Genomic evidence of an ancient Inland Temperate Rainforest

Megan Ruffley[1], Megan Smith[2], Anahi Espindola[3], Daniel Turck[1], Niels Mitchell[1], Bryan Carstens[4], Jack Sullivan[1], and David Tank[1]

[1]University of Idaho
[2]Indiana University Bloomington
[3]University of Maryland
[4]The Ohio State University

September 25, 2021

Abstract

The disjunct temperate rainforests of the Pacific Northwest of North America (PNW) are characterized by late-successional dominant tree species western redcedar (Thuja plicata) and western hemlock (Tsuga heterophylla). The demographics of these species, along with the PNW rainforest ecosystem in its entirety, have been heavily impacted by the geological and climatic changes the PNW has experienced over the last 5 million years, including mountain orogeny and repeated Pleistocene glaciations. These environmental events have ultimately shaped the history of these species, with inland segments potentially being extirpated during the Pleistocene glaciation. Here, we collect genomic data for both species across their ranges in order to develop multiple demographic models, each reflecting a different hypothesis on how the ecosystem dominant species may have responded to dramatic climatic change. Results indicate that inland and coastal populations in both species diverged an estimated ~2.5 million years ago and experienced a decrease in population size during glaciation, with a subsequent population expansion. Importantly, we found evidence for gene-flow between coastal and inland populations during the mid-Holocene. It is likely that intermittent migration in these species has prevented allopatric speciation. In conclusion, the combination of genomic data and population demographic inference procedures involving machine learning establish that populations of the ecosystem dominants Thuja plicata and Tsuga heterophylla persisted in refugia located in both the coastal and inland regions, with populations expanding and contracting in response to glacial cycles with occasional gene-flow.

Introduction

The old-growth cedar-hemlock forests of the Pacific Northwest of North America characterize one of the most diverse temperate rainforests in the world (Newmaster et al. 2003). This ecosystem includes disjunct coastal and inland temperate rainforest (ITR) elements, with the latter located in the Northern Rocky Mountains and separated from the larger coastal rainforest located in the Cascades and coastal ranges by approximately 200 km of xeric habitat. The entire Pacific Northwest region has been widely impacted by Pleistocene glacial/interglacial cycles (Waitt & Thorson 1983), with flora and fauna being strongly impacted by these climatic changes. The ITR has been of particular interest because of the dramatic implications of the alternative hypotheses that have been proposed to explain its history during and after the Pleistocene. One, the Recent Dispersal (RD) hypothesis, posits the recent (<5K years ago; Kya) establishment of the ITR, invoking a post-Pleistocene colonization of the inland areas from coastal populations, and implying that the ITR is a recent propagule of the coastal forest with little evolutionary novelty. The second hypothesis posits that the ITR represents an ancient disjunction between the inland and coastal forests (Brunsfeld *et al.* 2001) that occurred pre-Pleistocene (>2.5 million years ago; Mya). In this Ancient Vicariance (AV) hypothesis, while the onset of the glaciers caused massive contraction of the ITR, inland refugia persisted during cold periods and subsequently expanded *in-situ* post-Pleistocene. The AV hypothesis predicts that

allopatry may have led to speciation of some taxa in the ITR, and thus, that the inland region harbors a unique endemic flora and fauna. These two hypotheses broadly encapsulate the proposed modes of the formation of the disjunction. They are critical to understanding general biogeographic processes associated with the ecosystem and the region, and they also have broad implications for conservation and management of this diversity hotspot.

The range of the PNW temperate rainforest is defined by the distributions of the two late-successional dominant species, *Tsuga heterophylla* Raf. (Sarg.) (western hemlock) and *Thuja plicata*Donn ex. D. Don. (western redcedar). Pollen records from the central and southern ITR suggest these species have only been present inland for <3 Ky (Mehringer 1996, Rosenburg *et al.* 2003, Chase*et al.* 2008, Gavin *et al.* 2009), and represent the primary data point for the RD hypothesis. Suitable inland habitat for western hemlock is not completely occupied, suggesting the species range is still expanding (Gavin & Hu 2006). Similarly, Rosenburg et al. (2003) found no record of this species' pollen in southeastern BC prior to ~ 3500 ya. Most pollen records concur that western hemlock pre-dates evidence of western redcedar (Mehringer 1996, Whitlock 1992). This coincides with microsatellite molecular evidence for western redcedar samples across the disjunction (O'Connell *et al.* 2008), which supports one southern coastal refugium throughout the Pleistocene, with no evidence for ancient, disjunct inland refugia, nor for northern coastal refugia (e.g., the Haida Gwaii archipelago). O'Connell et al. (2008) further suggested that, given the lack of hierarchical structure in these three clusters, the divergence between them has been recent and rapid, which is congruent with post-glacial recolonization of the northern coast and ITR (O'Connell *et al.* 2008). While this inference was based on microsatellite loci, recent advances with reduced representation sequencing (Peterson *et al.* 2012, Andrews *et al.* 2016) provide enhanced power to test phylogeographic hypotheses regarding formation of disjunct populations (Carstens *et al.*2012, Garrick *et al.* 2015). Indeed, a recent study used clustering analyses for population assignment based on a genome-wide panel of SNPs in western redcedar to support the presence of an ITR refugium (Fernandez et al. 2021). However, they did not attempt to estimate population parameters, such as the timing of demographic events through model comparisons, which would enable the demographic hypotheses to be evaluated.

Though much of the available evidence supports the recent, post-glacial colonization of the ITR by these dominant tree species, a number of phylogeographic investigations have been conducted to evaluate the impact of the disjunction on other species in the PNW temperate rainforest (e.g., Soltis *et al.* 1997, Brunsfeld *et al.*2001). To date, eleven other species complexes with disjunct ranges in the PNW have been investigated in a phylogeographic framework including five amphibians (Nielson *et al* . 2001, Carstens *et al.*2004, Wilke & Duncan 2004, Steele *et al.* 2005, Metzger *et al.* 2015), one mammal (Carstens *et al.* 2005), two plants (Brunsfeld *et al.* 2006; Carstens *et al.* 2013, Ruffley*et al.* 2018), three mollusks (Smith *et al.* 2017, Smith*et al.* 2019, Rankin *et al.* 2019), and an arthropod (Espíndola *et al.* 2016). These species span the tree of life and, based on analyses of their genetic variation, they also span the possible phylogeographic histories for the PNW temperate rainforest. Some species, such as the tailed frogs (*Ascaphus* ; Neilson et al. 2001), salamanders (*Dicamptodon, Plethodon* ; Carstens *et al.* 2004; Steele et al. 2005), and jumping slugs (*Hemphilia* ; Rankin et al., 2019) show clear evidence of an ancient divergence between the ITR and coastal populations, indicating pre-Pleistocene divergence. Conversely, other species, such as *Salix melanopsis*(Carstens *et al.* 2013), *Microtus richardsoni* (Carstens*et al.* 2005), and taildropper slugs (*Prophysaon andersoni* ; Smith 2018) show evidence of post-glacial recolonization of the inland from the coast. Other phylogeographic models, such as pre-Pleistocene divergence with migration, have also been supported with genomic evidence (*Alnus rubra;* Ruffley *et al.* 2018). These results suggest that some ITR endemics might have been present before the ecosystem dominant species were established if these ecosystem dominants colonized the ITR more recently, as supported by pollen records and early molecular studies. Inferring the phylogeographic history of the ecosystem dominant species that establish the boundaries of the PNW temperate rainforest will provide a critical insight for the availability of suitable habitat for refugial populations in the ITR, and will be a central contribution towards the understanding of the biological history of the area.

The hypothesis that the ITR has an ancient (pre-Pleistocene) divergence from the coastal rainforest and has persisted throughout glacial cycles in refugia in the interior Northwest is compelling because its presence

would support the habitat requirements of other species that show evidence of ancient vicariance. Additionally, paleontologists have questioned the plausibility of the old-growth ITR becoming so established in less than 3500 years (Mehringer 1996). However, the persistence of the ITR throughout the Pleistocene has received no support from the pollen record (e.g., Mehringer 1996, Chase *et al.* 2008, Gavin *et al.* 2009). Whether or not the ITR persisted throughout the Pleistocene also has other implications for how the PNW disjunct community as a whole has adapted to the dramatic climatic changes, either in concert or individualistically (Davis 1981, Habeck 1987, Sullivan *et al.* 2000, Flessa and Jackson 2005). Common insight from paleoecology suggests that modern communities of PNW forest have assembled over a long history of individual responses to climate change (David 1981, Flessa and Jackson 2005), and the hypothesis of a recently assembled, rapidly diverse ITR poses a challenge to this insight.

This question is not novel to the PNW temperate rainforest, as the idea of the species responding individualistically, rather than in-concert, in response to climatic changes is an ecological notion dating back to the early 20$^{\text{th}}$ century (Gleason 1926), and has been demonstrated empirically (Burbrink *et al.* 2016). However, Kirkwood (1922), one of the first to characterize the ecology of species in the northern Rocky Mountains in general, emphasized how the understanding of the ITR would be dramatically improved when "the individualities of the constituent species were understood". Alternatively, there is evidence in other communities that species do, in fact, respond to climatic changes in concert (Chen *et al.*2014, Gehera *et al.* 2017). For plant communities specifically, the idea of community-wide concerted response to climatic change can be traced back to early 20$^{\text{th}}$ century plant ecologist Clements (1918), and his idea that communities are "superorganisms" whose interactions are interwoven and dependent on one another. Regardless of the organism or ecosystem, researchers have long been fascinated with the question of whether or not species in the same environment respond asynchronously or synchronously to climatic changes (Sullivan *et al.* 2000, Carstens *et al.* 2005, Hickerson*et al.* 2006).

In this study, we test predictions about the phylogeographic history of these two species, specifically with respect to whether or not they harbor cryptic diversity across the disjunction; that is, show evidence of pre-Pleistocene divergence and no subsequent migration. These predictions serve as a test of the predictive framework that was originally developed by Espíndola *et al.* (2016) and recently updated with life history traits by Sullivan *et al.* (2019). We then validate these predictions, and ultimately test whether the ITR persisted throughout the Pleistocene (Brunsfeld *et al* . 2001) by generating genomic data for individuals from these species throughout their ranges. For this, we rely on coalescent simulations, the joint site frequency spectrum, and machine learning inference procedures to develop and test our phylogeographic hypotheses. We also validate the power of predictive phylogeography in detecting the presence and absence of cryptic diversity. Finally, we evaluate the potential role of genomic data in uncovering the history of the past and explore how our inferences can be influenced by various datatypes and perspectives in genomics and paleontology.

Materials and Methods

Field Sampling and Sequencing

Field collections of plant material were made throughout the coastal and inland PNW temperate rainforest for western redcedar, *Thuja plicata,* and western hemlock, *Tsuga heterophylla,* between April and June of 2016 and 2018. Fresh tissue for specimens was dried and stored in silica gel. Voucher specimens of collections were preserved in the Stillinger herbarium and can be located on the Consortium of Pacific Northwest Herbaria data portal (http://pnwherbaria.org). Leaf tissue from 137 *Thuja plicata* individuals (Fig. 1, Table S1) and 48 *Tsuga heterophylla* (Fig. 1, Table S2) individuals were extracted using a modified CTAB protocol (Doyle & Doyle 1987), purified using Sera-Mag SpeedBeads (Thermo Fisher Scientific; Rohland & Reich 2012), and their DNA quantified using a Qubit 2.0 Fluorometer (Life Technologies).

Three double-digest restriction site associated DNA sequencing (ddRADseq) libraries (Peterson *et al.* 2012) were prepared. Two of these were for *Thuja plicata* , assigning half of the samples to one or the other, and one for *Tsuga heterophylla* . For the*Thuja plicata* libraries, the restriction enzymes used were*EcoRI* and

SbfI , while they were *SbfI* and*MspI* (New England Biolabs, USA) for *Tsugaheterophylla* . All libraries were size selected using a size window of 200-500 bp using a BluePippin (Sage Science). All digestion, ligation and PCR products were purified using Agencourt AMPure XP purification system (Beckman Coulter). For *Thuja plicata* , sequences were generated as 50 bp single end reads using an Illumina HiSeq 2500 at the Berkeley sequencing facility. For *Tsuga heterophylla* , sequences were generated as 150 bp paired-end reads using an Illumina HiSeq 4000 at The Ohio State University Wexner Medical Center. Raw sequences were processed using Ipyrad (Eaton 2014, Eaton & Overcast 2020) with a minimum coverage of 10 and clustering threshold of 0.80. Ipyrad includes Vsearch (Rognes *et al.* 2016) and Muscle (Edgar 2004) for sequence clustering. Though we had overlapping reads for *Tsuga heterophylla* , we opted to not merge them and only used single-end reads. Complete assembly procedures were performed and documented in Jupyter notebooks (https://github.com/ruffleymr/ThujaTsugaAnalyses)

Predictive Phylogeography

Prior to analyzing genomic data, we made predictions about whether or not *Thuja plicata* and *Tsuga heterophylla* are expected to harbor cryptic diversity using the random forest (RF) classifier developed for this system by Espíndola *et al.* (2016) and Sullivan*et al.* (2019). For the predictor variables, we gathered occurrence data previously used for predictive phylogeography of species in the PNW (Espíndola *et al.* 2016, Sullivan *et al.* 2019) and occurrence data from recently investigated species (Smith *et al.* 2017, Smith *et al.* 2018, Ruffley *et al.* 2018). These occurrence data are a combination of GBiF records and field collections and we used them to extract bioclimatic variables from WOLRDCLIM v2 (Fick & Hijmans, 2017). Along with these bioclimatic variables, taxonomic rank and discrete trait variables, such as life stage at dispersal, outcrosser or selfer, dispersal mechanism, and trophic level (Table S3, Sullivan *et al.*2019), were used as the predictor variables in the RF classifier.

The predicted trait (response variable) was harboring or not cryptic diversity ("cryptic" *vs.* "non-cryptic"). To date, twelve species complexes with disjunct ranges in the PNW have been investigated in a phylogeographic framework (Avise et al. 1987); *Ascaphus truei / A. montanus* (Nielson *et al* . 2001, Metzger *et al.* 2015),*Plethodon idahoensis / P. vandykei* (Carstens *et al.* 2004),*Prohphysaon coeruleum* (Wilke & Duncan 2004), *Microtus richardsoni* (Carstens *et al.* 2005), *Dicamptodon aterrimus*and *complex, and D. tenebrosus* (Steele *et al.* 2005),*Salix melanopsis* (Brunsfeld *et al.* 2006; Carstens *et al.* 2013), *Conaphe armata* (Espíndola *et al.* 2016),*Haplotrema vancouverense (* Smith *et al.* 2017*), Alnus rubra (* Ruffley *et al.* 2018*)* , *Prophysaon dubium/andersoni (* Smith *et al.* 2019*), Hemphillia sp.*complex (Rankin *et al.* 2019). Of these, eight species were classified as non-cryptic, according to the respective study, meaning the species does not harbor a deep divergence between populations and has also not experienced significant gene flow between population (Table S3). The remaining 6 species/complexes were classified as cryptic because the coastal and inland populations are deeply diverged and in some cases are described as different species.

We constructed four different RF classifiers using different combinations of the predictor variables we had available: 1) bioclimatic variables only, 2) bioclimatic variables and taxonomy, 3) bioclimatic variables and life history traits, and 4) bioclimatic variables, taxonomy, and life history traits. In all of these, we are predicting the probability of a species being cryptic. We reported the average out-of-the-bag error rates for these classifiers, which is the proportion of simulations that were misclassified out of all the simulations left out of the construction of the classifier, averaged across classes.

With each of these classifiers, we predicted the presence or absence of cryptic diversity for *Thuja plicata* and *Tsuga heterophylla* , separately. To do this, we gathered occurrence records for the species in question, *Thuja plicata* (791; 569 GBIF records and 222 field collections (Table S4) and *Tsuga heterophylla* (468; 346 GBIF records and 111 field collections (Table S5). We excluded all occurrence records from GBIF that fell outside of the PNW temperate rainforest (35° to 65° latitude, -160deg to -100deg longitude). We used these locality coordinates to extract values from 19 bioclimatic variables from WOLRDCLIM v2 on 5 Feb 2019 (Fick & Hijmans, 2017) at a resolution of 30 arc-secs (~1 km^2). We also assembled trait data to coincide with the trait data collected for PNW taxa for predictive phylogeography as in (Sullivan *et al.* 2019). Using these data, we use the four classifiers and followed the procedure of Sullivan et al. (2019) to predict the probability

of each species being cryptic. We ultimately aimed to validate these predictions using phylogeographic model testing describe below. After validation was successful, we included the newly classified species data gathered in this study to assess whether the classifier improved in overall accuracy with the addition of two plant species.

Population Structure

To assess population structure in each species, we explored the genome-wide SNP data using STRUCTURE v2.3.4 (Pritchard *et al.*2000). We ran STRUCTURE for K values 1 to 10 with 5 replicates per K , where each replicate is a different sample of unlinked SNPs, subsampled from the same linked SNP dataset. We ran STRUCTURE for 500,000 generations and discarded the first 10% as burn-in. The data were modeled assuming admixture and correlated allele frequencies between populations (Falush *et al.* 2003), while all other parameters were kept as their default. For the sake of comparison to results reported by Fernandez et al. (2021), Structure Harvester (Earl & vonHoldt 2012) was then used to assess K using the Evanno method (Evanno *et al.* 2005). We acknowledge that many empirical studies have applied Evanno's K and interpreted the results as a measure of model fit, but the popularity of a given methods should not be mistaken for its appropriateness. Evanno's K lacks properties that a parameter in evolutionary genetics should ideally possess, for example it is extremely susceptible to uneven sampling (Puechmaille 2016) and reproducible inference is challenging (Novembre 2016). For these reasons, we follow Pritchard *et al* . (2000) in treating STRUCTURE as a tool for data exploration rather than phylogeographic inference. Our inferences are derived from the results of the model-based analyses described below.

Joint Site Frequency Spectra

In the remainder of our analyses, we study our data using Joint Site Frequency Spectra (jSFS) because it summarizes much of the genomic data into one statistic that can be used for inference (Gutenkunst *et al.* 2009, Xu & Hickerson 2015). The jSFS is essentially the overlap in the distribution in frequency of alleles between two populations and the pattern of overlap can tell us a lot about demographic processes in the past, both analytically (Gutenkunst *et al.* 2009, Excoffier*et al.* 2013) and visually (Fig. 2). More specifically, a single SFS represents the distribution of the number of sites that are present at each of the N allele frequencies in the population, where$N$ is equal to the number of total chromosomes present in the population. For a diploid organism, this is twice the number of individuals. A joint SFS is then the combination of two SFS, one for each population, as a matrix that is (N_{pop1} + 1) by (N_{pop2} + 1) cells. Each row is one of the allele frequencies in the first population, beginning with 0 and then ranging from $\frac{1}{N_{pop1}}$ to N_{pop1} and each column is the allele frequencies in the second population, again beginning with 0 and ranging from $\frac{1}{N_{pop2}}$ toN_{pop2} . Each cell then indicates the number of sites at that corresponding allele frequency in both populations. If the entire jSFS is standardized by the total number of sites, each cell indicates the proportion of sites at the corresponding population allele frequencies. The first row and column correspond to the sites that are at given frequencies in one population while not present at all in the other population, referred to hereafter as the "0" rows and columns. Again, these indicate the variants present in one population and not the other, thus the cell at row "0" and column "0" indicates the sites that do not vary in either population. With SNP data and for demographic model selection, this cell is not typically considered because it is only relevant for scaling the proportion of invariant sites for parameter estimates. Thus, when estimating demographic parameters from these models, the monomorphic cell along with linked SNPs is needed to inform the composite likelihood of the models (Excoffier *et al.*2013).

There is a trade-off between the number of chromosomes that can be included from each population and the number of unlinked SNPs included in the jSFS because the jSFS cannot accommodate missing data. The missing data are generally due to random missing data and allelic dropout from reduced representation sequencing (Andrews *et al.*2016), where loci are not represented across all or even a majority of individuals in the population. Thus, the more samples per population included, the fewer SNPs there are to sample from to construct the jSFS. For this reason, we down-sampled the number of SNPs and alleles (chromosomes in the population) to construct three different jSFS datasets for each species using custom python scripts (github.com/isaacovercast/easySFS). We enforced a different number of alleles to be included per population,

which resulted in a different number of unlinked SNPs being sampled in each dataset (Table 1). These datasets thus represent a spectrum of genomic information ranging from more individuals in the population but fewer SNPs to fewer individuals represented from the populations with many more SNPs included. We used unlinked SNPs for model selection (see below) to satisfy the assumption that SNPs are independent. We subsampled 100 different observed jSFS for each of the sample sizes for each of the species (600 observed jSFS in total) and masked monomorphic sites in all jSFS. For parameter estimation using the jSFS, we use the full SNP dataset, meaning we included linked SNPs in the construction of the jSFS. We also considered the monomorphic cell in the jSFS when estimating parameters because this cell provides information important to scale the invariant sites in the genome. To calculate the monomorphic cell, we measured the ratio of monomorphic sites and polymorphic sites in our entire datasets for each species and then used those ratios, multiplied by the total number of biallelic SNPs used in the empirical jSFS.

Demographic Modeling

One of the most important recent advances in phylogeography is the development of model selection as a framework for inference of evolutionary processes (e.g., Carstens et al., 2013). After exploring our data using STRUCTURE, we constructed eleven demographic scenarios (Fig. 2) to test using a machine-learning model-selection framework (Smith & Carstens 2020). For each focal species, these alternative demographic hypotheses include divergence between the coastal and inland populations that occurred either before or after the Pleistocene glaciations. The pre-Pleistocene divergence scenarios (models B-G in Fig. 2) model the populations diverging at the time of the xerification of the Columbia Basin (Waring & Franklin 1979), which followed the uplift in the Cascades Mountains (Priest 1990). In the recent dispersal models (models H-K in Fig.2), post-Pleistocene divergence between the populations posits the ITR populations arising from the coastal populations only via dispersal of coastal migrants; these models imply a recent time of divergence, as the coastal migrants could only have recolonized the inland region after the last glacial retreat, ~ 10 kya (Waitt and Thorson 1983). The varying migration scenarios include divergence with migration, where migration eventually ends between the coast and ITR populations a substantial time after divergence. In divergence with secondary contact models, migration begins again between the coast and ITR populations, at the very earliest, after the retreat of the Cordilleran ice sheet, ~ 10 kya (Waitt and Thorson 1983). Thus, these ancient vicariance with secondary contact models (models C & G) encompass both the older "ancient vicariance" and "recent dispersal" scenarios of Brunsfeld *et al.* (2001). The bottleneck events that are modeled are those that hypothetically occurred in the populations at the onset and for the duration of the Pleistocene and the subsequent population expansion events occur after the retreat of the glaciers, more likely as recent as 3500 ya (Whitlock 1992, Mehringer 1996).

Before using our data to assess these models, we first assessed our statistical power using simulations. Specifically, we simulated genomic data similar to the empirical genomic data we generated for *Thuja plicata* and *Tsuga heterophylla* under each of the eleven models. Specifically, we used the R package delimitR (Smith & Carstens 2020) which relies on the multi-dimensional SFS (mSFS) and the machine learning algorithm abc-randomForests (Pudlo et al. 2015) for model selection. For this, we simulated jSFS under eleven demographic scenarios we deem plausible for both species (Fig. 2) using fastsimcoal2 (Excoffier 2011, Excoffier *et al.* 2013). We generated 10,000 simulated jSFS for each model. We then converted our jSFS into mSFS by flattening the matrix and binning the cells into a coarser representation of itself. In delimitR (Smith *et al.* 2020), we then used the mSFS of the simulated datasets as the predictor variables to train a RF (Breiman 2001, Liaw & Weirner 2002, Pudlo *et al.*2018) classifier to distinguish among the eleven demographic models. This allowed us to assess the limits of the inference we could make with respect to phylogeographic model selection and construct a confusion matrix to summarize this differentiability among models. It also generated the classifier used in model selection for our empirical datasets.

Following the constructing of a demographic model RF classifier, we assessed the demographic models (Fig. 2) for the empirical datasets for each of *Thuja plicata* and *Tsuga heterophylla* . These classifiers simultaneously self-cross-validate by testing the accuracy of the decision trees being constructed using out-of-bag error rates. For this, data that were not used to construct specific decision trees were then classified using those trees.

Thus, the data being tested are not included in the construction of the decision trees classifying it. This results in overall error rates for the classifier, as well as specific model misclassification rates. This is an error rate specific to the classifier and represents how often a model class is incorrectly identified, and as which model.

We constructed six different classifiers to mimic the six empirical jSFS, with differing coastal and inland sample sizes and unlinked SNPs (Table 1). We then used the appropriate classifier to make predictions for the 100 corresponding subsampled jSFS. We summarized the support for each model and each dataset by the number of votes for each model. We then estimated the posterior probability for the best model.

Parameter Estimation

Once the best model for each species was identified, we used Fastsimcoal2 to estimate its demographic parameters and their 95% confidence intervals using the full, linked SNP datasets for each species and the monomorphic cell of the jSFS. We also estimated an additional parameter not included in the prior modeling: the mutation rate (μ) in substitutions/site/million years. Fastsimcoal2 uses a modified expectation maximization algorithm, known as a conditional expectation maximization (ECM; Brent 1974, Meng and Rubin 1993) for maximum-likelihood optimization, which can get trapped in local optima of the likelihood surface. Therefore, we performed 100 independent parameter optimizations with different initial values, 100,000 simulations to estimate the expected folded jSFS, and 40 conditional EM cycles per optimization. Following the first optimization, we identified the global maximum likelihood and parameter estimates and performed an additional 100 independent optimizations using these maximum likelihood parameter estimates as the starting values.

To estimate confidence intervals, we simulated 100 parametric bootstrap replicates using the ML parameter estimates from the final optimizations of the empirical datasets. We then re-optimized parameters of the simulated datasets, initiating the parameters at the maximum-likelihood estimates from the original optimization. We used these parameter estimates to generated 95% high density confidence intervals for all parameters (Kruschke 2011).

All computational analyses were done using servers at the IBEST Computational Resources Core at the University of Idaho.

Results

Sequencing & jSFS

Following assembly of the ddRADseq data, we recovered 124,484 loci (214,183 SNPs) for *Thuja plicata* , and 142,804 loci (893,487 SNPs) for *Tsuga heterophylla* , all of which were shared across a minimum of four individuals per species. To construct the jSFS for these species, the data were downsampled such that each SNP was represented in all individuals included in the jSFS (Table 1). In using the jSFS to make our inference about demographic histories, we excluded a considerable amount of sequence data, although the models are nevertheless distinguishable (Table 1).

Predictive Phylogeography

Before assessing whether these species harbor cryptic diversity, we made phylogeographic predictions of cryptic and non-cryptic for both species following the procedure introduced by Espindola et al (2016) using Random Forest with bioclimatic variables associated with sample localities and taxonomic ranks. Following Sullivan et al. (2019), we also included trait values along with the bioclimatic and taxonomic variables. The error rates we recovered were congruent with those found by Sullivan et al. (2019) and thus these classifiers were used to make predictions about *Thuja plicata* and *Tsuga heterophylla* . Each classifier predicted neither species to harbor cryptic diversity (Table 2), with the only variation in the prediction being the less accurate classifier (bioclimatic data only).

Population Structure

We explored the population structure for both species using STRUCTURE (Pritchard *et al.* 2000) for possible K of 1 through 10. For *Thuja plicata* , we found three clusters (Fig. 3, Fig. S1). While two of the clusters are geographically restricted to the coast or the inland, the third has no clear geographic structure. At $K = 2$ (Fig. 3), we recover the geographically structured pattern observed in the first two clusters of $K = 3$ (albeit some coastal samples were sometimes present in the inland cluster).

For *Tsuga heterophylla,* we selected $K = 2$ as the optimal$K$ value (Fig. S2) We do not see a geographic association between the coastal and inland samples with the two clusters (Fig. 3). However, when we investigate $K = 3$, we do observe a strong geographic structure, with one cluster mostly restricted to the inland and the other present along the coast. Because the justification of delta$K$ is based on simulations with no hierarchical population structure (which is almost always present in nature), interpretation of multiple clustering scenarios is critical.

Demographic Modelling

The STRUCTURE results provided some guidance for deciding how many populations to model in the demographic models investigated. Since the primary goal of this study is to make inferences regarding the evolutionary history of the ecosystem dominants, we decided to model two populations on the basis of geography (i.e., samples from either the inland or coastal forests), which largely corresponds to the division between the two coastal and one inland population in the $K = 3$ STRUCTURE analysis. We combine the clusters that contain coastal samples because these are not structured in a geographically meaningful manner (Fig. 3), so organizing them by cluster would combine genetically similar populations that aren't necessarily geographically close.

We developed eleven demographic models (Fig. 2) to assess the phylogeographic history of each species. In these models, we considered both ancient and recent divergence events, and various migration scenarios, including divergence with and without migration and secondary contact. We also model possible bottleneck events associated with the onset of the Pleistocene (˜2.5 Mya). We modeled population expansion to be associated with population regrowth after glacial retreat ˜10 Kya (Fig. 2). We used fastsimcoal2 to simulate DNA sequence data, setting the number of loci and variable sites to match the empirical data, and then summarize that data using jSFS. We simulated 100 datasets for three different dataset sizes and for each species. No missing data can be included in the calculation of the jSFS, therefore for a particular locus to be included, it must be present in all individuals. Thus, there is a trade-off between numbers of individuals and SNPs considered in the analysis.

The analyses of model identifiability resulted in low error rates for classifying each of the eleven models (Table 1, Fig. S3, S4) Most models were classified correctly most of the time, with all of them having a classification accuracy above 0.72 (Fig. S3, S4), except for the models with a recent divergence event between coastal and inland populations. We therefore collapsed these models, which all varied in the presence/absence of migration and bottleneck and expansion events, into a single recent dispersal model (Fig. 4, Fig. S5). This decreased the overall error rate dramatically (Table 1); the identifiability of the recent dispersal class of models increased to 0.90.

The first classifier, with all eleven models, was used to make predictions using the observed jSFS for each species (Fig. 5). For each dataset size, we used 100 different jSFS that were constructed by subsampling unlinked SNPs randomly. For *Thuja plicata* , all datasets had the highest prediction probability for the same model: ancient divergence between coastal and inland populations, followed by a bottleneck in both populations, and then population expansion contemporaneous to secondary contact between populations due to migration ("AV + sc + bot/exp", Fig. 5; model G in Fig. 2). On average, each *Thuja plicata* dataset received 552/1000 votes for that model and had an average posterior probability of 0.72 (Fig. 5). Meaning some datasets, or certain subsampled jSFS, had higher and/or lower prediction probabilities for this model.

The results were different for *Tsuga heterophylla* in that each dataset did not have the same prediction probability. Those with most SNPs supported models similar to *Thuja plicata* ("AV + sc + bot/exp", Fig. 5; model G in Fig. 2). On average, this model received 564/1000 votes in the classifier for each observed

jSFS and had an average estimated posterior probability of 0.83 (Fig. 5). With fewer SNPs included in the jSFS, but more samples represented in the population, the model that had the highest prediction probability was model C, or ancient divergence between coastal and inland populations, followed by secondary contact between populations due to migration ("AV + sc", Fig. 2), which is very similar to model G. The difference being that model G includes the population bottleneck and expansion that are not modeled in the former model C ("AV + sc", Fig. 2). On average, this model received 532 /1000 votes in the classifier for each observed jSFS and had an average estimated posterior probability of 0.78. We suspect this lower model support could be due to the fact that, in comparison to *Thuja plicata* , there were fewer SNPs to inform parameters associated with the additional process – population bottleneck and expansion – as well as fewer individuals sampled from each population.

Parameter Estimation

The parameter estimates obtained in our analyses were in units of coalescent generations and for *Thuja plicata and Tsuga heterophylla* generally fit with most of our expectations based on the history of the bioregion. For both species, the population sizes estimated for the coastal population are slightly larger than those of the ITR (Table 3), consistent with their current distributions. For both species, the median divergence time estimates between the coastal and inland rainforests were approximately 252,000 generations ago (Table 3). The timing of the population bottleneck event for both species was estimated to be between 50 and 90 Kya (Table 3). The time of the population expansion for *Thuja plicata* was ˜1,050 generations ago, whereas the time of population expansion for *Tsuga heterophylla* was nearly twice that value (2,020 generations ago). The magnitude of the *Thuja plicata* coastal bottleneck was apparently slightly larger than that of the inland bottleneck. The opposite was true for *Tsuga heterophylla* , where the bottleneck in the inland was more severe than that on the coast (Table 3). Not surprisingly, the populations with the most severe bottleneck also had the largest population expansion rates (Table 3). Note that this expansion rate was modeled backwards in time; a negative rate indicates the population is getting smaller towards the past, thus expanding forward in time. In both species, migration rates from the coast to the ITR were larger than migration rates from the ITR to the coast (Table 3).

In order to interpret our time estimates, which in coalescent analyses are expressed in units of generations, we need to consider the generation length of each species. The generation length is essentially the average amount of time between consecutive generations in a population. For western redcedar, estimates of trees reaching maturity typically range from 20-30 years (Turner 1985), however trees can reach maturity as early as 10 years in some open grown areas (Minore 1990). The same is true for western hemlock, where most estimates suggest maturity is reached between 25-30 years (Owen et al. 1984) but with trees reaching maturity earlier in some cases (Tesky 1992). To be conservative, we assumed a relatively short generation length of 10 yrs / generation and a longer generation length of 25 yrs / generation for both species. In doing this, we can convert our estimates of the timing of the events from generation into years while accounting for uncertainty in generation length amongst the populations (Table 4). From these generation lengths, we calculate median divergence times between inland and coastal populations between 2.5 - 6.3 Mya (Table 4), where each estimate has an associated 95% CI (Table 4). This indicates the divergence between inland and coastal populations, for both species, was*before* the onset of the Pleistocene. Similarly, given the overlap in 95% CI for the estimates for both species, we cannot reject that these species' disjunct populations diverged in concordance.

Discussion

History of the ITR

The demographic modeling results suggest that the ITR represents expansion from pre-Pleistocene relictual inland temperate rainforests and that this forest periodically received migrants from coastal populations, presumably via wind dispersal. Genomic evidence from both western redcedar and western hemlock supports this ancient divergence between the ITR and the coastal rainforest, with the evidence apparent in the statistical model selection as well as the observed allele frequencies. The genomic signature of refugia (an

anciently diverged population) is an abundance of rare alleles not shared with other populations. O'Connell et al. 2008, while they acknowledge some genetic differentiation between interior and coastal populations, suggested the divergence was shallow enough to support recent divergence with an absence of subsequent migration (e.g., model H, Fig. 2). All of the recent dispersal models were not supported by our analysis (Fig. 5). Moreover, in a recent genomic study of western redcedars (and mountain hemlock, T*suga mertensiana* , a sub-alpine congener of western hemlock examined here), Fernandez et al. (2021) found that ITR populations of western redcedar are a mix of two genomic clusters, one restricted to the southern portion of the ITR and the other more prominent in the northern portion of the ITR and shared with the central Cascades. Thus, their genomic data are the first to suggest persistence of this late successional dominant tree species in the inland region throughout the Pleistocene. Here, we have collected thousands of loci across individuals from both ITR and coastal populations for western redcedar (*Thuja plicata* , see also Fernandez et al. 2021), and the other late successional dominant, western hemlock (*Tsuga herterophylla*). With these data, we have been able to examine the jSFS and observe the high frequency of rare alleles present in the coastal and ITR populations separately, which indicates their ancient divergence. More importantly, we have modeled coalescent processes to account for coalescent stochasticity and explicitly conducted statistical tests of multiple plausible evolutionary (phylogeographic) models. This has ultimately allowed us not only to infer the presence of ITR refugia for *both* species throughout the Pleistocene, but also to date the divergence times between inland and coastal populations for both species to before the Pleistocene, assess demographic histories of the two species, and estimate migration patterns between coastal and inland populations of each of the two species. These inferences illustrate the analytical power of explicit statistical phylogeographic modeling.

This has implications on how the fossil pollen record informs our understanding of the history of the PNW (Whitlock 1992). Given the difficulty of identifying cedar pollen (Faegri & Iverson 1992), the pollen classification of *Thuja plicata* is generally limited to being a member of the Cupressaceae family, while the identification of *Tsuga heterophylla* pollen is more reliable. Of this *Thuja-Tsuga* pollen record, the ITR is inferred to have been present at the southern and central ranges $<$ 4-6.3 Kya (Mehringer 1996, Rosenberg *et al* . 2003, Chase *et al.* 2008, Herring & Gavin 2015), and not detected in high levels at the northern range extent until roughly 2-3 Kya (Gavin *et al.* 2009). Before this, the *Thuja-Tsuga* pollen record is not recognized in the area prior to 100 Kya, suggesting a rather recent population expansion of both species throughout the range. However, given the genomic evidence showing an abundance of rare alleles in the ITR populations and our coalescent analysis, we infer that populations of *Thuja plicata* and *Tsuga heterophylla* must have been in the ITR during the Pleistocene earlier than 6 Kya, though in such small populations that the pollen was not abundant enough for current detection.

Analytical Considerations

The limitations to the use of the jSFS to summarize genomic data should be recognized. As described above, when we summarized our data into a single jSFS, we downsampled data so that every SNP was included in each individual in the jSFS. We note that doing this required us to forfeit a considerable amount of data (Table 2). We performed a sensitivity analysis on the use of the jSFS by constructing three different dataset sizes, of 100 jSFS each, for each species, *Thuja plicata* and *Tsuga heterophylla* (Table 2), which resulted in 100 model predictions per species, per dataset (Fig. 5). The biggest discrepancy in the entire inference is within the *Tsuga heterophylla* prediction, where a different demographic history is supported by the jSFS that included more individuals and fewer SNPs, and the jSFS with the most SNPs and fewest individuals (Fig. 5). While the two models supported are generally consistent with our overall inference of pre-Pleistocene divergence followed by secondary contact, they differ in the presence of a population bottleneck during the Pleistocene and subsequent population expansion after the last glacial retreat. The difference in the model support for western hemlock across downsampling regimes could be due to the dataset with more SNPs being able to estimate the bottleneck and expansion parameters more effectively, and therefore showing strong support for that model. Conversely, the information in the dataset with fewer SNPs may have just been insufficient to estimate those parameters. For the purposes of our study more data is not necessary, but for future and more precise demographic parameter inference, this may be the case.

Our model-selection procedure supports, for both western redcedar and western hemlock, a pre-Pleistocene divergence event, followed by secondary gene flow between the populations. The approach used here for model selection using Random Forest and the jSFS (Smith et al. 2017, Smith & Carsten 2020) had yet to be tested using plant species or demographic models of this complexity. This is a likelihood-free approach that is based on simulating allelic data while accounting for coalescent stochasticity and demographic processes. Model selection, both in general and when implemented with machine-learning as employed here, is as accurate as the data are distinct in model space. This means that we should be able to assess if the empirical data are insufficient to distinguish among these models, which is indicated by the classifier's error rates. Indeed, our simulations indicated that the genomic signatures of the class of four recent dispersal models (Models H-K, Fig. 2) are not differentiable from each other (Fig S4). This is most likely due to the recent divergence time between the coast and inland populations resulting in low resolution of the distinct migration patterns those models are simulated under. However, all hypotheses positing post-Pleistocene dispersal, regardless of migration pattern, are well-differentiated from those positing a pre-Pleistocene dispersal, or specifically the persistence of disjunct coastal and inland populations through the Pleistocene (Fig S4). Additionally, when we pool the recent dispersal models into a general recent dispersal model, our data show that the error rates in all of our classifiers are extremely low, indicating high confidence in our classifier and high information content in data with respect to distinguishing among the final eight demographic scenarios we propose (Fig. 4). Again, the power of the machine learning classifier depends on how distinct the data are in model space and models can be very simple or very complex, which all influences the power of the classifier.

The approach employed here provides flexibility to the demographic model designs and simulation of data, as well as computational efficiency. As is true for all inferences based on model selection, it remains possible that some as yet unexamined model may be a better description of the true evolutionary history of these taxa, perhaps specifically those that model more than two populations and therefore more complex divergence and migration scenarios. Nevertheless, the approach to inference that we have adopted here (*i.e.,* developing models that are derived from extrinsic information such as pollen records and climate data, collecting genomic data, ranking models, assessing their identifiability, and making inferences) is an extremely powerful framework for phylogeographic research. In contrast to the approach that bases inference on methods designed for data exploration, our approach to inference utilizes existing data to formulate hypotheses that can then be supported (or not) as new data are collected and analyzed.

Predictive Comparative Phylogeography of PNW Rainforest Taxa

Comparative phylogeographic studies have addressed disjunct mesic forest taxa in the PNW of North America for decades. These have largely focused on the rather coarse phylogeographic hypotheses synthesized by Brunsfeld et al. (2001). Several taxa, primarily amphibians (*Ascaphus truei* /*A. montanus* , Nielson et al, 2001; *Plethodon vandykei* /*P. idahoensis* , Carstens et al, 2005; *Dicamptodon copei* /*D. aterrimus* , Carstens et al. 2005), show evidence of an ancient vicariance and persistence of populations in inland refugia throughout the Pleistocene, with no evidence of gene flow. Red alder (*Alnus rubra*) genomics (Ruffley et al. 2018) show evidence of ancient vicariance but with consistent migration between coastal and inland populations through the Pleistocene. Still other disjunct rainforest endemics, including water voles (*Microtus richardsoni* , Carstens et al 2005) and several disjunct taildropper slugs (*Prophysaon andersoni* , *P. dubium* , *P. coeruleum* , and*P. vanattae* ; Smith et al., 2018), show no evidence of Pleistocene persistence in ITR refugia, but rather have attained the disjunct distribution via post-Pleistocene dispersal.

Because of this collection of complex evolutionary histories for mesic forest disjunct taxa, and the conservation and evolutionary implications, Espindola et al. (2016) and Sullivan et al. (2019) have developed a framework for predicting the presence or absence of cryptic divergence in this system. Our data on the two mesic forest climax community dominant tree species represent a critical refinement to this predictive framework. This is especially true because western hemlock and western redcedar show strong evidence for a pre-Pleistocene divergence as well as evidence of post-Pleistocene gene flow through the non-zero estimation of migration rates between the populations (Table 3), a pattern thus far seen in only one other disjunct taxon (*Alnus rubra*); this should increase our ability to identify other inland rainforest taxa that may demonstrate a

similar evolutionary history, especially those with wind-dispersed pollen and/or seeds/spores.

Conclusion

Using genomic evidence and modern demographic inference procedures with machine learning, we have shown evidence for the persistence of ITR populations of both ecosystem dominant tree species *Thuja plicata*(western redcedar) and *Tsuga heterophylla* (western hemlock) throughout the Pleistocene. This is critical because both other mesic forest disjuncts and ITR endemics (e.g., endemic jumping slugs such as*Hemphilia camelus* , *H. skadei* , and *H. danielsi* ; Rankin et al. 2019) are rainforest-dependent. Our results, as well as the recent results for western redcedar of Fernandez et al. (2021), indicate that these late-successional dominant tree species persisted throughout the Pleistocene, providing habitat for rainforest dependents, including numerous understory plant species (Bjork 2010) that form the ecosystem. Further, the refugial populations in the ITR were likely small, as we show support here for Pleistocene-related population bottleneck events in both species. This evidence does coincide with the paleontological record, which suggests that the temperate rainforest did not dominate the PNW landscape until the last 5000 years, and only in the last 3500 years did the expansion of ITR begin. Coupled with the recent population expansion, we also show evidence for secondary contact at this time between the coastal and ITR populations for both species. This recent gene flow has likely muddled other genetic inferences made about western redcedar previously, which suggested that the ITR populations were a result of coastal recolonization. While our data indicate that coastal migrants contributed to the genetic architecture of the current ITR populations, our data provide strong evidence that Pleistocene refugia contributed to that architecture as well. This is supported by the high proportion of rare alleles observed in the ITR populations for *Tsuga heterophylla* and *Thuja plicata* , rare alleles that could only be the result of an ancient vicariant event with the coastal population.

Acknowledgements

Support for this work comes from National Science Foundation grants no. DEB-1457519 and DEB-1457726, and the Institute for Bioinformatics and Evolutionary Studies (IBEST) at the University of Idaho, which is supported by NIH NCRR 1P20RR016454-01, NIH NCRR 1P20RR016448-01, and NSF EPS-809935. We thank members of the Tank, Sullivan, and Carstens' labs for their insight over the years on this system and phylogeography.

Data Accessibility Statement

All raw Sequence reads for each library are present here: *pending DOI for dryad*

Jupyter notebooks corresponding to all analyses in this study, including sequence assembly from raw reads, for each species are available here: https://github.com/ruffleymr/ThujaTsugaAnalyses

Statement of Authorship: MR, DCT, and JS developed research concept; we, in addition to MLS, AE and BC, contributed to study design and implementation. MR, MLS, AE, DT, MS, NM all collected samples in the field and performed lab work for sequencing. MR performed analysis and wrote the manuscript. All authors contributed to critiques of the analysis and subsequent revisions of the text.

References

Andrews, K., Good, J., Miller, M., Gordon L., Hohenlohe P.A. (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* **17,** 81–92. https://doi.org/10.1038/nrg.2015.28

Avise JC, Arnold J, Ball RM *et al.* (1987) Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics* ,**18** , 489–522.

Bjork CR (2010). Distribution patterns of disjunct and endemic vascular plants in the interior wetbelt of northwest North America. *Botany* . **88** (4): 409-428. https://doi.org/10.1139/B10-030

Brent RP (1974) Algorithms for Minimization Without Derivatives.*IEEE Transactions on Automatic Control* , **19** , 632–633.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5-32

Brunsfeld SJ, Sullivan J, Soltis DE, Soltis PS (2001) Comparative phylogeography of north- western North America : a synthesis. In:*Integrating ecological and evolutionary processes in a spatial context* , pp. 319–339.

Brunsfeld SJ, Miller TA, Carstens BC (2007) Insights into the Biogeography of the Pacific Northwest of North America: Evidence from the Phylogeography of Salix melanopsis. *Systematic Biology* ,**32** , 129–139.

Burbrink FT, Chan YL, Myers EA, Ruane S, Smith BT, Hickerson MJ. (2016) Asynchronous demographic responses to Pleistocene climate change in eastern Nearctic vertebrates. Ecol Lett. 19:1457–67.

Carstens BC, Stevenson AL, Degenhardt JD, Sullivan J (2004) Testing nested phylogenetic and phylogeographic hypotheses in the Plethodon vandykei species group. *Systematic biology* , **53** , 781–792.

Cartens, B.C., Brunsfeld, S.J., Demboski, J.R., Good, J.M. and Sullivan, J. (2005), INVESTIGATING THE EVOLUTIONARY HISTORY OF THE PACIFIC NORTHWEST MESIC FOREST ECOSYSTEM: HYPOTHESIS TESTING WITHIN A COMPARATIVE PHYLOGEOGRAPHIC FRAMEWORK. Evolution, 59: 1639-1652. doi:10.1111/j.0014-3820.2005.tb01815.x

Carstens BC, Brennan RS, Chua V *et al.* (2013) Model selection as a tool for phylogeographic inference: An example from the willow Salix melanopsis. *Molecular Ecology* .

Chan YL, Schanzenbach D, Hickerson MJ. (2014) Detecting concerted demographic response across community assemblages using hierarchical approximate Bayesian computation. Mol Biol Evol. 31:2501–15.

Chase, M., C. Bleskie, I. R. Walker, D. G. Gavin, and F. S. Hu. 2008. Midge-inferred Holocene summer temperatures in Southeastern British Columbia, Canada. Palaeogeography Palaeoclimatology Palaeoecology 257:244-259.

Clements, F.E. (1916) Plant succession: an analysis of the development of vegetation. Carnegie institution of Washington, Washington, USA.

Davis M.B. (1981) Quaternary History and the Stability of Forest Communities. In: West D.C., Shugart H.H., Botkin D.B. (eds) Forest Succession. Springer Advanced Texts in Life Sciences. Springer, New York, NY

De La Torre, A. R., Li, Z., Van de Peer, Y., & Ingvarsson, P. K. (2017). Contrasting Rates of Molecular Evolution and Patterns of Selection among Gymnosperms and Flowering Plants. *Molecular biology and evolution* , *34* (6), 1363–1377. https://doi.org/10.1093/molbev/msx069

Dettman JR, Taylor, JW. 2004. Mutation and evolution of microsatellite loci in Neurospora. Genetics, 168:1231-1248.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* , 19, 11–15.

Earl, DA. and vonHoldt, **B** M. (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources vol. 4 (2) pp. 359-361 doi: 10.1007/s12686-011-9548-7

Eaton DAR (2014) PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* , 30, 1844–1849.

Eaton DAR, Overcast I (2020) ipyrad: Interactive assembly and analysis of RADseq datasets, *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btz966

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* , 26 (19): 2460-2461. doi:10.1093/bioinformatics/btq461

Espindola A, Ruffley M, Smith ML *et al.* (2016) Identifying cryptic diversity with predictive phylogeography. *Proceedings of the Royal Society B: Biological Sciences* , **283** , 20161529.

Evanno, G., Regnaut, S. and Goudet, J. (2005), Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology, 14: 2611-2620. doi:*10.1111/j.1365-294X.2005.02553.x*

Excoffier L, Foll M (2011) fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* , **27** , 1332–1334.

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013) Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics* , **9** .

Faegri, K., and J. Iversen. 1992. Textbook of Pollen Analysis. Hafner, New York.

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

Fernandez, MC, Hu, FS, Gavin, DG, et al. (2021) A tale of two conifers: Migration across a dispersal barrier outpaced regional expansion from refugia. *Journal of Biogeography* , **00** , 1–11. https://doi.org/10.1111/jbi.14209

Fick, S.E. and Hijmans, R.J. (2017), WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol, 37: 4302-4315. doi:10.1002/joc.5086

Flessa, K. W., S. T. Jackson, J. D. Aber, M. A. Arthur, P. R. Crane, D. H. Erwin, R. W. Graham, J. B. C. Jackson, S. M. Kidwell, C. G. Maples, C. H. Peterson, and O. J. Reichman. 2005. The Geological Record of Ecological Dynamics: Understanding the Biotic effects of Future Environmental Change. National Academies Press, Washington, D.C.

Garrick, R. C., Bonatelli, I. A. S., Hyseni, C. Morales, A., Pelletier, T. A., Perez, M. F., Carstens, B. C. (2015). The evolution of phylogeographic datasets. *Molecular Ecology* , **24** , 1164–1171.

Gavin, D. G., L. B. Brubaker, J. S. McLachlan, and W. W. Oswald. 2005. Correspondence of pollen assemblages with forest zones across steep environmental gradients, Olympic Peninsula, Washington, USA. Holocene 15:648-662.

Gavin, D. G., and F. S. Hu. 2006. Spatial variation of climatic and non-climatic controls on species distribution: the range limit of Tsuga heterophylla. Journal of Biogeography 33:1384-1396.

Gavin DG, Hu FS, Walker IR, Westover K (2009) The Northern Inland Temperate Rainforest of British Columbia: Old Forests with a Young History? *Northwest Science* , **83** , 70–78.

Gehara, M, Garda, AA, Werneck, FP, et al. (2017) Estimating synchronous demographic changes across populations using hABC and its application for a herpetological community from northeastern Brazil. *Mol Ecol* . 26: 4756– 4771. https://doi.org/10.1111/mec.14239

Gleason, H. A. (1926). The Individualistic Concept of the Plant Association. *Bulletin of the Torrey Botanical Club* , *53* (1), 7–26. doi: 10.2307/2479933

Gottesfeld AS, Swanson FJ, Gottesfeld LMJ. (1981) A Pleistocene low-elevation subalpine forest in the wester cascades, Oregon. Northwest Sci. 55: 157-167

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

Habeck, J. R. (1987). Present-Day Vegetation in the Northern Rocky Mountains. *Annals of the Missouri Botanical Garden* , *74* (4), 804–840. doi: 10.2307/2399451

Hebda RJ, Mathewes RW. (1984) Holocene history of cedar and native indian cultures of the North American Pacific Coast. Science 225: 711-713.

Herring EM, Gavin DG (2015) Climate and vegetation since the Last Interglacial (MIS 5e) in a putative glacial refugium, northern Idaho, USA. *Quaternary Science Reviews* . 117: 82-95. https://doi.org/10.1016/j.quascirev.2015.03.028.

Hickerson MJ, Stahl EA, Lessios HA. (2006) Test for simultaneous divergence using approximate Bayesian computation. Evolution. 60:2435–53.

Janes, JK, Miller, JM, Dupuis, JR, et al. (2017) The $K = 2$ conundrum. *Mol Ecol* . 26: 3594 – 3602. https://doi.org/10.1111/mec.14187

Kirkwood JE (1922). Forest distribution in the northern Rocky Mountains. Univ. Montana Studies, Bull. 247: 1-180.

Kruschke, J. K. 2011. *Doing Bayesian data analysis: a tutorial with R and BUGS.* Burlington, MA: Academic Press.

Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. *R news* ., 2/3, 18-22.

Meng X-L, Rubin DB (1993) Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika* , **80** , 267–278.

Mehringer, P. J., Jr. (1996). Columbia River Basin Ecosystems: Late Quaternary Environments. Contract Report. Interior Columbia Basin Ecosystem Management Project.

Metzger G, Espindola A, Waits LP, Sullivan J (2015) Genetic structure across broad spatial and temporal scales: Rocky mountain tailed frogs (Ascaphus montanus; Anura: Ascaphidae) in the Inland Temperate Rainforest. *Journal of Heredity* , **106** , 700–710.

Minore, Don. 1990. *Thuja plicata* Donn ex D. Don western redcedar. In: Burns, Russell M.; Honkala, Barbara H., technical coordinators. Silvics of North America. Volume 1. Conifers. Agric. Handb. 654. Washington, DC: U.S. Department of Agriculture, Forest Service: 590-600. [13419]

Nielson M, Lohman K, Sullivan J (2001) Phylogeography of the Tailed Frog (*Ascaphus truei*): Implications for the Biogeography of the Pacific Northwest. *Evolution* , **55** , 147–160.

Newmaster, S. G., R. J. Belland, A. Arsenault, and D. H. Vitt. 2003. Patterns of bryophyte diversity in humid coastal and inland cedar-hemlock forests of British Columbia. Environmental Reviews 11:S159-S185.

Novembre, J. 2016. Pritchard, Stephens, and Donnelly on Population Structure. Genetics 204, 391-393.

O'Connell, L.M., Ritland, K. & Thompson, S.L. (2008). Patterns of post-glacial colonization by western redcedar (*Thuja plicata* , Cupressaceae) as revealed by microsatellite markers. *Botany* , 86, 194–203.

Owens, John N.; Molder, Marje. 1984. The reproductive cycles of western

and mountain hemlock. Victoria, BC: Ministry of Forests, Information

Services Branch. 32 p. [19144]

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* , 7.

Priest GR (1990) Volcanic and Tectonic Evolution of the Cascade Volcanic Arc , Central Oregon. *Journal of Geophysical Research* ,**95** , 19583–19599.

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

Puechmaille, S. J. 2016. The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. Molecular Ecology Resources 16, 608-627.

Pudlo P, et al. (2015) Reliable ABC model choice via random forests. Bioinformatics 32(6):859–866.

Rankin AM, Wilke T, Lucid M, Leonard W, Espindola AE, Smith ML, Carstens BC, Sullivan J (2019) Complex interplay of ancient vicariance and recent patterns of geographical speciation in north-western North American temperate rainforests explains the phylogeny of jumping slugs (*Hemphillia spp.*), *Biological Journal of the Linnean Society* , Volume 127, Issue 4, Pages 876–889, https://doi.org/10.1093/biolinnean/blz040

Rognes T, Flouri T, Nichols B, Quince C, Mahe F (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* , 4, e2584.

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research* , 22, 939–946.

Rosenberg, S.M. Walker, I.R. & Mathewes, RW (2003). Postglacial spread of hemlock (Tsuga) and vegetation history in Mount Revelstoke National Park, British Columbia, Canada. Canadian Journal of Botany. 81.

Ruffley M, Smith ML, Espındola A, Carstens BC, Sullivan J, Tank DC. (2018) Combining allele frequency and tree-based approaches improves phylogeographic inference from natural history collections. Mol Ecol. 2018;27:1012–1024. https://doi.org/10.1111/mec.14491

[dataset] Ruffley M, Smith ML, Turck D. 2021. Thuja plicata and Tsuga heterophylla ddRADseq raw data. *Pending DOI.*

Smith ML, Ruffley M, Espındola A, Tank DC, Sullivan J, Carstens BC. (2017) Demographic model selection using random forests and the site frequency spectrum. Mol Ecol. 2017;26:4562–4573. https://doi.org/10.1111/mec.14223

Smith ML, Ruffley M, Rankin AM, Espındola A, Tank DC, Sullivan J, Carstens BC. (2018) Testing for the presence of cryptic diversity in tail-dropper slugs (*Prophysaon*) using molecular data, *Biological Journal of the Linnean Society* , Volume 124, Issue 3, July 2018, Pages 518–532, https://doi.org/10.1093/biolinnean/bly067

Smith, M.L. and Carstens, B.C. (2020), Process-based species delimitation leads to identification of more biologically relevant species*. Evolution, 74: 216-229. doi:10.1111/evo.13878

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

Sullivan J, Arellano E, Rogers DS (2000) Comparative Phylogeography of Mesoamerican Highland Rodents: Concerted versus Independent Response to Past Climatic Fluctuations. *The American Naturalist* ,**155** , 755–768.

Sullivan, J, Smith, ML, Espíndola, A, Ruffley M, Rankin A, Tank DC, Carstens BC. (2019) Integrating life history traits into predictive phylogeography. *Mol Ecol* . 2019; 28: 2062– 2073. https://doi.org/10.1111/mec.15029

Steele CA, Carstens BC, Storfer A, Sullivan J (2005) Testing hypotheses of speciation timing in Dicamptodon copei and Dicamptodon aterrimus (Caudata: Dicamptodontidae). *Molecular Phylogenetics and Evolution* , **36** , 90–100.

Tesky, Julie L. 1992. Tsuga heterophylla. In: Fire Effects Information System, [Online]. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory (Producer). Available: https://www.fs.fed.us/database/feis/plants/tree/tsuhet/all.html [2020, April 14].

Turner, David P. 1985. Successional relationships and a comparison of biological characteristics among six northwestern conifers. Bulletin of the Torrey Botanical Club. 112(4): 421-428. [16784]

Waitt, R. B., and R. M. Thorson. 1983. The Cordilleran ice sheet in Washington, Idaho, and Montana. Pp. 53-70 in S. C. Porter, ed. Late-Quaternary environments of the United States. Minneapolis: University of Minnesota Press.

Waring RH, Franklin JF (1979) Evergreen coniferous forests of the pacific northwest. *Science (New York, N.Y.)* , **204** , 1380–1386.

Whitlock, C. 1992. Vegetational and climatic history of the Pacific Northwest during the last 20,000 years: implications for understanding present day biodiversity. Northwest Environmental Journal 8:5-28.

Wilke, T & Duncan, N. (2004). Phylogeographical patterns in the American Pacific Northwest: Lessons from the arionid slug Prophysaon coeruleum. Molecular ecology. 13. 2303-15. 10.1111/j.1365-294X.2004.02234.x.

Xue AT, Hickerson MJ. (2015) The aggregate site frequency spectrum for comparative population genomic inference. Mol Ecol. 24:6223–40.

Figures

Hosted file

`image1.emf` available at https://authorea.com/users/436532/articles/538782-genomic-evidence-of-an-ancient-inland-temperate-rainforest

Figure 1. Localities sampled for western redcedar (*Thuja plicata)* and western hemlock (*Tsuga heterophylla*). Locality information for each collection can be found in Table S1 and S2.

Hosted file

`image2.emf` available at https://authorea.com/users/436532/articles/538782-genomic-evidence-of-an-ancient-inland-temperate-rainforest

Figure 2. (A-K) Summarized folded jSFS (43 by 43 cells) for 10,000 simulations under each associated demographic model. Scale indicates the proportion of loci in each cell, with 0.001 being the maximum, meaning if the proportion is higher than this value, the color is that same as the maximum. Dashed lines represent the times of all events that can occur in a given model, including divergence (*div*), bottleneck (*bot*), expansion (*exp*), and migration initiation (*mig*) events. Migrations arrows indicate asymmetrical migration between populations, b is the magnitude of a bottleneck and r is the population growth rate during expansion for a given population.

Hosted file

`image3.emf` available at https://authorea.com/users/436532/articles/538782-genomic-evidence-of-an-ancient-inland-temperate-rainforest

Figure 3. Left panels: STRUCTURE results for *Thuja plicata* (western redcedar) and *Tsuga heterophylla* (western hemlock), where each bar indicates an individual in the population and the color indicates the proportion of genetic variation associated to a particular cluster. Clusters indicated by K values in the top right corner. Coastal samples are denoted with a C in the label and inland samples with an I. Right panels: Sampling localities plotted according to the proportion of genomic variation attributed to each cluster, with clusters at $K = 2$ and $K = 3$.

Hosted file

`image4.emf` available at https://authorea.com/users/436532/articles/538782-genomic-evidence-of-an-ancient-inland-temperate-rainforest

Figure 4. Confusion matrix depicting the prediction accuracies and inaccuracies for eight demographic models using delimitR for model selection, which involves the simulated mSFS and 'abcrf'. Rows indicate the model the data were simulated under and columns indicate the model that was predicted, each cell then indicates the proportion of simulated data under the true model that is classified as the predicted model. Thus, the diagonal cells of the matrix depict the proportion of correct model classifications, as the predicted model aligns with the true model, whereas the off-diagonal cells depict the proportion of model simulations that are incorrectly classified, and specifically which model the simulations are incorrectly classified as.

Hosted file

`image5.emf` available at https://authorea.com/users/436532/articles/538782-genomic-evidence-of-an-ancient-inland-temperate-rainforest

Figure 5. A. Barplots represent the proportion of observed jSFS, at the corresponding average number of SNPs in the jSFS, that are classified as a given model, which is indicated by the color of the barplot. Solid barplots (left side) represent *Tsuga heterophylla*predictions and diagonal stripped barplots (right side) represent predictions for *Thuja plicata.* B. Table indicating the average number of model votes for the most selected model, "AV + sc + bot/exp", for both species, along with the average estimated posterior probability (PP) for the same model. C. Corresponding observed jSFS at each SNP count (10000, 3000, and 1000) for *Tsuga heterophylla*(top row) and *Thuja plicata* (bottom row).

Table 1. The average number of unlinked SNPs used in the 100 empirical datasets, with the corresponding number of samples from the coastal and inland populations, where each sample represents an allele for an individual, most often both alleles are included, but in some cases only one allele from an individual is included in the construction of the jSFS. The error rate for all models represents the average error rate for all model classifications for the classifier constructed with the corresponding data size. The error rate with RD collapsed corresponds to the overall error rate for the classifier when the four Recent Dispersal models are collapsed into a single model, RD.

Hosted file

`image6.emf` available at https://authorea.com/users/436532/articles/538782-genomic-evidence-of-an-ancient-inland-temperate-rainforest

Hosted file

`image7.emf` available at https://authorea.com/users/436532/articles/538782-genomic-evidence-of-an-ancient-inland-temperate-rainforest

Table 2. Phylogeographic predictions of cryptic and non-cryptic nature for*Thuja plicata* and *Tsuga heterophylla* using random forest with specified predictor variables. Error rate indicates the error rate of the RF classifier used to make the predictions. PP: prediction probability. The updated error rate is the error rate of the new classifier constructed with the new data from *Thuja plicata* and*Tsuga heterophylla* .

Table 3. Parameter estimates for *Thuja plicata* and*Tsuga heterophylla* for the model selected more often for the data, "AV + sc + bot/exp". The population sizes, N inland and N coast, are in units of the number of alleles in the population. All of the events, T_{div}, T_{bot} andT_{exp} , are in units of coalescent generations. The magnitude of the bottleneck, btnmag, indicates the instantaneous shrinkage of the population by that proportion. The growth rates indicate population size change, backward in time, as the number of alleles removed from the population per generation. Thus, a negative rate indicates population expansion forward in time. The migration rates indicate the proportion of alleles moving to the other population per generation. The mutation rate is in substitutions per site per generation.

Hosted file

`image8.emf` available at https://authorea.com/users/436532/articles/538782-genomic-evidence-of-an-ancient-inland-temperate-rainforest

Table 4. Divergence time estimates and time of population expansion and secondary contact estimates for both species. Estimates are in years, calculated from multiplying the divergence time in generations by an estimate generation length of 10 years and 25 years for both species.

	MaxL Estimate (Mya)	low 95% CI	high 95% CI	MaxL Estimate (Mya)	low 95% CI	high 95% CI
Thuja plicata	10 yrs/generation			25 yrs/generation		
T_{div}	2,528,140	2,313,410	2,950,090	6,320,350	5,783,525	7,375,225
$T_{exp/sc}$	10,430	10,100	10,950	26,075	25,250	27,375
Tsuga heterophylla						
T_{div}	2,522,850	2,238,900	2,959,670	6,307,125	5,597,250	7,399,175
$T_{exp/sc}$	20,210	15,920	22,900	50,525	39,800	57,250