying selection may purge certain polymorphisms in some populations, but not others, further differentiating populations. Environmental heterogeneity can lead to varied directions of selection for specific polymorphisms, leading populations to become differentiated due to adaptation to their local environment (Wright 1931; Levene 1953; Deakin 1966; Christiansen 1974). Maintenance of polymorphisms can be influenced by demographic history and life history strategies (Karlin 1982). High gene flow can limit this differentiation across heterogeneous environments due to gene swamping, where even advantageous alleles are much less likely to become fixed (Slatkin 1987; Yeaman and Whitlock 2011). However, there are some cases where advantageous alleles are maintained across heterogeneous environments under high gene flow (Comeault et al. 2014; Laurent et al. 2016). The decades of evolutionary theory and empirical work suggest a species can arrive at a set of advantageous alleles through a variety of pathways and maintain these alleles in many ways. Therefore, untangling the interplay between selection and demographic history can help better understand a species' development and maintenance pathways of polymorphisms, which can allow inferences for what may happen to a species under new threats.

To better understand how theory relates to species-specific evolutionary histories, quality study systems need to be chosen, and modern tools need to be leveraged. Advancing in sequencing and computation capabilities has allowed us to test evolutionary theory more robustly than ever before, due to the increase in the number of polymorphisms investigators can generate. Ponderosa pine (*Pinus ponderosa*) offers a great opportunity to further our understanding of the influences of demographic events, selection and gene flow on current population structure and adaptation capability. Ponderosa pine is a long-generation, slow-growing, outcrossing, wind-dispersed conifer that is distributed from West

Texas to British Columbia and exists across extremely wide ecological amplitudes, e.g. populations can exist at 400 to 1100mm mean annual precipitation (Kerr et al. 2015; Dixit et al. 2022). Due to its wide distribution, Ponderosa pine provides biodiversity and ecosystem services, such as supporting rare or endangered species (Keith 1965; Beier et al. 2008), and maintaining a sustainable water yield (Folliott et al. 1989; Allen et al. 2002). While readily abundant in many areas of North America, Ponderosa pine has a very unclear population structure, with previous studies suggesting anywhere between 2-9 intra-specific varieties using SSRs (Conkle and Critchfield 1988; Callaham 2013; Potter et al. 2013, 2015; Willyard et al. 2017). More recently, a smaller scale study identified SNPs associated with environmental gradients in Sierra Nevada seed sources (Shu and Moran 2023). Species' range genome-wide assessments of population structure have never been conducted within Ponderosa, probably due to the sequencing and computational challenges of working with very large datasets required to understand huge and complex conifer genomes. The reference genome assembly is currently absent for the species, but Ponderosa is a diploid species with an estimated genome size of 20-22 Gbp. Limited work has examined evolutionary forces shaping population structure of Ponderosa pine, mainly due to an unclear fossil record (Cole 1983; Roberts and Hamann 2015; Shinneman et al. 2016, Willyard et al 2021), therefore the specific evolutionary factors that have led to such a complex population structure in Ponderosa pine have yet to be examined. Furthermore, negative physiological responses across the species range have been reported in the past decades (*e.g.* increased mortality, decreased regeneration), most likely due to changing climate regimes (Kolb et al. 2007, 2019; Petrie et al. 2023). Ponderosa pine's wide ecological amplitude, unclear evolutionary history and complex population structure suggests the high potential for unique selection strategies, demographic history and patterns of gene flow. Furthermore, the negative effects due to climate change seen in Ponderosa pine offer an opportunity to understand how past evolutionary factors influence future maladaptation risk.

To address these knowledge gaps, the driving questions of this study are (1) What population structure is best supported using genome-wide data? (2) Is there evidence that demographic history and/or selection has shaped current population structure? If so, what type of selection is occurring and what environmental factors act as selection pressures? (3) What populations are at higher risk of maladaptation in the near future due to past evolutionary forces and current genome x environment relationships?

Methods

Our study uses two independent genomic datasets to assess the project goals. Two datasets allow for better inferences of patterns as more samples are used in making final conclusions. Final conclusions were based on the synthesis of results from both datasets.

Dataset one (poolseq) sample collection

Needle tissue was collected from 12-20 mature individuals in 26 unique locations respectively across the species' natural range (Figure 1, Table 1, Supporting Information Table S1). Sampled individuals were

separated at least 30m from each other, to reduce chances of relatedness among samples. The terminal end of a branch was cut off and 5-10 fascicles of needles were collected for each individual tree. Latitude, longitude, and elevation were recorded for all individuals. Following collection, all needles were placed in a -20°C freezer. All 26 locations were on public lands managed by the United States Forest Service (USFS) or the United States Bureau of Land Management (BLM). Most collections were conducted by respective USFS, BLM or Cal Fire members; any collections not done by USFS or BLM members were done with the approval of each respective organization. All 26 locations represented novel ponderosa pine sampling locations with no known official seed source name and as such short-hand names were arbitrary.

Dataset two (common garden) sample collection

Seeds were gathered from 24 seed sources across the geographic range of Ponderosa and split into four regions: 1) Eastern Washington State and Idaho. 2) Arizona, Southwestern Utah and New Mexico. 3) Montana and Wyoming, and 4) California, with 6 seed sources within each region (Figure 1, Table 1, Supporting Information Table S2). Seed sources were chosen based on availability from previous collections maintained by Northern Arizona University (NAU), The Forest Service's Experimental Forest at Priest River, Idaho and the Forest Service's National Seed Laboratory. All seed source names used in our study came from previously established seed and such the names given by their original collectors were kept for consistency. A total of 50-60 seeds per seed source were planted in November 2022 in the Northern Arizona University Greenhouse. Seed was taken out of the freezer, planted and watered in the same day. Germinated seedlings were watered to saturation 3 times a week in 20cm pots for the duration of this experiment. Seedlings were also fertilized to maintain health and growth throughout the first year of their lives. Germination rates ranged from 10%- 90%. Tissue for DNA extraction was collected from 404 seedlings (101 individuals per region, 5-32 individuals per seed source).

DNA extraction and Sequencing

For both datasets, DNA was extracted using 60-100mg of needle tissue homogenized using a SPEX Tissue MiniG 1600 Prep Machine. DNA was extracted using either a MP Bio EZ Plant DNA extraction kit (MP Biomedicals, LLC, Irvine, CA), Nucleospin Plant DNA extraction kit (MACHEREY-NAGEL, Dueren Germany) or a CTAB DNA extraction protocol. Following extraction, each extract was quantified using a Qubit 4.0 Fluorometer (Thermo Fisher, MA, USA) and placed in a -20°C freezer. For dataset one DNA samples were pooled in sets of 4 with respect to their sampling location, creating 96 composite samples. The same amount of input DNA for each sample was added in the composite sample to make sure each individual sample had an accurate representation. Two samples failed to produce viable libraries leaving 94 composite samples to sequence. No pooling was conducted for dataset two; all 404 individuals were sequenced individually. All samples across the two datasets were then sequenced using the same 55,000 capture seq probes (40,000 probes for coding regions, 15,000 for intergenic/non-coding regions). Probes were designed using a draft Ponderosa pine transcriptome (Gilewski and Mattsson, unpublished data) and the loblolly pine reference genome v 2.0

(<https://treegenesdb.org/FTP/Genomes/Pita/v2.0/genome/>). Libraries were sequenced using 150x150bp on an Illumina platform conducted at Rapid Genomics (LGC Biosearch Technologies, Gainesville, FL). Probes generated 300-400 Mega-bp worth of data per sample.

Sequence analysis

The same software for sequencing analysis was used for both datasets, however each dataset was analyzed independently. Raw reads were quality controlled (QC'd) and trimmed of adapters using fastp (v0.23.2; Chen et al. 2018), FASTQC (Andrews 2010) and MultiQC (Ewels et al. 2016). Initial filtering criterion for raw reads was a phred score higher than 30. Following initial QC, reads were mapped to the loblolly pine reference genome (version 2.0) using bwa-mem (v0.7.10; Li and Durbin 2009) using default settings. Mapped reads were then sorted by coordinates and any duplicates were removed using samtools (v1.5; Li et al. 2009). Following mapping, single nucleotide polymorphisms (SNPs) were called using bcftools (Danecek et al. 2021). For dataset one, 111,576,152 SNPs were initially called. For dataset two, 284,913,192 SNPs were initially called. Initial filtering criteria for SNPs were quality score greater than 20, call rate greater than 0.75, minor allele frequency greater than 0.01. SNP filtering was conducted using vcftools (Danecek et al. 2011) and bcftools (Danecek et al. 2021). After initial filtering, 1,282,955 SNPs and 1,501,748 SNPs (datasets one and two, respectively) were retained for further analyses.

Current population structure

For each filtered SNP dataset, population structure was assessed using a principal component analysis (PCA) and a discriminant analysis of principal components (DAPC), with functions within plink (Purcell et al. 2007) and adegenet R package (Jombart 2008). Optimum number of PCs included in the DAPC were decided based on the optim.a.score function within adegenet. Twenty discriminant axes were included in Dataset one; the optimum number of PCs to include for this dataset was 8 with an average classification accuracy of 84%. Twenty discriminant axes were also included in Dataset two; the optimum number of PCs for dataset two in the DAPC was 12 with a classification accuracy of 76%. The effect of markers potentially under selection on population structure was assessed by re-running all analyses using putative neutral markers, as determined by Hardy-Weinberg equilibrium (HWE) using vcftools (Danecek et al. 2011). Due to the pooled nature of dataset one, a higher HWE alpha value was used (0.10) to account for any effect the pooled samples had on determining HWE. The markers potentially under selection did not have any effect on population structure and so the results of the full datasets are presented in the main text (Supporting Information Figure S1)

Current population genetic metrics

Wright's pairwise fixation index (F_{ST}) (Wright 1949), Θ_{π} (Tajima 1983), inbreeding coefficient (F_{IS}) and private alleles were calculated on putative-neutral markers. A total of 979,250 SNPs were retained after filtering. For dataset two, a HWE alpha value of 0.05 was used, leaving 1,018,650 SNPs. Pairwise F_{ST} was determined using StAMPP (Pembleton et al. 2013), an R package designed to accommodate more

than just diploid samples and visualized using a neighbor joining tree from ape (Paradis et al. 2004). Θ_{π} was calculated on 10kb windows of each scaffold using vcftools (Danecek et al. 2011). Θ_{π} and private alleles for each collection location were calculated using functions found within snpR (Hemstrom and Jones 2023). FIS was calculated using vcftools (Danecek et al. 2011). Vcftools has been used in actual polyploids to estimate FIS previously (Lee et al. 2020) because it takes a method of moments approach, using each SNP to calculate FIS, with $F_{\text{SNP}} = (O - E) / (N - E)$, where O is the observed number of homozygotes, E is the expected number of homozygotes (given population allele frequency), and N is the total number of genotyped loci. Because of the method moments basis using vcftools to calculate FIS is therefore robust even on pool seq data.

Isolation by distance (IBD), isolation by environment (IBE), Tajima's D (Tajima 1989) and site frequency spectra (SFS) were estimated on datasets using both neutral and non-neutral markers. IBD and IBE were conducted using Mantel tests with 9999 permutations with functions within the ade4 R package (Dray et al. 2007). Latitude and Longitude were used for IBD analysis, and 33 climate variables were used for IBE (Table 1, Supporting Information Table S1, S2). SFS was calculated using R scripts from Liu et al 2018. 1D SFS were generated for each genetic cluster within each dataset. 2D SFS were then generated for pairwise combinations of genetic clusters within each dataset.

Future maladaptation risk using genomic offset

Both full SNP datasets (neutral and non-neutral markers) were filtered using a higher MAF filter (0.05) for genomic offset. This higher MAF left 677,119 SNPs for dataset one and 649,507 SNPs for dataset two. Missing SNPs were imputed by filling in missing values for each site with the most common value for each site. Climate data was downloaded from Climate NA (Wang et al. 2016; Mahony et al. 2022). Genomic offset was estimated using gradient forest (Ellis et al. 2012). Model quality was assessed by seeing if the model could accurately project range-wide population structure after training. The final model for both datasets consisted of 33 climate variables (Supplementary Figure 1, Supporting Information Table S1, S2), 1000 trees and 101 bins. Genomic Offset was predicted for the SSP 245 Climate Scenario for 2041-2070 using the 8 GCM Ensemble Model projections from Climate NA (Wang et al. 2016; Mahony et al. 2022).

Genome x environment relationships using genome x environment association (GEA)

The same imputed SNP datasets from the genomic offset analysis were also used for GEA analysis. Selection of environmental variables included in the GEA was based on gradient forest importance measure. Being that Redundancy Analysis (RDA) is sensitive to collinearity among variables, a sub-selection of variables from the genomic offset were used with the RDA, based on variable importance metrics found within gradient forest. Gradient forest identified Summer Precipitation, Winter Precipitation and Yearly Continentality as three climate variables with highest importance in dataset one, and Summer precipitation, Winter Precipitation and Spring Relative Humidity as the three variables with the highest importance in dataset two (Supplementary Figure 1). To account for all potential

combinations of variables, dataset one and two were run using both sets of climate variables in separate tests. The RDAs were conducted, and overall model fit was assessed using functions within vegan R package (Dixon 2003).

Sparse partial Least Squares Regression (sPLS) was also used for GEA, using functions within the RGCCA R package (Girka et al. 2023). For both datasets, initially all SNPs and all environmental variables from the genomic offset analyses were tested. All environmental variables were used for the sPLS, because sPLS relies on an eigen decomposition of matrices to then create functions that maximize variance among matrices; it is therefore less sensitive to unstable parameter estimation among variables due to the effect of collinearity relative to the multiple linear regression used in the RDA. Following these full model tests, the loadings for the respective sPLS functions for the SNP matrix and the environmental matrix were assessed. The variables that were in the 1% quantile of loadings were chosen to be re-run to better understand which variables were contributing most. A final model of SNPs and environmental variables using that filtering criteria was then re-run a final time. Following the identification of candidate loci, both RDA and sPLS were then compared using variance explained (R^2), to see which model best explained the SNP data and thus gave the most robust representation of environmental candidate loci. A PCA was then rerun, using FactoMineR (Lê et al. 2008), on candidate loci and the most important environmental variables to illustrate how alleles changed with respect to changing climate.

Functional annotation of candidate loci and enrichment analysis

A custom database for the *Pinus taeda* reference genome was created with snpEff (Cingolani et al. 2012). The genes for SNPs identified from the GEA were identified using the Pita 2.01 annotation. Sequences for each candidate loci were then extracted from the reference genome using samtools (v1.5; Li et al. 2009). Functional annotation was first conducted on candidate loci using egglog (Cantalapiedra et al. 2021). To gain more information on functions of candidate loci, *A.thaliana* orthologs were identified using the blastx function on NCBI with search restricted to *A.thaliana*. Over-representation analysis was run on gene ontology (GO) and KEGG ortholog (KO) terms respectively, which were identified from the egglog functional annotation. Unique GO and KO terms were identified for candidate loci for both datasets. The counts of each GO and KO term were used to create a distribution of observed occurrences. Any term that was in the 95 percentile or greater was deemed to be over-represented. Networks for over-represented GO terms were then created using functions found at quickGO (<https://www.ebi.ac.uk/QuickGO/annotations>)

Demographic Modeling and Patterns of Selection

Demographic history was inferred using changes in effective population size (N_e) using Stairway plot 2 (Liu and Fu 2020). One-dimensional SFS with singletons excluded for each unique genetic cluster in dataset two were used. Stairway plot 2 is very sensitive to singletons. Only dataset two was used in this analysis as there was a concern that using pool-seq data may provide false inferences due to the very different distributions of 2D SFS in dataset one compared to dataset two. Stairway plot 2 has only been

used for diploid samples in previous studies. Twenty years was used as the estimated generation time and 1.0×10^{-9} was used as the estimate mutation rate based on a midpoint value of nuclear mutation rates of pines $0.70\text{--}1.31 \times 10^{-9}$ found in Willyard et al. 2007, which also falls within the evolutionary rate estimated for gymnosperms in De La Torre et al. 2017.

Signatures of selection were inferred using multiple methods; (1) $\omega = \text{dn/ds}$, (2) Transitions to transversion ratios (Ts/Tv), and (3) Tajima's D that was corrected for demographic history using simulations. All three methods were used on each dataset. Because identifying selection is so nuanced, patterns of all three methods were used to make final inferences on signatures of selection. $\omega = \text{dn/ds}$ has been shown to differentiate sequences maintained under purifying, balancing selection and neutral evolution, but has its own limitations (Nielsen and Yang 1998; Yang et al. 2000; Kryazhimskiy and Plotkin 2008; Fijarczyk and Babik 2015; Spielman et al. 2015). $\omega = \text{Dn/ds}$ was estimated by dividing non-synonymous mutations by synonymous mutations for each coding sequence identified using snpeff (Cingolani et al. 2012). This method may be coarser than many maximum likelihood methods but being that our reference genome is a different species, and thus we are comparing interspecific sequences to calculate dn/ds and aren't violating suggestions seen in previous literature (Kryazhimskiy and Plotkin 2008; Fijarczyk and Babik 2015). Ts/Tv was calculated on 10kb windows using vcftools (Danecek et al. 2011). High Ts/Tv bias has been related to genes under adaptive evolution in many species (Guo et al. 2017; Sane et al. 2023, 2025) but is by no means a perfect relationship to selection. Tajima's D is a classic test of neutrality that is extremely sensitive to demographic history (Tajima 1989). To correct Tajima's D for demographic history, first median estimates of Ne across generations for each genetic cluster were used in SLiM 3 (Haller et al. 2019) to create neutral simulations that took into account each genetic cluster's specific demographic history. Simulations were conducted on 100MB fragments using only fixed effect neutral mutations (selection coefficient = 0.0) and the same mutation rate used in stairway plot 2 models. Recombination rate of 10×10^{-4} was used for all simulations, which is based on a rounded lower bound of conifers' genome-wide recombination rates from Jaramillo-Correa et al. 2010. Ten independent simulations were run for each genetic cluster. Tajima's D was then calculated on simulated VCFs in 10kb windows using vcftools (Danecek et al. 2011) creating 100 Tajima's D windows that represented possible Tajima D values for neutral mutation; the minimum and maximum simulated values were used as upper and lower critical values for rejecting the null hypothesis of Tajima's D. Similar genetic clusters were seen across both datasets and as such the same critical values were used for both datasets. The critical Tajima's D values for the interior cluster were $-2.1451 < \text{neutral} < -0.951064$. The critical Tajima's D values for the northern interior cluster were $-2.27033 < \text{neutral} < -0.892826$. The critical Tajima's D values for the northern coastal cluster were $-1.52952 < \text{neutral} < 0.813934$. The critical Tajima's D values for the California cluster were $-2.29491 < \text{neutral} < -0.951064$. Following the simulations, observed Tajima's D values were then calculated on 10kb windows for each individual genetic cluster using vcftools (Danecek et al. 2011) for both datasets. Vcftools has been used to calculate Tajima's D using 10kb windows previously in polyploid organisms (Liu et al. 2024) and thus can perform robustly for our pool-seq data. Any observed Tajima's D value below the lower critical value obtained from simulations was considered to be under purifying selection, similarly any observed

Tajima's D value above the higher critical value for each genetic cluster was considered to be under balancing selection.

Results

Population structure

Our results across two datasets show strong support for 3 distinct genetic clusters within Ponderosa pine and moderate support for 4 distinct genetic clusters (Figure 1, Supporting Information Figure S2). The DAPC for both datasets show a strong split within the coastal portion of the US, with seed sources and collections sites in the Northwest being distinct from most of California (Figure 1, Supporting Information Figure S2). Our datasets suggest that the two clusters in the coastal portion differentiate in northern California or Southern Oregon (Figure 1); thereby suggesting a California cluster and northern coastal cluster. Our results also show differentiation in the interior west to a degree (Figure 1, Supporting Information Figure S2). The DAPC shows the southwest being distinct from Montana and Wyoming, but locations in the great plains show strong similarity to the southwest (Figure 1, Supporting Information Figure S2). This suggests that individuals in Montana and Wyoming are unique relative to the rest of the interior portion of the range; suggesting an interior cluster that exists from the southwest to the great plains and a northern interior cluster that exists in Montana and Wyoming.

Population genomic metrics

Population genomic metrics support broad scale patterns seen in population structure and suggest unique nucleotide diversity and private alleles occurring within genetic clusters (Figure 2), a potential effect of isolation by distance and environment. Nucleotide diversity and private alleles varied widely across sampled locations (Figure 2). Californian populations had consistently high nucleotide diversity and private alleles (Figure 2). Similarly, most populations in the northern coastal cluster areas had high nucleotide diversity and private alleles, although there were a few with lower levels of nucleotide diversity and private alleles. Populations in the interior cluster and northern interior cluster showed the strongest amount of within cluster variation in both nucleotide diversity and private alleles, with some populations in those clusters having some of the highest nucleotide diversity and private alleles and others the lowest magnitude in each metric. Both the California and northern coastal clusters show similar levels of inbreeding, with a large amount of excess heterozygosity (Figure 2). The interior and northern interior clusters also show generally similar levels of inbreeding with most sites showing an excess or an expected amount of heterozygosity (Figure 2). However, some areas in the four corner states and the plains (Southwestern Colorado, Central Colorado, South Dakota and Nebraska) show lower than expected levels of heterozygosity, suggesting higher levels of inbreeding (Figure 2).

Pairwise F_{ST} supports observed patterns in population structure found in the DAPCs for both datasets (Supporting Information Figures S3-S4, Supporting Information Table S3, S4). Pairwise F_{ST} also shows potential substructuring within genetic clusters, suggesting some collection locations or seed sources may exchange more alleles with certain areas within their cluster than others (Supporting Information

Figures S3-S4, Tables S3-S4), which supports the small geographic scale differences seen in nucleotide diversity, private alleles and FIS. These substructure clusters tend to mirror geographic distance, that is areas closer together tend to aggregate on the same split of the tree (Supporting Information Figures S3-S4, Tables S3-S4). Geographic clustering patterns seen in pairwise F_{ST} are further supported by significant isolation by distance correlation identified for each dataset (dataset one: $r = 0.7554588$, permuted p-value = $1e^{-04}$, dataset two: $r = 0.6606423$, permuted p-value = $1e^{-04}$) (Supporting Information Figure S5). Environment also plays a role in genome differentiation as seen by significant correlations found in isolation by environment tests for both datasets (dataset one: $r = 0.6465629$, permuted p-value = $1e^{-04}$, dataset two: $r = 0.234628$, simulated p-value = $1e^{-04}$) (Supporting Information Figure S5).

Two-dimensional site frequency spectra (2D SFS) support genome-wide differentiation seen in the DAPC and pairwise F_{ST} , suggesting unique allele makeups across genetic clusters. More specifically, our results suggest that many high frequency alleles in interior genetic clusters occur at much lower frequency or are absent in coastal clusters and vice versa (Supporting Information Figures S6-S7). This is seen via the many sites that occur at very different allele count classes between interior clusters and coastal clusters (Figures S6-S7). Less pronounced differences are seen when the northern coastal cluster is compared with the California cluster, and when the Interior cluster is compared to the Northern interior cluster (Supporting Information Figures S6 -S7). These 2D SFS (Supporting Information Figures S6-S7) still show differences in allele counts classes, supporting the pairwise F_{ST} and DAPC clustering (Figure 1, Supporting Information Figures S3-S4).

Demographic history

One-dimensional site frequency spectra (1D SFS) show an abundance of rare alleles within each genetic cluster. Stairway Plot 2 suggests numerous bottlenecks for all four genetic clusters (Figure 3). Stairway plot 2 was able to estimate N_e for all clusters dating back to 400-500 kya. Starting N_e was similar for all clusters (slightly greater than 40,000,000), however final N_e estimates and the demographic history varied for each cluster (Figure 3, Supporting Figures S8-S11). Even though the Interior cluster is similar to the Northern Interior cluster (based on the DAPC, FIS, and pairwise F_{ST}), the Interior genetic cluster shows a vastly different demographic history with multiple population bottleneck events occurring 4,000 to 100,000 years ago (Figure 3, Supporting Information Figures S8, S9). The Northern Interior genetic cluster shows the oldest population bottleneck with a long consistent decline in effective population size occurring 50,000 to 300,000 years ago (Figure 3, Supporting Information Figure S9). Even though the northern coastal cluster and the California cluster are genetically distinct, they show similar demographic histories, with population bottleneck events occurring from 1000 to 100,000 years ago (Figure 3, Supporting Information Figure S10, S11). The main difference between demographic histories between coastal clusters is that the northern coastal cluster shows one long consistent decline in effective population size whereas the California cluster shows multiple bottlenecks and plateaus over that time period (Figure 3, Supporting Information Figure S10, S11).

Patterns of Selection

All three methods used to make inferences on selection showed similar results across both datasets. General patterns for each metric were similar across genetic clusters, with the exception of Tajima's D for the northern coastal cluster. $\omega = dn/ds$ and Ts/Tv were the same across genetic clusters within each dataset. While Tajima's D values varied across genetic clusters within each dataset. For dataset one, 8542 unique coding sequences contained polymorphisms; 74.72 % of those had a $dn/ds > 1$, 18.06% had a $dn/ds < 1$ and 7.21% had a $dn/ds = 1$. For dataset two, 9171 unique coding sequences contained polymorphisms; 73.94% had a $dn/ds > 1$, 18.65 % had a $dn/ds < 1$ and 7.41% had a dn/ds equal to one. For Ts/Tv ratios, in dataset one 97.22% percent of SNPs existed in windows greater than the 0.5 equilibrium value, and in dataset two 97.54% of SNPs existed in windows greater than the equilibrium value.

In both datasets, the northern coastal cluster showed the highest number of SNPs that occur in windows with Tajima's D within the simulated critical bounds and below the lower critical value, followed by a smaller number of SNPs that occurred in windows above the higher critical value (Table S5). While the other three genetic clusters showed similar mixtures of SNPs occurring in windows above, within and below their respective critical values (Table S5).

Genomic offset

Our genomic offset analysis identified many climate variables that are related to population structure (Supporting Information Figures S1, S12, Tables S1, S2) and showed that even in an optimal climate scenario some areas are at high risk of maladaptation, whereas other areas are at very little risk. Current observed population structure was modeled using current climate data (Supporting Information Figure S12, Supporting Information Tables S1-S2), suggesting the climate variables used in the analyses (Supporting Information Figure S1, Tables S1-S2) are important for population structure. Modeling the current observed population structure also provided a robust initial G x E coordinate for calculating genomic offset, which helped support the robustness of the projections. Areas with higher risk of maladaptation are the inland northwest, British Columbia, portions of the southwest and patchy areas in the great plains and Montana/Wyoming based on the SSP 270 climate scenario in 2041-2070 (Figure 4).

Genome x environment association (GEA)

Our analysis identified many candidate loci that are strongly correlated to spring, summer and winter moisture. Genome and environment relationships were best shown by the sparse partial least squares regression (sPLS) compared to the redundancy analysis (RDA), suggesting stronger evidence of association of the loci identified using the sPLS. The candidate loci identified from the sPLS have biological functions related to water stress and developmental pathways, further supporting the role of these genes in environmental responses. For dataset one, 4711 SNPs were associated with Summer precipitation, Winter Precipitation, and Continentality (TD) and produced an adjusted R^2 of 0.102. From those SNPs, 1426 were associated with Summer Precipitation, 1159 with Winter Precipitation and 2126

with TD. When the RDA was rerun using only these 4711 SNPs and the three climate variables, the adjusted R^2 increased to 0.20. When using Summer precipitation, Winter Precipitation and Spring Relative Humidity, a similar adjusted R^2 of 0.102 was observed and significant 3288 SNPs were identified. From those SNPs, 1249 were associated with Summer Precipitation, 922 with Winter Precipitation and 1117 with Spring Relative Humidity. When the RDA was rerun using these 3288 SNPs and the second set of climate variables the adjusted R^2 increased to 0.17.

For dataset two, 3147 SNPs were associated with Summer precipitation, Winter Precipitation, and TD, produced an adjusted R^2 of 0.08 and identified 9762 candidate loci. A total of 2787 SNPs were significantly associated with summer precipitation, 3347 were significantly correlated with winter precipitation and 3628 SNPs were correlated with yearly TD. When Spring relative humidity, summer precipitation and winter precipitation were used in the RDA, an adjusted R^2 of 0.09 was observed and 8854 candidate SNPs were identified. A total of 2153 SNPs were correlated with summer precipitation, 2835 SNPs were correlated with winter precipitation, and 3866 SNPs were correlated with spring relative humidity.

For dataset one, when all SNPs and all environmental variables were used the sparse partial least squares regression derived functions that were to explain an average of 15.32% of variance in the SNP matrix and an average of 29.5% of variation in the environmental data. The top one percent contributions of loadings for the original sPLS components consisted of 6672 SNPs and thus were chosen as candidate loci. The sPLS was then rerun with these 6672 SNPs and all environmental variables. The new sPLS function was able to explain an average of 83.08% of variation in the SNP matrix and an average of 28.47% of variation in the environmental data. Elevation and Winter precipitation were the two variables in the 1% quantile of loadings for environmental data and thus chosen as the two variables with the most explanatory power for dataset one. When the sPLS was rerun with the 6672 SNPs and just elevation and winter precipitation, the sPLS functions were able to explain an average of 83.08% of variation in the SNP matrix and 83.46% of variation in the environmental matrix.

For dataset two, when the sPLS was run with all SNPs and all environmental variables the sPLS functions were able to explain 11.65% of variation in the SNP matrix and 19.17% of variation in the environmental data. 6246 SNPs in the second dataset were in the 1% quantile of loadings for this original sPLS function and thus were chosen as candidate loci. The sPLS was rerun with these 6246 SNPs and all environmental variables, resulting in an average variation explained of 73.78% in the SNP data and 19.00% in the environmental data. Spring relative humidity and summer precipitation were the two environmental variables in the 1% quantile for the environment data and therefore chosen as the two most explanatory environmental variables for dataset two. When the sPLS was rerun with the 6462 candidate loci and spring relative humidity and summer precipitation an average of 73.78% of variation was explained in the SNP matrix and 88.53% of variation in the environmental data. The much higher variance explained using the sPLS for both datasets indicates that the sPLS was a much better fit for our dataset leading to more robust associations between SNPs and environment variables. The candidate

loci identified from both datasets using the sPLS were used for further analysis via functional annotation.

PCA of candidate loci for each dataset along with their respective climate variables are shown in Figure 5. Each PCA shows stark differentiation among genetic clusters and some within genetic cluster variation. These show that the alleles that occur in the California cluster are generally associated with higher winter precipitation and spring relative humidity, but lower summer precipitation (Figure 5). The northern coastal cluster has alleles associated with increasing winter precipitation, summer precipitation and elevation but relatively similar spring relative humidity (Figure 5). The southwest and great plains cluster also has alleles that are associated with a gradient of winter precipitation, summer precipitation and elevation and an overall lower spring relative humidity compared to coastal clusters (Figure 5). Interestingly, CA 993 clusters closer to the Northern Coastal Cluster sources in these candidate loci, but when all SNPs are considered the seed source clusters with the California coastal cluster. This is a high elevation seed source and may have a climate more similar to the Northern Coastal Cluster rather than coastal California.

Functional annotations and enrichment analysis

For dataset one, snpEFF identified that 93.217% of candidate loci had a modifier effect, that is a change in protein structure, along with another 4.361% having a moderate effect. 65.622% of SNPs were missense variants with another 31.734% being silent variants. 84.58% of candidate loci were in introns with 15.42% being in exons. There was an overall Ts/Tv ratio of 2.4307. EggNog identified 1282 unique genes from gene prediction that were involved in 277 unique KO terms and 3124 unique GO terms. Twelve of those KO terms were over-represented, and 157 GO terms were over-represented (Table 2, Supporting Information Figure S13). Metabolic and environmental adaptation pathways were the common pathways of over-represented KO terms (Table 2). Over-represented GO terms included response to abiotic stress, defense pathways, signal transduction, transcript regulation, metabolism and cell/organelle development (Supporting Information Figure S13). For Dataset two, snpEFF identified 93.437% of candidate loci as having a modifier effect along with another 3.934% having a moderate effect. Missense variants accounted for 61.011% of variants and 36.18% were silent variants; 83.41% of variants were in introns and 16.59% were in exons. The overall Ts/Tv ratio was 2.4252. EggNog identified 948 genes from gene prediction that were involved in 212 unique KO terms and 2715 unique GO terms. 11 KO terms and 135 GO terms were over-represented (Table 2, Supporting Information Figure S14). Similar pathways and GO term functions seen with the first dataset were found with these over-represented terms for dataset two. Queries for all candidate loci, ortholog metrics, KO and GO terms are shown in Supporting Information table S6.

The most abundant *A. thaliana* orthologs for both datasets are shown in Table 3. Similar abundances of proteins are seen across both datasets, showing consistency of sPLS across datasets. Genes producing proteins such as non-specific serine/threonine protein kinase (EC 2.7.11.1), Protein kinase domain-

containing protein and RING-type E3 ubiquitin transferase (EC 2.3.2.27) were some of the most abundant orthologs identified.

Genes producing proteins such as Peroxidase (EC 1.11.1.7), ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase (EC 3.2.2.6) and P-loop containing nucleoside triphosphate hydrolases superfamily protein were also very common. Other interesting orthologs identified included genes that produce Leucine-rich repeat protein kinase family proteins, Kinesin-like proteins, 2-oxoglutarate (2OG) and Fe (II)-dependent oxygenase superfamily proteins. The explained variance across SNP and environmental datasets from the sPLS along with the level of effect in variants and the biological relevance of over-represented KO terms, GO terms and *A.thaliana* orthologs, all support a strong association of the loci with the environmental variables identified.

Discussion

Genome-wide analysis supports the presence of 3-4 genetic clusters currently in the species

The results of our study provide an update in the storied saga of population structure in Ponderosa pine. Our results support 3 or 4 distinct clusters within Ponderosa pine; a California cluster, a northern Coastal Cluster, an Interior cluster and potentially a Northern Interior cluster. Our results show limited support for the southwestern Ponderosa pine variety (*Pinus ponderosa* subsp. *brachyptera*) and the great plains variety (*Pinus ponderosa* subsp. *readiana*) (Conkle and Critchfield 1988; Potter et al. 2013; Callaham et al. 2013; Willyard et al. 2017); as dataset one shows individuals from the Southwest to South Dakota cluster together (Figure 1, Supporting Information Figures S2-S4). Montana and Wyoming appear to be slightly differentiated from the rest of the intermountain west which disagrees with Conkle and Critchfield 1988, Potter et al. 2013, Callaham et al. 2013 and Willyard et al. 2017. However, pairwise FST and ordination distance on the DAPC suggest gene flow is still occurring between the interior and northern interior clusters; which does provide some support for Little 1971's var *scopulorum*. Interestingly, our results somewhat mirror differentiation in the coastal portion of the Ponderosa pine seen in Conkle and Critchfield 1988, and Callaham et al. 2013; with the main difference in our results showing California is differentiated from Washington, Oregon and Idaho. While our study does not have an outgroup to confirm phylogenomic results and designate true variety differences; our study does suggest there is a need to update current variety distinctions. Future studies could use modern phylogenomic techniques to provide more concrete variety distinctions.

Bottleneck events have a large effect on current population structure

The results of Stairway Plot 2 suggest that broad and fine scale bottlenecks from the mid-Pleistocene to the Holocene played a major role in the strongly structured genetic clusters we see today. The extremely high effective population size that each genetic cluster had (40-50 million individuals) ca. 400-500 kya, would suggest much higher continuity in populations of Ponderosa pine in the mid-Pleistocene. This continuity in Ponderosa pine populations most likely lessened with each bottleneck; furthering differentiation across genetic clusters with each subsequent bottleneck due to random losses of alleles

and reduced corridors for gene flow (Nei et al. 1975, Hewitt 1996; Charlesworth 2002). As effective population size decreased over time, the effectiveness of selective forces most likely steadily rose (Ohta 1992, Lanfear et al. 2014; Chen et al. 2017); further differentiating genetic clusters.

The specific paleoecological events that match estimated population bottlenecks are as follows. The Mid-Bruhnes event occurred ca. 430 thousand years ago (kya) (Head and Gibbard 2015; Barth et al. 2018), our demographic modeling shows strong declines in the lower confidence bounds in all clusters around that time. A Milankovitch forcing event, a sudden cooling and warming based on gain and loss of ice, is believed to have occurred around 100kya (Andrews and Dyke 2007), around that time our demographic models show a strong decline in N_e followed by subsequent stability in the interior, California, northern coastal clusters and a smaller decrease in effective population sizes in the northern interior cluster. There was a major Yellowstone glacial advance estimated to have occurred 70-45 kya (Clark and Bartlein 1995), in that time period we see a sharp decline in effective population size of the northern interior cluster (the closest cluster to that glacier), a small decline in the California and Northern Coastal Clusters but stability in the Interior cluster. Another Yellowstone glacial advance was believed to occur from 47-54 kya (Clark and Bartlein 1995), which coincides with another small decline in N_e in the northern interior cluster and a steady decline the Northern Coastal cluster, but stability in the interior and California clusters, which are much further away from that glacial advance. Glacial advances across the Northern Rocky Mountains were also believed to occur 34-47 kya (Clark and Bartlein 1995), which coincides with a continuing decline in the northern coastal cluster but stability in all other clusters. The stability of the Northern interior cluster in this time, and from there on, may indicate populations were existing in refugia with high and consistent levels of protection from other glacial advances and climatic events due to the strong die off seen in previous events. From 22.5-30kya a glacial advance occurred in the Rocky Mountains (Clark and Bartlein 1995), in that time we see a sharp decline in N_e in the Interior cluster, a consistent continuing decline in the northern coastal cluster, the beginnings of a decline in the California cluster and stability in the northern interior cluster. A glacial advance then occurred between 15.5-19.5 kya (Clark and Bartlein 1995), where we see the end of a sharp decline in the interior cluster, a sharp decline in the California cluster, and the continued consistent decline in the Northern Coastal Cluster. A multi-millennia long drought is believed to have occurred from 14.5kya to 12.9kya (Polyak et al. 2012), in this time we see a drop in N_e in the lower confidence intervals of the interior cluster, along with a strong drop in N_e that occurs after this time period in the mean projection. The northern coastal cluster also shows the continuing consistent decline during that time period. Interestingly, we see relative stability in the California and northern interior clusters across all other clusters during this time, which potentially suggests not all portions of North America experienced that drought. A cold event is believed to occur around 8.2kya (Dean et al. 2002), during that time we see sharp die-offs in all clusters besides the northern interior cluster. From 4-8kya, a general decline in temperature was combined with general increases in moisture (Shuman and Marsicek 2016). During this time period, high variability in both temperature and moisture is still believed to have occurred (Shuman and Marsicek 2016). Decreases in N_e in the 4-8kya time period are observed in all clusters besides the northern interior cluster. During this cooler and wetter period, Ponderosa pine might have been outcompeted by other

faster-growing high elevation conifers, causing this bottleneck event. Bottlenecks due to competition may be seen again 2-4kya, as there was another large increase in moisture and cooler temperatures (Shuman and Marsicek 2016); we see relative stability in the interior and northern interior clusters but declines in both the California and northern coastal clusters. The stability seen in the interior clusters during this time period suggests that around 4 kya the cluster reached its current occurrence. Whereas the current occurrence for the California and northern coastal clusters may have not occurred until after this event ended 2 kya.

Spatially heterogeneous selection shapes population structure

Heterogeneous environments combined with moderate to high levels of gene flow that occur within each genetic cluster provide a basis for excess heterozygosity indicative of balancing selection. Our GEA analyses suggest that winter, spring and summer moisture availability act as the most important selection pressures that currently affect Ponderosa populations. Moisture availability varies greatly among and within genetic clusters (Supporting Information Table S2, S3) (Dixit et al. 2022). Pairwise F_{ST} (Supporting Information Figures S3 and S4) and DAPC (Figure 1) both suggest at least moderate levels of gene flow occurring within all genetic clusters, despite the stark environmental gradients (Supporting Information Table S2, S3). Certain alleles are most likely favored in specific environments (Figure 5) (Wright 1931; Levene 1953; Deakin 1966; Christiansen 1974), but can still be dispersed to other areas within genetic clusters where they may be less favored or selected against; helping to create an abundance of heterozygotes even with numerous bottleneck events (Slatkin 1987; Yeaman and Whitlock 2011; Comeault et al. 2014; Laurent et al. 2016). We saw one high elevation southern California seed source (CA 993) showing similar alleles across candidate loci as the northern coastal cluster (Figure 5). This could be due to similar climates independently favoring specific alleles (Wright 1931; Levene 1953; Deakin 1966; Christiansen 1974), creating similar populations despite minimal chances of gene flow, or it could be due to a remnant of when California and northern coastal states had higher continuity before the numerous bottlenecks occurred. Future studies should examine the occurrence of selection over time using simulation methods or fossil records.

While it is difficult to untangle whether demographic events or selection solely created differentiation in specific loci, our results show the importance of the interplay between both demographic history and selection in shaping what researchers see today out on the landscape (Lenormand 2002; Charlesworth 2003). Our results show signatures of balancing selection maintain the majority of SNPs identified, based on dn/ds , a high Ts:Tv bias and Tajima's D. Purifying selection and neutral evolution most likely play smaller roles in maintaining the SNPs we identified. The unique patterns of selection seen in the northern coastal cluster in the Tajima's D analysis highlights the sensitivity of that metric to demographic history; as similar start and end N_e 's are seen in that cluster but the magnitude of bottlenecks are much smaller creating a wider distribution for possible neutral values of Tajima's D. Furthermore, Our results echo the common sentiment of the importance of demographic history on making inferences of selection as our Tajima D results would be very different if we didn't create unique neutral Tajima's D

distributions for each genetic cluster and relied on the unit normal distribution, $N(0,1)$, that Tajima's D theoretically approximates (Tajima 1989).

Ecological implications of GEA Candidate Loci

The importance of moisture availability in tree health is common knowledge, and our results suggest moisture variables would be an important factor to keep in mind for assisted gene flow or assisted migration (Aitken and Whitlock 2013). The changes in allele frequency in response to changing climate variables (Figure 5) further supports that certain alleles are favored under specific moisture conditions (Wright 1931; Levene 1953; Deakin 1966; Christiansen 1974). Furthermore, the overrepresented functions (Table 3) on candidate loci suggest unique drought and stress responses occur among each genetic cluster specific to local moisture regimes. Our study provides a foundation for future studies to explore if these allele and climate associations offer robust indicators of seed source transfer performance in restoration settings. Specifically, if seed sources that exist currently in drier conditions can perform well in areas that are projected to be drier in the future. Future studies should also explore the plasticity of specific alleles seen in candidate loci to better determine if seed sources need to be transferred to new areas or not.

Increased risk of maladaptation evidenced by genomic offset

Our genomic offset reveals that areas across all genetic clusters are at-risk of maladaptation most likely due to drastic shifts in future climate. Portions of the range with the highest genomic offset (Southwest, Great Plains, portions of California, Pacific Northwest), and thereby predicted to be at higher risk of maladaptation coincide with recent studies showing altered survival, growth and phenology in response to changing climate (Martinez-Berdeja et al. 2019), increased mortality events (Klenner et al. 2009; Kolb et al. 2019), reduction in growth, regeneration, and seedling survival, as a consequence of climate change and other compounding factors such as bark-beetle outbreaks and higher intensity fires (Kolb et al. 2007, 2019; Petrie et al. 2023). Our results don't necessarily suggest that demographic events and selection have resulted in an overarching unfit species of Ponderosa pine. Risk of maladaptation is more likely due to climate projections suggesting rapid changes in a short amount of time (IPCC 2014). However, the evidence of spatially heterogeneous selection seen in each genetic cluster may be of benefit as advantageous alleles for future climates may already be out on the landscape.

Genomic offset is a relatively new method and a rapidly developing field of study that still needs field-based tests to confirm model projections, even with all due diligence conducted by investigators. The overlap in predicted maladaptation and negative effects seen in field surveys does support our projections (Klenner et al. 2009; Martinez-Berdeja et al. 2019; Kolb et al. 2007, 2019; Petrie et al. 2023). Higher genomic offsets have also been correlated with negative effects in common gardens in other systems (Lind et al. 2024), suggesting this method does provide robust inferences.

Conclusions

In summary, genome-wide data suggests the need to update varieties within Ponderosa pine. We found strong evidence of demographic events and spatially heterogeneous selection shaping population structure. Numerous bottlenecks combined with purifying selection, spatially heterogeneous selection and neutral events all contributed to the current population structure we see today. Moisture related variables were the most important selection pressures causing differentiation. We found evidence to support that evolutionary events that shape current population structure can also play a role in maladaptation risk in the future.

Declarations

Acknowledgements

We'd like to thank A. Willyard for sharing her extensive knowledge of Ponderosa pine, and J. Reinhardt (RMRS), K.Corella (Calfire), M. Battaglia (RMRS), J. Debell (USFS), R.Heckman (RMRS), S. Barga (RMRS), M.Vance (BC MOF), N.Eiten (BLM), M.Hightower (BLM), D.Harriman (BLM), E.Moran (UC-Merced), C.Jensen (USFS), M.Partain (USFS), G.Cline for collecting samples and sending seeds.

Funding

This project was supported by The Bureau of Land Management (BLM) - Colorado [grant #L21AC10398-00] and NAU-TRIF funds to A.R.D.L.T.

Ethics approval and consent to participate

No animal or human subjects were used in this study and as such no ethic approval was required nor was consent to participate.

Consent for publication

All authors consent to have this work published. No sensitive data from any persons was collected in this study and no portion of this study was conducted on private property.

Competing Interests

The authors declare no competing interests.

Author contributions

A.R.D.L.T. designed and supervised the project and obtained research funding; A.L collected samples and helped with experimental design; S.M.C performed all genomic and bioinformatic analyses; A.R.D.L.T and S.M.C wrote the manuscript; all authors approved the final version of the manuscript.

Data Availability

Raw sequence reads for all samples used in this study are uploaded at NCBI under the bioproject PRJNA1241888, <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1241888>.

References

1. Aitken SN, Whitlock MC. 2013. Assisted gene flow to facilitate local adaptation to climate change. *Annual review of ecology, evolution, and systematics*. 44(1):367-388.
2. Al Azawi TNI, Khan M, Mun BG, Lee SU, Lee DS, Rahim W, et al. 2022. Nitric Oxide-Induced Downregulation of a NAD (P)-Binding Rossmann-Fold Superfamily Gene Negatively Impacts Growth and Defense in *Arabidopsis thaliana*. In *Proceedings of the Korean Society of Crop Science Conference. The Korean Society of Crop Science*. 143
3. Allen CD, Savage M, Falk DA, Suckling KF, Swetnam TW, Schulke T, et al. 2002. Ecological restoration of southwestern ponderosa pine ecosystems: a broad perspective. *Ecological applications*. 12(5):1418-1433.
4. Alrayes L, Stout J, Schroeder, D. 2023. *Arabidopsis* RAD16 homologues are involved in UV tolerance and growth. *Genes*, 14(8):1552.
5. Andrews JT, Dyke AS. 2007. Late Quaternary in North America. *Science Letters*. 231:111-130.
6. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
7. Ariga H, Katori T, Tsuchimatsu T, Hirase T, Tajima Y, Parker J E, et al. 2017. NLR locus-mediated trade-off between abiotic and biotic stress adaptation in *Arabidopsis*. *Nature Plants*. 3(6): 1-8.
8. Barth A M, Clark P U, Bill NS, He F, Pisias NG. 2018. Climate evolution across the Mid-Brunhes transition. *Climate of the Past*. 14(12): 2071-2087.
9. Bakalova S, Nikolova A, Nedeva D. 2004. Isoenzyme profiles of peroxidase, catalase and superoxide dismutase as affected by dehydration stress and ABA during germination of wheat seeds. *Journal of plant physiology*. 30(1-2): 64-77.
10. Beier P, Rogan EC, Ingraldi MF, Rosenstock SS. 2008. Does forest structure affect reproduction of northern goshawks in ponderosa pine forests?. *Journal of Applied Ecology*. 342-350.
11. Brown AF, Kann ML, Rand DM. 2007. Gene flow versus local adaptation in the northern acorn barnacle *Semibalanus balanoides*: insights from mitochondrial DNA variation. *Evolution* 55:1972–1979.
12. Callaham RZ. 2013. *Pinus ponderosa*: Geographic races and subspecies based on morphological variation. Res. Pap. PSW-RP-265. Albany, CA: US Department of Agriculture, Forest Service, Pacific Southwest Research Station. 53:265.
13. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution*. 38(12): 5825-5829.
14. Charlesworth B. 2002. Effective population size. *Current Biology*. 12(21): R716-R717.

15. Clark PU, Bartlein PJ. 1995. Correlation of late Pleistocene glaciation in the western United States with North Atlantic Heinrich events. *Geology*. 23(6): 483-486.
16. Chae L, Sudat S, Dudoit S, Zhu T, Luan S. 2009. Diverse transcriptional programs associated with environmental stress and hormones in the Arabidopsis receptor-like kinase gene family. *Molecular plant*. 2(1):84-107.
17. Chauhan H, Singh K. 2023. Genome-wide identification of 2-oxoglutarate and Fe (II)-dependent dioxygenase family genes and their expression profiling under drought and salt stress in potato. *PeerJ*, 11:e16449.
18. Chen J, Glémin S, Lascoux M. 2017. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular biology and evolution*. 34(6): 1417-1428.
19. Christiansen FB. 1974. Sufficient conditions for protected polymorphism in a subdivided population. *The American Naturalist*. 108(960): 157-166.
20. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang, L, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *fly*. 6(2): 80-92.
21. Cole KL. 1983 .Late Pleistocene vegetation of Kings Canyon, California. *Quat Res*. 19:117–129.
22. Comeault AA, Flaxman SM, Riesch R, Curran E, Soria-Carrasco V, Gompert Z, et al. 2015. Selection on a genetic polymorphism counteracts ecological speciation in a stick insect. *Current Biology*. 25(15): 1975-1981.
23. Conkle MT, Critchfield WB. 1988. Genetic variation and hybridization of Ponderosa pine. In: *Ponderosa pine: the species and its management, Washington State University Cooperative Extension*. 1988:27-43.
24. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. 2011. The variant call format and VCFtools. *Bioinformatics*. 27(15): 2156-2158.
25. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience*. 10(2): giab008
26. Darwin C. 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life
27. Deakin MA.1966. Sufficient conditions for genetic polymorphism. *The American Naturalist*. 100(916):690-692.
28. Dean WE, Forester RM, Bradbury JP. 2002. Early Holocene change in atmospheric circulation in the Northern Great Plains: an upstream view of the 8.2 ka cold event. *Quaternary Science Reviews*. 21(16-17): 1763-1775.
29. del Campillo E, Gaddam S, Mettle-Amuah D, Heneks J. 2012. A tale of two tissues: AtGH9C1 is an endo- β -1,4-glucanase involved in root hair and endosperm development in Arabidopsis. *PloS One*. 7(11): e49363. doi: 10.1371/journal.pone.0049363

30. De La Torre AR, Li Z, Van de Peer Y, Ingvarsson PK. 2017. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Molecular biology and evolution*. 34(6): 1363-1377.
31. Dixit A, Kolb T, Burney O. 2022. Trade-off between growth rate and water use efficiency in southwestern Ponderosa pine provenances. *Forest Ecology and Management*. 515:120239.
32. Dixon P. 2003. VEGAN, a package of R functions for community ecology. *Journal of vegetation science*. 14(6): 927-930.
33. Dong X, Nou IS, Yi H, Hur, Y. 2015. Suppression of ASK β (AtSK32), a clade III Arabidopsis GSK3, leads to the pollen defect during late pollen development. *Molecules and cells*. 38(6):506-517.
34. Dray S, Dufour A. 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*. 22(4): 1–20. doi:10.18637/jss.v022.i04.
35. Ellis N, Smith SJ, Pitcher CR. 2012. Gradient forests: calculating importance gradients on physical predictors. *Ecology*. 93(1):156-168.
36. Ewels P, Magnusson M, Lundin S, Källér M. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*.doi: 10.1093/bioinformatics/btw3
37. Ffolliott PF, Gottfried GJ, Baker Jr MB. 1989. Water yield from forest snowpack management: research findings in Arizona and New Mexico. *Water Resources Research*. 25(9):1999-2007.
38. Fijarczyk A, Babik W. 2015. Detecting balancing selection in genomes: limits and prospects. *Molecular ecology*, 24(14), 3529-3545.
39. Fu W, Fan D, Zhang Y. 2024. Genome-wide identification and characterization of xyloglucan endotransglucosylase/hydrolase gene family in maize (*Zea mays* L.) and the function of ZmXTH30 in response to drought stress. *Environmental and Experimental Botany*. 222:105744.
40. Girka F, Camenen E, Peltier C, Gloaguen A, Guillemot V, Le Brusquet L, Tenenhaus, A. 2023. Multiblock data analysis with the RGCCA package. *Journal of Statistical Software*. 1-36.
41. Guo C, McDowell IC, Nodzenski M, Scholtens DM, Allen AS, Lowe WL, Reddy TE. 2017. Transversions have larger regulatory effects than transitions. *BMC genomics*, 18(1), 394.
42. Haller BC and Messer PW. 2019. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Molecular biology and evolution*. 36.3:632-637.
43. Han L, Shen B, Wu X, Zhang J, Wen YJ. 2024. Compressed variance component mixed model reveals epistasis associated with flowering in Arabidopsis. *Frontiers in Plant Science*. 14:1283642.
44. Head MJ, Gibbard PL. 2015. Early–Middle Pleistocene transitions: linking terrestrial and marine realms. *Quaternary International*. 389:7-46.
45. Hemstrom W, Jones M. 2023. snpR: User friendly population genomics for SNP data sets with categorical metadata. *Molecular Ecology Resources*. 23(4):962-973.
46. Hewitt GM. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological journal of the Linnean Society*. 58(3):247-276.

47. Higashi Y, Hirai MY, Fujiwara T, Naito S, Noji M, Saito, K. 2006. Proteomic and transcriptomic analysis of Arabidopsis seeds: molecular evidence for successive processing of seed proteins and its implication in the stress response to sulfur nutrition. *The Plant Journal*. 48(4):557-571.
48. Hok S, Danchin EG, Allasia V, Panabières F, Attard A, Keller H. 2011. An Arabidopsis (malectin-like) leucine-rich repeat receptor-like kinase contributes to downy mildew disease. *Plant, cell & environment*. 34(11):1944-1957.
49. Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*. 23:183–201.
50. IPCC. 2014. Summary for policymakers. In: Climate Change 2014: Impacts, Adaptation and Vulnerability - Contributions of the Working Group II to the Fifth Assessment Report (eds CB Field, VR Barros, DJ Dokken et al.). *Cambridge University Press*, Cambridge, UK. 1–32
51. Jadoski SO, Salvador ED, Klar AE, Lima JPP. 2004. Peroxidase enzyme (EC. 1.11. 1.7) activity as an indicator of water stress in sweet pepper plants. *In International Symposium on Soilless Culture and Hydroponics*. 697:517-522.
52. Jain M, Nijhawan A, Arora R, Agarwal P, Ray S, Sharma P, et al. 2007. F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. *Plant physiology*. 143(4):1467-1483.
53. Jaramillo-Correa JP, Verdú M, González-Martínez SC. 2010. The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. *BMC Evolutionary Biology*. 10(1):22.
54. Jiang H, Tang B, Xie Z, Nolan T, Ye H, Song GY. et al. 2019. GSK 3-like kinase BIN 2 phosphorylates RD 26 to potentiate drought signaling in Arabidopsis. *The Plant Journal*. 100(5):923-937.
55. Jiang L, Lin Y, Wang L, Peng Y, Yang M, Jiang Y, et al. 2023. Genome-wide identification and expression profiling reveal the regulatory role of U-box E3 ubiquitin ligase genes in strawberry fruit ripening and abiotic stresses resistance. *Frontiers in Plant Science*. 14:1171056.
56. Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. 24(11):1403-1405.
57. Karlin S. 1982. Classifications of selection-migration structures and conditions for a protected polymorphism. *Evol. Biol*, 14(61):204.
58. Keith JO. 1965. The Abert squirrel and its dependence on ponderosa pine. *Ecology*. 46(1-2):150-163.
59. Kerr KL, Meinzer FC, McCulloh KA, Woodruff DR, Marias DE. 2015. Expression of functional traits during seedling establishment in two populations of Pinus ponderosa from contrasting climates. *Tree Physiology*. 35(5):535-548.
60. Kim ST, Choi M, Bae SJ, Kim JS. 2021. The functional association of ACQOS/VICTR with salt stress resistance in Arabidopsis thaliana was confirmed by CRISPR-mediated mutagenesis. *International Journal of Molecular Sciences*. 22(21):11389.
61. Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*. 61:893–903

62. Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge UnivPress, Cambridge UK
63. Kittelson PM, Maron JL. 2001. Fine-scale genetically based differentiation of life-history traits in the perennial shrub *Lupinus arboreus*. *Evolution*, 55(12):2429-2438.
64. Klenner W, Arsenault A. 2009. Ponderosa pine mortality during a severe bark beetle (Coleoptera: Curculionidae, Scolytinae) outbreak in southern British Columbia and implications for wildlife habitat management. *Forest Ecology and Management*. 258:S5-S14.
65. Kolb TE, Agee JK, Fule PZ, McDowell NG, Pearson K, Sala A, Waring RH. 2007. Perpetuating old Ponderosa pine. *Forest Ecology and Management*. 249(3):141-157.
66. Kolb T, Dixit A, Burney O. 2019. Challenges and opportunities for maintaining Ponderosa pine forests in the southwestern United States. *Tree Planters' Notes*. 62(1&2): 104-112.
67. Kreps JA, Wu Y, Chang HS, Zhu T, Wang X, Harper JF. 2002. Transcriptome changes for *Arabidopsis* in response to salt, osmotic, and cold stress. *Plant physiology*. 130(4): 2129-2141.
68. Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS genetics*, 4(12), e1000304.
69. Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. *Trends in ecology & evolution*. 29(1):33-41.
70. Laurent S, Pfeifer SP, Settles ML, Hunter SS, Hardwick KM, Ormond L, et al. 2016. The population genomics of rapid adaptation: disentangling signatures of selection and demography in white sands lizards. *Molecular ecology*. 25(1):306-323.
71. Lê S, Josse J, Husson F. 2008. FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software*. 25(1): 1–18. doi:10.18637/jss.v025.i01.
72. Lee KM, Ranta P, Saarikivi J, Kutnar L, Vreš B, Dzhus M, et al. 2020. Using genomic information for management planning of an endangered perennial, *Viola uliginosa*. *Ecology and evolution*. 10(5):2638-2649.
73. Li C, Zhang B, Yu H. 2021. GSK3s: nodes of multilayer regulation of plant development and stress responses. *Trends in Plant Science*. 26(12):1286-1300.
74. Li H, Durbin R. 2009. Fast and accurate short read alignment with burrowswheeler transform. *Bioinformatics*. 25:1754–1760.
75. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer, N., et al. 2009. The sequence alignment/map format and SAMtools. *bioinformatics*. 25(16):2078-2079
76. Lind BM, Candido-Ribeiro R, Singh P, Lu M, Obreht Vidakovic D, Booker TR, et al. 2024. How useful is genomic data for predicting maladaptation to future climate?. *Global Change Biology*/ 30(4):e17227.
77. Liu S, Ferchaud AL, Grønkjær P, Nygaard R, Hansen MM. 2018. Genomic parallelism and lack thereof in contrasting systems of three-spined sticklebacks. *Molecular Ecology*. 27: 4725-4743.
78. Liu X, Fu YX. 2020. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome biology*. 21(1):280.

79. Liu Y, Zhou Y, Cheng F, Zhou R, Yang Y, Wang Y, et al. (2024). Chromosome-level genome of putative autohexaploid *Actinidia deliciosa* provides insights into polyploidisation and evolution. *The Plant Journal*. 118(1):73-89.
80. Lenormand T. 2002. Gene flow and the limits to natural selection. *Trends in ecology evolution*, 17(4):183-189.
81. Leonhardt N, Kwak JM, Robert N, Waner D, Leonhardt G, Schroeder, JI. 2005. Microarray Expression Analyses of Arabidopsis Guard Cells and Isolation of a Recessive Absciscic Acid Hypersensitive Protein Phosphatase 2C Mutant. *Plant Cell* 16:596–615. *Plant Cell*. 17(4): 1330-1333.
82. Lee D, Polisensky DH, Braam J. 2005. Genome-wide identification of touch-and darkness-regulated Arabidopsis genes: a focus on calmodulin-like and XTH genes. *New Phytologist*. 165(2):429-444.
83. Levene H. 1953. Genetic equilibrium when more than one ecological niche is available. *The American Naturalist*. 87(836):331-333.
84. Li P, Li YJ, Wang B, Yu HM, Li Q, Hou BK. 2017. The Arabidopsis UGT87A2, a stress-inducible family 1 glycosyltransferase, is involved in the plant adaptation to abiotic stresses. *Physiologia Plantarum*. 159(4):416-432.
85. Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences*. 104(suppl_1):8597-8604.
86. Mahony CR, Wang T, Hamann A, Cannon AJ, 2022. A CMIP6 ensemble for downscaled monthly climate normals over North America. *International Journal of Climatology*. 42(11): 5871-5891. DOI: <https://doi.org/10.1002/joc.7566>
87. Martínez-Berdeja A, Hamilton JA, Bontemps A, Schmitt J, Wright JW. 2019. Evidence for population differentiation among Jeffrey and Ponderosa pines in survival, growth and phenology. *Forest Ecology and Management*. 434:40-48.
88. McKay J, Bishop JG, Lin JZ, Richards JH, Sala A, Mitchell-Olds T. 2001. Local adaptation across a climatic gradient despite small effective population size in the rare sapphire rockcress. *Proc R Soc Lond B Biol Sci*. 268:1715–1721.
89. Mondal R, Biswas S, Srivastava A, Basu S, Trivedi M, Singh SK, Mishra Y. 2021. In silico analysis and expression profiling of S-domain receptor-like kinases (SD-RLKs) under different abiotic stresses in *Arabidopsis thaliana*. *BMC genomics*. 22(1):817.
90. Müller SM, Wang S, Telman W, Liebthal M, Schnitzer H, Viehhauser A, et al. 2017. The redox-sensitive module of cyclophilin 20-3, 2-cysteine peroxiredoxin and cysteine synthase integrates sulfur metabolism and oxylipin signaling in the high light acclimation response. *The Plant Journal*. 91(6):995-1014.
91. Nei M, Maruyama T, Chakraborty R. 1975. The bottleneck effect and genetic variability in populations. *Evolution*. 1-10
92. Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3), 929-936.

93. Ogawa-Ohnishi M, Yamashita T, Kakita M, Nakayama T, Ohkubo Y, Hayashi Y, et al. 2022. Peptide ligand-mediated trade-off between plant growth and stress response. *Science*. 378(6616):175-180.
94. Oidaira H, Sano S, Koshiha T, Ushimaru T. 2000. Enhancement of antioxidative enzyme activities in chilled rice seedlings. *J. Plant Physiol*. 156:811 - 813.
96. Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annual review of ecology and systematics*. 263-286.
97. Pan J, Li X, Li J, Liu H, Hou Z, Yang Z, Xu L. 2022. Genome-wide identification of U-box gene family and expression analysis under abiotic stresses in *Salvia miltiorrhiza*. *Biocell*. 46(7):1751.
98. Panigrahi GK, Sahoo A, Satapathy KB. 2021. Differential expression of selected Arabidopsis resistant genes under abiotic stress conditions. *Plant Science Today*. 8(4):859-864.
99. Paradis E, Claude J, Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*. 20(2):289-290.
100. Pembleton LW, Cogan NO, & Forster JW. 2013. StAMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular ecology resources*. 13(5):946-952.
101. Petrie MD, Hubbard RM, Bradford JB, Kolb TE, Noel A, Schlaepfer DR., et al. 2023. Widespread regeneration failure in Ponderosa pine forests of the southwestern United States. *Forest Ecology and Management*. 545:121208.
102. Polyak VJ, Asmerom Y, Burns SJ, Lachniet MS. 2012. Climatic backdrop to the terminal Pleistocene extinction of North American mammals. *Geology*. 40(11):1023-1026.
103. Potter K M, Hipkins VD, Mahalovich MF, Means RE. 2013. Mitochondrial DNA haplotype distribution patterns in *Pinus ponderosa* (Pinaceae): Range-wide evolutionary history and implications for conservation. *American Journal of Botany*. 100(8):1562-1579.
104. Potter KM, Hipkins VD, Mahalovich MF, Means RE. 2015. Nuclear genetic variation across the range of Ponderosa pine (*Pinus ponderosa*): Phylogeographic, taxonomic and conservation implications. *Tree Genetics & Genomes*. 11:1-23.
105. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*. 81(3):559-575.
106. Rehman HM, Khan UM, Nawaz S, Saleem F, Ahmed N, Rana IA, et al. 2022. Genome wide analysis of family-1 UDP glycosyltransferases in *Populus trichocarpa* specifies abiotic stress responsive glycosylation mechanisms. *Genes*. 13(9):1640.
107. Roberts DR, Hamann A. 2015. Glacial refugia and modern genetic diversity of 22 western North American tree species. *Proceedings of the Royal Society B: Biological Sciences*. 282(1804):20142903.
108. Sane M, Diwan GD, Bhat BA, Wahl LM, Agashe D. (2023). Shifts in mutation spectra enhance access to beneficial mutations. *Proceedings of the National Academy of Sciences*, 120(22), e2207355120.

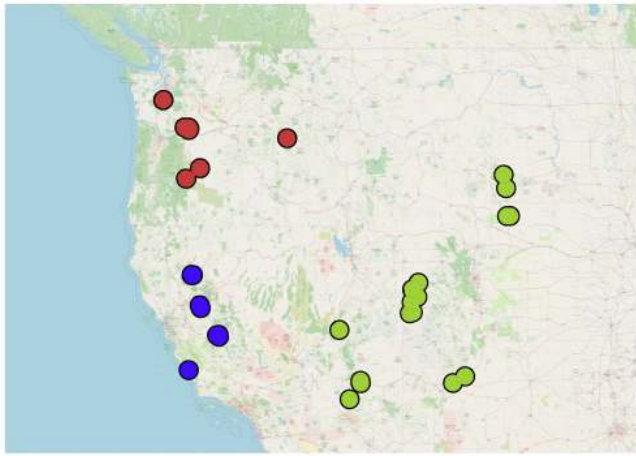
109. Sane M, Parveen S, Agashe D. 2025. Mutation bias alters the distribution of fitness effects of mutations. *PLoS Biology*, 23(7), e3003282.
110. Sambatti JB, Rice KJ. 2006. Local adaptation, patterns of selection, and gene flow in the Californian serpentine sunflower (*Helianthus exilis*). *Evolution*. 60(4):696-710.
111. Shinneman DJ, Means RE, Potter KM, Hipkins VD. 2016. Exploring climate niches of Ponderosa pine (*Pinus ponderosa* Douglas ex Lawson) haplotypes in the western United States: Implications for evolutionary history and conservation. *PLoS One*. 11(3):e0151811.
112. Shu M, Moran EV. 2023. Identifying genetic variation associated with environmental gradients and drought-tolerance phenotypes in Ponderosa pine. *Ecology and Evolution*. 13(10): e10620.
113. Shuman BN, Marsicek J. 2016. The structure of Holocene climate change in mid-latitude North America. *Quaternary Science Reviews*. 141:38-51.
114. Slatkin M. 1987. Gene flow and the geographic structure of natural populations. *Science*, 236(4803):787-792.
115. Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics*. 105(2):437-460.
116. Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 123(3):585-595.
117. Thibaud MC, Arrighi JF, Bayle V, Chiarenza S, Creff A, Bustos R, et al. 2010. Dissection of local and systemic transcriptional responses to phosphate starvation in *Arabidopsis*. *The Plant Journal*. 64(5):775-789.
118. Vekemans X, Lefèbvre C. 1997. On the evolution of heavy-metal tolerant populations in *Armeria maritima*: evidence from allozyme variation and reproductive barriers. *Journal of Evolutionary Biology*. 10(2):175-191.
119. Wakeley J. 1998. Segregating sites in Wright's island model. *Theoretical Population Biology*. 53(2):166-174.
120. Wakeley, J. 2000. The effects of subdivision on the genetic divergence of populations and species. *Evolution*. 54(4):1092-1101.
121. Wakeley J, Aliacar N. 2001. Gene genealogies in a metapopulation. *Genetics*. 159(2):893-905.
122. Wang T, Hamann A, Spittlehouse D, Carroll C. 2016. Locally Downscaled and Spatially Customizable Climate Data for Historical and Future Periods for North America. *PLoS ONE*. 11(6): e0156720. doi:10.1371/journal.pone.0156720
123. Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Molecular Biology and Evolution*. 24(1):90-101.
124. Willyard A, Gernandt DS, Potter K, Hipkins V, Marquardt P, Mahalovich MF, et al. 2017. *Pinus ponderosa*: A checkered past obscured four species. *American Journal of Botany*. 104(1):161-181.

125. Willyard A, Gernandt DS, Cooper B, Douglas C, Finch K, Karemera H, et al. 2021. Phylogenomics in the hard pines (Pinus subsection Ponderosae; Pinaceae) confirms paraphyly in Pinus ponderosa, and places Pinus jeffreyi with the California big cone pines. *Systematic Botany*. 46(3):538-561.
126. Wright S. 1931. Evolution in Mendelian populations. *Genetics*. 16(2): 97.
127. Wright S. 1949. The genetical structure of populations. *Annals of eugenics*. 15(1):323-354.
128. Xie S, Jiang H, Ding T, Xu Q, Chai W, Cheng B. 2018. *Bacillus amyloliquefaciens* FZB42 represses plant miR846 to induce systemic resistance via a jasmonic acid-dependent signalling pathway. *Molecular plant pathology*. 19(7):1612-1623.
129. Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A. M. K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1), 431-449.
130. Yeaman S. 2015. Local adaptation by alleles of small effect. *The American Naturalist*. 186(S1):S74-S89.
131. Yeaman S, Whitlock MC. 2011. The genetic architecture of adaptation under migration–selection balance. *Evolution*. 65(7):1897-1911

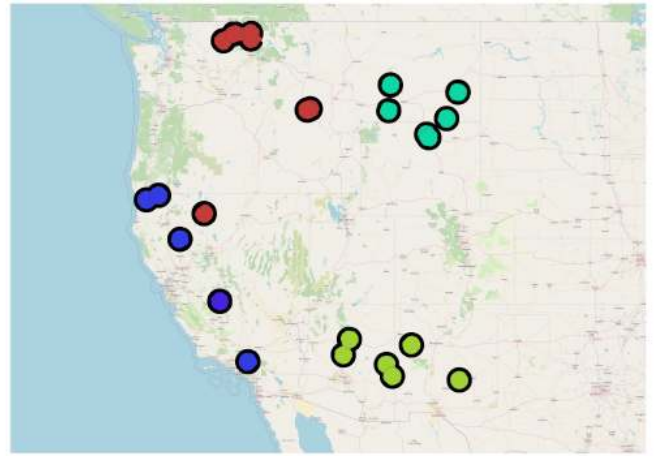
Tables

Tables 1 to 3 are available in the Supplementary Files section.

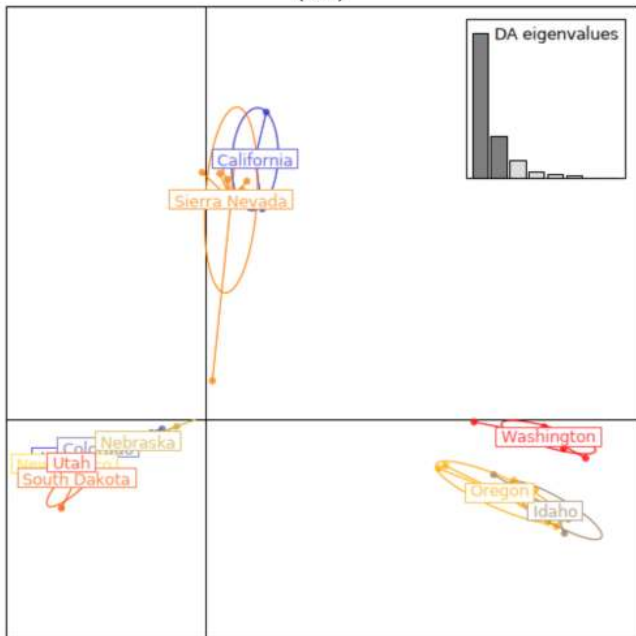
Figures



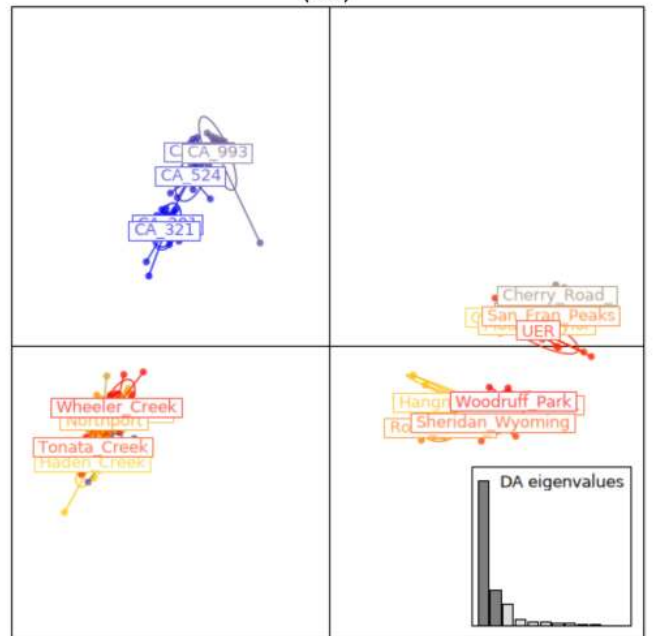
(a)



(b)



(c)



(d)

Figure 1

(a) Sampling locations of dataset one are colored by the DAPC genetic cluster. (b) Seed sources from dataset two are colored by genetic cluster from the DAPC results. (c) DAPC of dataset one with State as population label. Top left cluster is the interior West, bottom middle cluster is California, top right cluster is Pacific and Inland Northwest. (d) DAPC of dataset two with seed source as population label. Top left cluster is most of California seed sources, bottom left cluster is Eastern Washington, Idaho and Northeastern California seed sources, top right cluster is Arizona New Mexico seed sources, bottom right is Montana and Wyoming seed sources.

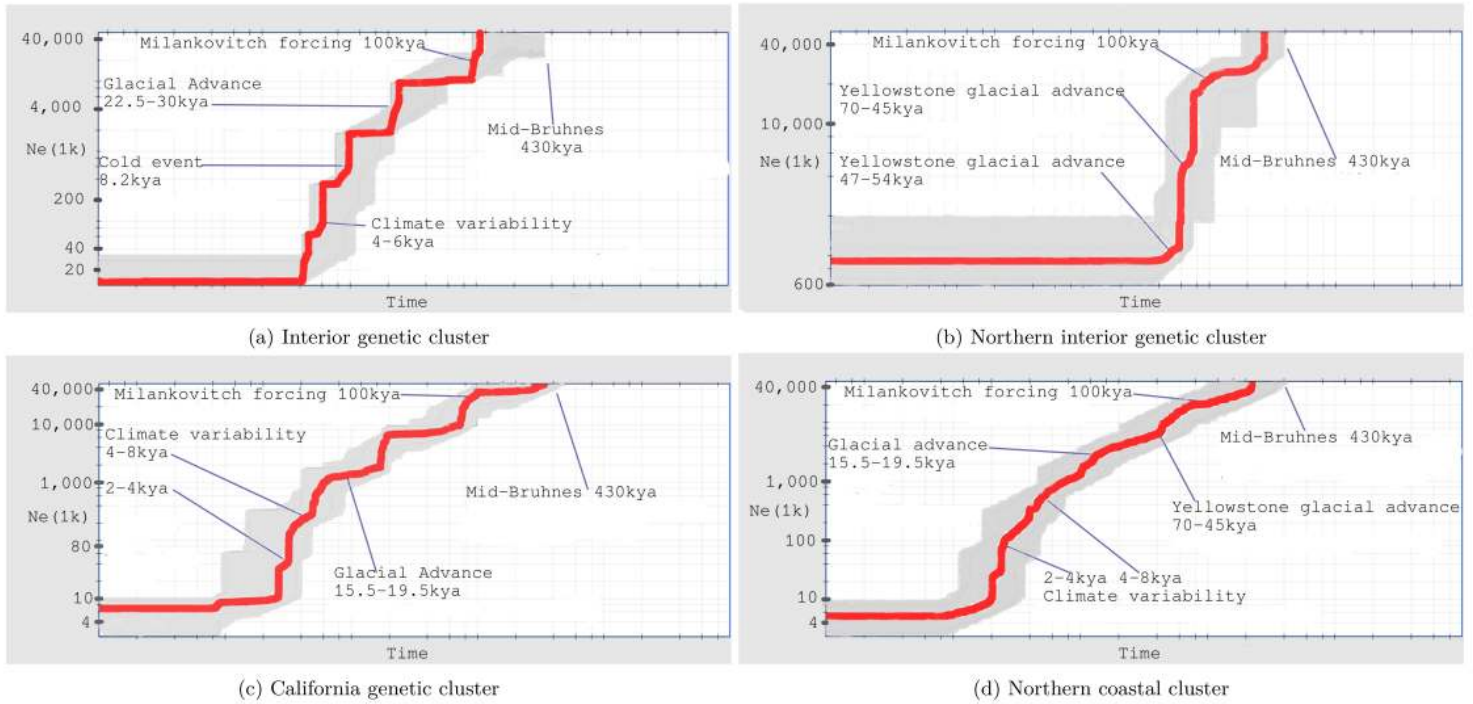


Figure 2

(a) Nucleotide diversity across sampling locations in dataset one. (b) Nucleotide diversity across seed sources in dataset two (c) Private alleles across sampling locations in dataset one (d) Private alleles across seed sources in dataset two (e) Inbreeding coefficient (FIS) across sampling locations ordered West to East in dataset one (f) Inbreeding coefficient (FIS) across seed sources ordered West to East in dataset two. Colors are based on genetic clusters determined by the DAPC in Figure 1.

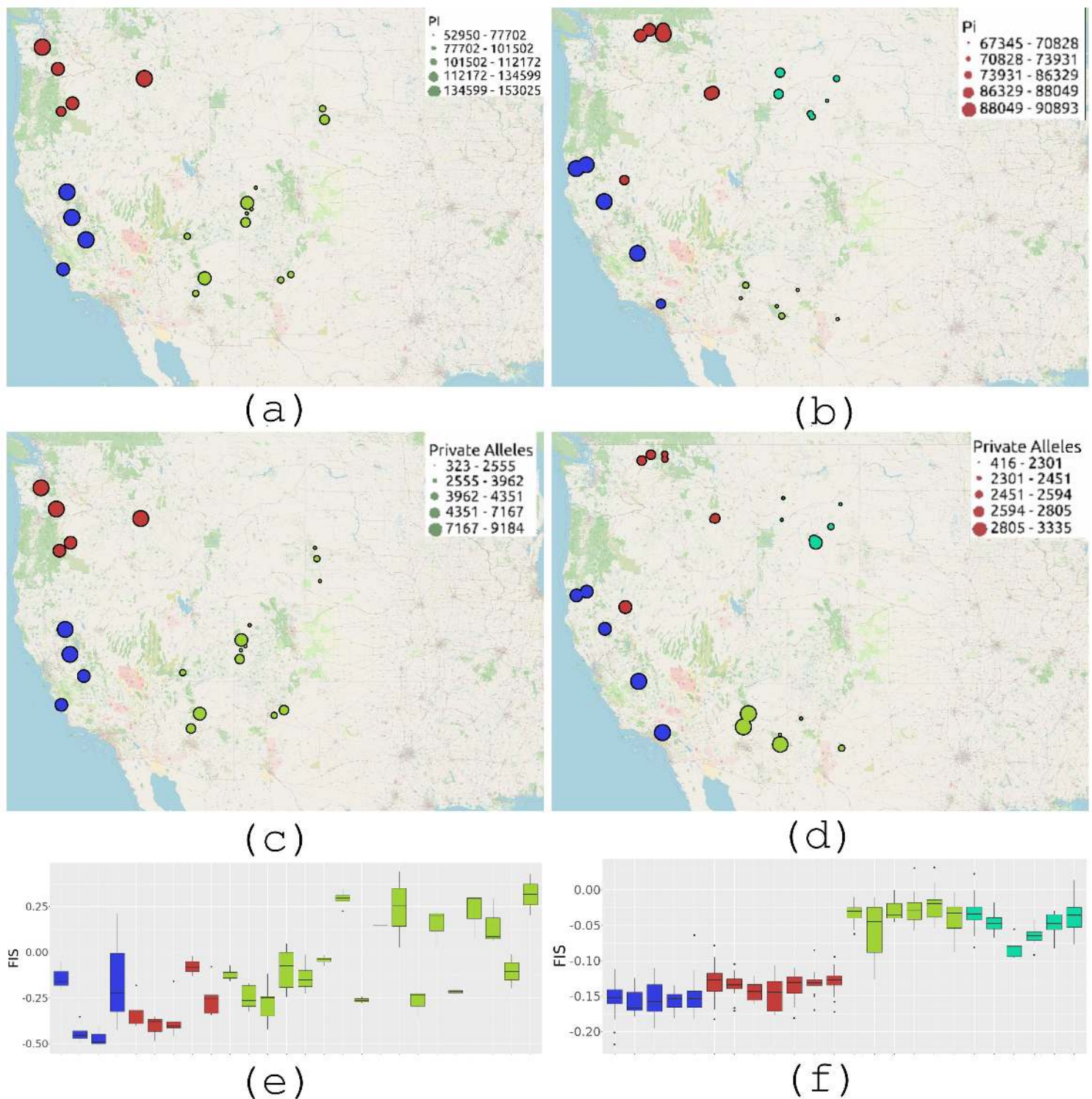


Figure 3

Modeled change in effective population size over time (log scale) from stairway plot for each genetic cluster found in dataset two. Red lines indicate mean estimates, while gray lines indicate 90,75,25,10 percent quantiles respectively. Overlaid are N_e estimates following major bottleneck events. Full detailed plots of demographic history for each genetic cluster are found in Figures S8-S11. (a) Starting median estimates of N_e for the interior cluster is ca. 40,000,000 while current median N_e is ca. 20,000. (b) Starting median estimates of N_e for the Northern interior cluster is 45,000,000 while current median N_e

is ca. 900,000. (c) Starting median estimates of N_e for the California cluster is ca. 40,000,000 while current median N_e is ca. 8,500. (d) Starting median estimates of N_e for the Northern Coastal Cluster is ca. 40,000,000 while current median N_e is ca. 5,500.

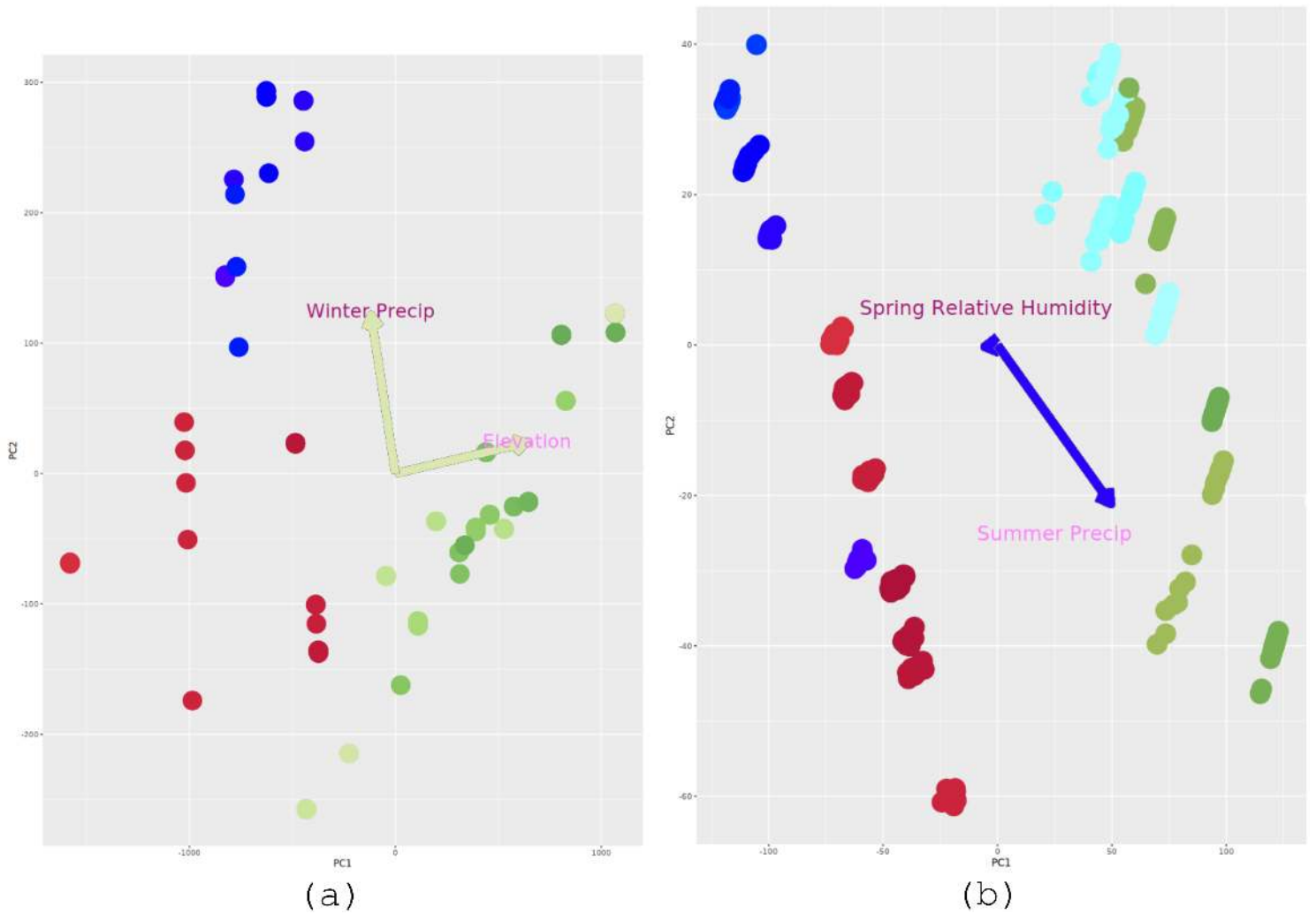
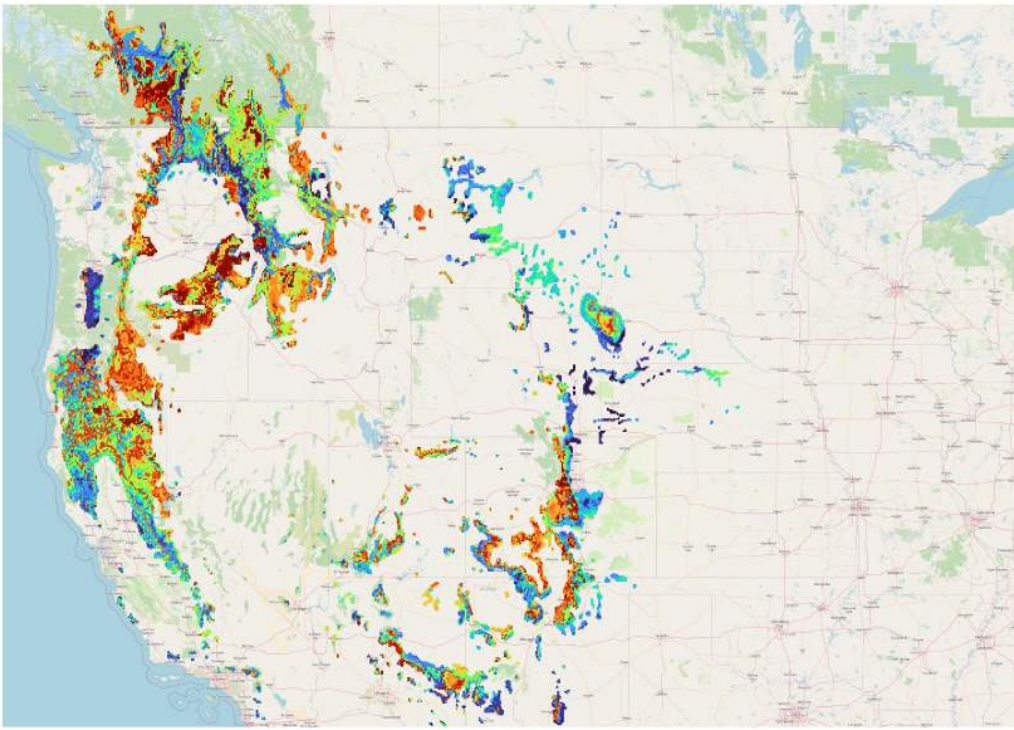
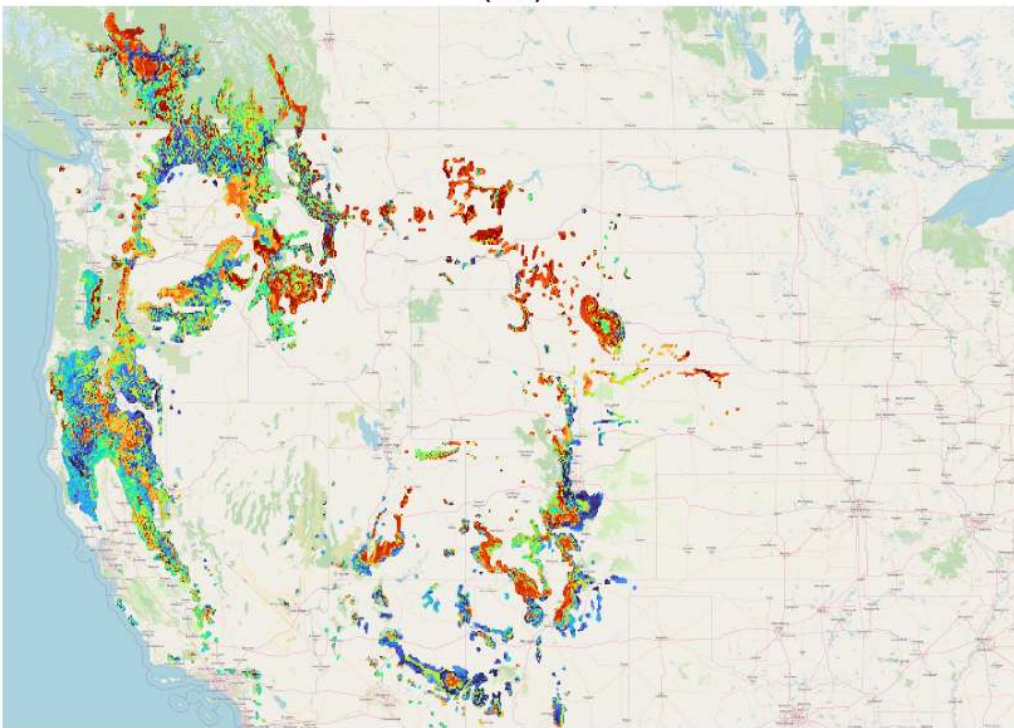


Figure 4

(a) Genomic offset predictions for 2041-2070 SSP 270 based on dataset one. (b) Genomic offset predictions for 2041-2070 SSP 270 based on dataset two. Darker Red pixels indicate higher genomic offset values, while cooler blue pixels indicate closer to zero genomic offset.



(a)



(b)

Figure 5

Biplot of candidate SNPs and important environmental variables denoted from the sPLS. Individuals are shown colored by the genetic cluster from Figure 1 and Figure 2. Greens represent the interior cluster, cyans represent the northern interior cluster, red represent the northern coastal cluster and dark blues represent the California cluster. In dataset one PC 1 accounts for 96.19 percent of variation and PC 2 accounts for 3.57 percent of cumulative variation. In dataset two PCA accounts for 73.94 percent of

variation, PC2 accounts for 11.25 percent of cumulative variation. Vectors represent where specific variables increase, that is, following the vector from 0 outward each variable goes up while on the other side of the line the variable goes down. (a) Allele changes in response to increasing elevation are seen left to right, while allele frequency changes in response to increasing winter precipitation are seen top to bottom; (b) shows how allele frequency changes in response to increasing spring relative humidity are seen right to left in the plot. Allele changes in response to increasing summer precipitation are seen along a diagonal top left to bottom right in the figure.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PipoBMCSuppTablesredo8125.xlsx](#)
- [Tables.docx](#)