Comparative Phylogeography and Community Ecology Reveal Evolutionary Processes That Contribute to Ecosystem Structure Through Time

A Dissertation
Presented in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
with a
Major in Bioinformatics and Computational Biology
in the
College of Graduate Studies
University of Idaho
by
Megan Rose Ruffley

Major Professors: David C. Tank, Ph.D. and Jack Sullivan, Ph.D.
Committee Members: Luke J. Harmon, Ph.D.; Bryan C. Carstens, Ph.D.
Department Administrator: David C. Tank, Ph.D.

May 2020

Authorization to Submit Dissertation

This dissertation of Megan R. Ruffley, submitted for the degree of Doctor of Philosophy with a Major in Bioinformatics and Computational Biology and titled "Comparative Phylogeography and Community Ecology Reveal Evolutionary Processes That Contribute to Ecosystem Structure Through Time," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ________________________________ Date: __________
David C. Tank, Ph.D.

Major Professor: ________________________________ Date: __________
Jack Sullivan, Ph.D.

Committee Members: ________________________________ Date: __________
Luke Harmon, Ph.D.

________________________________ Date: __________
Bryan C. Carstens, Ph.D.

Department
Administrator: ________________________________ Date: __________
David C. Tank, Ph.D.

Abstract

The research presented here lies at the intersection of phylogeography, phylogenetics, and community ecology and aims to understand how evolutionary processes have contributed to the diversity seen around the world. This diversity is investigated at the population, species, and community level in order to understand the various evolutionary processes that impact these levels. At the population and species level, phylogeographic inference is used to understand when and how populations diverge, maybe due to geographic barriers, and whether they experience gene-flow after said divergence. For this, we specifically investigate plant species in the Pacific Northwest temperate rainforest to understand how geologic and climatic history of the region has influenced the genetic structure of the populations throughout their range. While this is done at the population and species level, inferences across species can be aggregated to make inferences about the community and ecosystem as a whole. To more specifically understand the processes influencing community structure at the community level, we also model the process of many species assembling into a community at once, sometimes experiencing habitat pressure or interspecific competition for resources. We can then use these models to make inferences about the pressures influencing diversity patterns across communities. By investigating evolutionary and ecological processes at the population, species, and community level, we are able to interpret the effect of many different processes impacting diversity at various scales. With the myriad of threats facing all life on earth, being able to use what we know about the impacts of climatic, geographic, and historical processes on evolutionary trajectories will make for better predictions of future sustainability and ecosystem survival.

Acknowledgements

Support for this research comes from National Science Foundation (NSF) grants no. DEB-1457519, DEB-1457726, DEB-1208912, 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. Support also comes from the Genomics Resource Core through IBEST which is supported in part by NIH COBRE Phase III grant P30GM103324.

Dedication

To my Mother, Susan Ruffley, without whom this work would not exist.

Thank you for your love and support.

Table of Contents

Authorization to Submit Dissertation ii

Abstract iii

Acknowledgements iv

Dedication v

Table of Contents vi

List of Tables viii

List of Figures xii

Statement of Contribution xvi

Chapter 1: Combining Allele Frequency and Tree-Based Approaches Improves Phylogeographic Inference from Natural History Collections 1

- Abstract 1
- Introduction 2
- Methods and Materials 4
- Results 12
- Discussion 15
- Conclusions 18
- Literature Cited 20
- Tables 27
- Figures 42

Chapter 2: Identifying Models of Trait-Mediated Community Assembly Using Random Forests and Approximate Bayesian Computation 51

- Abstract 51
- Introduction 51
- Methods and Materials 54
- Results 60
- Discussion 62
- Conclusion 66
- Supplemental Information 66
- Literature Cited 71
- Tables 78
- Figures 96

Chapter 3: Genomic Evidence of an Ancient Inland Temperate Rainforest 102

- Abstract 102

Introduction 102

Methods and Materials 106

Results 111

Discussion 115

Conclusion 118

Literature Cited 120

Tables 127

Figures 131

Appendices 136

License Agreement for Chapter 1 136

License Agreement for Chapter 2 140

List of Tables

Table 1.1 *Alnus rubra* collections information. From the left: DNA extraction number, collector and collection number, latitude, longitude, locality description. All herbarium specimens are located in the Stillinger Herbarium at the University of Idaho and can be found online (http://pnwherbaria.org). 27

Table 1.2 Variables (mode fragment length, concentration of fragment length distribution, and year collected) examined for a relationship with the number of raw reads produced by a given sample. 28

Table 1.3 Grid search values for each of the four PHRAPL runs (applied to both datasets; 7 and 46 trees) for the collapse time and migration rate parameters. The first two runs are broad, while the final two runs narrow the grid search.. 29

Table 1.4 The average wAIC scores for the four core models evaluated in PHRAPL under both sets of trees, for all four runs (see Table S3 for grid search values)............................... 30

Table 1.5 The wAIC scores for all models averaged across fastsimcoal2 runs, along the selection frequency, within each downsampled dataset. 31

Table 1.6 Divergence times and their 95% Confidence Intervals estimated in fastsimcoal2 for the three highest ranked models, Ancient Vicariance (AV) with Migration, AV with Coastal Migration, and AV with Inland Migration, under two generation lengths, 6 and 8 years per generation... 32

Table 1.7. The average wAIC scores for all thirteen models evaluated in PHRAPL under 7 trees, for all four runs (see Table S3 for grid search values). 33

Table 1.8. The average wAIC scores for all thirteen models evaluated in PHRAPL under 46 trees, for all four runs (see Table S3 for grid search values). 34

Table 1.9. From the left; parameters for the Ancient Vicariance with Coastal Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. Migration is the migration probability for a lineage from the coastal population to migrate to the inland. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation. 35

Table 1.10. From the left; parameters for the Ancient Vicariance with Inland Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. Migration is the migration probability for a lineage from the inland population to migrate to the coast. Inland and

Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation. 36

Table 1.11. From the left; parameters for the Ancient Vicariance with Asymmetrical Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. The migration parameters are the migration probability for a lineage from one population to migrate to the other. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation. 37

Table 1.12. From the left; parameters for the Ancient Vicariance with Coastal Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. Migration is the migration probability for a lineage from the coastal population to migrate to the inland. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation. 38

Table 1.13. From the left; parameters for the Ancient Vicariance with Inland Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. Migration is the migration probability for a lineage from the inland population to migrate to the coast. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation. 39

Table 1.14. From the left; parameters for the Ancient Vicariance with Asymmetrical Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. The migration parameters are the migration probability for a lineage from one population to migrate to the other. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation. 40

Table 1.15. Power analysis results, showing the selection frequency for the four main models, and the average AICw across all 100 optimizations. 41

Table 2.1. Average error rates for model classification approaches in classifying each of the three community assembly models, as well as overall classification error. 78

Table 2.2. Community assembly model predictions from RF and model posterior probabilities from ABC for all local kipuka plant species and eight individual kipuka communities. All predictions were made with simulations using an OU model of trait evolution. 79

Table 2.3. Default parameter prior distributions used for simulating data in CAMI. 80

Table 2.4. Summary statistics calculated in CAMI and used for model selection in RF. 81

Table 2.5. The average proportion of misclassified simulations using the standard approach of phylogenetic dispersion metrics for all regional/local community sizes tested and for each model of community assembly. 82

Table 2.6. The average proportion of misclassified simulations using phenotypic dispersion metrics for all regional/local community sizes tested and for each model of community assembly. 83

Table 2.7. Average error rates, or proportion of incorrectly classified simulations, when classifying community assembly and trait evolution models using random forest for all sizes of the local community used. 84

Table 2.8. Average error rates, or proportion of incorrectly classified simulations, when classifying community assembly and trait evolution models using ABC for all sizes of the local community used. 85

Table 2.9. Results for empirical data when using dispersion metrics MPD and MNTD with phylogenetic and phenotypic information for the total kipuka plant community and the eight individual kipuka plant communities. The mean and standard deviation of the null distribution are included, as well as the p-value for the for the observed value's position in the null distribution. Significant p-values are indicated with **.observed value's position in the null distribution. Significant p-values are indicated with **. 86

Table 2.10. All species included in the regional and local community phylogeny for the kipukas od CRMO. Species that were present in the community but not in the Spermatophyta phylogeny were replaced in the phylogeny with a close relative, denoted in the table. The minimum, maximum, and range of vegetative height for each species is also denoted in cm. 87

Table 2.11. Supplement. Parameter estimates, log-likelihoods and AIC scores for BM and OU models of trait evolution. The BM model of trait evolution was optimized one time, with the resulting log-likelihood and AIC. For the OU model, we struggled to estimate the parameter values absolutely. However, we could estimate the quotient of $\sigma 2$and α and did so for 21 parameter combinations. Regardless of the $\sigma 2$and α though, OU was always a better fit model. 91

Table 2.12. Supplement. Parameter Presence/Absence matrix indicating which kipuka species is present on one of the eight specific kipukas investigated. The table contains all of the species that occur in the local kipuka community. 92

Table 2.13. Model probabilities for the total kipuka plant community and eight separate kipukas using RF and ABC. For this, the reference data was made up of data simulated under OU models of trait evolution and data simulated under BM models of trait evolution. 94

Table 2.14. Model probabilities for environmental filtering for the eight kipuka plant communities using RF and ABC, as well as the median tE estimates using ABC. 95

Table 3.1. Phylogeographic predictions of cryptic and non-cryptic 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 indication 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*. 127

Table 3.2. The average number of unlinked SNPs used in the 100 empirical data sets, with the corresponding number of samples from the coastal and inland populations, where each sample represents an allele for an individual, most often both allele areas 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. 128

Table 3.3. Parameter estimates for *Thuja plicata* and *Tsuga heterophylla* for the model selected most 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} and T_{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. 129

Table 3.4. Divergence time estimates and time of population expansion and secondary contact estimates for both species. Estimates are in years that were calculated from multiplying the divergence time in generations by an estimate generation length of 10 for both species. 130

List of Figures

Figure 1.1. Four major demographic models representing the hypothesized phylogeographic history of *Alnus rubra* in the PNW temperate rainforest. In total, 13 demographic models were designed and evaluated for both datasets. Names and abbreviations of all models are listed in the left panel. Parameters estimated include population size, Θ_C and Θ_I, migration rate, μ, divergence time, τ_{div}, and time of secondary contact, τ_S. The only additional parameter estimated but not included in the model design was mutation rate. 42

Figure 1.2. Number of raw reads recovered from ddRADseq experiments in relation to the mode fragment length found in the DNA extracts from 13 *Alnus rubra* individuals. 43

Figure 1.3. STRUCTURE results for K = 2 (a) and K = 3 (b) for 18 populations of *Alnus rubra*. Species distribution models for *A. rubra* in the PNW rainforest under current (c) and last glacial maximum (LGM; d) conditions. The stippled area in the bottom right panel shows the extent of the Cordilleran ice sheet at the LGM. 44

Figure 1.4 a) Model selection results between the four major demographic models. b) Averaged way values across runs in both fastsimcoal2 (FSC) and PRHAPL datasets. White: Recent Dispersal, dark grey: Ancient Vicariance (AV) with Migration, light grey: AV then Secondary Contact. Notes: AV not shown; average wAIC of 0. Because these are averaged wAIC values, they do not sum to one. 45

Figure 1.5. Model Selection results from fastsimcoal2 (FSC2) show three models with the highest wAIC densities out of all 13 demographic models evaluated across 400 independent optimizations of the likelihood for each model (20 subsampled datasets at a subsampling threshold of 20% (a) and 30% (b), each optimized 20 times). The three best models represent continuous migration after the ancient vicariance event with varying migration patterns. (c, d) Model adequacy results for the three best models in both FSC2 datasets. Note that in each dataset, the test statistic for all three models are very similar making it appear as if there is only one test statistic (black dashed line), however they are actually three overlapping values. In (d) the three test statistics are more variable, hence the appearance of a thicker line. All six evolutions of model adequacy were non-significant. 46

Figure 1.6. a) Distribution of DNA fragment lengths, with a mode fragment length of 7,853 bp, present in Sample 420. b) Distribution of DNA fragment lengths, with a mode fragment length of 2,058 bp, present in Sample 441. Results from a Fragment Analyzer (Advanced Analytical)... 47

Figure 1.7. A) Missing data quantified across all 49 samples and organized by relatedness using population assignment probabilities from STRUCTURE at K=3. B) Uniform distribution simulated in R using the *runif* function to generate 49 random variables from a uniform distribution with a maximum and minimum bound corresponding to the maximum and minimum missing data value observed across all individuals. 48

Figure 1.8. STRUCTURE results for K = 8 for 18 populations of *Alnus rubra*; no clear spatial genetic structure present. 49

Figure 1.9. STRUCTURE results for K = 4 (left) and K = 5 (right) for 18 populations of *Alnus rubra*; no clear spatial genetic structure present. 50

Figure 2.1. Outline of data simulation process. (1.1) Simulate the regional phylogeny. (1.2) Simulate trait evolution along the regional phylogeny. (1.3) Simulate the assembly of the local community by sampling species at random from the regional species pool and calculating the probability of persistence for each sampled species. These probabilities are calculated differently depending on the model of assembly being simulated, and if a species' probability of persistence is greater than a randomly generated probability, then that species survives in the local community. 96

Figure 2.2. Error rates, or proportion of incorrectly classified simulations, when classifying community assembly models compared to the size of the local community used. Four model identification approaches are summarized here. The first is the average error rate when using dispersion metrics (MPD and MNTD) from phylogenetic information (dotted). The second is the average error rate when using dispersion metrics from functional trait information (big dashed). The final two are model selection approaches employed in CAMI, ABC (gray), and RF (small dashed). 97

Figure 2.3. Estimation of tE and tC under their respective non-neutral models of community assembly, coupled with one of two models of trait evolution. In each graph, the individual boxplots represent the median values of either tE or tC from 100 independent attempts at parameter estimation, thus they are not posterior distributions, but rather a distribution of median parameter estimates. The x-axis denotes the true value of tE or tC simulated under. The light gray boxes represent datasets with regional/local community sizes of 200/100 and the dark gray boxes represent regional/local community sizes of 800/400. The dotted line in each plot represents a 1:1 correlation between estimated and true values of either tE or tC. A. Environmental filtering community assembly with a BM model of trait evolution. B. Competitive exclusion community assembly with a BM model of trait evolution. C. Environmental filtering community assembly with an OU model of trait evolution. D. Competitive exclusion community assembly with an OU model of trait evolution. 98

Figure 2.4. left) Regional phylogeny of species in the Craters of the Moon National Monument and Preserve, coupled with each species' maximum vegetative height in meters represented by the filled bar plots by each species. Species only present in the regional community have their trait bars colored white, while species that are also present in the local community have their trait bars colored black. The bars are truncated at 6 meters, as only the four trees in this study are larger than 6 meters, and those species and their heights are available in supplemental table 8. right) Nine panels displaying the prior (light gray) and posterior (dark gray) probability distributions of tE under an environmental filtering model and OU model trait evolution. The dotted line represents the median estimate of tE. A) Estimate from the entire local kipuka plant species pool. B-I) Estimates from the separate eight kipuka communities. 99

Figure 2.5. Supplement. Error rates, or proportion of incorrectly classified simulations, when classifying community assembly and trait evolution models compared to the size of the local community used. Square symbols indicate RF was used to classify data and circle symbols indicate ABC was used to classify data. 100

Figure 2.6. Supplement. Linear regression models between the model support for environmental filtering, as predicted from RF and ABC) and the tE median estimates from ABC. The correlation coefficient when comparing RF model support values and tE estimates was 0.65, and when comparing with ABC the coefficient was 0.89. 101

Figure 3.1. Localities sampled for western Red Cedar and western Hemlock. Locality information for each collection can be found in Supplemental Table 1. 131

Figure 3.2. (A-K) Summarized folded jSFS (43 by 43 cells) for 10,000 simulations under each associated demographic model. Scale indications 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 all events that can occur in a given model, including divergence (*div*), bottleneck (*bot*), and 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. 132

Figure 3.3. left panels: STRUCTURE results for *Thuja plicata* and *Tsuga heterophylla* 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 = 2. 133

Figure 3.4. Confusion matrix depicting the prediction accuracies and inaccuracies for 11 demographic models using delimitR for model selection, which involves the simulated and binned jSFS and 'abcrf'. The rows indicate the model the data were simulated under and the 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. 134

Figure 3.5. 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). 135

Statement of Contribution

Chapter 1

M.R. and Anahi Espindola (AE) generated and analyzed the data. MR, AE, Megan Smith (MLS), Bryan Carstens (BC), David C. Tank (DCT), and Jack Sullivan (JS) contributed in design of the study. M.R. performed genomic assembly and processing, population structure, and demographic modeling analyses. A.E. performed species distribution models. M.R. wrote the manuscript, with edits and comments from all authors. All authors approved the final version of the manuscript that was first submitted and the edits and responses to peer reviews. D.T., B.C., and J.S. acquired the funding for this research.

Chapter 2

MR, DCT and Luke Harmon (LH) developed research concept. Bob Weeks contributed to the creation of the non-neutral assembly models and Katie Peterson collected all empirical data. MR developed simulation models in CAMI, performed all analyses, and wrote the manuscript. All authors contributed to critiques of the analysis and subsequent revisions of the text.

Chapter 3

MR, DCT, and JS developed research concept; in addition to MLS, AE and BC contributed to study design and implementation. MR, MLS, AE, Dan Turck, and Niels Mitchel all collected samples in the field and performed lab work for sequencing. MR performed all analysis and wrote the manuscript. All authors contributed to critiques of the analysis and subsequent revisions of the text.

Chapter 1: Combining Allele Frequency and Tree-Based Approaches Improves Phylogeographic Inference from Natural History Collections

"Combining allele frequency and tree-based approaches improves phylogeographic inference from natural history collections" Ruffley, M, Smith, ML, Espíndola, A, Carstens, BC, Sullivan, J, Tank, DC. *Molecular Ecology*. Vol. 27, 2018, pp 1012– 1024. https://doi.org/10.1111/mec.14491

Abstract

Model selection approaches in phylogeography have allowed researchers to evaluate the support for competing demographic histories, which provides a mode of inference and a measure of uncertainty in understanding climatic and spatial influences on intraspecific diversity. Here, to rank all models in the comparison set, and determine what proportion of the total support the top-ranked model garners, we conduct model selection using two analytical approaches –allele frequency-based, implemented in fastsimcoal2, and gene tree-based, implemented in PHRAPL. We then expand this model-selection framework by including an assessment of absolute fit of the models to the data. For this, we utilize DNA isolated from existing natural history collections that span the distribution of red alder (*Alnus rubra)* in the Pacific Northwest of North America to generate genomic data for the evaluation of 13 demographic scenarios. The quality of DNA recovered from herbarium specimen leaf tissue was assessed for its utility and effectiveness in demographic model selection, specifically in the two approaches mentioned. We present strong support for the use of herbarium tissue in the generation of genomic DNA, albeit with the inclusion of additional quality control checks prior to library preparation and analyses with multiple approaches that incorporate various data. Analyses with allele frequency spectra and gene trees predominantly support *A. rubra* having experienced an ancient vicariance event with intermittent and frequent gene flow between the disjunct populations. Additionally, the data consistently fit the most frequently selected model, corroborating the model selection techniques. Finally, these results suggest that the *A. rubra* disjunct populations do not represent separate species.

Introduction

Understanding how conspecific populations evolve is central for identifying and quantifying diversity. Phylogeography aims to increase our understanding of these historical processes (e.g., Avise et al. 1987), and as the field has expanded, several approaches have been used. While the earliest investigations derived their inferences from qualitative interpretations of patterns evident in the genetic and geographic data, later studies began to more explicitly test hypotheses (e.g., Sullivan et al. 2000; Knowles 2001) and estimate parameters under explicit analytical models, such as isolation with migration in IMa2 (Hey 2010) and Migrate-n (Beerli & Felsenstein 2001). In hypothesis testing and model selection studies, models representing historical demographic scenarios are evaluated in a statistical framework, and the inferences are drawn from the results of the test (Knowles & Maddison 2002). Early examples used parametric simulation and frequentist statistics (e.g., DeChaine & Martin 2005), whereas later examples utilized Bayesian (e.g., Fagundes et al. 2007) or information theoretic (Carstens et al. 2009) approaches to consider and rank multiple models. Such approaches allow historical knowledge of the species or complex of study to be incorporated into the models that are assessed (Gutenkunst et al. 2009a). Phylogeographic model selection can be implemented through a variety of approaches and software, such as Approximate Bayesian Computation (ABC; Csilléry et al. 2012), ∂a∂I (Gutenkunst et al. 2009b), fastsimcoal2 (Excoffier & Foll 2011; Excoffier et al. 2013), and PHRAPL (Jackson et al. 2015). All incorporate coalescent theory (Kingman 1982) to model evolutionary processes that occur at the population level, such as genetic drift, migration, and population expansion and/or contraction over time. As opposed to hypothesis testing approaches that reject or fail to reject individual models (and thus experience difficulties with multiple comparisons), model selection frameworks can be designed to rank all models in the comparison set, and thus provide one measure of confidence in the form of what proportion of the total support is garnered by the top-ranked model. However, a potential shortcoming of such a framework is that there is no guarantee that a model that represents the true evolutionary history is included in the comparison set (Templeton 2008).

Phylogeographic inference is ideally drawn from multiple sources, including geographic information (e.g., Hugall et al. 2002) and descriptive summaries of the data (e.g,.

Petit & Grivet 2002). Analytical models that incorporate the coalescent process are valuable, particularly when they have a demonstrably good fit to the empirical data. While assessments of model adequacy and model fit have generally been lacking in phylogeographic research, they are vital components of inferences that are derived from statistical analysis (Gelman & Shalizi 2013). Here, we expand a model-selection framework such that it includes an assessment of model fit. We first conduct model selection using two analytical approaches – allele frequency-based, implemented in fastsimcoal2 (Excoffier et al. 2013), and gene tree-based, implemented in PHRAPL (Jackson et al. 2015) – and then assess the absolute fit of the models to the data.

Phylogeographic analysis depends on comprehensive sampling across the geographic range of a species or complex (Knowles & Maddison 2002; Pinceel et al. 2005). Herbarium and other natural history museum specimens are important sources for such sampling when specimens are available for DNA extraction. Plant tissue dried and preserved in silica gel can be used to recover high quality genomic data (Varma et al. 2007), even several years after collection (e.g. Eaton & Ree 2013). However, there still remain a large number of herbarium specimens without associated silica dried tissue, which results in one having to use tissue directly from the herbarium specimen sheet that was not dried strategically for DNA preservation. The use of genomic data in phylogeography has increased the resolution at which we can discern competing hypotheses, and thus improved our overall understanding of phylogeographic processes (Carstens et al. 2012). However, it is unclear if the DNA that can be extracted from herbarium specimens is sufficiently intact to serve as the source material for generating genome-scale datasets, particularly when systematically distributed missing data can result in implicitly biased inferences (Andrews et al. 2016).

In this work, in addition to extending model-based phylogeographic inference to incorporate model fit, we also aim to understand the quality of DNA needed to recover useful genomic data from herbarium-sampled leaf tissue. We further ask whether such genomic data are plagued with biased, or non-uniform, missing data. Finally, using genomic data from herbarium specimens, descriptive analyses, and two model selection approaches, we aim to understand the phylogeographic history of Alnus rubra (Bong) in the disjunct mesic forests of the Pacific Northwest of North America.

Methods and Materials

Study System

The Pacific Northwest (PNW) temperate rainforests form a disjunct ecosystem that ranges from the Cascade Mountain Range to the Pacific coast, extending from northern California to southern Alaska, and exists along the northern Rocky Mountains (NRM) in central to northern Idaho. The range of mesic forests in the region was likely continuous prior to the uplift of the Cascades (*ca*. 5 MYA, Waring & Franklin 1979; Priest 1990), which generated a rain shadow cast across the Columbia Basin and forced inland forests to retreat to suitable, wet habitat along the NRM. Because of this, the coastal and inland NRM forests became isolated by ~ 300 km of unsuitable habitat. The later onset of Pleistocene glaciations (*ca*. 1.5 MYA) led to the expansion of Cordilleran ice sheets, which covered much of the inland rainforests, further reducing the available habitat for rainforest species. As a consequence of these events, at least some rainforest species were unable to persist in the inland NRM forest throughout the Pleistocene (reviewed in Brunsfeld *et al.* 2001).

Due to its disjunct nature, the PNW rainforest has been the focus of several phylogeographic studies (Nielson *et al.* 2001; Carstens *et al.* 2004; Steele *et al.* 2005; Brunsfeld *et al.* 2007; Metzger *et al.* 2015), indicating that the history of the species was tightly associated to that of the biome. Studies showed that while some species harbor cryptic diversity (i.e., pre-Pleistocene divergences) across the disjunction, others do not. This led to the definition of two principal phylogeographic hypotheses for the biome, which explain the presence or absence of cryptic diversity along the disjunction. The first hypothesis explains the presence of cryptic diversity in a lineage, and is known as pre-Pleistocene vicariance or Ancient Vicariance (AV) (Brunsfeld *et al.* 2001). The uplift of the Cascade Mountain Range has been implicated as causing the disjunction. The AV hypothesis posits that conspecific populations were continuously present along the coast and in the inland NRM forests, but that the two areas were genetically isolated from each other. Following the end of the Pleistocene glacial cycles (~ 13 KYA), conspecific populations locally recolonized newly freed suitable areas. Because this hypothesis posits that there has been no gene flow between the inland and coastal populations since the initial vicariance event, the lineages would have been evolving independently for *ca.* 2 MY, leading to the presence of cryptic diversity.

The alternative hypothesis, post-Pleistocene dispersal or Recent Dispersal (RD) explains the absence of cryptic diversity in some taxa (Brunsfeld *et al.* 2001) . This hypothesis posits local Pleistocene extinction in the NRM. The current species presence in the inland NRM forest is thus due to dispersal from coastal populations to the inland following glacial retreat, though there is also evidence of dispersal occurring from the inland NRM to the coast (e.g Carstens *et al.* 2013). Ultimately, the RD hypothesis suggests absence of significant genetic differentiation between inland and coastal populations because dispersal to the inland happened after 13 KYA. However, RD is not the only phylogeographic scenario that could result in an absence of significant genetic differentiation between the inland and coastal populations. Due to the regions being exposed to cyclical glacial periods, there could also have been episodic, repeated migration since the Ancient Vicariance event. In this case, there were still inland populations throughout the Pleistocene, as in the AV hypothesis, but persistent gene flow prohibited any deep divergence between the coastal and inland populations. Likewise, secondary contact, where the inland and coastal lineages underwent AV and only since the end of the Pleistocene came back in contact, could also result in the lack of cryptic diversity.

The primary distribution of *A. rubra* is west of the Cascades in the coastal temperate rainforest from southeastern Alaska to central California, with disjunct populations in the inland NRM temperate rainforest of Idaho. Two studies have investigated the history of *A. rubra* using genetic data (Strenge 1994; Brumble 2008) and suggested recent dispersal to the inland rainforest. Strenge (1994; also see Soltis *et al.* 1997) characterized two cpDNA genotypes, a coastal southern type and a coastal northern type, and the inland individuals included were of the southern coastal genotype. Brumble (2008) identified a 17 bp indel in the chloroplast *psbA-trnH* spacer that was also present in some, but not all, closely related *Alnus sp*. Thus, the genetic data in these two studies were limited, and the inferences were perhaps driven by a single ancestral polymorphism. Recently, a predictive framework was developed to detect the presence of cryptic diversity from locality data by assessing geo-referenced climatic data and taxonomic ranks (Espíndola *et al*. 2016). This predictive framework also predicts that *A. rubra* should not contain cryptic diversity.

ddRAD Sequencing

Leaf tissue from 49 *A. rubra* herbarium specimens (18 localities, Table 1.1) were sampled from the Stillinger Herbarium at the University of Idaho. The sampled specimens were collected from the coastal and inland PNW rainforest between the years of 2000 and 2007 by various collectors (Table 1.1) and cover the complete range of the species. DNA was extracted using a modified CTAB protocol (Doyle & Doyle 1987), purified using Sera-Mag SpeedBeads (Thermo Fisher Scientific; Faircloth & Glenn 2012; Rohland & Reich 2012), and quantified using a Qubit 1.0 Fluorometer (Life Technologies). Genomic data were generated using double digest restriction site associated DNA sequencing (ddRADseq) (Peterson *et al.* 2012), with the restriction enzymes *EcoRI* and *SbfI* (New England Biolabs, USA), and size selection at a 650 (±50) bp window on a BluePippin (Sage Science). All digestion, ligation and PCR products were purified using Agencourt AMPure XP purification system (Beckman Coulter). Sequences were generated as 300 bp paired-end reads using an Illumina MiSeq in the Institute for Bioinformatics and Evolutionary Studies (IBEST) Genomics Resource Core at the University of Idaho. Raw sequences were processed using PyRad (Eaton 2014) under a minimum coverage of 7 and clustering threshold of 85% (see Dryad link for complete parameter file). PyRad includes Vsearch (Rognes *et al.* 2016) and Muscle (Edgar 2004) for sequence clustering. To merge overlapping reads, Paired-End AssembleR (PEAR) (Zhang *et al.* 2014) was used, and only sequences that merged with their paired end were used in subsequent analyses.

Data quality and effect on missing data

Before ddRADseq library preparation, DNA extracts from 13 of the 49 individuals were quantified using a Fragment Analyzer (Advanced Analytical), which describes the distribution of fragment sizes in a particular sample (Fig. 1.6). The mode of the fragment size distribution, concentration of the fragments in the distribution, year of collection, and each variable’s potential interactions (Table 1.2) were used for linear regression to predict the total number of raw reads for the 13 samples. This allowed us to evaluate our ability to predict data quality based on DNA quality and/or quantity descriptors.

To confirm the presence of unbiased missing data across sampled localities, missing data were quantified across all 49 samples, and organized by relatedness using population assignment probabilities from STRUCTURE (see below) at K=3 (Fig. 1.7a). This ordered distribution was compared to a uniform distribution of the same size using a two-sided Kolmogorov-Smirnov (KS) test (Panchenko 2006). The uniform distribution was simulated in R using the *runif* function to generate 49 random variables from a uniform distribution with a maximum and minimum bound corresponding to the maximum and minimum missing data value observed across all individuals (Fig. 1.7b).

Population Structure

STRUCTURE 1.3.4 (Pritchard 2010) was used to estimate population structure across all sampled localities. All unlinked SNPs from all samples were used in the analyses. Following Pritchard (2010), all parameters were kept as default, aside from the burn-in (set to 200,000 generations) and the MCMC length (set to 1,000,000 generations). The data were modeled assuming admixture and correlated allele frequencies between populations (Falush *et al.* 2003). We tested a range of K values from 1 to 10 and repeated each run 10 times to capture variation in the likelihood estimate of each K value. The individual and population level probabilities of belonging to a particular cluster K were visualized using STRUCTURE PLOT (Ramasamy *et al.* 2014).

Species Distribution Models

To gather more information about the potential range extent of the species during the Last Glacial Maximum (LGM; ~18Kya), we used a species distribution modeling (SDM) approach (Peterson *et al.* 2011). To do this, we gathered 772 unique observations of *A. rubra* from the Global Biodiversity Information Facility (GBIF) and the Consortium of Pacific Northwest Herbaria. We selected eight of the least correlated bioclimatic variables from the 19 total Worldclim bioclimatic variables (r_2<0.7; *i.e.*, bio1, bio2, bio3, bio5, bio7, bio12, bio15, and bio17) at a 30 seconds resolution (Hijmans *et al.* 2005), and used them to adjust SDMs. To do this, we used the package biomod2 (Thuiller *et al.* 2009) and applied an ensemble approach in R. For this, we adjusted the final ensemble model by using the AUCs (area under the curve) of nine modeling methods, including generalized linear model (GLM), generalized additive model (GAM), classification tree analysis (CTA), artificial neural network (ANN), surface range envelop (SRE), flexible discriminant analysis (FDA), multiple adaptive regression splines (MARS), random forest (RF), and Maxent as weighting units, and we selected 10,000 pseudo-absences from the background area (a polygon encompassing the entire range of the species and the totality of the PNW region). We then projected the ensemble model into geographic space, using both current and paleo-climatic data obtained from Worldclim corresponding to current and LGM climatic conditions.

Demographic Inference from Allele Frequency Spectra

Alleles were grouped based on geography into two populations: a coastal population and an inland population. Folded joint allele frequency spectra (jAFS) were then constructed to summarize bi-allelic frequencies across both populations. AFS is a commonly used statistic for population genetic inference (Wakeley 2008; Nielsen *et al.* 2009), and because of this, jAFS, as well as multi-dimensional AFS, have been increasingly used for demographic inference (Keinan *et al.* 2007; Gutenkunst *et al.* 2009a; Smith *et al.* 2017). An AFS cannot accommodate any missing data, and RADseq data is commonly plagued with missing data due to allelic dropout. Therefore, we constructed two sets of jAFS by subsampling SNPs at two different missing data thresholds: 20% and 30%. The threshold value indicates the percentage of individuals from each population that must contain a given SNP for it to be included in the jAFS. To account for variation in the subsampling technique, we constructed 20 jAFS in each subsampling category, for a total of 40 observed jAFS. All jAFS were made using custom Python scripts developed by J. Satler (https://github.com/jordansatler/SNPtoAFS). The first dataset, subsampled at a 20% threshold, included jAFS from ten inland and nine coastal alleles ranging in 65 – 73 SNPs. The second dataset, subsampled at a 30% threshold, included jAFS from 15 inland alleles and 14 coastal alleles ranging in 26 – 34 SNPs.

Model selection was performed on each observed jAFS using fastsimcoal2 (Excoffier & Foll 2011; Excoffier *et al.* 2013). Under this approach, we estimated the composite likelihood of a jAFS between two populations, given a particular demographic model, and for each model parameter. The optimization of each parameter and the composite likelihood was done using the Expectation-Conditional Maximization (ECM) algorithm (Meng & Rubin 1993). In ECM, the E-step consisted of 100,000 coalescent simulations to estimate the expected jAFS under the current demographic parameters to approximate the composite likelihood, as in Excoffier *et. al* (2013). The CM-step consisted of a series of conditional maximizations (Brent 1974) corresponding to the number of parameters included in the model being investigated. The minimum and maximum number of ECM cycles was set to 10 and 30. The optimization process ended when the maximum number of cycles was complete, or when the difference in the composite likelihood under the current parameters compared to the likelihood under the proposed parameters was less than 0.001. Thirteen different demographic models were evaluated (Figure 1) with 20 independent optimizations from different starting parameters (Excoffier *et al.* 2013), and the maximum likelihood parameter estimates resulting from each independent optimization were used as the starting parameters in a final maximization of the composite likelihood. We then calculated Akaike Information Criterion (AIC) values (Akaike 1974) using the maximum composite likelihood estimated from this run, and compared the models using Akaike weights, wAIC (Johnson & Omland 2004). First, we compared just the four principal demographic models (AV, RD, AVwM, AVtS; Figure 1) using wAIC, then we compared all 13 models using wAIC. We assume that because the collection of unlinked SNPs are randomly distributed across the genome (Excoffier *et al.* 2013), the composite likelihood is a good approximation of the true maximum likelihood and can thus be used in AIC calculations for model comparison.

For each model, we estimated τ_{div} as divergence time in generations, τ_{SC} as time of secondary contact in generations, *m* as various probabilities of migration to and from coastal and inland populations, Θ_0 and Θ_1 as $\Theta=4Ne\mu$ where Ne is the number of genes in each deme, and the mutation rate μ (only parameter not shown in model design; Figure 1) in substitutions/site/generation. The initial values for parameters Θ_0 and Θ_1 were drawn from a log uniform distribution between 0.01 and 10, and the parameter space explored was constrained only by the minimum bound of the prior distribution. The mutation rate, μ, was estimated from a minimum bounded log uniform distribution between 1e-9 and 1e-7. Divergence time, τ_{div}, was drawn from a minimum bounded uniform prior distribution with a minimum of 50,000 and maximum of 1,000,000 generations for all models involving an AV event, whereas τ_{DIV} from recent dispersal models was drawn from a fully bounded, uniform distribution with a minimum of 500 and a maximum 50,000 generations. Divergence time estimates were converted from generations to years using a generation length of 6-8 years per generation (Orwa *et al.* 2009). The secondary contact models included the time of the gene exchange event (τ_{SC}) as a parameter. The prior distribution for τ_{SC} was uniform with a minimum in 500 and maximum in 50,000 generations. Migration was defined as the probability of a given lineage to move from one population to the other and was drawn from a log uniform prior distribution with min 1e-10 and max 0.1. Migration was considered as either unidirectional (only from the inland, or only from the coast), bidirectional (a separate migration rate parameter was estimated for each direction), or symmetrically bidirectional (only one migration rate parameter is estimated; Figure 1).

To investigate model adequacy, we performed a goodness-of-fit test, which evaluates whether the observed data fit a particular model. The goodness-of-fit test is done using a likelihood ratio G-statistic, $CLR = \log_{10}(CL_0/Cl_E)$, where CL_0 is the observed maximum composite likelihood and CL_E is the estimated maximum composite likelihood (Excoffier *et al.* 2013). To represent the expected distribution of data given the best model, we performed parametric bootstrapping with the Maximum Likelihood (ML) parameter estimates of the selected model to generate 100 simulated jAFS that had an equal number of alleles per population as the empirical data. We then optimized the likelihood of each of these datasets given the model and used these maximum likelihoods to calculate the null distribution for the G-statistic. This process of parametric bootstrapping using the ML parameter estimates was done in fastsimcoal2 and repeated for the three best models in each of the subsampling threshold categories, 20% and 30%. The *p*-value for each goodness-of-fit test was calculated as the proportion of simulated G-values that were greater than the observed test-statistic over all the total number of G-values, in this case 100. This simulated data also permitted calculating 95% confidence intervals for parameters of interest under a model of interest.

Demographic Inference from Gene Trees

Phylogeographic inference using approximate likelihoods, or PHRAPL (Jackson *et al.* 2015), is conducted using gene tree topologies without branch lengths as input. While the gene trees can be constructed using either linked or unlinked SNPs, we opted for the former because we were unable to use linked SNPs in the analysis with allele frequency spectra. For this, all loci that were present in at least four individuals from the coast and four individuals from the inland were used to construct a total of 63 gene trees in PAUP* (Swofford 2003). Before constructing trees, we used DT-ModSel (Minin *et al.* 2003) to select an appropriate model of sequence evolution for each locus. A total of 42 models were evaluated, and the best model was selected using decision theory. The model and corresponding parameter values were used to construct a maximum likelihood tree in PAUP*. A heuristic search started with a neighbor joining tree and performed tree bisection and reconnection (TBR) as the branch swapping mechanism for a maximum time limit of five minutes, at which time only the most optimal tree was saved. Two datasets were assembled; in the first dataset, only gene trees produced with 20 or more SNPs were included, which resulted in seven trees total. In the second dataset, only gene trees made with two or more SNPs were included, which resulted in 46 trees total. We performed the same model selection procedure that is described below, on both datasets.

All thirteen models that were designed in fastsimcoal2 were also designed and evaluated in PHRAPL. All the PHRAPL models have one less parameter than the models in fastsimcoal2, because PHRAPL does not estimate mutation rate. PHRAPL uses a grid search to investigate parameter space and optimize approximate likelihoods under user-specified "grid values." Runs began with a broad range of grid values for all parameters and move toward specific values once likelihood peaks are identified. We ran a total of four grid searches on each dataset of trees, with 6-8 grid values investigated for the coalescent time parameter and the migration parameter(s) each, on every grid search. The first two grid searches investigated broad ranges in the coalescent time parameter, while the final two grid searches narrowed these values considerably (Table 1.3). All secondary contact models included an additional parameter representing the timing of the secondary contact, this event time was set to occur prior to the coalescent event, at a relative time of 0.25 for all runs. We first compared only the four core models (Figure 1.1) using wAIC, and then compared all thirteen models at once using wAIC. All computational analyses were done using servers at the IBEST Computational Resources Core at the University of Idaho.

Power Analysis

Data were simulated under the Ancient Vicariance with Asymmetrical Migration (AVwM) in fastsimcoal2 using parameters drawn from a prior distribution. The parameters of the AVwM model that were drawn from a prior are Θ_0 and Θ_1, as $\Theta=4Ne\mu$ where Ne is the number of genes in each deme, m_{12} as the migration rate to the coast, m_{21} as the migration rate to the inland, τ_{div}, as the divergence time for inland and coastal populations, and the mutation rate μ in substitutions/site/generation. The prior distributions for all of the parameters are the same distributions that are used to draw the starting parameter values for optimization of the likelihood (see main text Methods, Demographic Model Selection, Allele Frequency Approach). 100 sets of parameters were drawn to simulate 100 different jAFS. The jAFS were simulated to emulate the empirical dataset constructed from a 20% missing data threshold. We used the same number of individuals per population, 10 inland and 9 coastal, and mirrored the range of unlinked SNPs. In the empirical data, for the 20% jAFS dataset, the range of SNPs is between 65-73, in the simulated data the range is between 62-80.

For each dataset, we optimized the likelihood of the data given the four main models; Ancient Vicariance (AV), Recent Dispersal (RD), AVwM, and Ancient Vicariance with Secondary Contact (AVwSC). We did not perform multiple, independent optimizations per model due to computational constraints. For each dataset, we compared the maximum likelihoods for the four models using AIC, and the model with the highest AICw was classified as the model selected.

Results

ddRAD sequencing, data quality, and effect on missing data

We recovered 648 loci with 614 unlinked biallelic SNPs, 5,494 total variable sites, and 79% missing data. We expected to recover ca. 6,000 loci, following approximate calculations (Peterson *et al.* 2012) given a genome size of roughly 5Mbp (MD & IJ 2012), 8-cutter and 6-cutter restriction enzymes, 70% of a half MiSeq lane (approx. 6 million reads), and expected coverage of 20x. Recovery of fewer loci could be due to many reasons, a few of which include protocol modifications, restriction enzyme selection, suboptimal size selection window, or not enough sequencing power (Peterson *et al.* 2012). Here, the quality of the genomic DNA, or average fragment size, is a primary factor in explaining the variation in the number of reads recovered (Figure 1.2). Our linear regression analysis showed that the only significant predictor variable was the mode of the fragment size distribution (Figure 1.2), explaining around 60% of the variation observed in the total number of raw sequence reads.

The KS test reported a p-value of 0.465 at $\alpha = 0.05$, indicating there is not a significant difference between the simulated uniform distribution and the observed distribution of missing data (Figure 1.7). Thus we can conclude that the missing data are uniformly distributed across all individuals, that is, there are no individuals that have an extremely high amount of missing data relative to any other individuals, which indicates the data can be subsampled (*i.e.*, missing data discarded) without biasing estimates (Wiuf 2006).

Population Structure

Because the analyses of data with missing data can be suspected to involve an overestimation of K (Pritchard 2010), we visualized all STUCTURE results for K = 2 to 8. When K = 2 (Figure 1.3), there was apparent spatial genetic structure separating coastal from inland populations. This result agrees with the expectations under an AV scenario. When K = 3 (Figure 1.3), two of the clusters were restricted to either inland or coastal populations, and the third included individuals from both areas, suggesting gene flow between the disjunct populations. Results for K = 4, 5, and 8 showed no geographic population structure (Figure 1.8, 1.9), additionally suggesting the presence of gene flow between coastal and inland populations.

Species Distribution Models

Our ensemble SDM could successfully recover the current range of the species (Figure 1.3c). Part of the projected range of the species at the LGM (Figure 1.3d) is substantially different from the current range. Specifically, the coastal area appeared to display high habitat suitability, while the suitable inland areas are more restricted than the current inland range. This suggests that during the LGM, the coastal area may have harbored large extents of continuous habitat for the species, whereas inland populations were likely more restricted.

Demographic Model Selection

When the four core models were compared, AVwM always had the highest wAIC, on average, regardless of the downsampling technique or whether fastsimcoal2 or PHRAPL was used (Figure 1.4). AVwM was consistently selected using fastimcoal2, with fairly high average AIC values (Figure 1.4). In PHRAPL, with seven trees produced from loci with 20 or more SNPs, the AVwM model was selected 75% of the time (Table 1.4) with an average wAIC of 0.389, and the RD model was selected 25% of the time with an average wAIC also of 0.362

(Figure 1.4). In PHRAPL, with 46 trees produced from loci with 2 or more SNPs, AVwM and RD were both selected 50% of the time (Table 1.4) with an average wAIC of 0.484 and 0.414, respectively (Figure 1.4).

When all 13 models were compared using wAIC from fastsimcoal2 estimated likelihoods, two models were selected consistently, AVwCM and AVwIM. In the first dataset (20% subsample threshold), AVwCM was selected around 50% of the time, while AVwIM was selected 43% of the time (Table 1.5). The remainder of selected models includes AVwM, AVtSC, and AVtSI, all of which were selected approximately 1.5% of the time. In the second fastimcoal2 dataset (30% subsample threshold), AVwIM was selected 56% of the time, and AVwCM was selected 42% of the time. Only two other models (AVwMsym and AVtSI) were selected, each at very low rates (1% of the time). In both datasets, the three models with the highest average wAIC were AVwIM, AVwCM, and AVwM (Figure 1.5a-b) and for all three models, the mode divergence time estimate was between 5.8 and 6.9 MYA (mean 6.3 MYA) (Table 1.6).

Model adequacy was evaluated for the three best models for both datasets assessed in fastsimcoal2 (Figure 1.5c-d). In both datasets, the *p*-values were non-significant, indicating that the data fit all three models. In the first dataset, the *p*-values for AV with asymmetrical migration, AV with Coastal Migration, and AV with Inland Migration were 0.69, 0.77, and 0.79, respectively (Figure 1.5c). In the second dataset, the *p*-values for AV with asymmetrical migration, AV with Coastal Migration, and AV with Inland Migration were 0.93 for each (Figure 1.5d).

In PHRAPL, there were a handful of models amongst the 13 compared that carried a majority of wAIC support in any particular run. In all of the runs, the average wAIC for the best model was 0.16 – 0.26 (Table 1.7, 1.8), indicating not particularly strong support for any one model. In the dataset with seven trees, the AVwIM model was selected 100% of the time, however the average wAIC for the AVwIM model was only 0.167. In all four runs with the seven-tree dataset, there were at least five models that carried more than 10% of the model weight (i.e. wAIC > 0.10) (Table 1.7). In the dataset with 46 trees, AVwMsym and RDsym were both selected 50% of time with an average wAIC of 0.22 and 0.214 (Table 1.8). In all four runs with the 46-tree dataset, there were at least four models with more than 10% of the model weight. In both PHRAPL datasets, the RD models collectively occupied over a third of the model weight in any given grid search.

Power Analysis

Out of the 100 datasets, 86 of them were correctly classified as AVwM with an average AIC of 0.703. The remaining 14 datasets were classified as AVwSC with an average AIC of 0.296 (Table 1.15).

Discussion

Allele Frequency and Gene Tree Approaches

fastsimcoal2 and PHRAPL each have unique properties that make them useful in performing model selection. While fastsimcoal2 is appealing because it summarizes unlinked SNPs in an allele frequency spectrum, PHRAPL uses topologies that can be generated from SNP data or entire loci. In this study, the number of unlinked SNPs that could be used in fastimcoal2 was limited because AFS does not accommodate missing data, and therefore the unlinked SNPs must be downsampled. Alternatively, PHRAPL accommodates missing data because not all individuals need be present in each gene tree, only a minimum total number of individuals from each population need be present. Depending on the missing data and number of linked and unlinked SNPs present in one's dataset, either approach could be viable.

fastsimcoal2 results were more consistent than PHRAPL results in selecting the same model with high support, especially when only comparing the four core models AVwM (Fig. 1.4). However, even when comparing all 13 models in fastsimcoal2, AVwCM and AVwIM maintained a majority of the wAIC support. Additionally, model adequacy in fastsimcoal2 supported that the data fit these three best models, all of which support an ancient vicariance event with some intermittent gene flow. When comparing the four core models in PHRAPL, there was consistency in selecting the AVwM model, but with relatively low support from wAIC. Because no single model had an overwhelming majority of the model weight, there was overall less support in which model was selected when comparing all 13 models in PHRAPL.

As for the performance of each analytical approach, it is unreasonable to conclude on which is more accurate given this study, because each approach performs model selection differently, i.e. uses a different approximate likelihood to evaluate model parameters. Specifically, fastimcoal2 uses an observed AFS to optimize an approximate likelihood based on an expected AFS under coalescent simulations (Excoffier *et al.* 2013), while PRHAPL uses gene tree topologies to optimize an approximate likelihood that is based on concordant gene tree topologies under coalescent simulations (Jackson *et al.* 2015). Both incorporate coalescent theory and whether one approach is more accurate than the other may depend on whether information in the available data is better captured in an AFS or gene trees.

Genomic data from herbarium specimens

While *ddRADseq* can be successful in recovering thousands of loci and SNPs (Peterson *et al.* 2012), this was not the case in our study. Our analyses (Figure 1.2) show that this is mostly due to lower quality (fragment size) DNA used in the protocol. No other variable, or interaction of variables, including year collected or concentration, recovered any significant relationship with the number of reads recovered. The ~10-year-old herbarium specimens used in this study, were presumably dried and stored in various ways that resulted in some specimens having better quality DNA. Our results indicate that as long as high quality DNA can be identified (*e.g.*, with a fragment analyzer or bio-analyzer), the method used here is a cost-effective approach for generating genomic data. Specifically, our results strongly suggest that the mean of the distribution of fragment sizes is a strong predictor of the number of reads that will be recovered in the *ddRADseq* approach. Examining the fragment length as an additional step for standardization can identify degraded samples, *i.e.* samples with an average fragment length < 5000 bp, when concentration alone cannot. Practically, we recommend verifying the quality of individual samples with fragment length identification to ensure that no highly degraded samples are used in library preparation. In addition, characterizing missing data as uniformly missing was crucial for implementing our subsampling strategies. If the distribution of missing data were systematically structured, subsampling could have drastically biased our likelihood estimates. Taken together, these results suggest that generating genomic data using DNAs obtained from herbarium specimens is possible, but the average fragment size of resulting DNAs, and the distribution of missing data should be considered for both the experimental and analytical approaches employed here.

Phylogeographic history of red alder

Overall, the Ancient Vicariance with Migration (AVwM) model was the strongest and most consistently supported model across datasets and approaches. Although this model selection approach was not intended to make inferences from the parameter estimates, but rather from the overall model selected, the divergence times estimated seem to indicate congruence with the PNW history of Pre-Pleistocene divergence. The divergence time estimates ranged 5.8 – 6.9 MYA and are older than the Pacific Northwest rainforest disjunction (~ 3-5 MYA; but see the 95% CI in Table 1.9, 1.10, 1.11, 1.12, 1.13, 1.14). Though we cannot precisely determine divergence times, we can characterize the timing of divergence as Pre-Pleistocene.

The AVwM model involves four possible migration scenarios. Two of these four scenarios are highly supported with allele frequency data and include migration in only one direction, AVwIM and AVwCM. When comparing all models, gene trees supported AVwIM as the best model. These results suggest that migration has been predominantly unidirectional, although the actual direction still remains unclear in light of these results. Disentangling which direction is the most likely requires investigating recolonization and/or specific migration route models, potentially, with the inclusion of more genetic data. Due to the lack of genetic structure within the coast or the inland, we do not think that expanding the geographic breadth of our samples is necessary (Figures 1.8 & 1.9). That said, model adequacy results suggest that more genetic data is required to distinguish between the three best migration scenarios for the AVwM model. These results also highlight the limit to the phylogeographic inference that we are able to make given these data, which is something that we feel should always be identified.

The SDMs and population structure results provide further evidence for the Ancient Vicariance with Migration scenario. The climatic niche projections using LGM conditions (Figure 1.3d) indicate that the expected inland range of *A. rubra* shifted into southern Idaho. This is consistent with the hypothesis that species ranges in this area were displaced south during the Pleistocene (Sullivan *et al.* 2000). Although not apparent in the SDM because of its later occurrence, pollen records indicate the presence of *Alnus* in the NRM of Canada throughout the Holocene (Gavin *et al.* 2009). This could indicate surviving populations of *A. rubra* in nunatak refugia as far north as Canada, or a rapid colonization of the area following glacial retreat. During the Holocene, the climate was considerably warmer than today (Wagner *et al.* 2000), which could explain why the inland forest extended much farther North than it does currently. Ultimately, the persistence of *A. rubra* in the inland during the Pleistocene is supported by our results, though whether southern inland populations, or migrants from the northern Cascades, or both, colonized the inland forest remains unclear. The strong genetic divide between the inland and coastal populations (Figure 1.3a-b) also corroborate that inland populations of *A. rubra* likely persisted through the Pleistocene.

Given the prominent role that gene flow has played in the phylogeographic history of *A. rubra*, we conclude, in agreement with Espíndola *et al.*'s (2016) prediction, that coastal and inland populations of red alder do not harbor cryptic diversity, and thus do not represent incipient sister species. Previously, in this system, non-cryptic taxa were often considered to be the result of recent dispersal. However, we show here that an alternative phylogeographic hypothesis – specifically, ancient vicariance with periods of gene flow – can also explain why some lineages in the disjunct mesic forests of the PNW may not harbor cryptic diversity, despite evidence of ancient population structure. We also show that the inclusion of more intraspecific data, genetic and geographic, does in fact increase our phylogeographic understanding of *Alnus rubra*, specifically because the Ancient Vicariance with Migration and Recent Dispersal models could not have been distinguished using the cpDNA from Strenge (1994; Soltis *et al.* 1997) or Brumble (2008). Finally, we acknowledge that the inclusion of more data would allow for the evaluation of more complex models, therefore, if in the future more *Alnus rubra* data is generated, we recommend the evaluation of more complex models that include population expansion and contraction.

Conclusions

In this study, we compared two approaches in phylogeographic model selection, allele frequency-based (Excoffier *et al.* 2013) and gene tree-based (Jackson *et al.* 2015), and used the results from both to draw phylogeographic inference in an emerging model system in comparative phylogeography. Importantly, both approaches resulted in a ranking of models that was useful in gauging relative support for all competing models. The most overwhelming indicator of successful model selection comes from the review of model adequacy, where we see the data consistently fit the models that were most frequently selected. Because assessing model fit is a critical component of any statistical inference, we feel that future phylogeographic studies should include explicit tests of model adequacy, as performed here. Further, we also conclude that mean fragment length is an effective measure of sample quality that will help in identifying samples that may be problematic for RAD-based genomic reduction sequencing strategies – samples where the concentration alone is not enough to indicate levels of degradation – early on in library preparation. We also demonstrate that genomic data obtained from DNA isolated from herbarium specimens do not necessarily result in systematically missing amounts of data, which allows for downsampling without the fear of drastically biasing the data. Finally, we were successful in using DNA from herbarium specimens to gather the genomic data necessary to make inferences regarding the phylogeographic history of *A. rubra*, where the combination of descriptive and model selection based tools was invaluable in recovering a meaningful phylogeographic inference that is supported by multiple, independent lines of evidence.

Literature Cited

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

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.

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.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4563–8.

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

Brumble AJ. 2008. Phylogeography of *Alnus rubra*. Masters Thesis. University of Idaho

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

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.

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*.

Carstens BC, Knowles LL (2010) Navigating the unknown: Model selection in phylogeography. *Molecular Ecology*, **19**, 4581–4582.

Carstens B, Lemmon AR, Lemmon EM (2012) The promises and pitfalls of next-generation sequencing data in phylogeography. *Systematic Biology*, **61**, 713–715.

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.

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

Csilléry K, François O, Blum MGB (2012) Abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, **3**, 475–479.

DeChaine EG, Martin AP (2005) Marked genetic divergence among sky island populations of Sedum lanceolatum (Crassulaceae) in the Rocky Mountains. *American Journal of Botany*, **92**, 477–486.

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

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

Eaton DAR, Ree RH (2013) Inferring phylogeny and introgression using RADseq data: An example from flowering plants (Pedicularis: Orobanchaceae). *Systematic Biology*, **62**, 689–706.

Edgar RC (2004) MUSCLE User Guide. *Nucleic Acids Research*, **32**, 1–15.

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

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

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

Fagundes NJR, Ray N, Beaumont M *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences*, **104**, 17614–17619.

Faircloth B, Glenn T (2012) AMPure XP substitute cost savings protocol. *Genome research*, **22**, 939–46.

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.

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.

Gelman A, Shalizi CR (2013) Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, **66**, 8–38.

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

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009b) Diffusion Approximations for Demographic Inference : ∂ a ∂ i. *Nature Precedings*, 2009.

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

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hugall A, Moritz C, Moussalli A, Stanisic J (2002) Reconciling paleodistribution models and comparative phylogeography in the Wet Tropics rainforest land snail Gnarosophia bellendenkerensis (Brazier 1875). *Proceedings of the National Academy of Sciences*, **99**, 6112–6117.

Jackson, N.D., Morales, A.E., Carstens, B.C., & O'Meara, B.C. (2017). PHRAPL: Phylogeographic Inference Using Approximate likelihoods. Systematic Biology, **66**, 1045–1053.

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19**, 101–108.

Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature genetics*, **39**, 1251–5.

Kingman J (1982) The coalescent. *Stoch. Proc. Appl.*, **13**, 235–48.

Knowles LL (2004) The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology*, **17**, 1–10.

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

Lacey Knowles L (2001) Genealogical portraits of speciation in montane grasshoppers (genus Melanoplus) from the sky islands of the Rocky Mountains. *Proceedings of the Royal Society B: Biological Sciences*, **268**, 319–324.

MD B, IJ L (2012) Angiosperm DNA C-values database (release 8.0, Dec. 2012). *http://www.kew.org/cvalues/homepage.html*.

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

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.

Minin V, Abdo Z, Joyce P, Sullivan J (2003) Performance-based selection of likelihood models for phylogeny estimation. *Systematic biology*, **52**, 674–683.

Nielsen R, Hubisz MJ, Hellmann I *et al.* (2009) Darwinian and demographic forces affecting human protein coding genes. *Genome Research*, **19**, 838–849.

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.

Orwa C, Mutua A, Kindt R, Jamnadass R, Anthony S (2009) Agroforestree Database: a tree reference and selection guide version 3.0. *http:www.worldagroforestry.org/sites/treedbs/treedatabases.asp)*, **0**, 1–5.

Panchenko P (2006) Kolmogorov-Smirnov Test. In: *Statistics for Applications*, pp. 83–90.

Peterson AT, Soberón J, Pearson RG *et al.* (2011) *Ecological niches and geographic distributions*.

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**.

Petit RJ, Grivet D (2002) Optimal randomization strategies when testing the existence of a phylogeographic structure [1]. *Genetics*, **161**, 469–471.

Pinceel J, Jordaens K, Pfenninger M, Backeljau T (2005) Rangewide phylogeography of a terrestrial slug in Europe: Evidence for Alpine refugia and rapid colonization after the Pleistocene glaciations. *Molecular Ecology*, **14**, 1133–1150.

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.

Ramasamy RK, Ramasamy S, Bindroo BB, Naik VG (2014) STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. *SpringerPlus*, **3**, 431.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ Preprints*, **4**, e2409v1.

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

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.

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.

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.

Strenge D. 1994 The intraspecific Phytogeography of *Polystichum munitum* and *Alnus rubra*. Masters Thesis. Washington State University.

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.

Swofford DL (2003) PAUP*: phylogenetic analysis using parsimony, version 3.0, b10. *21 Libro*, 11pp.

Templeton AR (2008) Nested clade analysis: An extensively validated method for strong phylogeographic inference. *Molecular Ecology*, **17**, 1877–1880.

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD - A platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Varma A, Padh H, Shrivastava N (2007) Plant genomic DNA isolation: An art or a science. *Biotechnology Journal*, **2**, 386–392.

Wagner B, Melles M, Hahne J, Niessen F, Hubberten HW (2000) Holocene climate history of Geographical Society ??, East Greenland - Evidence from lake sediments. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **160**, 45–68.

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

Wiuf C (2006) Consistency of estimators of population scaled parameters using composite likelihood. *Journal of Mathematical Biology*, **53**, 821–841.

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, **30**, 614–620.

Tables

Table 1.1 *Alnus rubra* collections information. From the left: DNA extraction number, collector and collection number, latitude, longitude, locality description. All herbarium specimens are located in the Stillinger Herbarium at the University of Idaho and can be found online (http://pnwherbaria.org).

Sample ID (DNA number)	Collector and Collection Number	Lat	Long	Locality
2015-411	AJ Brumble 06-86	46.13943	-115.667216	Idaho Co., ID. Along Hwy12 on Lochsa River
2015-412	AJ Brumble 05-39	48.18773	-116.435	Bonner Co., Id. Garfield Bay Rd.
2015-413	AJ Brumble 05-37	48.18788	-116.44015	Bonner Co., Id. Garfield Bay Rd.
2015-414	AJ Brumble 05-30	48.20281	-116.4503	Bonner Co., Id. Garfield Bay Rd.
2015-415	AJ Brumble 05-38	48.18773	-116.435	Bonner Co., Id. Garfield Bay Rd.
2015-416	AJ Brumble 05-71	47.1746948	-115.9104676	Shoshone Co., ID, Banks of St. Joe River
2015-417	AJ Brumble 06-94	46.226316	-115.49328	Idaho Co., ID. Along Hwy12 on Lochsa River
2015-418	AJ Brumble 06-106	45.98981	-118.06008	Umatilla Co., OR. Mill Crk. Rd.
2015-419	AJ Brumble 06-105	45.98786	-118.071	Umatilla Co., OR. Mill Crk. Rd.
2015-420	AJ Brumble 06-103	45.992183	-118.105216	Umatilla Co., OR. Mill Crk. Rd.
2015-421	Bradtke 1556	45.33166	-118.71666	Umatilla Co., OR. Umatilla National Forest, FR 54
2015-422	F.D. Johnson 0120	46.0645809	-118.3430209	Walla walla Co., Wa. Mill Creek
2015-423	AC Zack 0612	47.592128	-117.286658	Iller Creek, Spokane Co., Wa.
2015-424	AJ Brumble 05-57	45.90695	-123.961066	Cannon Beach, Clatsop Co., Oregon
2015-425	Gage 6816	44.43566	-120.34933	Forest Road 22, Crook Co., Or
2015-426	AJ Brumble 06-81	38.44933	-123.114616	Jenner, Sonoma Co., Ca
2015-427	AJ Brumble 06-68	41.491516	-124.0468	Hwy 101, Del Norte Co., Ca
2015-428	AJ Brumble 06-76	39.488316	-123.78815	Hwy 1, Mendocino Co., Ca
2015-429	AJ Brumble 06-15	54.30228	-130.339316	1600 Park Avenue, BC, Canada
2015-430	AJ Brumble 06-12	54.29583	-130.351583	Prince Rupert, BC, Canada
2015-431	AJ Brumble 06-16	54.30583	-130.335616	Prince Rupert, BC, Canada
2015-432	AJ Brumble 06-14	54.30228	-130.339316	1600 Park Avenue, BC, Canada
2015-433	AJ Brumble 06-71	41.37286	-124.013783	Newton Drury Scenic Parkway, Humbolt Co., Ca
2015-434	AJ Brumble 06-67	41.5137	-124.029016	Hwy 101, Del Norte Co., Ca
2015-435	AJ Brumble 06-17	54.305833	-130.33561	Prince Rupert, BC, Canada
2015-436	AJ Brumble 06-80	54.30583	-130.335616	Hwy 1, Sonoma Co., Ca
2015-437	AJ Brumble 05-21	47.771666	-122.29472	Lake Forest park, King Co., Washington
2015-438	AJ Brumble 05-25	47.771666	-122.29472	Lake Forest park, King Co., Washington
2015-439	AJ Brumble 05-13	47.94995	-124.392883	Forks, Clallam Co., wa
2015-440	AJ Brumble 05-43	46.54271	-124.03023	Long Beach Peninsula, Pacific Co., wa
2015-441	AJ Brumble 05-47	46.302183	-124.062616	Long Beach Peninsula, Pacific Co., wa
2015-442	AJ Brumble 05-52	45.8956	-123.9599	Cannon Beach, Clatsop Co., Oregon
2015-443	AJ Brumble 05-12	47.94995	-124.392883	Forks, Clallam Co., wa
2015-444	AJ Brumble 06-78	39.392516	-123.809383	Off Hwy 1, Mendocino Co., Ca
2015-445	AJ Brumble 06-11	54.295833	-130.351583	Prince Rupert, BC, Canada
2015-446	AJ Brumble 05-51	45.8956	-123.9599	Cannon Beach, Clatsop Co., Oregon
2015-447	AJ Brumble 05-50	45.8956	-123.9599	Cannon Beach, Clatsop Co., Oregon
2015-448	AJ Brumble 05-41	46.6071	-124.043116	Long Beach Peninsula, Pacific Co., wa
2015-379	AJ Brumble 06-97	46.332074	-115.076678	Hwy 12 along Lochsa River, Idaho Co., ID
2015-380	AJ Brumble 06-93	46.22815	-115.495716	Idaho Co., ID. Along Hwy12 on Lochsa River
2015-381	AJ Brumble 06-95	46.226316	-115.49326	Idaho Co., ID. Along Hwy12 on Lochsa River
2015-382	AJ Brumble 16-91	46.22815	-115.445716	Idaho Co., ID. Along Hwy12 on Lochsa River
2015-383	P.Brunsfeld 5800-2	48.248116	-116.293766	Storm Creek, Bonner Co., Id
2015-384	F.D. Johnson, S.J. Brunsfeld 7211	47.258751	-115.93699	3.6 mi downstream from avery, St. Joe River, Shoshone Co., Id
2015-385	F.D. Johnson, S.J. Brunsfeld 7212	47.529108	-116.546488	Killarny Lake, Kootenai Co, Idaho
2015-386	F.D. Johnson, S.J. Brunsfeld 7213	47.64289	-116.648197	Beauty Crk Rd., Kootenai Co, ID
2015-387	F.D. Johnson, S.J. Brunsfeld 7210	47.252097	-116.037179	Marble Creek, Shoshone Co., Id
2015-388	F.D. Johnson 0510	48.261184	-116.280881	Strong Creek, Bonner Co., Id
2015-389	P. Brunsfeld 5106-1	48.45996	-116.323483	Kanisku National Forest, Bonner Co., ID
2015-390	AJ Bumble 05-70	47.252578	-115.79239	Banks of St. Joe River, Shoshone Co., ID

!

Table 1.2 Variables (mode fragment length, concentration of fragment length distribution, and year collected) examined for a relationship with the number of raw reads produced by a given sample.

Sample ID	Raw Reads	Mode Fragment Length (bp)	Concentration of Fragment Length Distribution (ng/μL)	Year Collected
411	323,341	7914	8.5239	2006
415	407,328	5847	11.7729	2005
417	182,615	8175	7.1783	2006
418	417,292	7718	6.0023	2006
420	485,504	7653	7.0889	2006
436	110,015	5539	14.8469	2006
437	25,489	2943	3.0378	2005
441	2,637	2058	1.9202	2005
442	101,437	5000	7.1077	2005
443	11,429	2229	2.4967	2005
422	73,001	2390	4.0885	2001
425	59,637	2362	4.6983	2000
439	82,828	4000	4.5448	2005

Table 1.3 Grid search values for each of the four PHRAPL runs (applied to both datasets; 7 and 46 trees) for the collapse time and migration rate parameters. The first two runs are broad, while the final two runs narrow the grid search.

	Model Family	**Collapse Time (Coalescent Units)**	**Migration Rate (4*Nm*)**
1	AV	1.55, 2.7, 4.07, 6.7, 8.5, 10.1, 12.5	N/A
	RD	0.005, 0.01, 0.10, 0.362, 0.76	0.10,0.46,1.00,2.15,3.4, 4.5
	AVwM	1.55, 2.7, 4.07, 6.7, 8.5, 10.1, 12.5	0.10,1.00,2.15,3.4, 4.5, 6.3, 7.6
	AVtS	1.55, 2.7, 4.07, 6.7, 8.5, 10.1, 12.5	0.10,0.46,1.00,2.15,3.4, 4.5
2	AV	8.0, 9.4, 11.0, 13.6, 15.0, 16.3, 18.0	N/A
	RD	0.3, 0.5, 0.6, 1.0, 1.4, 1.8, 2.1	2.15, 4.0, 5.4, 6.7, 8.2, 10.0
	AVwM	8.0, 9.4, 11.0, 13.6, 15.0, 16.3, 18.0	2.15, 4.0, 5.4, 6.7, 8.2, 10.0
	AVtS	8.0, 9.4, 11.0, 13.6, 15.0, 16.3, 18.0	2.15, 4.0, 5.4, 6.7, 8.2, 10.0
3	AV	5.5, 6.2, 6.8, 12.8, 13.4, 13.9	N/A
	RD	0.01, 0.2, 0.362, 0.67, 0.92	5.5, 6.7, 7.6, 8.2, 8.6, 9.0
	AVwM	5.5, 6.2, 6.8, 12.8, 13.4, 13.9	5.5, 6.7, 7.6, 8.2, 8.6, 9.0
	AVtS	5.5, 6.2, 6.8, 12.8, 13.4, 13.9	5.5, 6.7, 7.6, 8.2, 8.6, 9.0
4	AV	5.5, 5.9, 6.2, 6.8, 7.1, 7.4	N/A
	RD	0.01, 0.2, 0.362, 0.67, 0.92	7.0, 7.2, 7.5, 7.8, 8.2, 8.5, 8.7
	AVwM	5.5, 5.9, 6.2, 6.8, 7.1, 7.4	7.0, 7.2, 7.5, 7.8, 8.2, 8.5, 8.7
	AVtS	5.5, 5.9, 6.2, 6.8, 7.1, 7.4	7.0, 7.2 ,7.5, 7.8, 8.2, 8.5, 8.7

Table 1.4 The average wAIC scores for the four core models evaluated in PHRAPL under both sets of trees, for all four runs (see Table S3 for grid search values).

		1	2	3	4
7 trees	AV	0.000	0.000	0.000	0.000
	AVwM	**0.416**	**0.389**	**0.380**	0.370
	RD	0.346	0.357	0.366	**0.380**
	AVtS	0.239	0.254	0.255	0.250
46 trees	AV	0.002	0.001	0.001	0.001
	AVwM	**0.567**	0.417	0.428	**0.524**
	RD	0.313	**0.427**	**0.518**	0.397
	AVtS	0.119	0.155	0.053	0.078

Table 1.5 The wAIC scores for all models averaged across fastsimcoal2 runs, along the selection frequency, within each downsampled dataset.

	20% Missing Data Threshold		30% Missing Data Threshold	
Model	**Average wAIC**	**Selection Frequency**	**Average wAIC**	**Selection Frequency**
AV	0.000	0.0	0.000	0.0
RD	0.000	0.0	0.000	0.0
RDsym	0.000	0.0	0.000	0.0
RDfC	0.000	0.0	0.000	0.0
RDfI	0.000	0.0	0.000	0.0
AVwM	0.113	2.8	0.118	0.8
AVwMsym	0.000	0.0	0.000	0.0
AVwCM	0.345	49.8	0.327	42.0
AVwIM	0.330	43.0	0.389	56.3
AVtS	0.035	0.0	0.029	0.0
AVtSsym	0.000	0.0	0.000	0.0
AVtSC	0.087	2.3	0.051	0.0
AVtSI	0.092	2.3	0.084	1.0

Table 1.6 Divergence times and their 95% Confidence Intervals estimated in fastsimcoal2 for the three highest ranked models, Ancient Vicariance (AV) with Migration, AV with Coastal Migration, and AV with Inland Migration, under two generation lengths, 6 and 8 years per generation.

		6 years / generation				**8 years / generation**			
	Model	**Original MLE Estimate**	**Median Estimate**	**95% CI**		**Original MLE Estimate**	**Median Estimate**	**95% CI**	
20%	AVwM	7.6E+06	6.5E+06	9.6E+05	1.3E+07	1.0E+07	8.7E+06	1.3E+06	1.7E+07
	AVwCM	1.1E+07	6.2E+06	4.1E+05	1.1E+07	1.5E+07	8.3E+06	5.5E+05	1.5E+07
	AVwIM	6.7E+06	6.9E+06	5.8E+05	1.3E+07	9.0E+06	9.3E+06	7.8E+05	1.7E+07
30%	AVwM	6.6E+06	6.5E+06	1.1E+06	1.3E+07	8.8E+06	8.6E+06	1.5E+06	1.7E+07
	AVwCM	1.1E+07	5.8E+06	4.4E+05	1.3E+07	1.4E+07	7.7E+06	5.9E+05	1.7E+07
	AVwIM	7.4E+06	6.0E+06	4.0E+05	1.3E+07	9.9E+06	8.1E+06	5.3E+05	1.7E+07
	mean		**6.3E+06**	**6.5E+05**	**1.3E+07**		**8.4E+06**	**8.6E+05**	**1.7E+07**

Table 1.7. The average wAIC scores for all thirteen models evaluated in PHRAPL under 7 trees, for all four runs (see Table S3 for grid search values).

	1	2	3	4
AV	0.000	0.000	0.000	0.000
AVwCM	0.011	0.020	0.017	0.015
AVwIM	**0.162**	**0.170**	**0.177**	**0.160**
AVwMsym	**0.109**	**0.101**	**0.106**	**0.101**
AVwM	0.063	0.067	0.050	0.042
RDfC	**0.140**	0.020	0.017	**0.106**
RDfI	**0.133**	**0.170**	**0.179**	**0.150**
RDsym	**0.133**	**0.101**	**0.107**	**0.112**
RD	0.052	0.062	0.048	0.043
AVtSI	0.002	0.018	0.016	0.015
AVtSC	**0.100**	**0.145**	**0.163**	**0.151**
AVtSsym	0.059	0.083	0.086	0.077
AVtS	0.036	0.044	0.033	0.028

Table 1.8. The average wAIC scores for all thirteen models evaluated in PHRAPL under 46 trees, for all four runs (see Table S3 for grid search values).

AV	0.000	0.000	0.000	0.000
AVwCM	0.000	0.005	0.003	0.003
AVwIM	**0.023**	0.081	0.067	0.061
AVwMsym	**0.218**	**0.197**	**0.240**	**0.228**
AVwM	**0.100**	**0.117**	**0.128**	**0.118**
RDfC	**0.156**	0.006	0.003	**0.107**
RDfI	**0.153**	0.079	0.059	0.090
RDsym	**0.168**	**0.258**	**0.241**	**0.192**
RD	0.055	**0.120**	**0.155**	0.089
AVtSI	0.000	0.000	0.002	0.002
AVtSC	0.016	0.024	0.041	0.034
AVtSsym	0.089	0.068	0.044	0.059
AVtS	0.021	0.044	0.016	0.018

Table 1.9. From the left; parameters for the Ancient Vicariance with Coastal Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. Migration is the migration probability for a lineage from the coastal population to migrate to the inland. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation.

jAFS20: Ancient Vicariance with Coastal Migration

Parameters	**MLE**	**Median Estimate**	**95% CI**	
Migration C → I	1.5E-02	1.1E-03	8.8E-09	4.1E-02
Inland *Ne*	2.6E+05	2.6E+05	1.1E+05	2.9E+06
Coastal *Ne*	7.8E+05	2.5E+06	1.2E+05	2.9E+07
Mutation Rate	1.0E-07	9.9E-08	3.9E-09	2.0E-07

Table 1.10. From the left; parameters for the Ancient Vicariance with Inland Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. Migration is the migration probability for a lineage from the inland population to migrate to the coast. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation.

jAFS20: Ancient Vicariance with Inland Migration

Parameters	MLE	Median Estimate	95% CI	
Migration I → C	1.3E-02	5.0E-03	4.1E-08	4.5E-02
Inland *Ne*	1.7E+05	4.7E+05	1.1E+05	9.9E+06
Coastal *Ne*	1.4E+05	2.3E+05	1.2E+05	2.4E+06
Mutation Rate	1.8E-07	1.1E-07	3.7E-09	1.9E-07

Table 1.11. From the left; parameters for the Ancient Vicariance with Asymmetrical Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. The migration parameters are the migration probability for a lineage from one population to migrate to the other. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation.

jAFS20: Ancient Vicariance with Asymmetrical Migration

Parameters	**MLE**	**Median Estimate**	**95% CI**	
Migration I → C	1.1E-06	2.6E-05	2.2E-09	3.6E-02
Migration C → I	5.1E-02	1.4E-06	1.6E-09	1.8E-02
Inland *Ne*	1.7E+05	3.2E+05	1.1E+05	3.6E+06
Coastal *Ne*	1.5E+05	3.8E+05	1.1E+05	2.8E+07
Mutation Rate	1.7E-07	9.9E-08	1.2E-08	2.0E-07

Table 1.12. From the left; parameters for the Ancient Vicariance with Coastal Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. Migration is the migration probability for a lineage from the coastal population to migrate to the inland. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation.

jAFS30: Ancient Vicariance with Coastal Migration

Parameters	**MLE**	**Median Estimate**	**95% CI**	
Migration C → I	1.5E-02	8.7E-04	4.4E-06	3.8E-02
Inland *Ne*	2.6E+05	2.7E+05	1.5E+05	6.9E+06
Coastal *Ne*	7.8E+05	2.3E+06	1.5E+05	5.2E+07
Mutation Rate	1.0E-07	9.5E-08	1.5E-09	1.7E-07

Table 1.13. From the left; parameters for the Ancient Vicariance with Inland Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. Migration is the migration probability for a lineage from the inland population to migrate to the coast. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation.

jAFS30: Ancient Vicariance with Inland Migration

Parameters	MLE	Median Estimate	95% CI	
Migration I → C	1.3E-02	1.3E-03	1.2E-08	4.1E-02
Inland *Ne*	1.7E+05	5.3E+05	1.3E+05	5.8E+06
Coastal *Ne*	1.4E+05	2.5E+05	1.1E+05	3.2E+06
Mutation Rate	1.8E-07	1.0E-07	5.7E-09	1.8E-07

Table 1.14. From the left; parameters for the Ancient Vicariance with Asymmetrical Migration model, the ML parameter estimate used in constructing the 100 parametric bootstraps, the Median ML estimate from the parametric bootstraps, and the 95% Confidence Interval for the ML parameter estimates from the parametric bootstraps. The migration parameters are the migration probability for a lineage from one population to migrate to the other. Inland and Coastal *Ne* are the effective population size estimates for the Inland and Coastal populations. The mutation rate is in substitutions/site/generation.

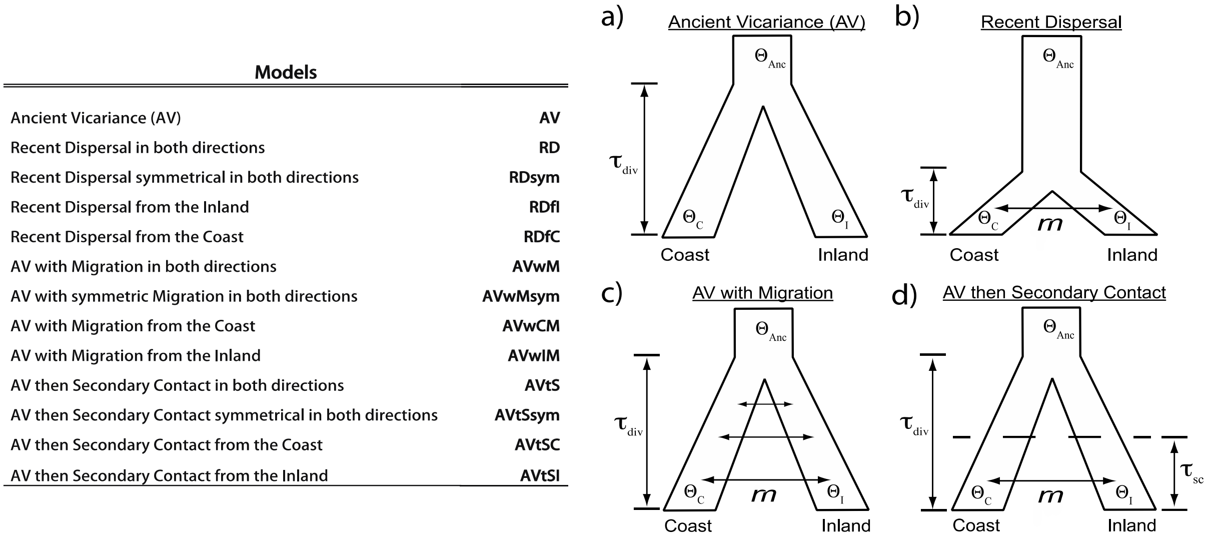
jAFS30: Ancient Vicariance with Asymmetrical Migration

Parameters	MLE	Median Estimate	95% CI	
Migration I → C	2.1E-03	9.0E-05	1.3E-09	4.4E-02
Migration C → I	3.9E-08	1.9E-06	9.2E-10	3.2E-02
Inland *Ne*	1.9E+05	3.0E+05	1.2E+05	4.0E+06
Coastal *Ne*	7.6E+05	4.5E+05	1.2E+05	1.0E+07
Mutation Rate	1.4E-07	1.0E-07	2.0E-08	2.0E-07

Table 1.15. Power analysis results, showing the selection frequency for the four main models, and the average AICw across all 100 optimizations.

	AV	RD	AVwM	AVwSC
Selection Frequency	-	-	86%	14%
Average AICw	-	-	0.703	0.296

Figures



Models	
Ancient Vicariance (AV)	AV
Recent Dispersal in both directions	RD
Recent Dispersal symmetrical in both directions	RDsym
Recent Dispersal from the Inland	RDfI
Recent Dispersal from the Coast	RDfC
AV with Migration in both directions	AVwM
AV with symmetric Migration in both directions	AVwMsym
AV with Migration from the Coast	AVwCM
AV with Migration from the Inland	AVwIM
AV then Secondary Contact in both directions	AVtS
AV then Secondary Contact symmetrical in both directions	AVtSsym
AV then Secondary Contact from the Coast	AVtSC
AV then Secondary Contact from the Inland	AVtSI

Figure 1. Four major demographic models representing the hypothesized phylogeographic history of *Alnus rubra* in the PNW temperate rainforest. In total, 13 demographic models were designed and evaluated for both datasets. Names and abbreviations of all models are listed in the left panel. Parameters estimated include population size, Θ_C and Θ_I, migration rate, μ, divergence time, τ_{div}, and time of secondary contact, τ_s. The only additional parameter estimated but not included in the model design was mutation rate.

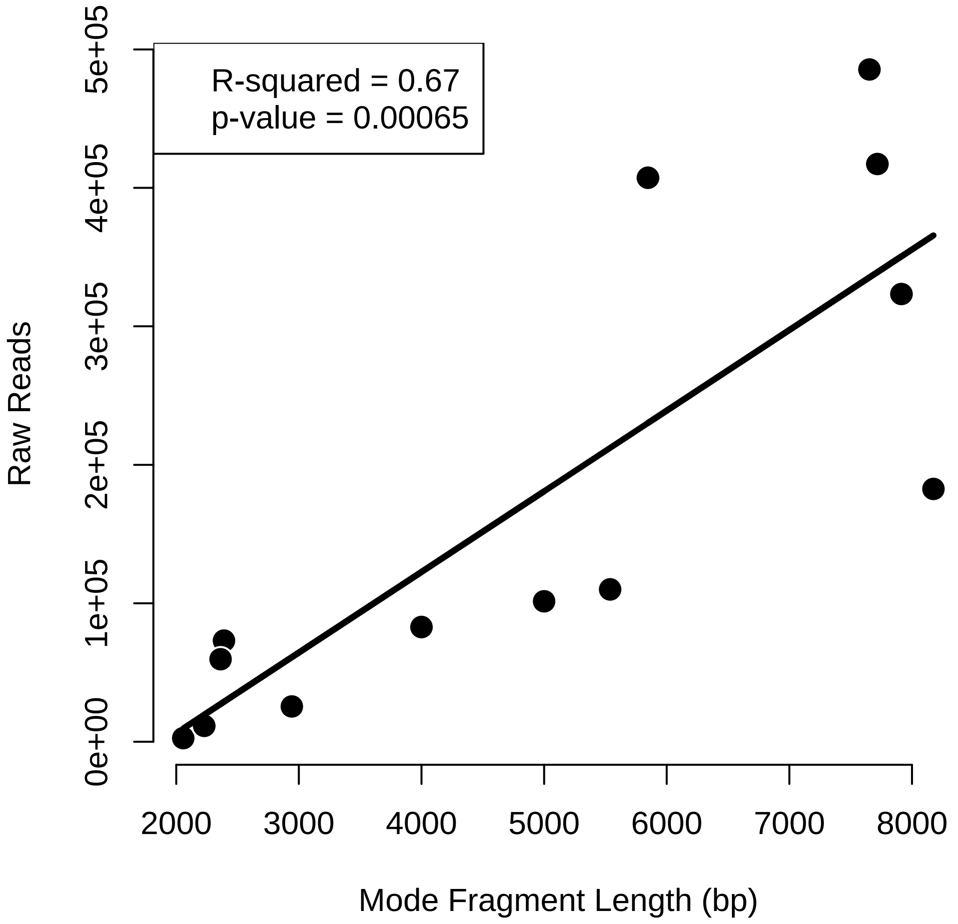


Figure 1.1. Number of raw reads recovered from ddRADseq experiments in relation to the mode fragment length found in the DNA extracts from 13 *Alnus rubra* individuals.

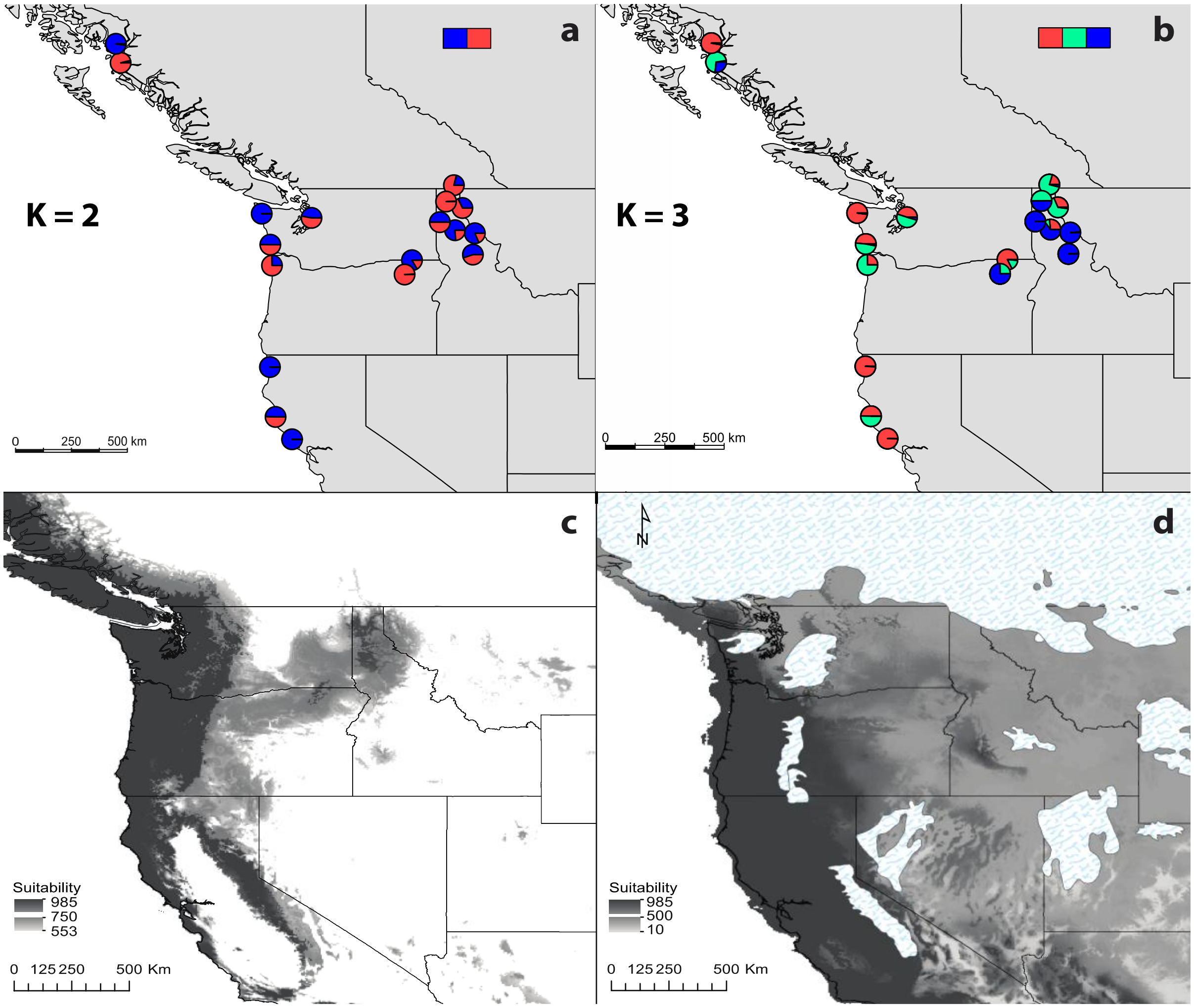


Figure 1.2. STRUCTURE results for K = 2 (a) and K = 3 (b) for 18 populations of *Alnus rubra*. Species distribution models for *A. rubra* in the PNW rainforest under current (c) and last glacial maximum (LGM; d) conditions. The stippled area in the bottom right panel shows the extent of the Cordilleran ice sheet at the LGM.

a

Program	Data	Ancient Vicariance	Recent Dispersal	AV with Migration	AV then Secondary Contact
fastsimcoal2	20%	0.000	0.000	**0.745**	0.276
	30%	0.000	0.000	**0.784**	0.215
PHRAPL	7 trees	0.001	0.362	**0.389**	0.250
	46 trees	0.001	0.414	**0.484**	0.101

b

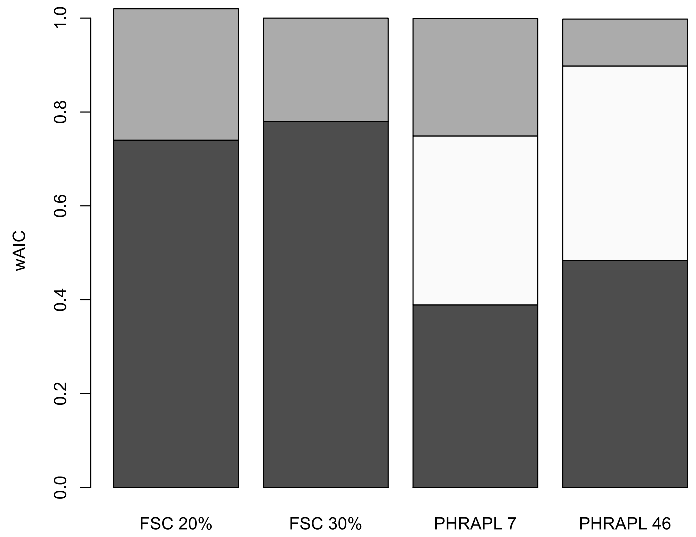


Figure 1.3 a) Model selection results between the four major demographic models. b) Averaged way values across runs in both fastsimcoal2 (FSC) and PRHAPL datasets. White: Recent Dispersal, dark grey: Ancient Vicariance (AV) with Migration, light grey: AV then Secondary Contact. Notes: AV not shown; average wAIC of 0. Because these are averaged wAIC values, they do not sum to one.

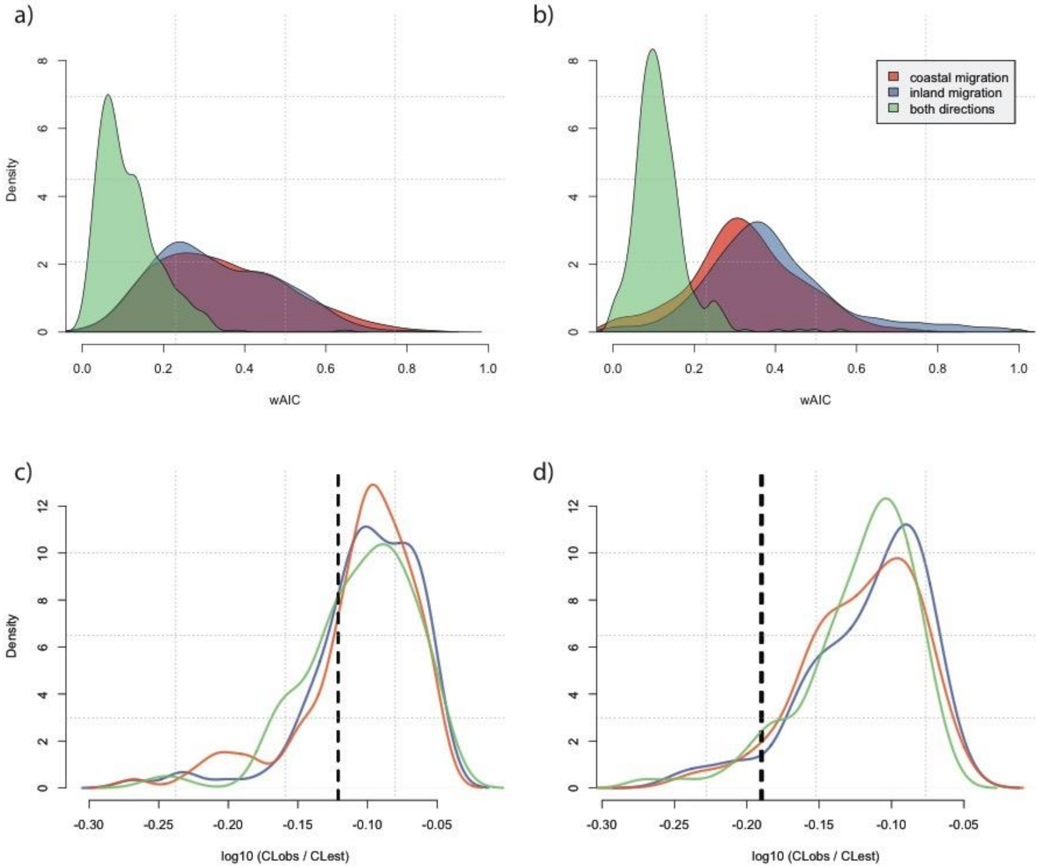


Figure 1.4. Model Selection results from fastsimcoal2 (FSC2) show three models with the highest wAIC densities out of all 13 demographic models evaluated across 400 independent optimizations of the likelihood for each model (20 subsampled datasets at a subsampling threshold of 20% (a) and 30% (b), each optimized 20 times). The three best models represent continuous migration after the ancient vicariance event with varying migration patterns. (c, d) Model adequacy results for the three best models in both FSC2 datasets. Note that in each dataset, the test statistic for all three models are very similar making itappear as if there is only one test statistic (black dashed line), however they are actually three overlapping values. In (d) the three test statistics are more variable, hence the appearance of a thicker line. All six evolutions of model adequacy were non-significant.

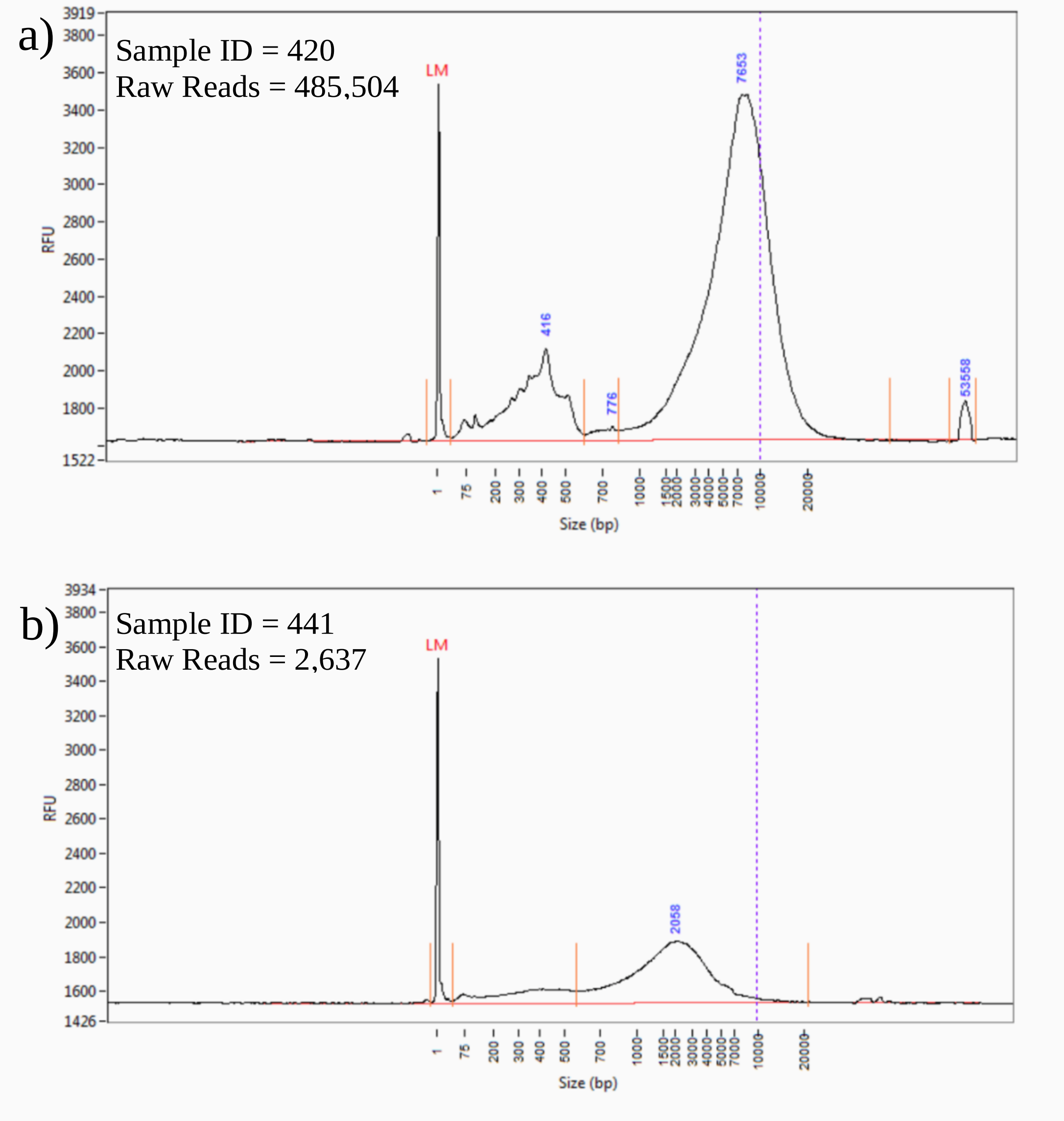


Figure 1.5. a) Distribution of DNA fragment lengths, with a mode fragment length of 7,853 bp, present in Sample 420. b) Distribution of DNA fragment lengths, with a mode fragment length of 2,058 bp, present in Sample 441. Results from a Fragment Analyzer (Advanced Analytical).

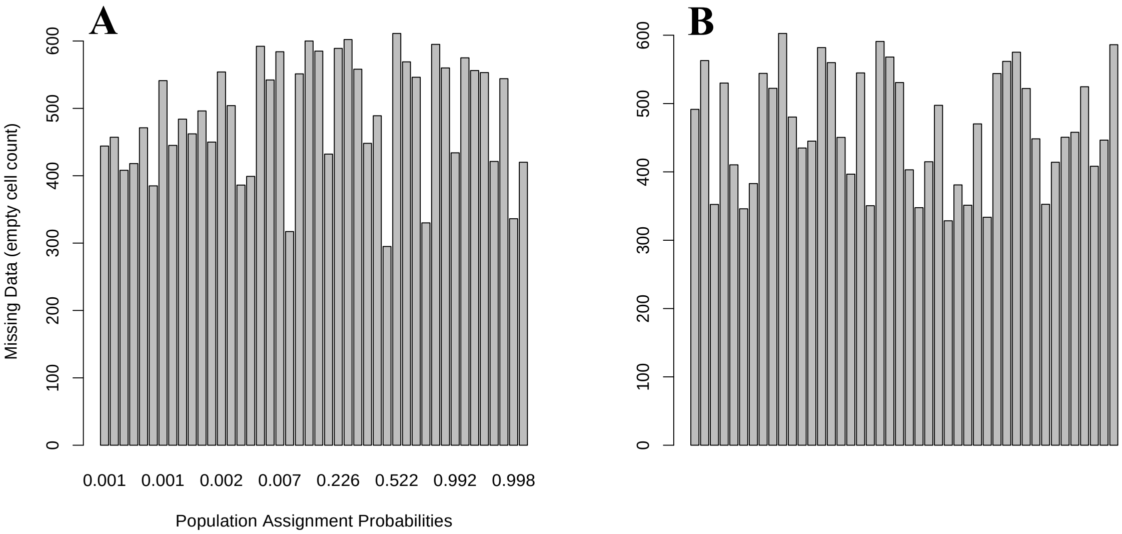


Figure 1.6. A) Missing data quantified across all 49 samples and organized by relatedness using population assignment probabilities from STRUCTURE at K=3. B) Uniform distribution simulated in R using the *runif* function to generate 49 random variables from a uniform distribution with a maximum and minimum bound corresponding to the maximum and minimum missing data value observed across all individuals.

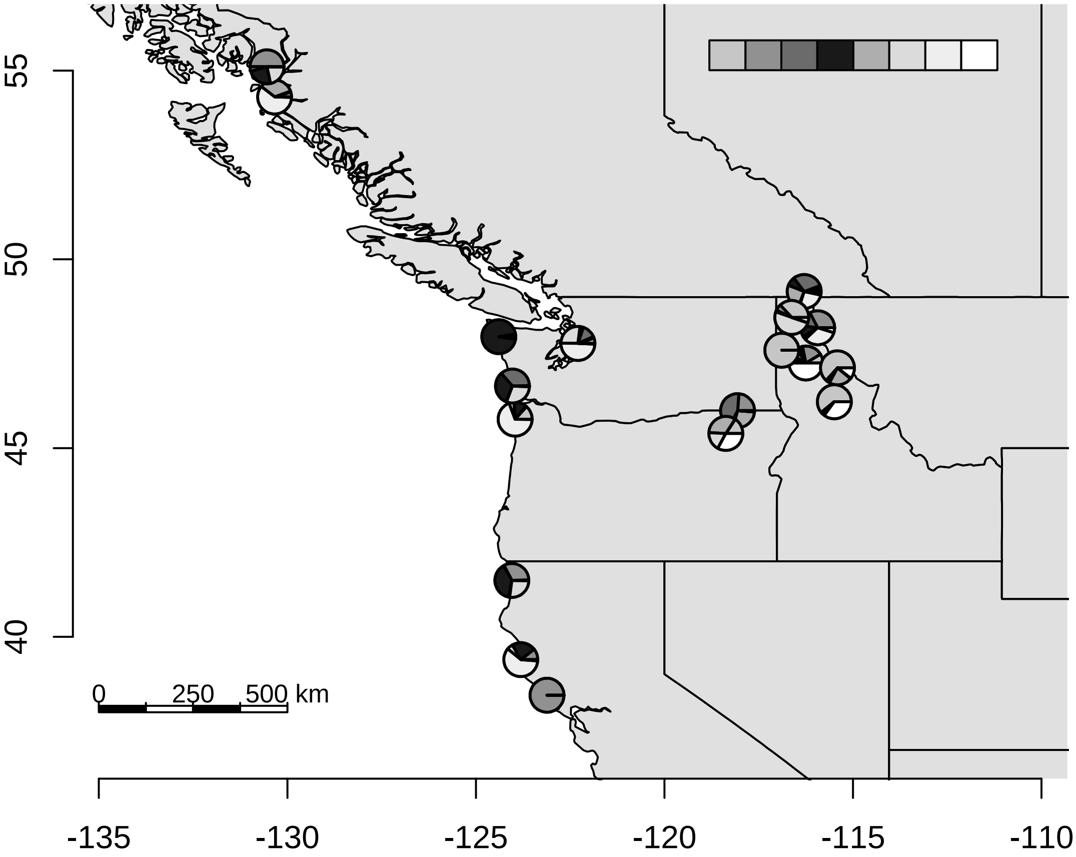


Figure 1.7. STRUCTURE results for K = 8 for 18 populations of *Alnus rubra*; no clear spatial genetic structure present.

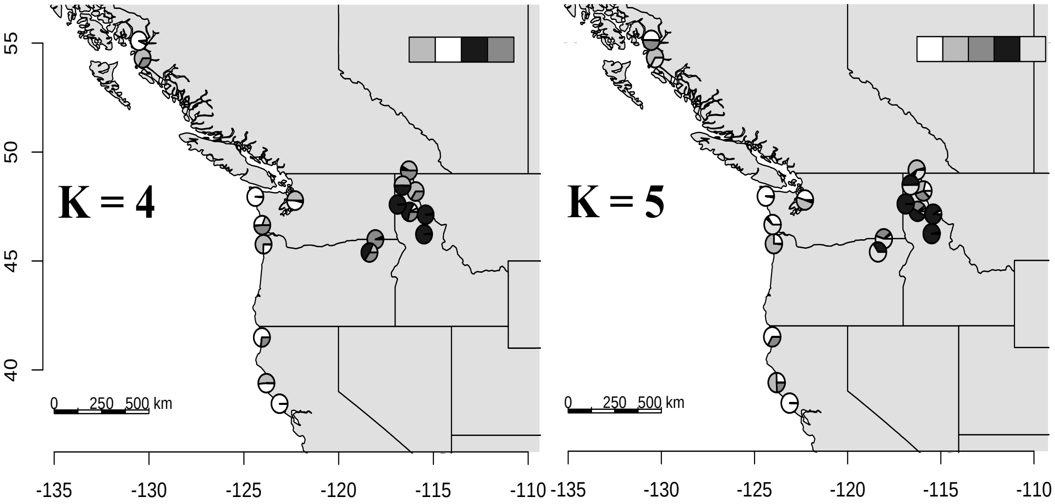


Figure 1.8. STRUCTURE results for K = 4 (left) and K = 5 (right) for 18 populations of *Alnus rubra*; no clear spatial genetic structure present.

Chapter 2: Identifying Models of Trait-Mediated Community Assembly Using Random Forests and Approximate Bayesian Computation

"Identifying models of trait-mediated community assembly using random forests and approximate Bayesian computation" Ruffley, M, Peterson, K, Week, B, Tank, DC, Harmon, LJ. *Ecology and Evolution*, vol. 9, 2019, pp 13218– 13230. https://doi.org/10.1002/ece3.5773

Abstract

Ecologists often use dispersion metrics and statistical hypothesis testing to infer processes of community formation such as environmental filtering, competitive exclusion, and neutral species assembly. These metrics have limited power in inferring assembly models because they rely on often-violated assumptions. We adapt a model of phenotypic similarity and repulsion to simulate the process of community assembly via environmental filtering and competitive exclusion, all while parameterizing the strength of the respective ecological processes. We then use random forests and approximate Bayesian computation to distinguish between these models. We find that our approach is more accurate than using dispersion metrics and accounts for uncertainty. We also demonstrate that the parameter determining the strength of the assembly processes can be accurately estimated. This approach is available in the R package CAMI; Community Assembly Model Inference. We demonstrate the effectiveness of CAMI using an example of plant communities living on lava flow islands.

Introduction

Though methods to infer community assembly vary, many approaches share a central idea based on phylogenetics: the pattern of shared evolutionary history between species that coexist provides insight into the historical processes that assembled the community (Brooks and McLennan 1991; Losos 1996; Grandcolas 1998; Webb 2000; Thompson *et al.* 2001; Webb *et al.* 2002). To gain insight into the assembly process, a collection of metrics have been used to characterize the patterns of diversity in a community using species/genus ratios and other higher taxonomic diversity metrics (Magurran 1988; Faith 1992; Weiher & Keddy 1995; Gotelli & Colwell 2001). Though informative, these patterns often provide little information about the processes that generated them (Peters 1991). Functional traits provide information about diversity and niche space within a community (Macarthur & Levins 1967; Weiher *et al.* 1999; McGill *et al.* 2006), and have long been used to understand resource partitioning between species, as well as coexistence (Cornwell *et al.* 2006; Kraft *et al.* 2007, 2015; de Bello *et al.* 2009). Though the collection and dimensionality of trait data is at times insurmountable, turning to phylogenetic information as a proxy for functional traits was, and is, a viable alternative. Measures of phylogenetic diversity and dispersion, which carry more information than higher taxonomic categories and hopefully, encompass trait information, have become widely used in community ecology to infer community assembly processes (Webb 2000; Webb *et al.* 2002a, 2008; Cavender-Bares *et al.* 2006; Kembel *et al.* 2010; Miller *et al.* 2017). These metrics focus on identifying alternative models of community assembly, environmental filtering and competitive exclusion. Environmental filtering occurs when the abiotic properties of an environment physically keep a species from existing there (Bazzaz 1991). Competitive exclusion describes when species that share the same or similar niche space compete for resources resulting in some species being excluded from the community altogether, also referred to as limiting similarity (Macarthur & Levins 1967). To determine whether non-neutral processes have predominantly influenced assembly patterns, phylogenetic dispersion metrics, such as mean pairwise distance (MPD) and mean nearest-taxon distance (MNTD) – which can be calculated using phylogenetic branch lengths, number of nodal distances, and phenotypic distances – are used to compare observed community dispersion to null expectations (Webb 2000; Gotelli & Colwell 2001; Webb *et al.* 2002a, 2008; Kembel *et al.* 2010).

More specifically, inferences of the assembly process using dispersion metrics are determined in a statistical hypothesis testing framework using several randomly generated null models (Conner & Simberloff 1979; Gotelli & Graves 1996). Commonly, the standard effect size of dispersion metrics, known as net relatedness index (NRI) for MPD and nearest taxon index (NTI) for MNTD (Webb 2000), are used as the test statistic to measure significance of the observed community dispersion compared to null expectations of community dispersion if the community were assembled randomly. However, inference is conditional on the assumption that the relevant phenotypes for the environment or competition are phylogenetically conserved amongst the species in the community, or harbor strong phylogenetic signal within the community of focus. If this assumption is true, and environmental filtering has predominately impacted the assembly process, the phylogenetic data are expected to be significantly clustered, or under-dispersed, in the local community. Likewise, when considering a community assembled by competitive exclusion, we expect to see significantly less shared evolutionary history as compared to null expectations, or significant phylogenetic over-dispersion (Weiher & Keddy 1995; Webb 2000; Cavender-Bares *et al.* 2006).

The dubious assumption of strong phylogenetic signal between the phylogeny and phenotypes is a main critique of these approaches. Kraft *et al.* (2007) showed via simulations that when the assumption of phylogenetically conserved traits was even mildly violated, phylogenetic dispersion metrics were inadequate to infer community assembly processes. Furthermore, this violation of assumptions can, in fact, lead to patterns contrary to those expected for a given assembly process (Weiher & Keddy 1995; Cavender-Bares *et al.* 2009; Mayfield & Levine 2010; HilleRisLambers *et al.* 2012; Gerhold *et al.* 2015). To circumvent this issue, one can assess whether or not functional traits of interest for the community are phylogenetically conserved, and then use that information to guide the inference procedure (Kraft 2007, Kembel *et al.* 2010). Though, if functional trait information is available, it is typically used in consort with phylogenetic information because using phenotypic information alone relies on expectations for how the phenotypes should be distributed in the community to infer non-neutral processes (de Bello *et al.* 2009; Graham *et al.* 2012). While in many instances both phylogenetic dispersion and phenotypic dispersion are measured and analyzed in a similar framework (HilleRisLambers *et al.* 2012), an approach that integrates both to simultaneously estimate support for alternative assembly models is lacking.

Finally, the inference procedure using dispersion metrics relies on statistical hypothesis testing, and therefore, on how well the null model represents neutral expectations. Currently, there exists an extensive number of null models that can be used to infer assembly processes, ranging from simple null models based on random shuffling of taxon labels (Gotelli & Graves 1996; Webb *et al.* 2002; Cornwell *et al.* 2006; Kembel *et al.* 2010), to incredibly dynamic null models (Pigot & Etienne 2015) and analytical frameworks (Stegen *et al.* 2013) that incorporate macroevolutionary processes such as speciation, dispersal, and extinction. There also exist simulation software (Münkemüller & Gallien 2015) to simulate the process of assembly with trait information mediating which species enter the community. However, even with more dynamic null models and simulation power, relying on statistical hypothesis testing and passing a significance threshold to infer an assembly processes is problematic. In part due to the sensitivity between p-values and sample size and how we interpret "significance", but also because each analysis of a particular data type and test statistic result in a measure of significance. Researchers are then responsible for integrating across a suit of hypothesis tests, some that may be significant while others are not, in order to draw an inference. Arguably, a model-based inference procedure is necessary to incorporate all data at once, rank models of community assembly by their relative support, and importantly, incorporate uncertainty in model inference. In this model-based inference procedure, we can simultaneously weigh the support for each community assembly model while also considering both phylogenetic and phenotypic data in the regional and local community. When each model garners a portion of support given the data, we are able to understand when a dominant signal of non-neutral or neutral assembly is present in the data (*i.e.* strong support for one model), when two process are acting simultaneously (*i.e.* strong support for two models), and when the data lack signal to identify a dominant process (*i.e.* equal support across all models).

Several approaches have implemented model-based inference procedures for community assembly already (Van Der Plas *et al.* 2015; Munoz *et al.* 2018; Pontarp *et al.* 2019), paving the way to measuring the relative impact of different processes on community assembly. However, we still lack a method that integrates both phylogenetic and phenotypic information in a species-based model where the strength of the non-neutral processes can be estimated. Here, we develop a stochastic algorithm to simulate communities assembled under environmental filtering and competitive exclusion processes by adapting coevolutionary phenotypic matching and repulsion models. In doing this, we avoid having to make any assumptions about how the traits have evolved along the phylogeny. Our approach simultaneously considers the phylogenetic and phenotypic information from species in the local and regional communities and parameterizes the relative strength of the assembly processes realizing strong to mild non-neutral assembly. Finally, we implement a model-selection inference procedure by using two approximate approaches, random forests (RF; Breiman 2001; Breiman & Cutler 2007) and approximate Bayesian computation (ABC; Csilléry *et al.* 2010). We acknowledge that while these assembly processes are often happening simultaneously in nature, when investigating a targeted trait hypothesized to play a role in the non-neutral assembly of a particular community, the model selection inference procedure holds power to detect the most conspicuous process, if applicable. We are using both model selection approaches because, though RF has been used for model selection in other contexts, it has not been used to distinguish between community assembly models like ABC has (Van Der Plas *et al.* 2015; Munoz *et al.* 2018; Pontarp *et al.* 2019); thus we document a comparison and collaboration of the two approaches here.

We make our approach available as an R package, CAMI, Community Assembly Model Inference (github.com/ruffleymr/CAMI). To demonstrate the effectiveness of CAMI, we use power analyses to show that our approach more accurately infers models of community assembly compared to hypothesis testing using dispersion metrics. We also show that the parameter governing the strength of the assembly processes can be accurately estimated using ABC. Finally, we demonstrate community assembly model inference and parameter estimation using CAMI, with an empirical example from the plant communities that exist on lava flow islands in Craters of the Moon National Monument and Preserve.

Methods and Materials

Community Assembly Models

We focus on three community assembly models: neutral, environmental filtering, and competitive exclusion. For all models, we assume communities are assembled from a regional pool of species where each species in the regional pool is equally likely to colonize the local community. We also assume the phylogenetic relationship between all species is known and that there is continuous trait information for all species. We simulate the assembly of a local community from the regional species pool under one of the three models. Under the neutral model of assembly, all species in the regional community have an equal probability of persisting in the local community (Hubbell 2001; Rosindell *et al.* 2012). The probability that a given species survives, or persists, in a non-neutrally assembled community is not equal for all species, and these varying probabilities of persistence drive the alternative models of community assembly.

To model environmental filtering, we adapt an approach from coevolutionary models (Nuismer *et al.* 2013; Nuismer & Harmon 2015) to relate trait interactions between species and their environment with the probability of surviving in a community. For interactions between species and their environment, we implement a phenotypic matching mechanism where the probability, $P(z_i, z_E)$ of a species persisting in the local community increases when the phenotype of the species z_i and the optimal phenotype of the environment z_E are more similar:

$$P(z_i, z_E) = Exp\left[-\frac{1}{t_E}(z_i - z_E)^2\right] \tag{1}$$

The probability a species with phenotype, z_i, persists in an environment with a phenotypic optimum, z_E, also depends on the strength of the environmental filtering, t_E. When t_E is large, filtering has a mild effect in that species are less penalized for having phenotypes dissimilar to the environmental optimum; whereas when t_E is small, the filtering effect is stronger because species are heavily penalized for phenotypes dissimilar to the optimum.

To model competitive exclusion, the probability, $P(z_i, \bar{z})$, of a species persisting in the local community increases as the phenotype of the species z_i and the mean phenotype of the local community $\bar{z}$ are more dissimilar.

$$P(z_i, \bar{z}) = 1 - Exp\left[-\frac{1}{t_C}(z_i - \bar{z})^2\right] \tag{2}$$

Here, the probability a species with phenotype, z_i, persists in a community with mean phenotypic, $\bar{z}$, depends on the strength of competition between species, t_C. When t_C is large, competition has a strong effect in that species are heavily penalized for having phenotypes similar to the mean phenotype of the local community. When t_C is small, competition is weaker in that species are less penalized for having a phenotype similar to the mean phenotype of the community.

2.0 Data Simulation

For a single simulation of community assembly, first, a regional community phylogeny is simulated under a constant birth-death process with speciation, λ, and extinction, μ, parameters, until the desired number of regional species, N, is reached (Figure 2.1; Stadler 2011). Traits are evolved on the regional phylogeny, one for each species, (Revell 2012) under either a Brownian Motion (BM; Felsenstein 1985) or Ornstein-Uhlenbeck (OU) model of trait evolution (Hansen 1997; Butler & King 2004) characterized by the rate of character change, σ^2, and, for OU models, the "strength of pull" to the trait optimum, α (Figure 2.1). Traits evolve under BM in a way that mimics drift over macroevolutionary timescales and OU does the same only it includes a selective regime in which traits are "pulled" toward a phenotypic optimum. We simulate under these different models of trait evolution because they do not enforce the assumption that trait differences are correlated to phylogenic differences and create more variability in how the data behave under the assembly models. Once the regional community exists with phylogenetic relationships and trait information, the assembly of the local community can begin.

The assembly process uses the probabilities of species persisting in local communities, $P(z_i, z_E)$ for environmental filtering and $P(z_i, \bar{z})$ for competitive exclusion, and a rejection algorithm to stochastically assemble the local community. When simulating under a competition model, the strength of competition between species, t_C, parameterizes the assembly process. Likewise, under an environmental filtering model, the strength of the environmental filter, t_E, along with the environmental phenotypic optimum, z_E, parameterizes the assembly process. For the investigative simulations, the phenotypic optimum is determined by a random draw from the simulated trait distribution of the regional community, and it remains constant throughout an entire simulation.

When a species colonizes the community, the probability of persistence is calculated, and the species is included in the local community if that probability is greater than a uniform random number between 0 and 1 (Figure 2.1). Otherwise, the species is rejected from being in the local community. This stochasticity included in the algorithm is more apparent in the emergent data when the ecological strength parameter is imposing weak non-neutral assembly. When a species is rejected from entering the community, it remains in the regional pool and is still able to colonize the local community again. In this case, the probability of persistence is recalculated, and the species has another chance to pass the rejection algorithm. As in the neutral model, the assembly process ends when the local community has reached species richness capacity, n.

All parameters mentioned are either fixed or drawn from a prior distribution. Information regarding the default prior distributions and fixed values for each parameter can be found in Table 2.3 or in the help documentation for the R package 'CAMI' (github.com/ruffleymr/CAMI). Any parameter mentioned, along with prior distributions, can also be set by the user. In simulations described here, the default prior distributions were used unless otherwise stated.

3.0 Inference Procedure

For a single simulation of community assembly, a regional and local phylogeny and a regional and local distribution of trait values is returned. This information is summarized in 30 different summary statistics that capture information about the phylogeny, trait distributions, and phylogenetic signal within the traits of the local community (Komsta & Novomestky 2015, Janzen *et al.* 2015; Pennell *et al.* 2015; Deevi *et al.* 2016, Kendall *et al.* 2018, Paradis & Schliep 2018; Table 2.4). These summary statistics are then used for model selection and parameter estimation.

To predict model probabilities from empirical data, we used two model selection approaches. The first approach uses a machine learning classification algorithm, random forests (RF; Breiman 1999; Liaw & Wiener 2002), to build a 'forest' of classification trees using the simulated summary statistics as predictor variables and the community assembly models as response variables. As a classifier is being built, RF is simultaneously measuring the 'Out of Bag' (OoB) error rates of the classifier by cross validating each classification tree with a subset of the original data that was not used to make the tree in question. The OoB error rates measure how often the data are incorrectly classified. Additionally, RF quantifies the effect of including each summary statistic on the accuracy of the classifier through two variable importance measures, Mean Decrease in Accuracy (MDA) and Mean decrease in Gini Index (GINI) (Breiman 2002).

RF is generally robust to noisy and/or overpowering predictor variables because each tree in the forest is constructed with a random subset of the data and predictor variables (Breiman & Cutler 2007), which reduces the correlation amongst the trees while still improving the overall predictive power of the forest. The second approach, ABC, when using the rejection algorithm, relies on the Euclidean distance between observed and simulated summary statistics to accept simulations into the posterior probability distribution of the models given the data (Csilléry *et al.* 2010). The support for each model then comes from the proportion of simulations from each model accepted into the posterior probability distribution. If there are summary statistics included that add a lot of noise to the classification process, ABC will lose power in distinguishing support between models. As mentioned, RF is able to measure which summary statistics are the most influential in distinguishing between the models, through importance measures such as MDA and GINI. We used this information to select a subset of 10 summary statistics to be used in ABC model selection, along with a tolerance of 0.001 (Csilléry *et al.* 2012). The performance of ABC in classifying the data can be measured using a cross validation approach for model selection which results in model misclassification rates for each model.

4.0 Power Analyses

We compared the accuracy of three approaches in identifying community assembly models from the data simulated under the three community assembly models in CAMI. The first approach follows previous work and uses dispersion metrics, such as MPD and MNTD (standardized as NRI and NTI), in statistical hypothesis testing to infer the community assembly process from phylogenetic and phenotypic information, separately (Webb 2000; Cornwell *et al.* 2006; Kembel *et al.* 2010; Kraft & Ackerly 2010). For MNTD calculated using phenotypic information, the nearest neighbor is the species closest in trait space (Ricklefs & Travis 1980; Graham *et al.* 2012; Swenson *et al.* 2012).

The second and third inference approaches are the approximate model selection techniques used in CAMI, RF (Breiman 1999; Liaw & Wiener 2002) and ABC (Toni *et al.* 2009; Csilléry *et al.* 2010, 2012). We measured the power of each approach in correctly classifying community assembly data (see sections 1.0 and 1.0) through the OoB error rates for RF and model cross validation for ABC. We performed these power analyses for a range of community sizes to assess whether the power of any of the approaches increased with sample size of the regional/local community, which in this case is species richness. For data to classify, we simulated 1,000 datasets in CAMI under each community assembly model for 20 different regional community sample sizes ranging from 50 to 1000, increasing by increments of 50, with the local community always half the size of the regional. For more details on each of the model identification techniques, refer to supplemental methods section 1.

We also investigated whether RF and ABC can be used to accurately infer the model of community and trait evolution simultaneously. For this, we performed the power analysis as described above, only here we classified six models (neutral, filtering, and competition models under both BM and OU models of trait evolution) rather than just the three community assembly models.

5.0 Parameter Estimation

We measured the ability of the ABC approach to estimate the strength of the assembly process, t_E and t_C, under non-neutral models of community assembly, environmental filtering and competitive exclusion. For both models, we attempted parameter estimation when the traits were simulated under a BM and an OU model of trait evolution. We also attempted parameter estimation for two sizes of regional communities, 200 and 800, with corresponding local community sizes of 100 and 400. We simulated 50,000 community assembly datasets under each condition to serve as the reference dataset for parameter estimation. For details on these simulations, reference the supplemental methods section 2.

We simulated 100 datasets each for 13 different values of t_E and t_C, ranging from 1 to 60 in increasing increments of 5 (see supplemental methods section 3 for other parameter details). These simulated datasets would serve as the "observed" datasets to use for parameter estimation, in which case we know what the true value of t_E and t_C are. To measure not only how accurately t_E and t_C are estimated, but whether all values can be estimated accurately, we performed parameter estimation in ABC for each of the simulated datasets with a rejection algorithm and a tolerance of 0.001. For this, we assumed that data simulated under environmental filtering and competitive exclusion models were correctly classified as those models. We repeated this procedure increasing the sample size of the regional and local community to measure whether t_E and t_C estimates improved with increased sample size.

6.0 Empirical System

Craters of the Moon National Monument and Preserve (CRMO) is a volcanic landscape in southern Idaho. The overlapping basalt lava flows formed along vents in the Great Rift between 2 – 15 KYA (Kuntz *et al.* 1982, 1986). Within the lava flows are kipukas – islands of vegetation that are completely surrounded by uninhabitable lava (Vandergast & Gillespie 2004). Given their isolated nature and recent colonization, the plants on kipukas are an ideal system for studying community assembly. We opted to use maximum vegetative height as our functional trait of interest because it is known to be an important proxy for resource partitioning and competitive ability in plants (Westoby 1998; Weiher *et al.* 1999; Cornwell *et al.* 2014).

The regional phylogeny was constructed for 113 species that occur in the CRMO by dropping non-CRMO species (79,768) from a Spermatophyta phylogeny (Smith & Brown 2017). Likewise, the local community phylogeny was constructed by dropping non-kipuka community species from the regional phylogeny, resulting in 63 local species (Table 2.10). If a particular species needed was not in the Spermatophyta phylogeny, we used a random relative in the same genus as a replacement (Qian & Jin 2016). In addition to the total local species pool on the kipukas, we also investigated eight kipukas individually, kipukas that consisted of 18-20 species from the local community (Table 2.10). Maximum vegetative height data for all species in the regional and local community were gathered using a combination of herbarium records, species descriptions, and floras (*e.g.* Hitchcock & Cronquist 2018).

To assess whether an assembly process has structured the plant community on kipukas, we used NRI and NTI calculated from both phylogenetic and phenotypic (maximum vegetative height) information, separately, and CAMI using RF and ABC to perform model selection. We also performed parameter estimation using ABC to understand what the influence of t_E or t_C was on the assembly processes in either the filtering or competition models, should they be highly supported. For more details regarding the empirical data analysis, including plant collections and data simulated for the analysis, refer to the supplemental methods sections 3.

Results

4.0 Power Analysis

The average proportion of misclassified simulations using the standard approach of phylogenetic dispersion metrics for all regional/local community sizes was 56 % (Table 2.1), decreasing from 63.3 to 52.9 % with increasing sample size (Figure 2.2, Table 2.5). For each of the community assembly models, the average misclassification rate for each model was consistent between MPD and MNTD (Table 2.1) when using phylogenetic information. When calculating these metrics from phenotypic information, the average misclassification rate varied depending on whether MPD or MNTD was being used, with MPD having a very low error rate, 3.9 %, and MNTD a high error rate, 48 % (Table 2.6).

Average error rates for both of our model selection approaches were substantially lower. The average random forests OoB error rate when classifying community assembly models was 2.6 %, ranging from 16.7% for small communities to 1.5 % for large communities (Figure 2.2). The average OoB error rates for each community assembly model with RF were 3.8%, 2.0 %, and 1.9 % for neutral, filtering, and competition models, respectively (Table 2.1). The average ABC model misclassification rate was 8.47 % (Table 2.1), ranging from 20.9 % for small communities to 5.9 % at large communities (Figure 2.2). The average ABC error rates for each community assembly model were 5.4%, 13.6%, and 6.32 % for neutral, filtering, and competition models, respectively (Table 2.1).

Using RF and ABC to classify models of community assembly and trait evolution simultaneously resulted in overall higher error rates compared to inferring community assembly alone (Fig. 2.5). On average, the average OoB error rate for RF was 23.2%, ranging between 45.7% and 16.2% from small to large communities (Table 2.7), and the overall error rate for ABC was 30.7 %, ranging between 50.8 % and 23.5 % from small and large communities (Table 2.8).

5.0 Parameter Estimation

For all models, the simulations with larger community sizes better estimated the true value of t_E and t_C compared to communities of smaller size (Figure 2.3). Regardless of sample size, t_C was overestimated when of smaller value. In both filtering and competition models, t_E and t_C are slightly underestimated when of larger value – though this is due to the true value of t_E and t_C being at the upper bound of the prior distribution, which if extended would not be apparent.

6.0 Empirical System

Several dispersion metrics used from phylogenetic and phenotypic information identified significant under-dispersion, or clustering, amongst plant species in the kipukas, suggesting a community assembly pattern of environmental filtering. When calculating NRI and NTI using phylogenetic information from all plants in the kipukas, the resulting p-value was 0.02 for MPD and 0.29 for MNTD. When calculating the same metrics from phenotypes, the resulting p-value for each test statistic was 0.03 and 0.01, respectively (Table 2.9). For the eight separate kipuka communities, only MPD using phylogenetic information identified two other community as significantly under-dispersed (Table 2.9).

We constructed two RF classifiers to make predictions about empirical data. One classifier was built with simulations from both trait models and the other classifier was built with data simulated only under an OU trait model. This OU models-only RF classifier was built because the trait data for the kipuka plants better fit an OU model of trait evolution compared to a BM model (see supplemental methods 4). The OoB error rates for these two classifiers were 25.50 and 23.61 %, respectively. We also estimated the error rate when using ABC in the same way as with RF. For these, the error rate for each cross-validation was 33.20 and 30.40 %. Using these data and approaches, we predicted the model of community assembly for the empirical data with RF and ABC, and saw a majority of support for environmental filtering, with the second highest support for the neutral model (Table 2.2 OU model-only prediction, Table 2.13 for OU and BM model predictions).

We performed parameter estimation of t_E for the environmental filtering model for each dataset under an OU model of trait evolution (Table 2.14). Each time 100 simulations were accepted as from the posterior distribution of t_E (Figure 2.4). We also compared the amount of model support for the environmental filtering models with the median estimate of t_E (Figure 2.6, Table 2.14) to show the relationship between the strength of the filtering process and the model support received.

Discussion

Performance of CAMI

Using CAMI, we can correctly classify models of community assembly and importantly, quantify the uncertainty associated with community assembly model inference. This approach improves upon current methods in community phylogenetics by harnessing the critical information present in the phenotypic and phylogenetic data that directly relate observed patterns to processes. Our approach is successful, in part, because over and under-dispersion in the phylogenetic and trait data are emergent properties of the community assembly models described. Through our method, we can control the processes that directly impact the amount of over and under-dispersion in the phenotypic data, along with their degree of association with the phylogenetic information. Furthermore, our inference pipeline is unique in allowing users to gauge or rank evidence for both neutral and non-neutral assembly processes.

The performance of RF and ABC are comparable in that they both accurately classify the community assembly models. A benefit to using RF is that all of the summary statistics from the simulated data can be used without compromising the power or computational speed of the method. Additionally, RF measures how important each summary statistic is for classifying the data accurately. While we don’t use this information for any additional community assembly inferences here, there is potential to ask which summary statistics play an important role in these assembly processes, and further, whether there are any biological implications to gain from that information. The main advantage of using ABC is that parameter estimation is straight forward using simulated data, and this is particularly relevant for estimating the strength of non-neutral assembly via t_E and t_C, though parameter estimation using RF is increasingly common.

The predictive approaches outlined here are not meant to replace dispersion metrics, but rather to be used as an additional tool in making inferences about community assembly. We have shown here, as others have (Kraft *et al.* 2007), that dispersion metrics are not reliable in determining models of community assembly with phylogenetic information alone. When using phenotypic data though, MPD proved to be comparable in accuracy at distinguishing community assembly models to RF and ABC; MNTD still had very high error rates (Table 2.1).

Though CAMI is currently implemented using one trait, the analyses do not necessarily need to be limited to one trait. If there are several traits of interest in a particular community, data dimension reduction techniques could be used, such as principle components or linear discriminate analysis, to associate each species with a singular value representing where they fall in trait space with respect to other species in the community. Though we do not explore the power of inferring models of community assembly from several traits defined in one composite dimension through simulations, we expect, to some degree, that the method will behave as presented above in the single-trait case. Using multiple traits in a true multivariate framework, which we have not implemented, could make for an even more powerful inference, as many factors influencing community structure could be measured at once (Weiher *et al.* 1998; Herben & Goldberg 2014; Kraft *et al.* 2015). However, if multiple traits are being considered, there also need be the consideration that there could be multiple phenotypic optima or complex routes of competition between species, and here we consider the presence of only a single optimum and equal competition amongst species (Weiher *et al.* 1998).

While we feel CAMI will continue to make progress in advancing our understand of community ecological patterns globally, there are still many aspects of community ecological theory yet to be incorporated (Belyea & Lancaster 1999; Weiher *et al.* 2011). The assembly models defined here could be made more powerful by considering other community dynamics such speciation, colonization, and extinction during the assembly process (Rosindell & Harmon 2013), as well as co-occurring and structured non-neutral processes (Keddy & Shipley 1989) where the relative importance of these processes can be measured (as in Van Der Plas *et al.* 2015; Munoz *et al.* 2018). These aspects may be more or less relevant depending on the taxonomic scale of the community being investigated (Weiher *et al.* 2011). Furthermore, the inference power could expand by making CAMI an individual-based model of community assembly (Rosindell *et al.* 2015; Pontarp *et al.* 2019), where individuals can diverge to speciate and harbor intraspecific diversity amongst phenotypes (Jung *et al.* 2010, 2014), all while abundance distributions and population demographics are being tracked (HilleRisLambers *et al.* 2012; Overcast *et al.* 2019). A spatially explicit model (see Pontarp *et al.* 2019) could allow for the exploration of how geography, or even local topography, impacts the assembly process. Ultimately, we believe this approach has the capability of being extended to incorporate many more complexities known to influence and emerge from the assembly process.

Inferring the Strength of the Assembly Process

Parameterizing the strength of the assembly process provides an additional mode of inference for the relative strength of the non-neutral community assembly processes, environmental filtering, t_E, and competitive exclusion, t_C. We have shown that ABC can be an appropriate tool to estimate both t_E and t_C accurately (Figure 2.3) for their respective community assembly models. We have also shown that empirical data, from different communities, do indeed bear some signal to indicate different magnitudes of t_E (Figure 2.4). Additionally, we show that the estimate of t_E has a relationship with the amount of support the corresponding non-neutral model receives, in this case, the environmental filtering model. We know that for filtering models, the smaller the value of t_E, the stronger the effects of filtering, thus the smaller the estimate of t_E , the greater the model support for environmental filtering (Figure 2.6). Having this measure that can quantify the influence of the assembly process at play opens the door for comparisons of communities globally that have been assembled by the same mechanism (Götzenberger *et al.* 2012). Prior to now, if multiple communities were inferred to be assembled via environmental filtering, there was no way to ask whether one environment's pressure was stronger relative to the other, while t_E and t_C now permits these questions.

Models of Trait Evolution

Identifying models of community assembly alone was much more successful than when trying to simultaneously identify models of trait evolution, as shown by the increase in error rates (Figure 2.5). When the model of trait evolution is identifiable, as in many BM and OU cases, simulating under both models is not necessary as it drastically increases the amount of simulations needed. Information about the best fit trait model, including parameter estimates, can be used to directly inform parameters used to simulate community assembly data in CAMI (as in the empirical study here). However, we do show that considering both models of trait evolution simultaneously versus only one at a time does not drastically change the community assembly inference (Table 2.13). Thus, should one be unable to properly, or with confidence, estimate the true model of trait evolution, the combined inference procedure in CAMI is appropriate, and this may be especially useful for early-burst or multi-optima OU models of trait evolution (Slater & Pennell 2013; Uyeda & Harmon 2014). We should note here that a model of trait evolution fit to community data, phylogenetic and phenotypic, involves excluding many taxa from the tree and trait distributions that would otherwise be included in phylogenetic comparative methods. This means the parameter estimates cannot be tied to the entire evolution of a particular trait, but rather its evolution amongst a certain set of species within a community.

Empirical Inference

When using CAMI to distinguish models of community assembly, a majority of support reliably goes to the environmental filtering model when considering the entire local kipuka community, with some support garnered by the neutral model (Table 2.2). When looking at the eight separate kipuka communities, the environmental filtering model still receives a majority of the support, but there is quite a lot of support garnered for the neutral model as well, and sometimes even for the competitive exclusion model (Table 2.2). Conveniently though, when comparing the model probability estimates with the t_E estimates, we get a better understanding of why the model support is where it is for a particular kipuka and that the t_E parameter is being estimated appropriately (Figure 2.6). Essentially, when t_E is representing weaker filtering effects, which corresponds to higher values of t_E, we see lower support for the filtering models.

When using dispersion metrics to distinguish models of community assembly, the reliability is less apparent. Many of the observed dispersion metrics fall at the lower ends of the random distribution of dispersion indices, and subsequently result in low p-values. However, one of the caveats of hypothesis testing is that there is a sort of arbitrary cutoff between when something is significant and when it is not that is predetermined by the user. In this case, technically the cutoff is 0.025 and so only four out of 36 metrics were significant. These issues are generally overcome with intuition because it is obvious some of the p-values are still very low, but they do highlight problems with hypothesis testing and relying on p-values for marks of biological significance.

For each kipuka species pool, the strength of the filtering process was estimated quite differently. For the entire species pool of the kipukas, the t_E estimate was a relatively moderate value, 15.4, given the prior range of 1 to 60, where values near 1 imply strong filtering, and values closer to 60 imply weak filtering. For other kipuka communities though, t_E was often a moderate estimate, falling somewhere in the middle of the prior distribution, though sometimes the estimate was very low (Figure 2.4D-E) and other times, quite high (Figure 2.4I). We recognize though that any interpretation of t_E is challenging because the parameter has never before been measured using any community or trait before. Thus, we expect with continued investigations of community data using CAMI we will decipher a sharper picture on how t_E behaves across many natural communities. These estimates are a start to that investigation given their correspondence with the model probabilities (Figure 2.6). We should note that in the case of these t_E estimates, the rate of character change is so low that a strong effect of filtering with that little phenotypic variation may be harder to detect than if more variation were present. Similarly, the estimates of t_E are less reliable when the community size is small (Figure 2.3), which is true in the case of these kipukas.

One anecdotal explanation for the support for the environmental filtering assembly model lies in the structure of the kipukas. Lava flow builds up on the edges of the habitable land on the kipuka forming a sort of "bowl," with the plant community inside the bowl. Species that generally grow taller than the bowl edges are less protected from heavy wind speeds common in the area and are more likely to be filtered from the environment. Likewise, with high wind speed comes a likely increase in dispersal ability for some species in the regional pool, which may explain the support of the neutral model. However, even though we can speculate on the cause for the support of an environmental filtering model acting on height in the kipukas, we still lack evidence of the true cause of the support, or mechanism of filtering.

While vegetative height has been hypothesized to play an important role in community structure, as a functional phenotype and a proxy for other important traits (Cornwell *et al.* 2014), because we only take into account a single functional trait, we recognize the potential limitations to these inferences. The CAMI framework permits testing multiple traits independently and comparing the evidence across how each trait influenced community assembly to better understand the historical and contemporary assembly processes (Herben & Goldberg 2014). Additionally, each trait, if influencing community assembly in a non-neutral way, will be associated with an estimate of t_E or t_C, which will also provide insight into the degree that each trait influences the assembly process for a particular community.

Conclusion

CAMI is a new approach able to estimate the probability of neutral and non-neutral community assembly models given observed phylogenetic and phenotypic information. By harnessing the power of simulations and approximate approaches for model selection, such as RF and ABC, we can quantify uncertainty in community assembly inferences. Additionally, new parameters described here, t_E and t_C, govern the strength of environmental filtering and competition models, respectively, and are estimable with empirical data. Defining the non-neutral assembly models and parametrizing the processes to mimic strong to mild assembly dynamics will add to what we know about communities that have been assembled via the same mechanisms. While there are other approaches that infer community assembly in a model-based framework (Van Der Plas *et al.* 2015; Munoz *et al.* 2018; Pontarp *et al.* 2019), CAMI offers a unique opportunity to use information that is readily available in phylogenetic community ecology. Given these data are common for community assembly studies, this framework could be readily applied to many existing systems and ultimately provide information about the patterns of community

Supplemental Information

Supplemental Methods

1.0 Prior Ranges for Data Simulation

For the data used in each model classification approach, for the power analyses, we simulated 1,000 datasets under each community assembly model and trait evolution model (6 models), for 20 different regional community sample sizes ranging from 50 to 1000, increasing by increments of 50, for a total of 120,000 simulations. In each of these simulations, the local community size, n, was exactly half the size of the regional community, N. The speciation, λ, extinction fraction (equal to μ/λ), rate of character change, σ^2, and, for OU models, strength of phenotypic pull, α, parameters were all drawn from a default prior distribution that was consistent for all simulations (Table 2.3). Default priors were selected because we believed they were reasonable representations of where community phylogenetic and functional trait data exist currently. Simulations of non-neutral models included the strength of assembly, t, parameter that is drawn from a uniform prior distribution between 1 and 60. This distribution roughly bounds the probability of persistence, P_{z_i,z_E} for environmental filtering and $P_{z_i,\bar{z}}$ for competitive exclusion, at a minimum of 0.001 and a maximum of 0.99. These minimum and maximum bounds will vary slightly for each simulation depending on σ^2 and, in the case of OU models, α.

2.0 Power Analyses: model identification techniques

2.2 Dispersion Metrics – For the first model identification technique, NRI and NTI (Webb et al. 2002, 2008; Kembel et al. 2010) were used to measure phylogenetic dispersion. Null distributions for each test statistic were generated by shuffling the taxon labels of the phylogenetic distance matrix (Webb et al. 2008; Kembel et al. 2010). We assumed the traits were phylogenetically conserved, thus if either test statistic were significantly greater than expected, i.e. p-value ≥ 0.975 and phylogenetic over-dispersion, competitive exclusion was inferred. If either test statistic were significantly lower than expected, i.e. p-value ≤ 0.025 and phylogenetic under-dispersion, environmental filtering was inferred. If both test statistics were non-significant, a neutral community assembly model was inferred. We totaled the number of correctly classified simulations to calculate the proportion of misclassified simulations and report the overall error rate of the approach.

We also calculated the standard effect size of MPD and MNND (in the same way NRI and NTI are calculated) to measure phenotypic dispersion. For this, instead of using phylogenetic distances, distances in phenotypic space were used to calculate MPD and MNND, along with the null distributions (Cornwell et al. 2006; Kraft & Ackerly 2010). We made the same assumptions about the traits being phylogenetically conserved and therefore followed the same inference procedure as with phylogenetic information.

2.3 Random Forests

When performing model selecting using RF, all 30 summary statistics (Table 2.4) we used in constructing a ‘forest’ of 1000 decision trees. Each tree classifies the simulated data as from either a neutral, environmental filtering, or competitive exclusion model of community assembly, essentially each tree casts a “vote” for a model. The proportion of votes a model receives is considered the probability of that model, meaning the model with the most votes, or highest probability, is the model selected. Each decision tree is constructed using the summary statistics from a random 2/3 of the data, while the remaining 1/3 of data, or out-of-bag (OOB) data, is used to cross-validate the accuracy of each decision tree. Through this, the error rate of each model is estimated and summarized in the OOB error rates. These error rates indicate the overall accuracy of the forest in distinguishing between the models.

2.4 Approximate Bayesian Computation – To determine the overall model misclassification rate when using ABC for model selection, we performed 500 cross-validation attempts for each community assembly model for each of the 20 regional community sample sizes. For each cross-validation, the rejection algorithm was used with a tolerance of 0.01. The total proportion of simulations that were incorrectly classified for each model were averaged as an overall error rate for each regional community sample size. Mean posterior probabilities were calculated amongst the correctly classified models as an additional measure of how much support the correct model received.

3.0 Reference Data Simulation for Parameter Estimation

We simulated 50,000 community assembly datasets under each condition to serve as the reference dataset for parameter estimation. In these simulations, the regional and local community sizes were fixed using the values stated above. A uniform prior was set on σ^2 to be between 2 and 4, which is narrower than the default uniform prior between 1 and 10. This was done because if one has the empirical phenotypic data, they are typically able to estimate a rate of trait evolution under BM or OU models of trait evolution, and would thus have more information for this parameter and not need to rely on the default uniform distribution. The remaining parameters $t, \lambda, \mu,$ and α were drawn from their default uniform prior distributions.

4.0 Empirical System

Voucher plant specimens were collected during the summer of 2016 and 2017 across 25 kipukas throughout the CRMO. In total, 63 unique plant species were documented as occurring in these CRMO kipukas, and these species represented the total local plant community. To assemble the regional community species list, we referenced a checklist of the vascular plants that occur in the CRMO (Popovich 2006). We selected an additional 50 species on the basis of their abundance and whether they had phylogenetic representation (see below) from this checklist to be included as species in the regional community. In addition to analyzing the total local species pool as a community, we also investigated eight kipuka communities separately that had a subset of 18-20 species from the local species pool.

To assess whether an assembly process has structured the plant community on kipukas, we used the standardized effect size of MPD and MNTD using phylogenetic information, MPD and MNND using phenotypic (maximum vegetative height) information, and CAMI using RF and ABC to make model predictions. We also performed parameter estimation using ABC to understand what the influence of $\mathrm{t_E}$ or $\mathrm{t_C}$ was on the assembly processes in either the filtering or competition models that were supported. Again, this was done on the total local species pool in the kipukas, as well as the separate eight kipuka plant communities (Table 2.12).

When using trait information, the maximum vegetative plant heights were log transformed because the data were strongly right skewed, which did not coincide with the simulated data. For analysis using dispersion metrics, we performed 1000 iterations of a random community to include in the null distribution and each random community was constructed by shuffling the taxon labels of either the phylogenetic or phenotypic distance.

Before simulating data for RF and ABC model selection, we determined the best fit model of trait evolution for the empirical data, given we have the regional phylogeny and all phenotypic (maximum vegetative height) information for each individual. This information was used to constrain the model of trait evolution and σ^2, and potentially α, that we simulated community assembly data under. The OU model of trait evolution was always selected over a BM model (Table 2.11), though the estimation of precise values for $\sigma^2$and α proved challenging because we could only estimate the quotient of $\sigma^2$and α (Hansen 1997; Butler & King 2004) (Table 2.11).

For analysis using CAMI, we simulated 60,000 community assembly datasets under all models of community assembly and trait evolution, 10,000 per model, to use for model selection. The number of regional and local species in each simulation was fixed to 113 and 63, respectively, to mimic the empirical data. We opted to use the combination of $\sigma^2$and α, where α was a reasonably estimate at 0.02 and $\sigma^2, 0.77$; both parameters were fixed to these values. The other parameters $\mathrm{t}, \lambda,$ and μ were drawn from their default uniform prior distributions (Table 2.3).

These data and the empirical data, from the total kipuka community and the eight separate kipukas, were summarized into 30 summary statistics (Table 2.4) to be used for RF and ABC. For RF, we constructed a classification forest, or classifier, of 5000 decision trees using all 60,000 simulations and 30 summary statistics. We then used this classifier to predict which model of community assembly structured the kipuka plant community through vegetative height. For ABC, we used the top 10 summary statistics from the RF classifier (as in sections 2.4 and 2.5) and all 60,000 simulations to estimate the posterior probability of each model given the data. For both RF and ABC, we predicted the model probabilities while only considering the community assembly models (neutral, filtering, competition), but then also while the model of trait evolution (BM and OU).

We performed parameter estimation using ABC to understand what the influence of t_E or t_C was on the assembly processes in either the filtering or competition models that were supported for each kipuka dataset. For the supported models, we simulated 50,000 community assembly datasets under the default uniform priors for parameters t, λ, μ, and α, with a narrower prior distribution on σ^2, centering the empirically estimated σ^2. We always accepted 100 simulations as from the posterior for parameter estimation.

Supplemental Results

1.0 Importance of Summary Statistics

The summary statistics that RF determined to be most informative were the difference in variance of trait values between the local and regional communities, the variance of the local traits, the kurtosis of the local traits, the variance of the regional traits, the bimodal coefficient calculated from the local traits, the difference in the mean of the trait values between the local and regional community, the difference between the normalized Lineage-Through-Time statistic (Janzen et al. 2015) calculated for both the local community phylogeny and regional community phylogeny, the slope of a linear model fitted to the absolute value of the phylogenetic independent contrasts against their expected variances (following Garland et al. 1992), mode length from the local trait distribution, and finally, the slope of a linear model fitted to the absolute value of the contrasts against node depth (after Purvis & Rambaut 1995).

Literature Cited

Bazzaz, F.A. (1991). Habitat Selection in Plants. *Am. Nat.*137, S116-S130.

Belyea, L.R. & Lancaster, J. (1999). Assembly within a contingent rules ecology. *Oikos*., 86, 402-416.

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

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. *Stat. Dep. Univ. Calif. Berkeley, CA, USA*.

Breiman, L. & Cutler, A. (2007). Random forests — Classification description: Random forests. *http//stat-www.berkeley.edu/users/breiman/RandomForests/cc_home. htm*.

Brooks, D.R. & McLennan, D.A. (1991). *Phylogeny, Ecology, and Behavior. A Research Program in Comparative Biology.* Univ. Chicago Press, Chicago, USA.

Butler, M.A. & King, A.A. (2004). Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *Am. Nat.,*164, 683-695.

Cavender-Bares, J., Keen, A. & Miles, B. (2006). Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. *Ecology*. 87, 109-122.

Cavender-Bares, J., Kozak, K.H., Fine, P.V.A. & Kembel, S.W. (2009). The merging of community ecology and phylogenetic biology. *Ecol. Lett.*, 12, 693–715.

Conner E., Simberloff, D. (1979) The assembly of species communities: chance or competition? *Ecology*., 60, 1132-1140.

Cornwell, W.K., Schwilk, D.W. & Ackerly, D.D. (2006). A trait-based test for habitat filtering: Convex hull volume. *Ecology*., 87, 1465–1471.

Cornwell, W.K., Westoby, M., Falster, D.S., Fitzjohn, R.G., O'Meara, B.C., Pennell, M.W., *et al.* (2014). Functional distinctiveness of major plant lineages. *J. Ecol.*, 102, 345–356.

Csilléry, K., Blum, M.G.B., Gaggiotti, O.E. & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.*, 25, 410-418.

Csilléry, K., François, O. & Blum, M.G.B. (2012). Abc: An R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.*, 3, 475–479.

de Bello, F., Thuiller, W., Leps, J., Choler, P., Clement, J.C., Macek, P., *et al.* (2009). Partitioning of functional diversity reveals the scale and extent of trait convergence and divergence. *J. Veg. Sci.*, 20, 475-486.

Deevi, S. & 4D Strategies (2016). modes: Find the Modes and Assess the Modality of Complex and Mixture Distributions, Especially with Big Datasets. R package version 0.7.0. https://CRAN.R-project.org/package=modes.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *Am. Nat.* 125, 1-15.

Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv.*, 61, 1-10.

Garland, T., Harvey, P.H. & Ives, A.R. (1992). Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.* 41, 18-32.

Gerhold, P., Cahill, J.F., Winter, M., Bartish, I. V. & Prinzing, A. (2015). Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). *Funct. Ecol.*, 29, 600–614.

Gotelli, N.J. & Graves, G.R. (1996). Null Models in Ecology. Smithsonian Inst. Press, Washington, DC.

Gotelli, N.J. (2000). Null model analysis of species co-occurrence patterns. *Ecology*., 83, 2091-2096.

Gotelli, N.J. & Colwell, R.K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.*, 4, 379-391.

Götzenberger, L., de Bello, F., Bråthen, K.A., Davison, J., Dubuis, A., Guisan, A., *et al.* (2012). Ecological assembly rules in plant communities-approaches, patterns and prospects. *Biol. Rev.* 87, 111-127.

Graham, C.H., Parra, J.L., Tinoco, B.A., Stiles, F.G. & McGuire, J.A. (2012). Untangling the influence of ecological and evolutionary factors on trait variation across hummingbird assemblages. *Ecology*., 93, S99-S111.

Grandcolas, P. (1998). Phylogenetic Analysis and the Study of Community Structure. *Oikos*. 82, 397-400.

Hansen, T.F. (1997). Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution.* 51, 1341-1351.

Herben, T. & Goldberg, D.E. (2014). Community assembly by limiting similarity vs. competitive hierarchies: Testing the consequences of dispersion of individual traits. *J. Ecol.*, 102, 156-166.

HilleRisLambers, J., Adler, P.B., Harpole, W.S., Levine, J.M. & Mayfield, M.M. (2012). Rethinking Community Assembly through the Lens of Coexistence Theory. *Annu. Rev. Ecol. Evol. Syst.* 43, 227-248.

Hitchcock, C. L. & Cronquist A.C. (2018). Flora of the pacific northwest: an illustrated manual, 2nd Edition. D. E. Giblin, B. S. Legler, P. F. Zika, and R. G. Olmstead (eds.). University of Washington Press. Seattle, WA.

Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. *Monogr. Popul. Biol.* Princeton University Press. Princeton, NJ.

Janzen, T., Höhna, S. & Etienne, R.S. (2015). Approximate Bayesian Computation of diversification rates from molecular phylogenies: Introducing a new efficient summary statistic, the nLTT. *Methods Ecol. Evol.*, 6, 566–575.

Jung, V., Albert, C.H., Violle, C., Kunstler, G., Loucougaray, G. & Spiegelberger, T. (2014). Intraspecific trait variability mediates the response of subalpine grassland communities to extreme drought events. *J. Ecol.*, 102, 45-53.

Jung, V., Violle, C., Mondy, C., Hoffmann, L. & Muller, S. (2010). Intraspecific variability and trait-based community assembly. *J. Ecol.*, 98, 1134-1140.

Keddy, P.A. & Shipley, B. (2006). Competitive Hierarchies in Herbaceous Plant Communities. *Oikos*. 54, 234-241.

Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D., *et al.* (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, 26, 1463–1464.

Kendall, M., Boyd, M., & Colijn, C. (2018). phyloTop: Calculating Topological Properties of Phylogenies. R package version 1.1. https://CRAN.R-project.org/package=phyloTop

Komsta, L., & Novomestky, F. (2015). moments: Moments, cumulants, skewness, kurtosis and related tests. R package version 0.14. https://CRAN.R-project.org/package=moments

Kraft, N.J.B. & Ackerly, D.D. (2010). Functional trait and phylogenetic tests of community assembly across spatial scales in an Amazonian forest. *Ecol. Monogr.*, 80, 401-422.

Kraft, N.J.B., Cornwell, W.K., Webb, C.O. & Ackerly, D.D. (2007). Trait evolution, community assembly, and the phylogenetic structure of ecological communities. *Am. Nat.*, 170, 271–283.

Kraft, N.J.B., Godoy, O. & Levine, J.M. (2015). Plant functional traits and the multidimensional nature of species coexistence. *Proc. Natl. Acad. Sci.,* 112, 797-802.

Kuntz, M.A., Champion, D.E., Spiker, E.C. & Lefebvre, R.H. (1986). Contrasting magma types and steady-state, volume-predictable, basaltic volcanism along the Great Rift, Idaho (USA). *Geol. Soc. Am. Bull.*, 97, 579-594.

Kuntz, M.A., Champion, D.E., Spiker, E.C., Lefebvrelsd, R.H. & Mcbroomes, L.A. (1982). The Great Rift and the Evolution of the Craters of the Moon Lava Field , Idaho. *Cenezoic Geol. Idaho Idaho Bur. Mines Geol. Bull.*, 26, 423-437.

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

Losos, J.B. (1996). Phylogenetic perspectives on community ecology. *Ecology*. 77, 1344-1354.

Macarthur, R. & Levins, R. (1967). The Limiting Similarity, Convergence, and Divergence of Coexisting Species. *Am. Nat.*, 101, 377-385.

Magurran, A. E. (1988). Diversity Indices and Species Abundance Models. In: Ecological Diversity and Its measurment. Springer, D. pp. 7-75.

Marks & Lechowicz. (2017). Alternative Designs and the Evolution of Functional Diversity. *Am. Nat.,* 167:55-66.

Mayfield, M.M. & Levine, J.M. (2010). Opposing effects of competitive exclusion on the phylogenetic structure of communities. *Ecol. Lett.*, 13, 1085–1093.

McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends Ecol. Evol.*, 21, 178-185.

Miller, E.T., Farine, D.R. & Trisos, C.H. (2017). Phylogenetic community structure metrics and null models: a review with new methods and software. *Ecography.* 40, 461-477.

Munoz, F., Grenié, M., Denelle, P., Taudière, A., Laroche, F., Tucker, C., *et al.* (2018). ecolottery: Simulating and assessing community assembly with environmental filtering and neutral dynamics in R. *Methods Ecol. Evol.*, 9, 693-703.

Münkemüller, T., Gallien, L. (2015) VirtualCom: a simulation model for eco-evolutionary community assembly and invasion. *Methods Ecol. Evol.*, 6, 735-743.

Nuismer, S.L. & Harmon, L.J. (2015). Predicting rates of interspecific interaction from phylogenetic trees. *Ecol. Lett.*, 18, 17-27.

Nuismer, S.L., Jordano, P. & Bascompte, J. (2013). Coevolution and the architecture of mutualistic networks. *Evolution.*, 67, 338-354.

Overcast, I., Emerson, B.C. & Hickerson, M.J. (2019). An integrated model of population genetics and community ecology. *J. Biogeogr.* 46, 816-829.

Paradis, E. & Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*., 35, 526-528.

Pennell, M.W., FitzJohn, R.G., Cornwell, W.K. & Harmon, L.J. (2015). Model Adequacy and the Macroevolution of Angiosperm Functional Traits. *Am. Nat.*, 186, E33–E50.

Peters. R.H. (1991) A critique for ecology. Cambridge University Press. New York, NY.

Pigot, A.L. & Etienne, R.S. (2015). A new dynamic null model for phylogenetic community structure. *Ecol. Lett.*, 18, 153-163.

Pontarp, M., Brännström, Å. & Petchey, O.L. (2019). Inferring community assembly processes from macroscopic patterns using dynamic eco-evolutionary models and Approximate Bayesian Computation (ABC). *Methods Ecol. Evol*., 10, 450-460.

Purvis, A. & Rambaut, A. (1995). Comparative analysis by independent contrasts (CAIC): An apple macintosh application for analysing comparative data. *Bioinformatics*.

Qian, H. & Jin, Y. (2016). An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. *J. Plant Ecol.*, 9, 233-239.

Revell, L.J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, 3, 217–223.

Ricklefs, R. & Travis, J. (1980). A Morphological Approach to the Study of Avian Community Organization. *Auk*. 97, 321-338.

Rosindell, J. & Harmon, L.J. (2013). A unified model of species immigration, extinction and abundance on islands. *J. Biogeogr.*, 40, 1107-1118.

Rosindell, J., Harmon, L.J. & Etienne, R.S. (2015). Unifying ecology and macroevolution with individual-based theory. *Ecol. Lett.*, 18, 472-482.

Rosindell, J., Hubbell, S.P., He, F., Harmon, L.J. & Etienne, R.S. (2012). The case for ecological neutral theory. *Trends Ecol. Evol.*, 27, 203-208.

Slater, G. & W Pennell, M. (2013). Robust Regression and Posterior Predictive Simulation Increase Power to Detect Early Bursts of Trait Evolution. *Syst. Biol.*, 63, 293-308.

Smith, S.A. & Brown, J.W. (2017). Constructing a broadly inclusive seed plant phylogeny. *Am. J. Bot.* 105, 302-314.

Stadler, T. (2011). Simulating trees with a fixed number of extant species. *Syst. Biol.*, 60, 676–684.

Stegen, J.C., Lin, X., Fredrickson, J.K., Chen, X., Kennedy, D.W., Murray, C.J., Rockhold, M.L., Jonopka, A. (2013). Quantifying community assembly processes and identifying features that impose them. *ISME J.*, 7, 2069-2079.

Swenson, N.G., Enquist, B.J., Pither, J., Kerkhoff, A.J., Boyle, B., Weiser, M.D., *et al.* (2012). The biogeography and filtering of woody plant functional diversity in North and South America. *Glob. Ecol. Biogeogr.*, 21, 798-808.

Thompson, J.N., Reichman, O.J., Morin, P.J., Polis, G.A., Power, M.E., Sterner, R.W., *et al.* (2001). Frontiers of Ecology. *Bioscience*. 51, 15-24.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M.P.H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, 6, 187–202.

Uyeda, J.C. & Harmon, L.J. (2014). A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Syst. Biol.*, 63, 902-918.

Vandergast, A.G. & Gillespie, R.G. (2004). Effects of Natural Forest Fragmentation on a Hawaiian Spider Community. *Environ. Entomol.*, 33, 1296-1305.

Van Der Plas, F., Janzen, T., Ordonez, A., Fokkema, W., Reinders, J., Etienne, R.S., *et al.* (2015). A new modeling approach estimates the relative importance of different community assembly processes. *Ecology.*, 96, 1502-1515.

Webb, C. (2000). Exploring the Phylogenetic Structure of Ecological Communities: An Example for Rain Forest Trees. *Am. Nat.*, 156, 145–155.

Webb, C.O., Ackerly, D.D., McPeek, M.A. & Donoghue, M.J. (2002). Phylogenies and Community

Ecology. *Annu. Rev. Ecol. Syst.*, 33, 475–505.

Webb, C.O., Ackerly, D.D. & Kembel, S.W. (2008). Phylocom: Software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, 24, 2098–2100.

Weiher, E. & Keddy, P.A. (1995). Assembly Rules, Null Models, and Trait Dispersion: New Questions from Old Patterns. *Oikos*., 74, 159-164.

Weiher, E., Clarke, G.D.P. & Keddy, P.A. (1998). Community Assembly Rules, Morphological Dispersion, and the Coexistence of Plant Species. *Oikos*. 81, 309-322.

Weiher, E., van der Werf, A., Thompson, K., Roderick, M., Garnier, E. & Eriksson, O. (1999). Challenging Theophrastus: A common core list of plant traits for functional ecology. *J. Veg. Sci.*, 10, 609–620.

Weiher, E. & Keddy, P. (1999). Assembly rules as general constraints on community composition. In: *Weiher E, Keddy P (eds) Ecological assembly rules: perspectives, advances, retreats.* Cambridge University Press, Cambridge, UK, pp 251-271.

Weiher, E., Freund, D., Bunton, T., Stefanski, A., Lee, T. & Bentivenga, S. (2011). Advances, challenges and a developing synthesis of ecological community assembly theory. *Philos. Trans. R. Soc. B Biol. Sci.*, 366, 2403-2413.

Westoby, M. (1998). A leaf-height-seed (LHS) plant ecology strategy scheme. *Plant Soil*., 199, 213-227.

Tables

Table 2.1. Average error rates for model classification approaches in classifying each of the three community assembly models, as well as overall classification error.

		Neutral	Filtering	Competition	Mean
Phylogenetic	MPD	4.810	72.590	90.845	56.082
	MNTD	4.930	66.000	99.390	56.773
Phenotypic	MPD	4.741	7.940	2.130	4.937
	MNTD	4.911	39.855	99.465	48.077
RF		4.845	3.013	2.855	3.571
ABC		5.440	13.640	6.320	8.467

Table 2.2. Community assembly model predictions from RF and model posterior probabilities from ABC for all local kipuka plant species and eight individual kipuka communities. All predictions were made with simulations using an OU model of trait evolution.

	RF			**ABC**		
	Competition	**Filtering**	**Neutral**	**Competition**	**Filtering**	**Neutral**
ALL	-	0.64	0.36	-	0.82	0.18
B	0.06	0.54	0.4	-	0.35	0.65
C	0.06	0.6	0.34	-	0.5	0.5
D	0.07	0.61	0.32	-	0.92	0.08
E	0.06	0.58	0.36	-	0.67	0.33
F	0.02	0.46	0.52	-	0.47	0.53
G	0.05	0.52	0.43	-	0.6	0.4
H	0.04	0.52	0.44	0.02	0.47	0.52
I	0.08	0.48	0.45	0.32	0.25	0.43

Table 2.3. Default parameter prior distributions used for simulating data in CAMI.

Parameter	Prior
Lambda (λ)	uniform(0.05, 2.)
epsilon ($\frac{\mu}{\lambda}$)	uniform(0.2, 0.8)
rate of trait evolution (σ^2)	uniform(1.0, 10.0)
Pull to Optimum in OU models (α)	uniform(0.01, 0.2)
Tau.Environment (t_E)	uniform(1, 60)
Tau.Competition (t_C)	uniform(1, 60)

Table 2.4. Summary statistics calculated in CAMI and used for model selection in RF.

Summary Statistics	Information	Citation/Package in R
Mean BL	mean of community phylogeny branch lengths	base r package
Var BL	variance of community phylogeny branch lengths	
Mean Reg BL	mean of regional phylogeny branch lengths	
var reg bl	variance of regional phylogeny branch lengths	
mean bl dif	difference between regional and community mean branch lengths	
var bl dif	difference between regional and community variance of branch lengths	
Mean tr	mean of community traits	
Var tr	variance of community traits	
Mean Reg tr	mean of regional traits	
var reg tr	variance of regional traits	
mean tr dif	difference between regional and community mean of traits	
var tr dif	difference between regional and community variance of traits	
Moran I	Moran's I	Paradis E., Claude J. & Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.
Age	age of community tree	
Colless	Colless' index of a tree	Michelle Kendall, Michael Boyd and Caroline Colijn (2018). phyloTop: Calculating Topological Properties of Phylogenies. R package version 2.1.1.https://CRAN.R-project.org/package=phyloTop
Sackin	The Sackin's index is computed as the sum of the number of ancestors for each tips of the tree.	
nLTT	This function takes two ultrametric phylogenetic trees, calculates the normalized Lineage-Through-Time statistic for both trees and then calculates the exact difference between the two statistics.	Janzen,T. Hoehna,S., Etienne,R.S. (2015) Approximate Bayesian Computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT. Methods in Ecology and Evolution. doi: 10.1111/2041-210X.12350
Msig	mean of squared contrasts	Pennell MW, FitsJohn RG, Cornwell WK, Harmon LJ. 2015. Model Adequacy and the macroevolution of Angiosperm functional traits. The American Naturalist.
Cvar	The coefficient of variation (standard deviation/mean) of the absolute value of the contrasts.	
Svar	The slope of a linear model fitted to the absolute value of the contrasts against their expected variances (following Garland et al. 1992).	
Shgt	The slope of a linear model fitted to the absolute value of the contrasts against node depth (after Purvis and Rambaut 1995).	
Dcdf	The D statistic obtained from a Kolmolgorov-Smirnov test from comparing the distribution of contrasts to that of a normal distribution with mean 0 and standard deviation equal to the root of the mean of squared contrasts	

Table 2.5. The average proportion of misclassified simulations using the standard approach of phylogenetic dispersion metrics for all regional/local community sizes tested and for each model of community assembly.

	MPD			MNTD			
Local Community Size	**Neutral**	**Filtering**	**Competition**	**Neutral**	**Filtering**	**Competition**	**Mean**
25	0.036	0.864	0.941	0.046	0.936	0.976	0.633
50	0.050	0.833	0.934	0.046	0.876	0.984	0.621
75	0.048	0.797	0.924	0.049	0.841	0.987	0.608
100	0.055	0.770	0.929	0.046	0.801	0.989	0.598
125	0.051	0.755	0.907	0.050	0.759	0.997	0.587
150	0.051	0.731	0.903	0.045	0.755	0.992	0.580
175	0.047	0.732	0.907	0.059	0.700	0.996	0.574
200	0.067	0.713	0.908	0.051	0.690	0.992	0.570
225	0.052	0.705	0.917	0.036	0.667	0.995	0.562
250	0.042	0.697	0.912	0.057	0.651	0.994	0.559
275	0.042	0.713	0.908	0.050	0.635	0.994	0.557
300	0.040	0.695	0.900	0.060	0.604	1.000	0.550
325	0.053	0.692	0.898	0.059	0.565	0.996	0.544
350	0.060	0.685	0.901	0.057	0.589	0.998	0.548
375	0.047	0.680	0.890	0.042	0.562	0.997	0.536
400	0.044	0.699	0.907	0.052	0.541	0.999	0.540
425	0.035	0.696	0.902	0.043	0.506	0.997	0.530
450	0.048	0.690	0.882	0.043	0.514	0.999	0.529
475	0.049	0.685	0.894	0.053	0.512	0.999	0.532
500	0.045	0.686	0.905	0.042	0.496	0.997	0.529

Table 2.6. The average proportion of misclassified simulations using phenotypic dispersion metrics for all regional/local community sizes tested and for each model of community assembly.

	MPD			MNTD			
Local Community Size	**Neutral**	**Filtering**	**Competition**	**Neutral**	**Filtering**	**Competition**	**Mean**
25	0.053	0.474	0.150	0.034	0.747	0.988	0.408
50	0.042	0.229	0.099	0.051	0.618	0.992	0.339
75	0.042	0.156	0.043	0.055	0.539	0.998	0.306
100	0.039	0.130	0.018	0.054	0.515	0.993	0.292
125	0.059	0.102	0.021	0.045	0.453	0.998	0.280
150	0.053	0.082	0.015	0.050	0.449	0.996	0.274
175	0.041	0.071	0.016	0.054	0.429	0.989	0.267
200	0.045	0.055	0.009	0.042	0.407	0.993	0.259
225	0.048	0.048	0.010	0.042	0.369	0.993	0.252
250	0.050	0.035	0.006	0.054	0.338	0.996	0.247
275	0.060	0.040	0.004	0.051	0.385	0.995	0.256
300	0.047	0.020	0.005	0.040	0.356	0.999	0.245
325	0.053	0.031	0.010	0.040	0.322	0.994	0.242
350	0.048	0.025	0.005	0.056	0.323	0.993	0.242
375	0.057	0.023	0.003	0.059	0.315	0.997	0.242
400	0.041	0.016	0.002	0.053	0.318	0.994	0.237
425	0.053	0.014	0.002	0.050	0.286	0.995	0.233
450	0.036	0.012	0.003	0.047	0.262	0.997	0.226
475	0.043	0.011	0.003	0.047	0.277	0.997	0.230
500	0.038	0.014	0.002	0.058	0.263	0.996	0.229

Table 2.7. Average error rates, or proportion of incorrectly classified simulations, when classifying community assembly and trait evolution models using random forest for all sizes of the local community used.

Local Community Size	**BM Neutral**	**BM Filtering**	**BM Competition**	**OU Neutral**	**OU Filtering**	**OU Competition**	**Out-of-Bag Error Rate**
25	50.00	44.53	45.10	49.90	48.60	36.10	45.70
50	39.40	34.40	37.90	38.50	40.64	31.20	37.01
75	33.00	28.16	31.50	33.20	33.00	31.70	31.76
100	29.80	26.70	25.40	29.70	34.00	27.00	28.77
125	28.50	23.40	23.80	27.40	31.50	27.20	26.97
150	22.20	24.30	24.50	23.00	29.70	24.60	24.72
175	22.50	21.32	19.00	24.90	28.50	26.00	23.70
200	22.50	21.50	20.30	23.90	25.03	22.50	22.62
225	19.40	20.20	20.50	23.30	25.00	21.20	21.60
250	18.20	18.90	17.10	22.10	23.90	24.80	20.83
275	18.70	17.80	17.80	20.20	22.90	25.10	20.42
300	17.70	16.50	16.20	17.80	21.50	21.20	18.48
325	15.40	16.30	18.70	17.70	22.90	22.90	18.98
350	15.90	15.90	16.30	19.30	23.80	21.10	18.72
375	15.90	15.30	17.40	18.40	23.60	20.10	18.45
400	14.70	15.80	16.10	19.50	22.80	22.50	18.57
425	14.80	14.50	14.80	16.70	22.90	19.20	17.15
450	13.30	14.10	13.70	15.70	21.30	21.30	16.57
475	11.80	15.80	13.40	14.50	22.40	19.00	16.15
500	12.90	14.50	12.50	16.00	22.90	18.60	16.23

Table 2.8. Average error rates, or proportion of incorrectly classified simulations, when classifying community assembly and trait evolution models using ABC for all sizes of the local community used.

Local Community Size	BM Neutral	BM Filtering	BM Competition	OU Neutral	OU Filtering	OU Competition	Mean
25	59.20	56.20	60.80	38.40	53.80	36.40	50.80
50	57.80	57.80	48.80	30.60	39.80	29.80	44.10
75	48.00	52.00	40.40	23.80	41.80	33.20	39.87
100	43.60	50.00	38.80	28.60	38.60	24.00	37.27
125	41.80	51.60	33.80	17.80	37.00	26.00	34.67
150	40.00	48.60	29.20	19.40	37.40	25.40	33.33
175	31.00	44.40	26.20	20.40	32.80	26.00	30.13
200	33.40	42.40	24.00	20.00	34.00	27.20	30.17
225	28.20	45.00	25.20	19.80	33.60	20.00	28.63
250	28.60	40.40	24.60	17.80	32.80	26.00	28.37
275	26.60	43.60	25.80	20.00	31.60	23.40	28.50
300	24.60	40.80	22.80	14.60	34.00	21.40	26.37
325	22.60	36.40	24.60	15.20	37.00	25.20	26.83
350	22.80	36.80	23.80	20.00	33.40	24.40	26.87
375	22.00	35.60	25.40	18.00	32.00	21.40	25.73
400	23.80	41.00	21.20	15.60	35.20	21.00	26.30
425	21.80	34.20	19.00	15.40	32.80	22.80	24.33
450	22.80	34.40	19.80	17.80	34.80	19.20	24.80
475	19.00	36.80	17.60	17.00	37.20	22.00	24.93
500	14.60	33.80	20.00	14.80	36.00	21.60	23.47

Table 2.9. Results for empirical data when using dispersion metrics MPD and MNTD with phylogenetic and phenotypic information for the total kipuka plant community and the eight individual kipuka plant communities. The mean and standard deviation of the null distribution are included, as well as the p-value for the for the observed value's position in the null distribution. Significant p-values are indicated with **.observed value's position in the null distribution. Significant p-values are indicated with **.

		MPD				**MNTD**			
	Kipuka	**Observed**	**Null Mean**	**Null SD**	**p-value**	**Observed**	**Null Mean**	**Null SD**	**p-value**
Phylogenetic	**All**	231.29	253.23	10.33	**0.02****	64.27	67.91	7.26	0.29
	1	218.75	254.93	28.1	**0.03**	125.27	125.98	24.58	0.5
	2	221.29	255.31	29.7	**0.05**	135.32	129.5	26.51	0.6
	3	233.74	255.15	30.99	0.28	100.85	132.55	29.17	0.14
	4	217.67	254.72	27.8	**0.02****	141.03	126.26	26.99	0.71
	5	213.2	255.18	27.9	**0.01****	97.2	129.75	26.85	0.12
	6	225.89	255.5	29.64	0.1	102.67	130.89	25.87	0.14
	7	228.18	256.04	27.97	0.12	108.32	125.6	25.23	0.26
	8	217.51	255.18	27.95	0.02	107.12	126.28	24.99	0.23
Phenotypic	**All**	0.92	1.13	0.12	**0.03**	0.02	0.08	0.02	**0.01****
	1	0.76	1.13	0.3	0.1	0.07	0.19	0.09	0.06
	2	0.87	1.13	0.3	0.21	0.09	0.19	0.09	0.14
	3	0.74	1.1	0.31	0.11	0.06	0.2	0.1	**0.05**
	4	0.85	1.11	0.28	0.18	0.12	0.19	0.09	0.3
	5	1.13	1.1	0.31	0.58	0.15	0.19	0.09	0.37
	6	1.25	1.13	0.3	0.67	0.12	0.19	0.09	0.28
	7	1.15	1.12	0.28	0.58	0.12	0.18	0.08	0.29
	8	1.28	1.12	0.28	0.74	0.12	0.19	0.09	0.25

Table 2.10. All species included in the regional and local community phylogeny for the kipukas od CRMO. Species that were present in the community but not in the Spermatophyta phylogeny were replaced in the phylogeny with a close relative, denoted in the table. The minimum, maximum, and range of vegetative height for each species is also denoted in cm.

Genera	specific epithet	Species from Smith Tree (if replacement is needed)	species min height (cm)	species max height (cm)	species range height (cm)
Achnatherum	lemmonii	Achnatherum richardsonii	20	70	20-70
Achnatherum	thurberianum	Achnaterum nelsonii	30.5	70	30.5-70
Acnatherum	occidentale		25	45	25-45
Agoseris	aurantiaca		10	60	10-60
Agoseris	glauca	Agoseris grandiflora	10	70	10-70
Allium	acuminatum		10	30	10-30
Alyssum	desertorum		10	25	10-25
Artemisia	tridentata		40	200	40-200
Artemisia	tripartita		20	60	20-60
Astragalus	filipes	Astragalus calycosus	30	90	30-90
Astragalus	lentiginosus	Astragalus canadensis	10	40	10-40
Astragalus	purshii	Astragalus newberryi	5	10	5-10
Balsamorhiza	sagittata		15	80	15-80
Boechera	divaricarpa		25	80	25-80
Calochortus	macrocarpus		20	70	20-70
Carex	filifolia	Carex geyeri	5	35	5-35
Chaenactis	douglasii		10	60	10-60
Chamaebatiaria	millefolium		100	200	100-200
Chenopodium	leptophyllum		20	60	20-60
Chrysothamnus	viscidiflorus		20	100	20-100
Cordylanthus	ramosus		10	90	10-90
Crepis	acuminata		20	70	20-70
Delphinium	andersonii		30	60	30-60
Descurainia	incana		15	120	15-120
Descurainia	pinnata		10	70	10-70
Diplacus	nanus		1.3	10	1.3-10
Elymus	elymoides		10	60	10-60
Eriastrum	sparsiflorum		5	30.5	5-30.5
Ericameria	nauseosa		20	280	20-280
Erigeron	pumilus		5	50	5-50
Erigeron	subtrinervis		15	80	15-80
Eriogonum	caespitosum		3	10	3-10
Erythranthe	suksdorfii		3	10	3-10

Festuca	viridula	Festuca idahoensis	40	80	40-80
Galium	bifolium		5	20	5-20
Gayophytum	ramosissimum	Gayophytum heterozygum	15	40	15-40
Hackelia	floribunda		30	100	30-100
Lappula	redowskii		5	40	5-40
Lathrocasis	tenerrima		3	35	3-35
Leymus	cinereus		100	200	100-200
Lithophragma	tenellum		10	25	10-25
Lithospermum	ruderale		20	60	20-60
Lomatium	ambiguum		10	80	10-80
Lomatium	foeniculaceum		3	40	3-40
Lomatium	multifidum	Lomatium dissectum	50	150	50-150
Lomatium	simplex	Lomatium idahoense	20	80	20-80
Mentzelia	albicaulis		10	40	10-40
Penstemon	cyananthus		30	70	30-70
Penstemon	procerus		5	40	5-40
Phacelia	heterophylla		20	120	20-120
Phlox	aculeata		61	122	61-122
Phlox	hoodii		3	6	3-6
Phlox	longifolia		10	50	10-50
Purshia	tridentata		100	200	100-200
Ranunculus	glaberrimus		5	20	5-20
Ribes	aureum		100	300	100-300
Senecio	sphaerocephalus	Senecio trinagularis	30	80	30-80
Sisymbrium	altissimum		30	150	30-150
Stephanomeria	tenuifolia		20	70	20-70
Thinopyrum	intermedium	Thinopyrum junceum	91.5	122	91.5-122
Thlaspi	arvense		10	50	10-50
Toxicoscordion	paniculatum		30	50	30-50
Viola	nuttallii	Viola glabella	3	12	3-12
Achillea	millefolium		20	40.5	20-40.5
Agastache	urticifolia		40	150	40-150
Agropyron	cristatum		30.5	91.5	30.5-91.5
Alnus	incana		460	2500	460-2500
Amsinckia	tessellata		15	60	15-60
Antennaria	microphylla		5	40	5-40
Arnica	cordifolia		10	60	10-60
Artemisia	arbuscula		10	40	10-40

Calochortus	eurycarpus		10	50	10-50
Carex	douglasii	Carex backii	15	46	15-46
Castilleja	chromosa		30.5	90	30.5-90
Castilleja	miniata		30.5	80	30.5-80
Chrysothamnus	viscidiflorus		20	100	20-100
Collinsia	parviflora		5	50	5-50
Collomia	linearis		10	40.5	10-40.5
Cornus	sericea		20	60	20-60
Cymopterus	glaucus		2	15	2-15
Delphinium	nuttallianum		15	40	15-40
Draba	verna	Draba alpina	5	20	5-20
Elymus	glaucus		50	100	50-100
Elymus	lanceolatus		30.5	91.5	30.5-91.5
Epilobium	ciliatum		30	100	30-100
Erigeron	speciosus		15	80	15-80
Fritillaria	pudica		10	30	10-30
Galium	bifolium		5	20	5-20
Gilia	inconspicua		8	32	8-32
Heuchera	parvifolia		23	30.5	23-30.5
Hydrophyllum	capitatum	Hydrophyllum canadense	10	46	10-46
Juniperus	scopulorum		100	1000	100-1000
Koeleria	macrantha	Koeleria macrantha	30	60	30-60
Lepidium	perfoliatum		20	60	20-60
Lomatium	grayi		15	50	15-50
Lomatium	nudicaule		20	90	20-90
Lupinus	argenteus		10	40.5	10-40.5
Mimulus	nanus	Mimulus ringens	1.25	10	1.25-10
Montia	chamissoi		5	20	5-20
Oenothera	caespitosa	Oenothera nuttallii	7.5	23	7.5-23
Opuntia	polyacantha		10	30.5	10-30.5
Orobanche	fasciculata		3	15	3-15
Penstemon	deustus	Penstemon attenuatus	20	60	20-60
Phacelia	hastata		25.5	76	25.5-76
Philadelphus	lewisii		150	250	150-250
Pinus	flexilis		396	1524	396-1524
Polygonum	douglasii	Polygonum aviculare	10	40	10-40
Potentilla	gracilis	Potentilla newberryi	40	80	40-80
Prunus	virginiana		100	500	100-500
Pseudotsuga	menziesii		2133.5	9144	2133.5-9144

Ribes	cereum		50	150	50-150
Symphoricarpos	oreophilus		50	150	50-150
Thalictrum	occidentale		40	100	40-100
Trifolium	variegatum		10	60	10-60
Viola	purpurea		0.5	1.5	.5-1.5

Table 2.11. Supplement. Parameter estimates, log-likelihoods and AIC scores for BM and OU models of trait evolution. The BM model of trait evolution was optimized one time, with the resulting log-likelihood and AIC. For the OU model, we struggled to estimate the parameter values absolutely. However, we could estimate the quotient of $\sigma^2$and α and did so for 21 parameter combinations. Regardless of the $\sigma^2$and α though, OU was always a better fit model.

	alpha	**sig.sq**	**Lik**	**AIC**	**AICc**
BM	-	0.33	-280.43	564.87	564.98
	0.01	0.34	-256.45	518.90	519.12
	0.05	0.43	-216.53	439.07	439.29
	0.10	0.54	-199.37	404.75	404.97
	0.15	0.65	-191.13	388.26	388.48
	0.20	0.77	-186.32	378.65	378.87
	0.25	0.88	-183.21	372.41	372.63
	0.30	1.00	-181.04	368.07	368.29
	0.35	1.12	-179.45	364.90	365.12
	0.40	1.24	-178.24	362.49	362.71
	0.45	1.37	-177.29	360.59	360.81
OU	0.50	1.49	-176.53	359.06	359.28
	0.55	1.61	-175.90	357.81	358.03
	0.60	1.73	-175.38	356.76	356.98
	0.65	1.85	-174.93	355.86	356.08
	0.70	1.97	-174.55	355.10	355.32
	0.75	2.10	-174.22	354.43	354.65
	0.80	2.22	-173.92	353.85	354.07
	0.85	2.34	-173.67	353.33	353.55
	0.90	2.46	-173.44	352.87	353.10
	0.95	2.58	-173.23	352.47	352.69
	1.00	2.71	-173.05	352.10	352.32

Table 2.12. Supplement. Parameter Presence/Absence matrix indicating which kipuka species is present on one of the eight specific kipukas investigated. The table contains all of the species that occur in the local kipuka community.

Species (Smith Tree Replacements Included)	**Kipuka 16**	**Kipuka 11**	**Kipuka 671**	**Kipuka 425**	**Kipuka 3**	**Kipuka 620**	**Kipuka Pratt1**	**Kipuka 18**
Figure Reference	**B**	**C**	**D**	**E**	**F**	**H**	**I**	**J**
Achnatherum_richardsonii	1	1	0	0	0	0	0	0
Achnatherum_nelsonii	0	0	1	1	0	0	0	0
Achnatherum_occidentale	0	0	0	0	0	0	0	0
Agoseris_aurantiaca	0	0	0	0	0	0	0	0
Agoseris_grandiflora	1	1	1	1	0	0	0	0
Allium_acuminatum	0	0	1	1	1	0	0	0
Alyssum_desertorum	0	0	0	0	0	1	0	0
Artemisia_tridentata	1	0	1	1	1	1	1	1
Artemisia_tripartita	1	1	0	0	0	0	1	1
Astragalus_calycosus	1	1	0	0	0	0	0	0
Astragalus_canadensis	0	0	0	1	0	0	0	0
Astragalus_newberryi	0	0	0	0	0	1	1	0
Balsamorhiza_sagittata	1	1	0	1	1	0	0	0
Boechera_divaricarpa	1	0	0	0	1	1	1	1
Calochortus_macrocarpus	0	0	0	1	0	0	0	0
Carex_geyeri	0	0	0	0	0	0	1	0
Chaenactis_douglasii	0	0	0	0	1	0	0	0
Chamaebatiaria_millefolium	0	0	0	0	1	0	0	1
Chenopodium_leptophyllum	0	0	0	0	0	0	0	0
Chrysothamnus_viscidiflorus	0	0	1	1	1	1	1	0
Cordylanthus_ramosus	1	0	1	0	1	0	0	0
Crepis_acuminata	0	0	1	1	0	0	0	0
Delphinium_andersonii	1	1	1	1	0	0	0	0
Descurainia_incana	0	0	0	0	1	1	1	0
Descurainia_pinnata	0	0	1	0	0	0	0	0
Diplacus_nanus	1	1	0	0	0	0	0	0
Elymus_elymoides	0	0	1	0	0	0	0	0
Eriastrum_sparsiflorum	0	0	0	0	1	0	0	0
Ericameria_nauseosa	0	1	1	1	1	0	0	1
Erigeron_pumilus	0	0	0	1	0	0	0	0
Erigeron_subtrinervis	0	0	0	1	0	0	0	0
Eriogonum_caespitosum	0	0	0	1	0	0	0	0
Erythranthe_suksdorfii	0	0	0	0	0	0	0	1
Festuca_baffinensis	0	0	0	0	0	0	0	0

Galium_bifolium	0	0	0	0	0	0	0	1
Gayophytum_heterozygum	1	0	0	0	0	0	0	0
Hackelia_floribunda	0	0	0	0	0	0	0	0
Lappula_redowskii	0	1	0	0	1	1	1	1
Lathrocasis_tenerrima	0	0	0	0	0	1	1	1
Leymus_cinereus	0	0	1	0	0	0	0	0
Lithophragma_tenellum	0	0	0	0	0	1	1	1
Lithospermum_ruderale	0	0	0	0	0	0	0	1
Lomatium_ambiguum	0	1	0	0	0	0	0	0
Lomatium_foeniculaceum	0	1	0	0	1	1	1	1
Lomatium_dissectum	1	0	0	0	1	0	0	1
Lomatium_idahoense	1	0	0	0	0	1	1	1
Mentzelia_albicaulis	1	1	0	0	0	0	0	0
Penstemon_cyananthus	0	0	1	1	0	0	0	0
Penstemon_procerus	0	0	0	0	0	0	1	0
Phacelia_heterophylla	1	0	0	0	0	0	0	0
Phlox_aculeata	0	1	0	0	0	0	0	0
Phlox_hoodii	0	0	0	0	1	1	1	1
Phlox_longifolia	1	0	0	1	0	1	1	0
Purshia_tridentata	0	1	1	1	0	1	1	1
Ranunculus_glaberrimus	0	0	0	0	0	0	0	0
Ribes_aureum	0	0	0	0	0	1	1	0
Senecio_triangularis	1	1	0	0	1	1	1	1
Sisymbrium_altissimum	0	0	1	1	0	0	0	0
Stephanomeria_tenuifolia	1	1	0	0	0	0	0	0
Thinopyrum_junceum	0	0	0	0	0	0	0	0
Thlaspi_arvense	0	0	0	0	0	0	0	0
Toxicoscordion_paniculatum	0	1	1	0	0	0	0	0
Viola_glabella	1	1	0	0	1	1	1	1

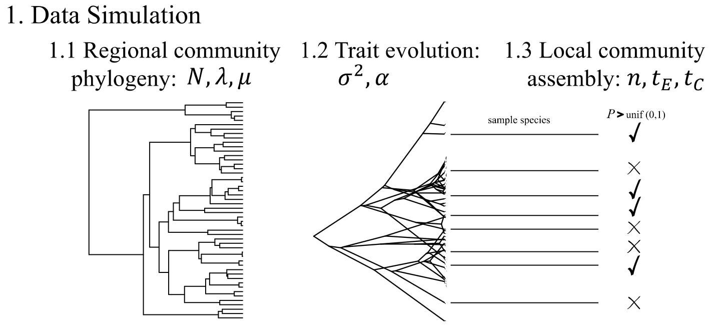
Table 2.13. Model probabilities for the total kipuka plant community and eight separate kipukas using RF and ABC. For this, the reference data was made up of data simulated under OU models of trait evolution and data simulated under BM models of trait evolution.

		BM			**OU**		
		Competition	**Filtering**	**Neutral**	**Competition**	**Filtering**	**Neutral**
RF	**ALL**	-	0.02	0.01	-	0.62	0.35
	B	0.02	0.19	0.16	0.06	0.33	0.23
	C	0.01	0.10	0.07	0.06	0.53	0.23
	D	0.01	0.14	0.05	0.07	0.46	0.27
	E	0.01	0.11	0.07	0.05	0.54	0.23
	F	0.01	0.06	0.05	0.06	0.41	0.42
	G	0.01	0.07	0.05	0.06	0.44	0.36
	H	0.02	0.14	0.17	0.05	0.33	0.30
	I	0.03	0.12	0.12	0.09	0.32	
ABC	**ALL**	-	0.03	0.02	-	0.57	0.38
	B	-	0.28	0.52	-	0.07	0.13
	C	-	-	-	-	0.63	0.37
	D	-	-	-	-	0.80	0.20
	E	-	-	-	-	0.72	0.28
	F	-	-	-	-	0.50	0.50
	G	-	-	-	-	0.58	0.42
	H	0.02	0.33	0.28	0.03	0.13	0.20
	I	0.15	0.20	0.17	0.25	0.12	0.12

Table 2.14. Model probabilities for environmental filtering for the eight kipuka plant communities using RF and ABC, as well as the median t_E estimates using ABC.

	kipuka	RF prob	ABC prob	Median Tau
Kipuka_671	D	0.61	0.92	7.52
Kipuka_425	E	0.58	0.67	15.99
Kipuka_11	C	0.60	0.50	21.00
Kipuka_16	B	0.54	0.35	23.89
Kipuka_3	F	0.46	0.47	26.72
Kipuka_620	I	0.52	0.60	32.77
Kipuka_Pratt1	G	0.52	0.47	36.15
Kipuka_18	H	0.48	0.25	37.78

Figures


1. Data Simulation
1.1 Regional community phylogeny: N, λ, μ
1.2 Trait evolution: σ^2, α
1.3 Local community assembly: n, t_E, t_C
sample species
$P >$ unif (0,1)
✓
×
✓
✓
×
×
✓
×

N	regional species pool
λ	speciation rate
μ	extinction rate
σ^2	rate of character change
α	strength of constraints (OU only)
n	local species pool
t_E	effect of environmental filtering
t_C	effect of competitive exclusions

Figure 2.1. Outline of data simulation process. (1.1) Simulate the regional phylogeny. (1.2) Simulate trait evolution along the regional phylogeny. (1.3) Simulate the assembly of the local community by sampling species at random from the regional species pool and calculating the probability of persistence for each sampled species. These probabilities are calculated differently depending on the model of assembly being simulated, and if a species' probability of persistence is greater than a randomly generated probability, then that species survives in the local community.

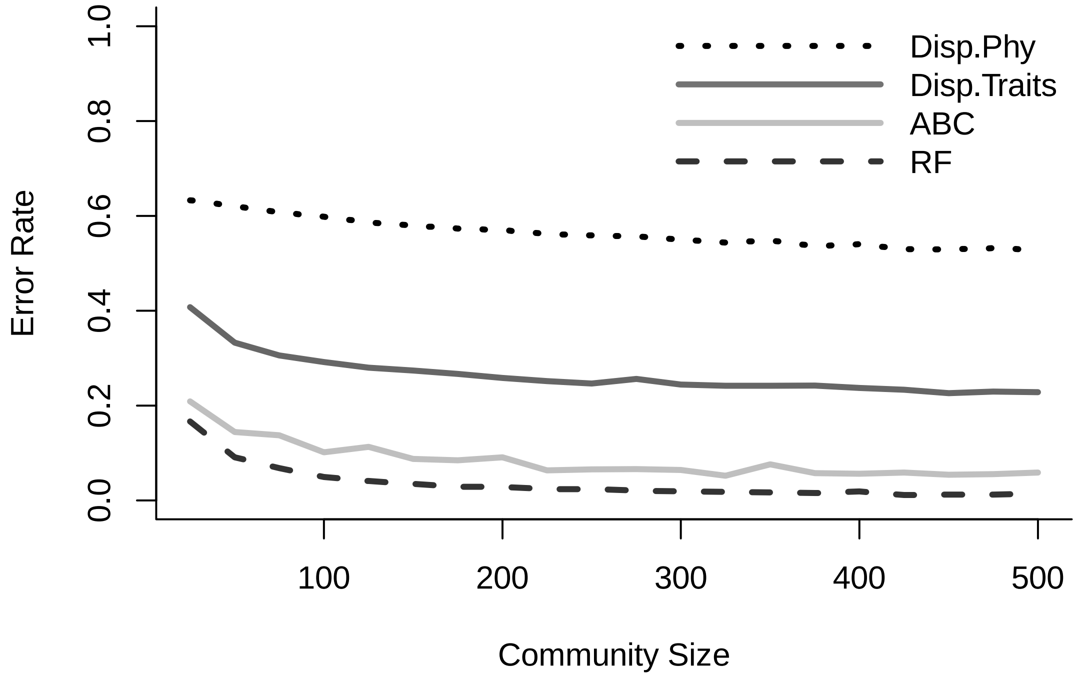


Figure 2.2. Error rates, or proportion of incorrectly classified simulations, when classifying community assembly models compared to the size of the local community used. Four model identification approaches are summarized here. The first is the average error rate when using dispersion metrics (MPD and MNTD) from phylogenetic information (dotted). The second is the average error rate when using dispersion metrics from functional trait information (big dashed). The final two are model selection approaches employed in CAMI, ABC (gray), and RF (small dashed).

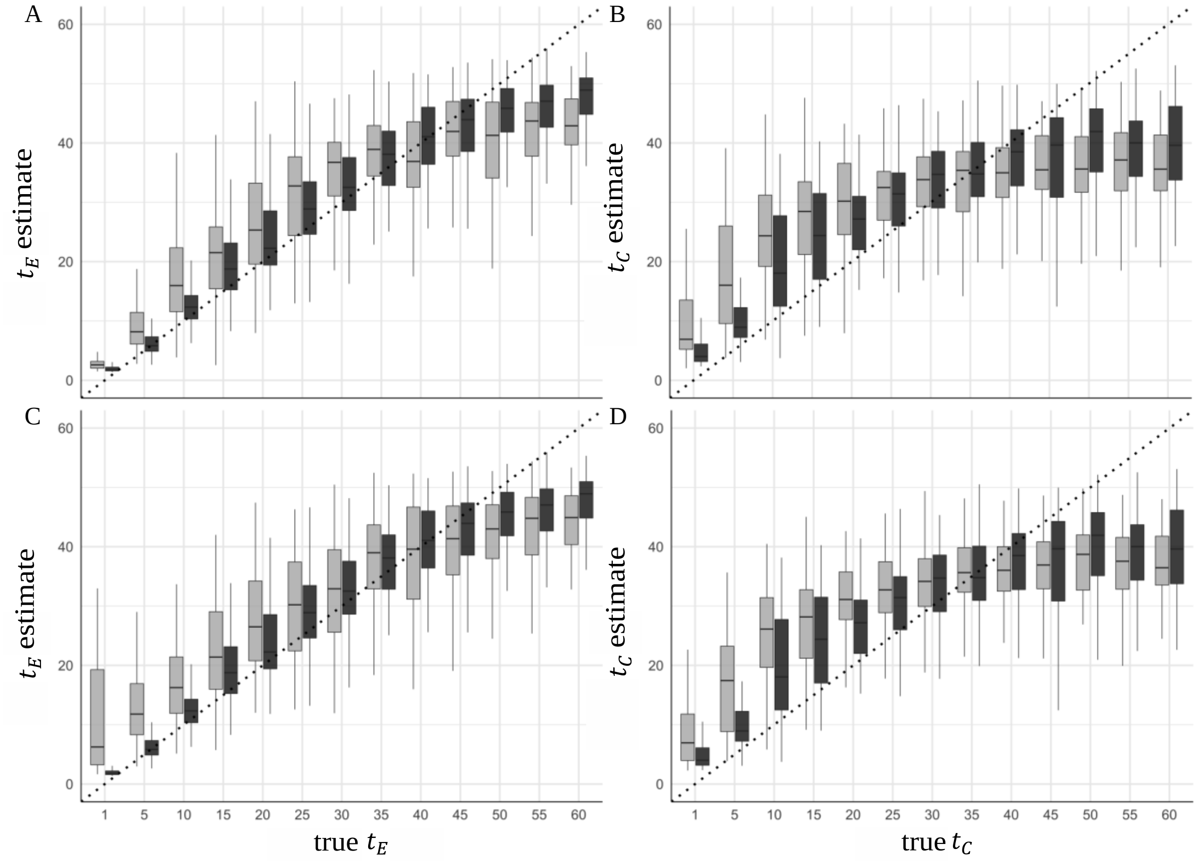


Figure 2.3. Estimation of t_E and t_C under their respective non-neutral models of community assembly, coupled with one of two models of trait evolution. In each graph, the individual boxplots represent the median values of either t_E or t_C from 100 independent attempts at parameter estimation, thus they are not posterior distributions, but rather a distribution of median parameter estimates. The x-axis denotes the true value of t_E or t_C simulated under. The light gray boxes represent datasets with regional/local community sizes of 200/100 and the dark gray boxes represent regional/local community sizes of 800/400. The dotted line in each plot represents a 1:1 correlation between estimated and true values of either t_E or t_C. A. Environmental filtering community assembly with a BM model of trait evolution. B. Competitive exclusion community assembly with a BM model of trait evolution. C. Environmental filtering community assembly with an OU model of trait evolution. D. Competitive exclusion community assembly with an OU model of trait evolution.

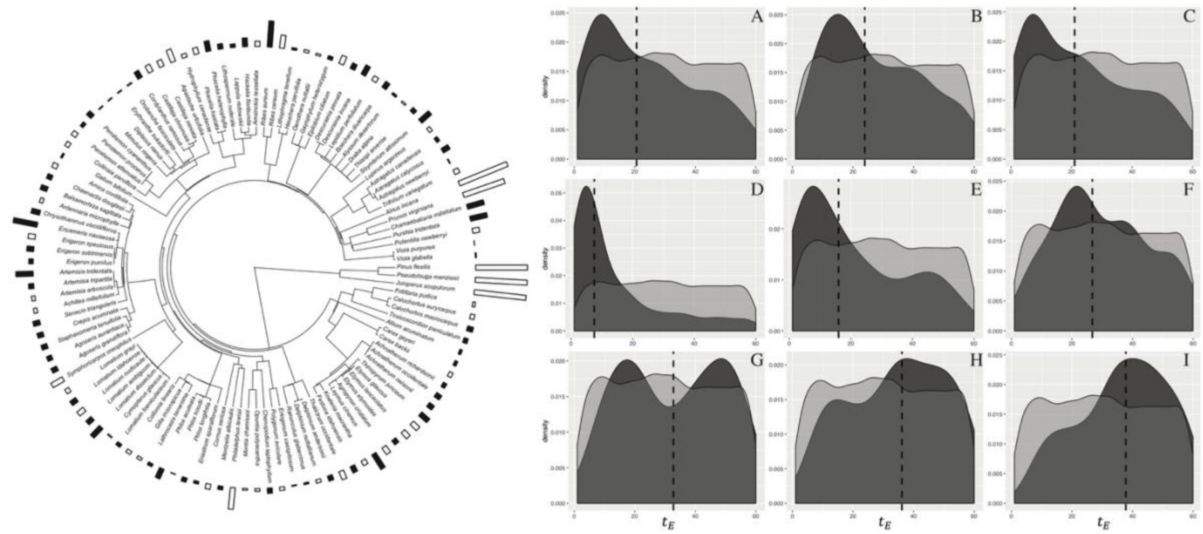


Figure 2.4. left) Regional phylogeny of species in the Craters of the Moon National Monument and Preserve, coupled with each species' maximum vegetative height in meters represented by the filled bar plots by each species. Species only present in the regional community have their trait bars colored white, while species that are also present in the local community have their trait bars colored black. The bars are truncated at 6 meters, as only the four trees in this study are larger than 6 meters, and those species and their heights are available in supplemental table 8. right) Nine panels displaying the prior (light gray) and posterior (dark gray) probability distributions of t_E under an environmental filtering model and OU model trait evolution. The dotted line represents the median estimate of t_E. A) Estimate from the entire local kipuka plant species pool. B-I) Estimates from the separate eight kipuka communities.

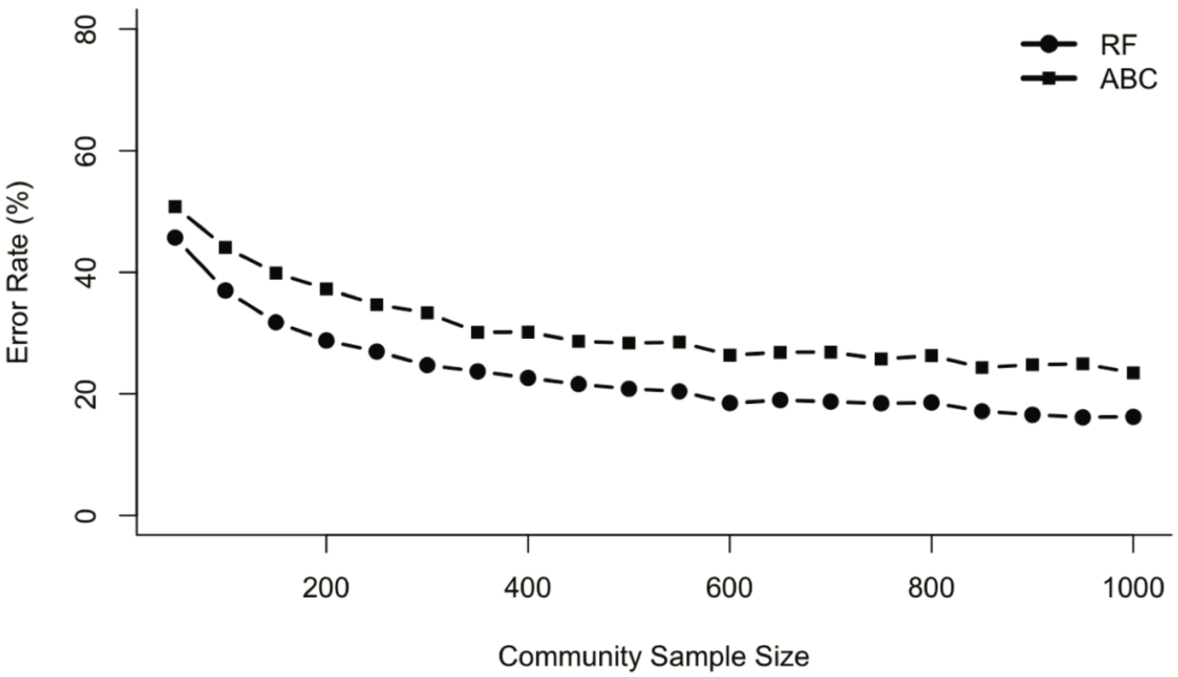


Figure 2.5. Supplement. Error rates, or proportion of incorrectly classified simulations, when classifying community assembly and trait evolution models compared to the size of the local community used. Square symbols indicate RF was used to classify data and circle symbols indicate ABC was used to classify data.

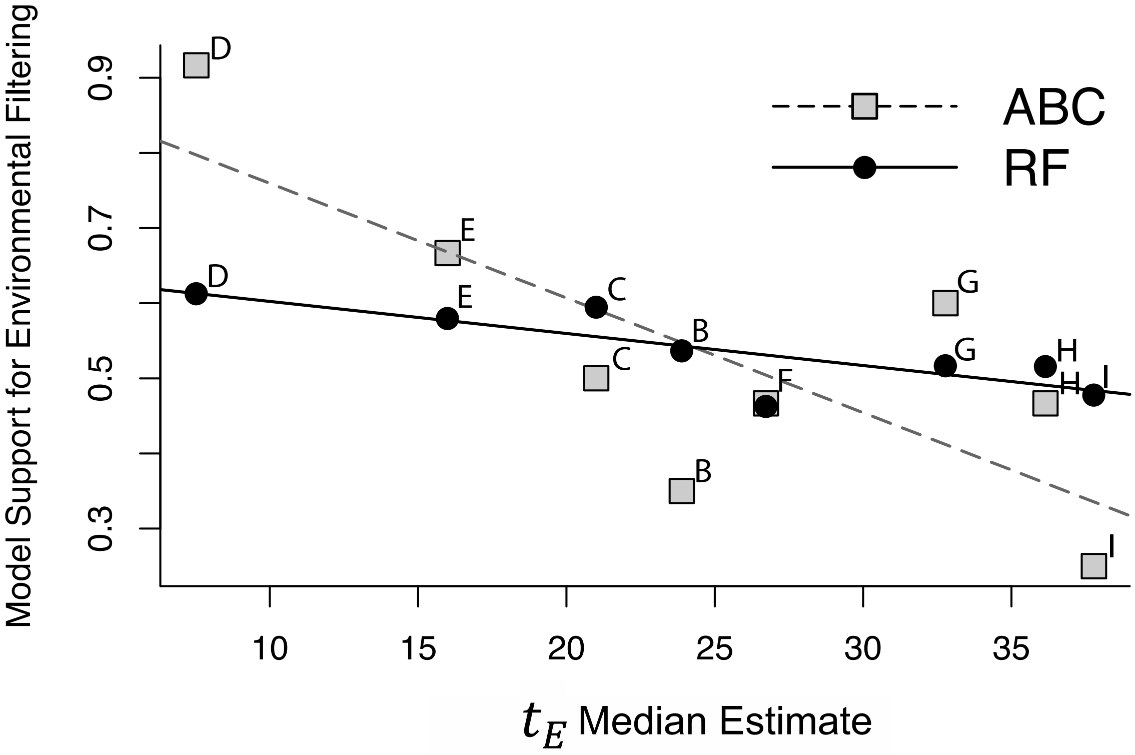


Figure 2.6. Supplement. Linear regression models between the model support for environmental filtering, as predicted from RF and ABC) and the t_E median estimates from ABC. The correlation coefficient when comparing RF model support values and t_E estimates was 0.65, and when comparing with ABC the coefficient was 0.89.

Chapter 3: Genomic Evidence of an Ancient Inland Temperate Rainforest

Abstract

The disjunct temperate rainforests of the Pacific Northwest are characterized by their dominant tree species Western Red Cedar (*Thuja plicata*) and Western Hemlock (*Tsuga heterophylla*). The demographics of these species, and thus the PNW rainforest, has been heavily impacted by the geological and climatic changes the PNW has experienced over the last 5 million years. The changes have ultimately shaped the history of these species, with the Pleistocene glaciation having a huge impact on our understanding of how long the inland temperate rainforest has be persisting in the inland. Paleontological records indicate a recent expansion of the coastal and inland temperate rainforest mid-Holocene, albeit the inland forest sometime after the coast. Here, we collect genomic data for both species across their range using reduced representation sequencing – a low cost approach to generating a large amount of genomic data that spans the entire genome. With this genomic data, we are able to design and assess the support for competing demographic scenarios. These scenarios alter in the divergence times between inland and coastal population, as well as migration and population size changing parameters, and are based off of the coalescent process that models alleles in a population backward in time. In these models, we can include these various demographic processes. We find support for the best demographic scenario using the genomic data and a machine learning inference procedure using the randomForests algorithm. We then optimized the parameters for the best models for both species. We show strong support that both species' inland and coastal populations diverged ~250,000 generations ago, followed by a decrease in population size, followed by population expansion and migration between populations coinciding with the mid-Holocene. These results suggest the populations of these species did expand in the ITR in the last 3,500 years, but also that populations were present in the inland throughout the Pleistocene. Through the use of genomic data and sophisticated inference procedures involving machine learning, we can unlock the history of the PNW temperate rainforest.

Introduction

The disjunct old-growth cedar-hemlock forests of the Pacific Northwest characterize one of the most diverse temperate rainforests in the world (Newmaster et al. 2003). The inland temperate rainforest (ITR) is approximately 200 km disjunct from the much larger ranging coastal rainforest. The Pacific Northwest region as a whole has been widely impacted by the Pleistocene's continuous glacial/interglacial cycles (Waitt & Thorson 1983), with flora and fauna being massively shaped by these climatic changes. The ITR has been of particular interest because of the dramatic implications of the alternative hypothesis proposed to explain the history of the ITR throughout and after the Pleistocene. While Recent Dispersal (RD) hypothesis posits the recent existence of the ITR (<5000 ka), having only colonized the inland from the coastal populations recently. The Ancient Vicariance (AV) hypothesis posits an ancient disjunction between the inland and coastal forest (Brunsfeld *et al.* 2001) that occurred Pre-Pleistocene (> 1.5 mya), and while the onset of the glaciers caused massive devastation to the ITR, refugia persisted and recolonized the ITR post-Pleistocene. These two hypotheses broadly encapsulate the dominant modes of the formation of the disjunction and are critical to understanding biogeographic processes.

The range of the PNW temperate rainforest is defined by the range of the two most prominent species that make up the forest habitat, *Tsuga heterophylla* Raf. (Sarg.) (western hemlock) and *Thuja plicata* Donn ex. D. Don. (western red cedar). According to pollen records from the central and southern ITR, these forests have only been present in the area for < 3500 ya (Mehringer 1996, Rosenburg *et al.* 2003, Chase *et al.* 2008, Gavin *et al.* 2009). Studies of *Tsuga heterophylla* have shown that suitable habitat in the inland has not been completely exhausted by western hemlock yet, suggesting the species could still be expanding post-glaciation (Gavin & Hu 2006). Rosenburg et al. (2003) only found record of western hemlock pollen in southeastern BC at ~ 3500 ya. Most pollen records concur that western hemlock pre-dates evidence of western red cedar (Mehringer 1996, Whitlock 1992). This coincides with molecular evidence for *Thuja plicata* samples across the disjunction (O'connel *et al.* 2008) that supports one southern coastal refugia throughout the Pleistocene, with no evidence for ancient, disjunct inland refugia or northern coastal refugia, such as one proposed on Haida Gwaii. It is also inferred 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'connel *et al.* 2008). While this inference was based on eight microsatellite loci, recent genetic advances with reduced representation sequencing (Peterson *et al.* 2012, Andrews *et al.* 2016) provide enhanced power to infer population genetic and phylogeographic processes amongst the disjunct populations (Carstens *et al.* 2012, Garrick *et al.* 2015).

Though most of the evidence supports the recent, post-glacial existence of the ITR, that has not prevented phylogeographers from studying the history and impact of the disjunction on other species in the PNW temperate rainforest (Soltis *et al.* 1997, Brunsfeld *et al.* 2001, Gavin *et al*. 2006). To date, eleven 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,* including *D. copei, D. ensatus, 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). These species span the tree of life and, according to their genetic variation, they also span the possible phylogeographic histories for the PNW temperate rainforest. Some species, such as the tailed frogs (Neilson et al. 2001) and *Plethodon* salamanders (Carstens *et al.* 2004) show evidence of an ancient divergence between the ITR and coastal populations, indicating Pre-Pleistocene divergence, while other species such as *Salix melanopsis* (Carstens *et al.* 2013) and *Microtus richardsoni* (Carstens *et al.* 2005) show evidence of post-glacial recolonization of the inland from the coast. Other phylogeographic scenarios such as pre-Pleistocene divergence with migration have also been supported with genomic evidence (*Alnus rubra;* Ruffley *et al.* 2018). The history of the entire ecosystem is complex and has thus garnered appropriate attention from biologists curious about the impact of the distribution on all of the species in the ecosystem, especially the plants (Soltis *et al.* 1997). Inferring the phylogeographic history of the species that established the boundaries of the PNW temperate rainforest will provide a critical insight for the availability of suitable habitat for refugia populations in the ITR.

The idea of the ITR having an ancient, or pre-Pleistocene, divergence from the coastal rainforest and persisting throughout glaciation in refugia in the interior Northwest is compelling because it would support the habitat requirement of other species that show evidence of ancient vicariance. Additionally, paleontologists have even questioned the plausibility of the old-growth ITR becoming so established in less that 3500 years (Mehringer 1996). However, this idea that the ITR persisted through the Pleistocene remains unsupported by paleontological data, specifically the pollen record in the ITR (Mehringer 1996, Chase *et al.* 2008, Gavin *et al.* 2009). Whether or not the ITR persisted throughout the Pleistocene has other implications for whether the PNW disjunct community as a whole has adapted to the dramatic climatic changes 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 theory of a recently assembled, rapidly diverse ITR poses a challenge to this insight.

This idea is not necessarily novel to the PNW temperate rainforest, as the idea of the species responding individualistically, instead of in-concert, in response to climatic changes is an ecological theory dating back to the early 20th century (Gleason 1926) and has been shown empirically (Burbrink *et al.* 2016). Though J. E. 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, this idea of community-wide concerted response to climatic change can be traced back to early 20th century plant ecologist, Clements (1918) and his idea that communities are “super-organisms” whose interactions are interwoven and dependent on one another. Regardless of the organism or ecosystem though, 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 first made predictions about the phylogeographic history for these two species, specifically with respect to whether or not they harbor cryptic diversity across the disjunction, i.e. show evidence of pre-Pleistocene divergence and no subsequent migration. These predictions serve as a test to the predictive framework that was originally developed by Espindola *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. After assessing population structure amongst the data, we construct eleven demographic scenarios to test using a machine-learning model selection framework (Smith & Carstens 2020). These alternative demographic hypotheses include divergence between the coastal and inland populations of western red cedar and western hemlock that occur either before or after the Pleistocene glaciations. The pre-Pleistocene divergence scenarios are meant to model the populations diverging at the time of the forest disjunction (Waring & Franklin 1979), which follows the uplift in the cascade mountain range (Priest 1990). The recent, post-Pleistocene divergence between the populations is meant to model the ITR diverging from the coastal populations only after the ITR was recolonized by coastal migrants, meaning the time of divergence between the populations would be very recent, as the coastal migrants could only have recolonized the inland, at the very earliest, after last glacial retreat, ~ 10 ka (Waitt and Thorson 1983). The varying migration scenarios include divergence with migration, where migration eventually ends between the coast and ITR populations a significant time after divergence. Divergence with secondary contact indicated migration begins again between the coast and ITR populations, at the very earliest, after the retreat of the Cordilleran ice sheet, ~ 10 ka (Waitt and Thorson 1983). The bottleneck events that are model are those that in theory occurred in the populations at the onset and for the duration of the Pleistocene and the following population expansion events occur after the retreat of the glaciers, more likely as recent as 3500 ya (Whitlock 1992, Mehringer 1996).

To test these models, we simulate genomic data under them, data that is similar to the genomic data we have generated for *Thuja plicata* and *Tsuga heterophylla*, and then use that simulated data to train a randomForest (Breiman 2001, Liaw & Weirner 2002, Pudlo *et al.* 2018) classifier to distinguish between the models, as in delimitR (Smith *et al.* 2020). The benefit to using delimitR is that the simulated data are in the form of a folded site frequency spectrum (SFS), or in the case of two populations, a joint folded site frequency spectrum (jSFS), and even in the case of multiple populations, a multidimensional folded site frequency spectrum (mSFS). For demographic model selection with genomic data, the jSFS is beneficial because it summarizes much of the genomic data into one statistic that can be used for inference (Gutenkunst *et al.* 2009, Xu & Hickerson 2015). Following the constructing of a demographic model RF classifier, we make predictions for the demographic histories of *Thuja plicata* and *Tsuga heterophylla*. Given the model with the highest prediction probability, we then estimate the parameters of the model, with confidence intervals and a focus on the divergence time between the coast and ITR populations, followed by an assessment of model fit. We specifically assess whether the divergence times indicate pre- or post-Pleistocene divergence between the ITR and the coastal forest, and whether the confidence intervals of the divergence times overlap between *Tsuga heterophylla* and *Thuja plicata* indicating a synchronous divergence between ITR and the coast.

Ultimately, we've generated genomic data for *Thuja plicata* and *Tsuga heterophylla* species across their coastal and inland populations' disjunction and used that genomic data to investigate whether the ITR is a result of pre-Pleistocene divergence from the coast, or a result of post-glacial recolonization from the coast. For this, we rely on coalescent simulations, the jSFS, 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. Additionally, we explore the role of genomic data in uncovering the history of the past and how our inferences can be influenced by various datatypes and perspectives in genomics and paleontology.

Methods and Materials

Field Sampling and Sequencing

Field collections were made throughout the coastal and inland PNW temperate rainforest for western red cedar, *Thuja plicata*, and western hemlock, *Tsuga heterophylla*, between April and June of both 2016 and 2018. Fresh tissue for specimens were dried and stored in silica gel. Voucher specimens of collections were preserved in the Stillinger herbarium and can be located on the PNW consortium (http://pnwherbaria.org). Leaf tissue from 137 *Thuja plicata* individuals (Figure 3.1) and 50 *Tsuga heterophylla* (Figure 3.1) individuals were extracted using a modified CTAB protocol (Doyle & Doyle 1987), purified using Sera-Mag SpeedBeads (Thermo Fisher Scientific; Rohland & Reich 2012), and quantified using a Qubit 1.0 Fluorometer (Life Technologies).

Three double digest restriction site associated DNA sequencing (ddRADseq) libraries (Peterson *et al.* 2012) were prepared: two for *Thuja plicata*, splitting the total number of samples between them, and one for *Tsuga heterophylla*. For both *Thuja plicata* libraries, the restriction enzymes used were *EcoRI* and *SbfI* (New England Biolabs, USA), along with a size selection window of 200-500 bp For Tsuga, the restriction enzymes used were *SbfI* and *MspI* (New England Biolabs, USA) with a size selection window of 200-500 bp. 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 *Tsgua 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, though the average coverage was 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 use single end reads. Complete assembly procedures were performed and documented in Jupyter notebooks and can be accessed at github.com/ruffleymr/ThujaTsugaAnalysis/IpyradNotebooks.

Predictive Phylogeography

To make predictions about whether or not *Thuja plicata* and *Tsuga heterophylla* harbor cryptic diversity we constructed a random forest classifier. For the predictor variables, we gathered occurrence data previously used for predictive phylogeography of species in the PNW (Espindoal *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). This occurrence data is a combination of GBiF records and field collections, and it was used to gather bioclimatic variables from WOLRDCLIM version 2 (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 (Sullivan *et al.* 2019), were used as the predictor variables in the RF classifier. The response variables, *i.e.* what we want to predict, was the index of “cryptic” or “non-cryptic”. We build four different classifiers using different combinations of the predictor variables we had available: bioclimatic variables only, bioclimatic variables and taxonomy, bioclimatic variable and life history traits, bioclimatic variables and taxonomy and life history traits. We reported the overall error rates for these classifiers.

With each of these classifiers, we predicted the presence of cryptic diversity for *Thuja plicata* and *Tsuga heterophylla*, separately. We gathered occurrence records for the species in question, *Thuja plicata* (791; 569 GBIF records, 222 field collections) and *Tsuga heterophylla* (468; 346 GBIF records, 111 field collections), also compiled from both GBiF records and field collections. We excluded all occurrence records from GBiF that fell outside of the PNW temperate rainforest (35° to 65° latitude, −160° to −100° longitude). We used these locality coordinates to download 19 bioclimatic variables from WOLRDCLIM version 2 on 5 Feb 2019 (Fick & Hijmans, 2017) at a resolution of ~1 km_2. We also assembled trait data to coincide with the trait data collected for PNW taxa for predictive phylogeography with life history information (Sullivan *et al.* 2018). Using this data, we followed the procedure of Sullivan et al. (2019) to predict the presence and absence of cryptic diversity using the four RF classifiers we constructed with different combinations of predictor variables. Finally, we included the new data gathered here for both species to assess how well each classifier improved in overall accuracy with the addition of two plant species.

Population Structure

To identify possible population structure, we explored the *ddRADseq* data from both species 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 with the first 10% discarded 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. Structure Harvester (Earl & vonHoldt 2012) was then used to evaluate the best K using the Evanno method (Evanno *et al.* 2005).

joint Site Frequency Spectra

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 chromosomes in the population. For a diploid organism, this is twice the number of individuals. A jSFS is then the combination of two SFS 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 $1/_{N_{pop1}}$ to N_{pop1} and each column is the allele frequencies in the second population, again beginning with 0 and ranging from $1/_{N_{pop2}}$ to N_{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 estimated. Thus, when estimating demographic parameters from these models though, 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 is due to the common problem of 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 downsampled the number of SNPs and alleles (chromosomes in the population) to construct three different jSFS data sets for each species. 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 data set (Table 3.2). These data sets thus represent a spectrum of genomic information ranging from more individuals in the population but less SNPs and fewer individuals represented from the populations, but a lot more SNPs included. We used unlinked SNPs for model selection to satisfy the assumption that each SNP is independent of each other. 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 data sets for each species and then used those ratios, multiplied by the total number of biallelic SNPs used in the empirical jSFS.

Demographic Modelling

For demographic inference, we used the R package delimitR (Smith & Carstens 2020) which relies on used jSFS and machine learning algorithm, abc-randomForests (Pudlo et al. 2015) for model selection. For this, we simulate jAFS under eleven demographic scenarios we deem plausible for both species (Figure 3.2) using fastsimcoal2 (Excoffier 2011, Excoffier *et al.* 2013). In delimitR, the simulated jSFS is summarized by flattening the matrix and binning the cells into a more-course representation of the jSFS. Using this array of binned, joint site frequencies as the predictor variables, we construct a randomForest classifier to delimit between the eleven demographic models, or the response variables. The classifier will simultaneously cross-validate itself by testing the accuracy of the decision trees being constructed. For this, data that are not used to construct specific decision trees are then used to make predictions on using those trees. Thus, the data being tested is not included in the construction of the decision tree 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 3.2). We then used the appropriate classifier to make predictions for the 100 corresponding subsampled jSFS. We summarized the support for each dataset in the number of votes for the best model and the estimated posterior probability for the best model.

Parameter Estimation

Once the best model was identified for each species, we used Fastsimcaol2 to estimate the demographic parameters of the model and their 95% confidence intervals. For this, we considered 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, known as a conditional expectation maximization (ECM; Brent 1974, Meng and Rubin 1993) algorithm for maximum likelihood optimization, which is considered an algorithm that can get stuck in local optima of the likelihood surface with not-optimal parameter estimates. Therefore, we performed 100 independent parameter optimizations with different initial values, 100000 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 simulations using the maximum likelihood 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).

We also used maximum likelihood estimates from the parametric bootstrap simulation to perform an assessment of model adequacy. For this we perform a hypothesis test using the likelihood ratio G-statistic, which is calculated as $CLR = \log_{10}(CL_0/Cl_E)$, where CL_0 is the relative observed maximum composite likelihood and CL_E is the estimated maximum composite likelihood (Excoffier *et al.* 2013). We calculated this test statistic for all of the parametric bootstrap simulations, and this served as the null distribution for the hypothesis test. We calculated the p-value as the number of test statistics in the null distribution that were greater than the observed G-statistics.

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 had a total of 124,4484 loci with 214,183 SNPs for *Thuja plicata* and 142,804 loci with 893,487 SNPs for *Tsuga heterophylla*, all of which were shared across a minimum of 4 individuals per species. When constructing the jSFS for these species from this data, the data was downsampled considerably such that each SNP was represented in all individuals included in the jSFS (Table 3.2). From this, we see the number of inland and coastal chromosomes we include in the jSFS is, at a maximum, half of the total samples in each population. In using the jSFS to make our inference about demographic histories, we are excluding a considerable amount of sequence data that we have generated. Albeit the models are distinguishable with the data used (Figure 3.4), but what does this mean about the amount of data we are collecting to ask phylogeographic questions.

Predictive Phylogeography

Before assessing whether these species truly 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 randomForest 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. We assessed how well the classifiers worked with each combination of the input data (bioclimatic, trait, and taxonomy 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 that both species do not harbor cryptic diversity (Table 3.1), with the only variation in the prediction being with the classifier that only used bioclimatic data, which also happens to be the classifier with the highest error rate. Whether or not this means both species are in fact non-cryptic still needs to be verified using genomic data as these classifiers, however accurate, are still only based on a handful of taxa, 12 total species/complexes, of which only two are plants (*Alnus rubra* and *Salix melanopsis*).

Population Structure

We investigated the population structure for both species using STRUCTURE (Pritchard *et al.* 2000) for possible K's of 1 through 10. For *Thuja plicata*, we found that best K value, according to Evanno's delta K method (Evanno *et al.* 2005), was K = 3, indicating a model of three distinct genetic clusters best fit the data. When we visualize which samples belong to each of the three clusters (Figure 3.3), we see no geographic association with the samples that belong to that third cluster. We do see that the other two clusters appear to be associated with the coast and inland. When we look at the sampling localities colored by K = 2 clusters (Figure 3.3C), we see this same pattern between the two clusters being associated to the coast and inland, albeit the quite a bit of the coastal samples associating more with the inland samples.

For *Tsuga heterophylla,* we found that best K value, according to Evanno's delta K method was K = 2, indicating two genetic clusters fit the data the best. Visualizing where those two clusters occur for the Tsuga samples (Figure 3.3), we do not see a geographic association between the coastal and inland samples with the two clusters. However, when we investigate K = 3, or three genetic clusters, we do see a heavy association with one cluster in the inland and one cluster along coast, albeit with some mixing amongst the cluster. Of course, there is much speculation with respect to the power of the Evanno method in truly determining the best "K" (Janes et al. 2017).

Demographic Modelling

The population structure results provided a good basis for deciding how many populations to model in the demographic models investigated. Ultimately, we decide to model two populations, where we group the samples based on whether they were sampled in either the inland or coastal forest. The potential third cluster contained samples that were not clustered geographically with one another and as such these could not be considered or treated as a third population in any demographic model. (Figure 3.3).

We developed eleven demographic models (Figure 3.2) to explain the phylogeographic history of each species. In these models, we considered both ancient and recent divergence events, altering migration scenarios, including divergence with and without migration and secondary contact. We also model possible bottleneck events associated with when the Pleistocene began, ~2.5 mya. We modeled population expansion as well, to be associated with population regrowth at the very earliest, after glacial retreat ~10 ka (Figure 3.2). We used fastsimcoal2 to simulate DNA sequence data, using 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 data set sizes for each species. When using jSFS, no missing data can be included, therefore, to use a particular locus, it must be present in all individuals. Thus, the more individuals you include, the fewer SNPs are typically shared across all of them. The three data set sizes had roughly 10,000, 3,000, and 1,000 unlinked SNPs (Table 3.2) and we did not consider the monomorphic cell for the jSFS for demographic model selection.

To distinguish the simulated jSFS data based on the model they were simulated under, we used delimitR (Smith et al. 2019), which uses the flattened jSFS matrix binned into a fewer number of cells as the predictor variables in a randomForest classifier, with the model identifier as the response. We used 10,000 simulated jSFS for each model when building the classifiers. A different classifier was constructed for each data set size, meaning we repeated all simulations per model for each data set size. This first resulted in the error rates of the classifiers for classifying each of the eleven models (Table 3.2). Most models were classified correctly a majority of the time, with all of them have a classification accuracy above 0.72, except for the models with a recent divergence event. These models with the lower classification rates were all of the models with a recent divergence between the coast and the inland populations. We collapsed these recent dispersal models, which all varied in the presence/absence of migration and bottleneck and expansion events, into a single recent dispersal model (Figure 3.4). When we do this, the overall error rate decreases dramatically (Table 3.2) and the accuracy of the recent dispersal model is 0.90.

The first classifier, with all eleven models, was used to make predictions using the observed jSFS for each species (Figure 3.5). For each data set size, we use 100 different jSFS that were constructed by subsampling unlinked SNPs randomly. For *Thuja plicata*, all data sets had the highest prediction probability for the same model, a model with an ancient divergence event between the coast and inland population, followed by a bottleneck in both populations, and then population expansion happening at the same time as secondary contact between populations ("AV + sc + bot/exp", Figure 3.5). On average each *Thuja* dataset received 552 votes for that model and had an average posterior probability of 0.72 (Figure 3.5).

The results were different for *Tsuga heterophylla* in that each data set did not receive the same prediction probability. Those with more SNPs supported the same model as *Thuja plicata* (Figure 3.5), one with ancient vicariance between the coastal and inland population, followed by a bottleneck in both populations and then population expansion with secondary contact at the time of the glacial retreat (model G, Figure 3.2). On average, this model received 564 votes in the classifier for each observed jSFS and had an average estimated posterior probability of 0.83 (Figure 3.5). With fewer SNPs included in the jSFS, but more samples represented in the population, the model that has the highest prediction probability is model C (Figure 3.2), which is a very similar model to model G, only the population bottleneck and expansion are not included in the model, rather there is just an ancient vicariance event followed by secondary contact, "AV + sc". On average, this model received 532 votes in the classifier for each observed jSFS and had an average estimated posterior probability of 0.78. We expect this model support could be because there are less SNPs to inform the parameters associated with the additional process, bottleneck and expansion, as well as less individuals from the population than *Thuja plicata*.

Parameter Estimation

The parameter estimates for western hemlock and western red cedar generally fit with most of our expectations for 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 3.3), as we know is generally true given their current distributions. All of the events that were dated are in units of coalescent generations. The first event data was the divergence between the ITR and the coastal rainforest. For both species, the median divergence time estimates were approximately 252,000 generations ago (Table 3.3). The time of the population bottleneck event for both species was between 50 and 90 ka (Table 3.3). The time of the population expansion for *Thuja plicata* was ~1050 generations ago, while the time of population expansion for *Tsuga heterophylla* was nearly twice that at 2020 generations ago. The magnitude of the bottleneck on the coast was apparently slightly more sever for *Thuja plicata*, than the inland bottleneck. The opposite was true for *Tsuga heterophylla*, where the bottleneck in the ITR is more severe than that of the one on the coast (Table 3.3). Not surprisingly, the populations with the more sever bottleneck also had the larger population expansion rate (Table 3.3). Note that this expansion rate is modeled backward in time, meaning a negative rate indicates the population is getting smaller as it goes backward in time, 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 3.3).

In order to convert the units of the timed events, which were in units of generations, we needed 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 red cedar, estimates of trees reaching maturity typically range from 20-30 years (Turner 1985), however trees have reached maturity as early as 10 years in some open grown areas (Minore 1990). The same is roughly true for western hemlock, where most estimates suggest maturity is reached between 25-30 years (Owen et al. 1984) but trees have reach maturity much quicker in some cases (Tesky 1992). To be conservative, we assume a generation length of 10 years per generation for both species. In doing this, we can convert our estimates of the time events into years (Table 3.4). Following this, the divergence event between the ITR and coastal population, for both species, is estimated as ~2.5 mya (Table 3.4).

Discussion

History of the ITR

The implication of the demographic modeling suggests that the ITR represents and ancient relic of the PNW temperate rainforest pre-Pleistocene. Genomic evidence from both western red cedar and western hemlock support this ancient divergence between the ITR and the coastal rainforest, with the evidence apparent in the model predictions and the observed allele frequencies. While many studies agree, the identification of refugia or an anciently diverged populations is an abundance of rare alleles not shared with the disjunct population. In previous genetic evidence for western red cedar (O'Connel *et al.* 2008), while they acknowledge some differentiation between interior and coastal populations, it was shallow enough to suggest recent divergence with an absence of subsequent migration. Here though, we've collected thousands of loci across individuals in both ITR and coastal populations. With these data, we've been able to appropriately model coalescent processes, that account for stochasticity and varying demographic events, and ultimately lead us to the inference that there have been refugia ITR populations of western red cedar and western hemlock throughout the Pleistocene that contributed to the genetics of the current ITR populations. Using the jSFS, we can observe the high frequency of rare alleles harbored by the coastal and ITR populations separately, that indicates their ancient divergence.

This has implications for how the paleontological record of pollen informs our understanding on the history of the PNW (Whitlock 1992). In the ITR, these species are not been abundant in the pollen record until < 3500 years ago (Mehringer 1996, Chase *et al.* 2008, Gavin *et al.* 2009). Given the genomic evidence of an abundance of rare alleles in the ITR populations, we propose populations of *Thuja plicata* and *Tsuga heterophylla* in the ITR during the Pleistocene were at low population sizes (Table 3.4), that were potentially spread out in patchy refugia within the interior. Could it be simple enough to suggest the location of these refugia did not produce a significant amount of pollen to be detected? Or maybe the populations were just not where subsequent pollen cores were taken? The former is more probable than the later given the fact that all cores in the PNW, including coastal core samples, indicate that only recently, < 5,000 years ago, did the southern coastal temperate forest begin to recolonize the coast and become more abundant, with the northern coastal expansion following in the coming millennia (Hebda & Mathewes 1984, Whitlock 1992, Meheringer 1996). Similar evidence suggests a very recent, < 2000, northern ITR recolonization as well (Gavin *et al.* 2009). Some have questioned the validity of the pollen identification, given the difficulty of identifying cedar pollen (Faegri & Iverson 1992) and its suggested unreliability in indicating presence of cedars nearby (Gavin *et al.* 2005), but this cannot explain the lack of western hemlock in the pollen record until these recent times.

One of the most compelling pieces of evidence for a southern coastal refugia for western red cedar during the Pleistocene is a macrofossil of *Thuja plicata* identified from the late Pleistocene (~ 35 ka) from the western Cascade Mountains south of the glacial extent in Oregon (Gottesfeld *et al.* 1981). This suggests the location of a western red cedar refugia that potentially existed for some time during the Pleistocene and could be extended as a hypothesized refugium for western hemlock was well. No such fossil evidence exists for the ITR to locate potential refugia, though many locations have been hypothesized as potential refugia. The lack of evidence, however, does not preclude the possibility that these refugia populations did exist in the ITR. Specifically, since the pollen record does not detect the ancient populations that persisted along the coast, we also do not expect them to identify the ancient populations in the inland. The paleontological record of the area has provided insight into the recolonization of the ITR suitable habitat following glacial retreat, and the timing of population expansion. The role now of our genomic data could be to identify the possible location of refugia in the inland in order to guide our research toward areas of paleontological research.

Modern demographic inference

Here, we have gathered thousands of loci from *Thuja plicata* and *Tsuga heterophylla* individuals throughout their disjunct range in order to assess the demographic histories of these species. By using the jSFS, we are able to summarize all of these loci into a single statistic to infer the history from our data with, a feat that is constantly changing with new data acquisition methods and statistical analyses (Carstens *et al.* 2012, Garrick *et al.* 2015). The benefit to using the jSFS is that this statistic can capture information about shared, and not shared, allele frequencies across the populations, which is exactly what we are interested in interpreting given our question of ancient or recent divergence between the coastal and ITR populations. When we visualize the jSFS of our empirical systems, we see the rare alleles harbored by the inland and coastal populations in the high proportion of loci location in the first row and column of the jSFS (Figure 3.5).

Visualizations aside, our model selection procedure has produced consistent results that both western red cedar and western hemlock support a pre-Pleistocene divergence event, albeit with some secondary contact amongst the populations. The approach used here for model selection using randomForests and the jSFS (Smith et al. 2017, Smith & Carsten 2020), has yet to be tested using plant species or demographic models of this complexity. This approach is a likelihood-free approach based on simulating allelic data under coalescent stochasticity and demographic processes. The model selection with machine learning is only as accurate as your data are distinct in model space. This means that should the data be insufficient to distinguish between these models; you would see this in the error rates of your classifier. Our data show that the error rates in all of our classifiers are extremely low, indicating high confidence in our classifier and our data's ability to distinguish between the eleven demographic scenarios we propose. This approach provides flexibility to the demographic model designs and simulation of data, as well as computational efficiency.

The use of the jSFS to summarize genomic data does have its limitations. As mentioned, when we summarize our data into a single jSFS, we have to downsample data so that every SNP is included in each individual in the jSFS. We note that doing this requires us to forfeit a considerable amount of the data we have generated (Table 3.2). We performed a sensitivity analysis on the use of the jSFS by constructing three different data set sizes, of 100 jSFS each, for each species, *Thuja plicata* and *Tsuga heterophylla* (Table 3.2), which resulted in 100 model predictions per species, per data set (Figure 3.5). The biggest discrepancy in the entire inference is within the *Tsuga heterophylla* prediction where a different demographic history is support by the jSFS that included more individuals from the population and less SNPs and the jSFS with the most SNPs and fewest individuals (Figure 3.5). 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. While the difference in the model support could come down to the data set with more SNPs being able to estimate the bottleneck and expansion parameters more effectively, and therefore show strong support for the model. Whereas the data with fewer SNPs may have just been lacking data for those processes. This does produce a larger question though about the consistency of our results given the level of courses, or resolution, we allow in the construction of jSFS.

Predictive Phylogeography

Prior to the addition of *Thuja plicata* and *Tsuga heterophylla* into the predictive framework for identifying cryptic diversity (Espindola *et al.* 2016, Sullivan *et al.* 2019), the error rates for the classifier were already extremely low. Thus, adding these species does not necessarily drastically improve our ability to make accurate predictions, though it does not decrease them either (Table 3.1). We made the predictions of non-cryptic for both species, prior to assessing their true phylogeographic history with genomic data, and for both species we see that the predictions were correct. While western hemlock and western red cedar do show evidence for a pre-Pleistocene divergence, they also show evidence of post-Pleistocene gene flow through the non-zero estimation of migration rates between the populations (Table 3.3). While adding these species to the predictive framework does not change the accuracy in the prediction, we are still adding information to the classifier about plant species specifically that will contribute to the accuracy in other predictions for plant species with disjunct ranges in the ITR.

Conclusion

Using genomic evidence and modern demographic inference procedures with machine learning, we are able to show evidence of ancient ITR populations for *Thuja plicata* (western red cedar) and *Tsuga heterophylla* (western hemlock) that persisted throughout the Pleistocene. The recent expansion of ITR populations in these species and colonization of newly suitable habitat within the past 3500 years, does not contradict this finding. The refugia populations in the ITR were likely of small population sizes, as we show support here for Pleistocene-related population bottleneck events in both species. Likewise, we show support for the recent population expansion of these species in the ITR within the last 10000 years. This evidence does coincide with the paleontological record that the temperate rainforest did not dominate the PNW landscape until the last 5000 years, and only in the last 3500 years did the ITR begin recolonization. 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 red cedar previously, suggesting the ITR populations are a result of coastal recolonization. While we agree coastal migrants contributed to the genetic architecture of the current ITR populations, we also argue that ancient 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.

Literature Cited

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.

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.

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.

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

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

Espíndola 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.

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-Sánchez 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.

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.

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.

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.* Elsevier, Amsterdam, section 2.3.5.

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.

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.

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, Espíndola 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.

Rognes T, Flouri T, Nichols B, Quince C, Mahé 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

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.

Tables

Table 3.1. Phylogeographic predictions of cryptic and non-cryptic 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 indication 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*.

		Thuja plicata		*Tsuga heterophylla*		
Predictor variables	Error Rate	Non-cryptic PP	Cryptic PP	Non-cryptic PP	Cryptic PP	Updated Error Rate
Bioclim	0.205	0.713	0.287	0.669	0.331	0.155
Bioclim,Taxonomy	0.0	1.0	0.0	1.0	0.0	0.0
Bioclim,Traits	0.0	1.0	0.0	1.0	0.0	0.0
Bioclim, Taxonomy, Traits	0.0	1.0	0.0	1.0	0.0	0.0

Table 3.2. The average number of unlinked SNPs used in the 100 empirical data sets, with the corresponding number of samples from the coastal and inland populations, where each sample represents an allele for an individual, most often both allele areas 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.

		SNPs	coastal	inland	Error Rate All Models	Error Rate RD collapsed
Thuja plicata	1	9036	11	10	27.90	12.93
	2	2484	26	24	30.89	15.83
	3	1041	42	42	33.87	19.23
Tsuga heterophylla	1	7929	13	8	27.10	12.26
	2	2698	19	12	31.21	15.94
	3	1195	27	16	34.78	19.88

Table 3.3. Parameter estimates for *Thuja plicata* and *Tsuga heterophylla* for the model selected most 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} and T_{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.

	Thuja plicata			*Tsuga heterophylla*		
	MaxL Estimate	min 95% CI	max 95% CI	MaxL Estimate	min 95% CI	max 95% CI
N inland	1317431	1118973	1611659	1574048	1165534	1938174
N coast	1514468	1301011	1673645	3361225	2865576	3708440
Tdiv	252814	231341	295009	252285	223890	295967
Tbot	53436	50545	62430	59417	51439	92440
Texp	1043	1010	1095	2021	1592	2290
BtnMag inland	0.5930	0.5391	0.6706	0.4792	0.4403	0.4992
BtnMag coast	0.4791	0.4336	0.4982	0.5888	0.5230	0.7201
Gro inland	-1.7E-04	-3.1E-04	1.9E-05	-4.4E-04	-5.0E-04	-3.9E-04
Gro2 coast	-6.8E-04	-8.0E-04	-5.9E-04	-1.7E-04	-2.1E-04	-1.1E-04
MIG inland -> coast	1.4E-05	1.0E-05	1.9E-05	1.2E-05	6.2E-06	2.0E-05
MIG coast -> inland	1.1E-04	9.5E-05	1.2E-04	4.5E-05	3.4E-05	5.7E-05
mutation rate	5.7E-09	5.0E-09	6.1E-09	4.6E-09	4.2E-09	5.1E-09

Table 3.3. Divergence time estimates and time of population expansion and secondary contact estimates for both species. Estimates are in years that were calculated from multiplying the divergence time in generations by an estimate generation length of 10 for both species.

	MaxL Estimate	min 95% CI	max 95% CI
Thuja plicata			
T_{div}	2,528,140	2,313,410	2,950,090
T_{exp}	10,430	10,100	10,950
Tsuga heterophylla			
T_{div}	2,522,850	2,238,900	2,959,670
T_{exp}	20,210	15,920	22,900

Figures

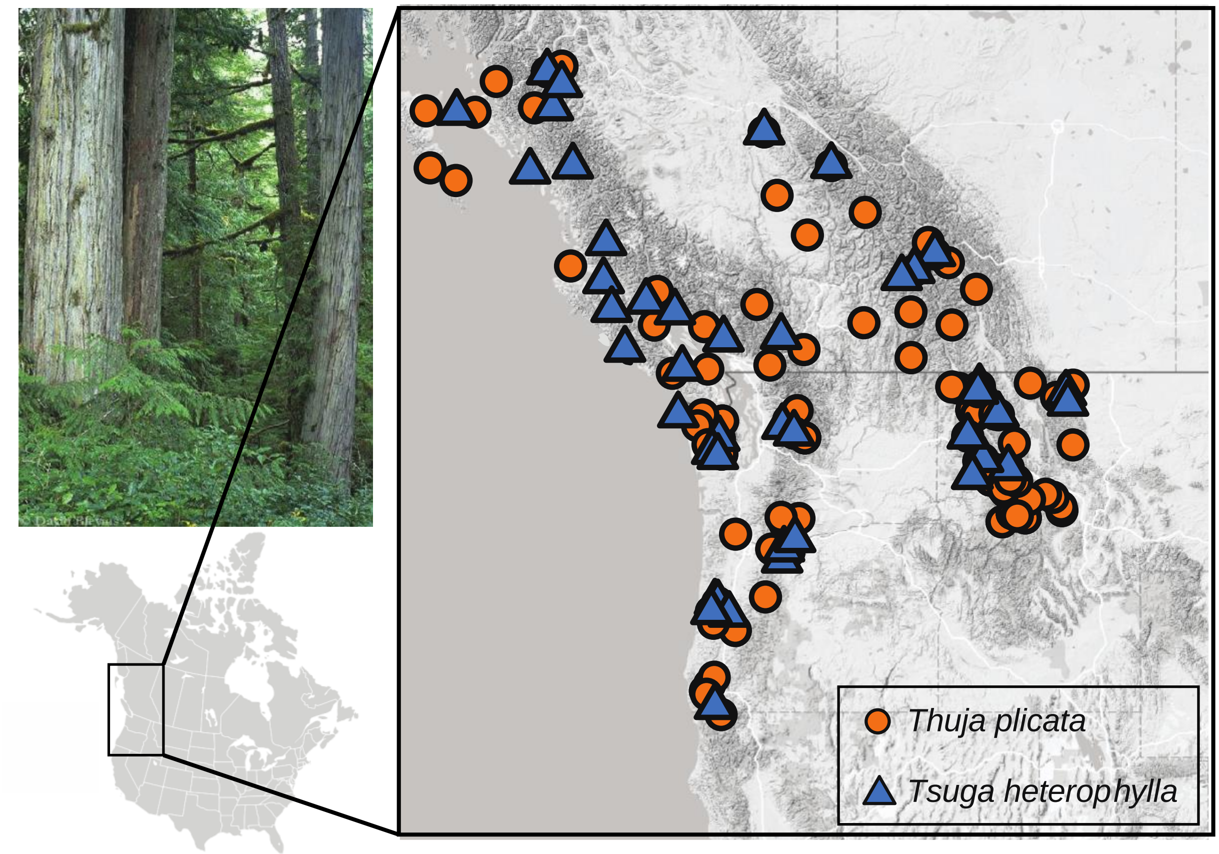


Figure 3.1. Localities sampled for western Red Cedar and western Hemlock. Locality information for each collection can be found in Supplemental Table 1.

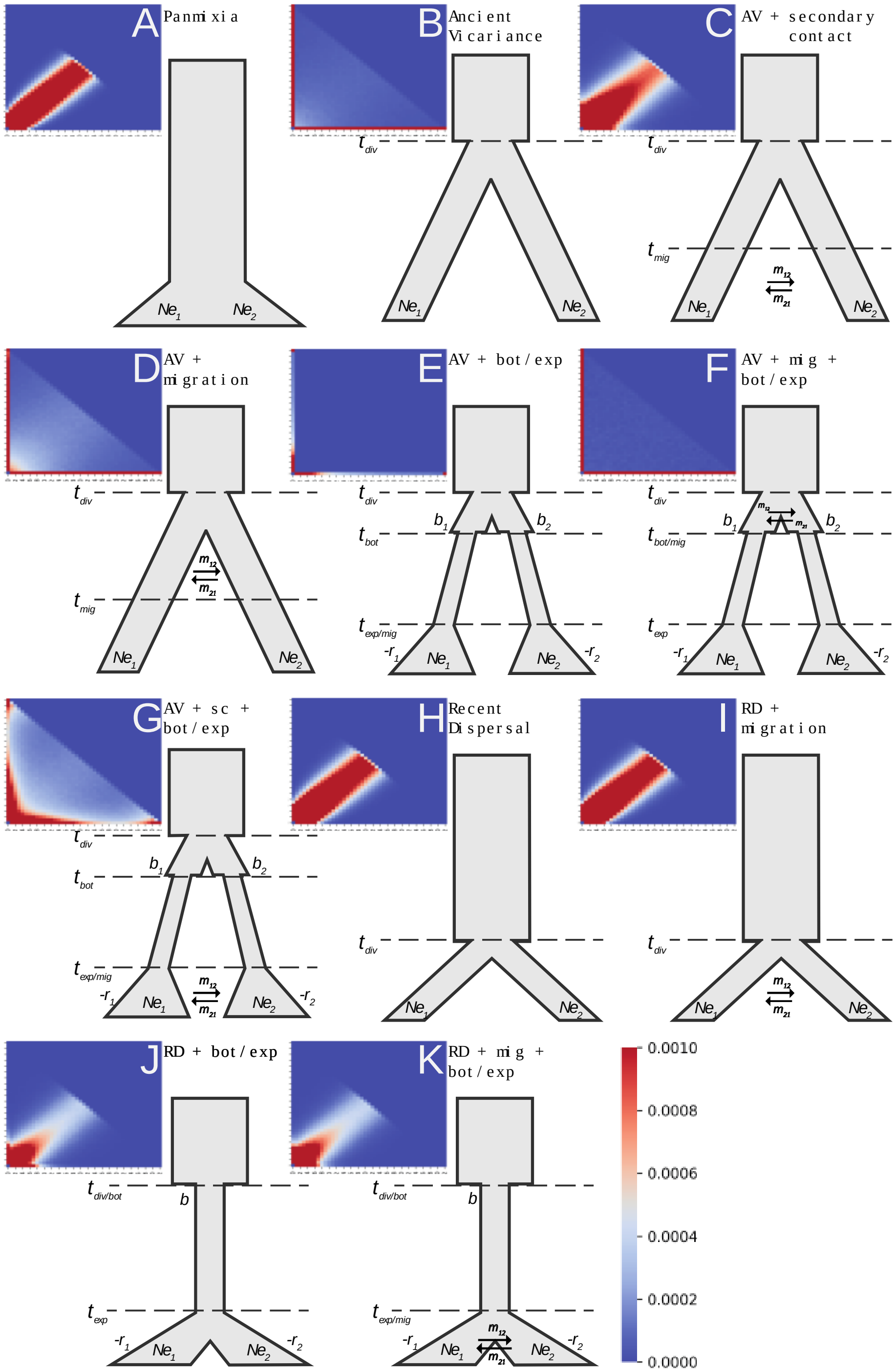


Figure 3.2. (A-K) Summarized folded jSFS (43 by 43 cells) for 10,000 simulations under each associated demographic model. Scale indications the proportion of loci in each cell, with 0.001 being the maximum, meaning if the proportion is higher than this value, the color is that same as the maximum. Dashed lines represent all events that can occur in a given model, including divergence (*div*), bottleneck (*bot*), and 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.

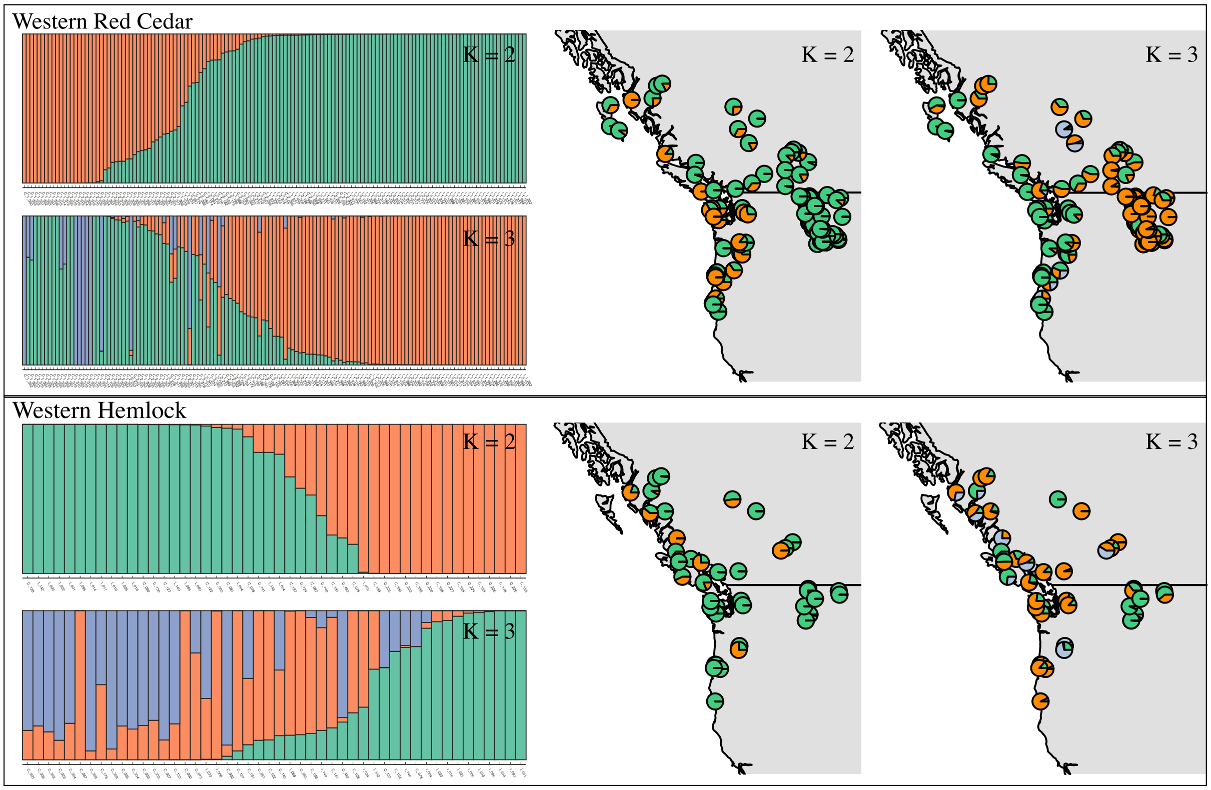


Figure 3.3. left panels: STRUCTURE results for *Thuja plicata* and *Tsuga heterophylla* where each bar indicates an individual in the population and the color indicates the proportion of genetic variation associated to a particular cluster. Clusters indicated by K values in the top right corner. Coastal 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 = 2.

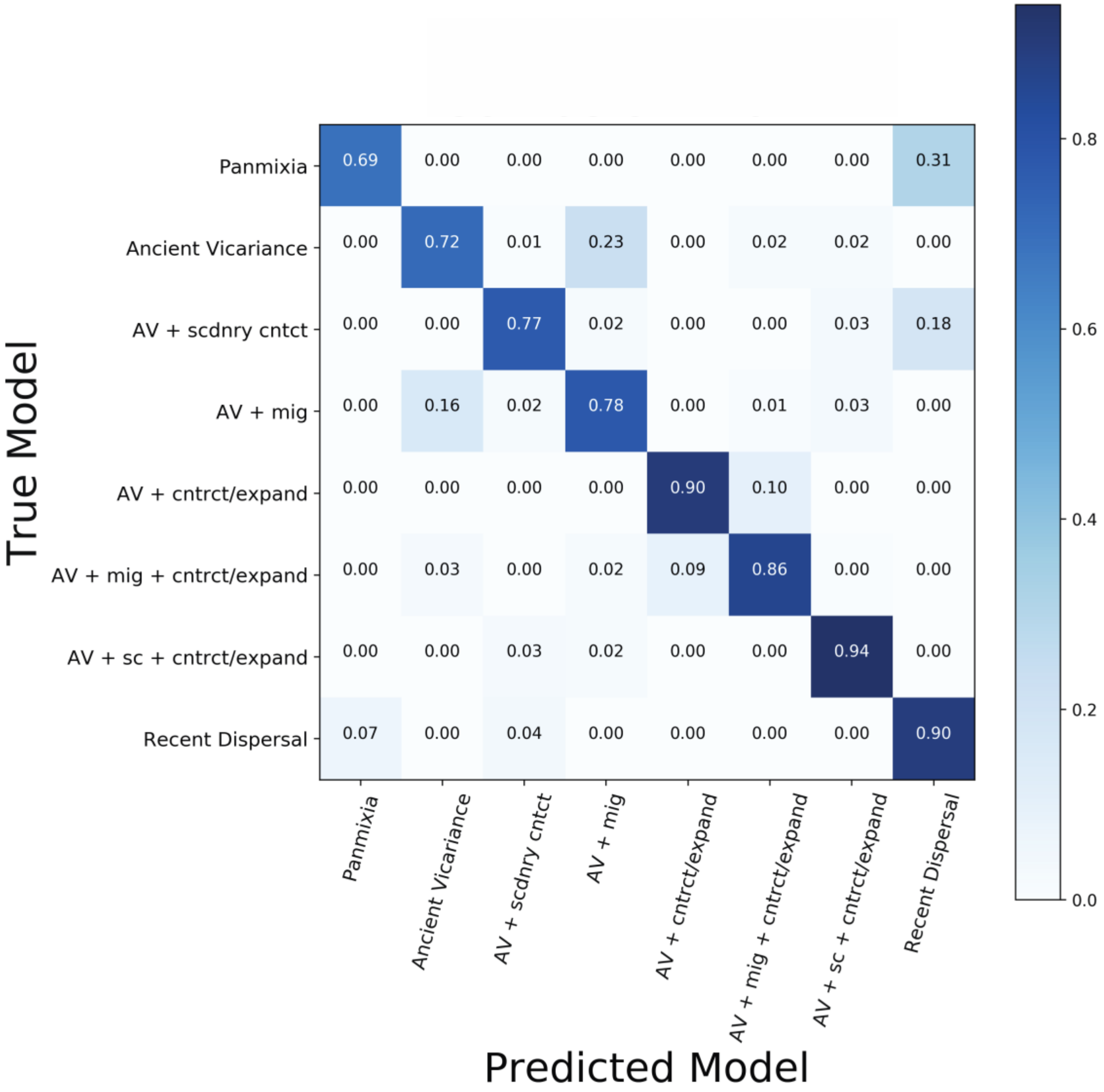
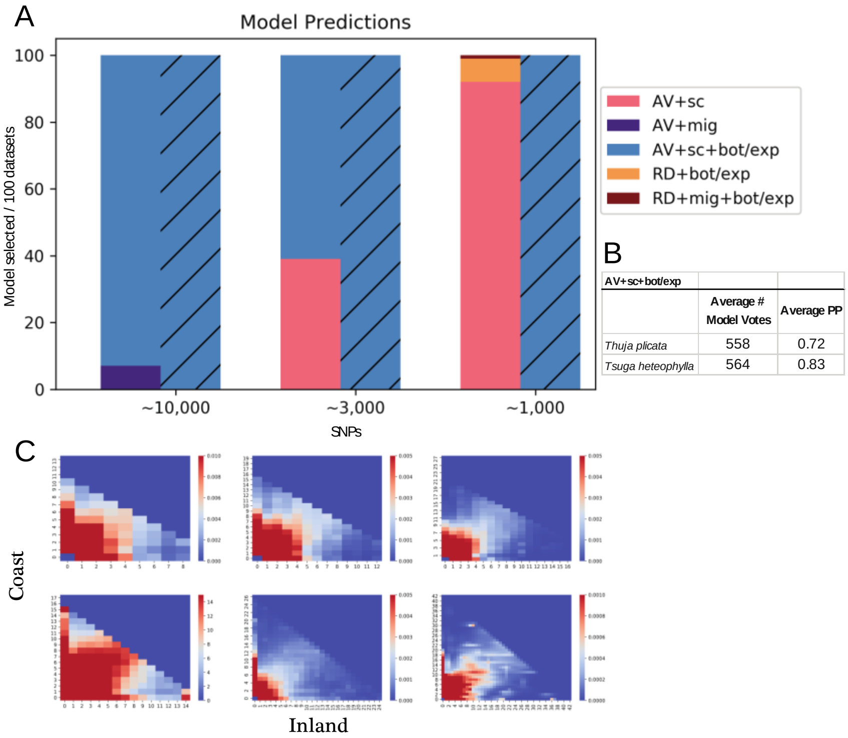


Figure 3.4. Confusion matrix depicting the prediction accuracies and inaccuracies for 11 demographic models using delimitR for model selection, which involves the simulated and binned jSFS and 'abcrf'. The rows indicate the model the data were simulated under and the columns indicate the model that was 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.



AV+sc+bot/exp		
	Average # Model Votes	Average PP
Thuja plicata	558	0.72
Tsuga heteophylla	564	0.83

Figure 3.5. 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).

Appendices

License Agreement for Chapter 1

This Agreement between 875 Perimeter Drive ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Copyright Clearance Center.

License Number	4840981447484
License date	Jun 02, 2020
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Molecular Ecology
Licensed Content Title	Combining allele frequency and tree-based approaches improves phylogeographic inference from natural history collections.
Licensed Content Author	Megan Ruffley, Megan L. Smith, Anahi Espindola, Bryan C. Carstens, Jack Sullivan, David C. Tank
Licensed Content Date	Feb 11, 2018
Licensed Content Volume	27
Licensed Content Issue	4
Licensed Content Pages	13
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Electronic
Portion	Full article
Publisher Tax ID	EU826007151

Terms and Conditions

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand- alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, and any CONTENT (PDF or image file) purchased as part of your order, is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts, You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY

NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

License Agreement for Chapter 2

Open Access Article
This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

For an understanding of what is meant by the terms of the Creative Commons License, please refer to Wiley's Open Access Terms and Conditions.