doi:10.1111/evo.1387

Process-based species delimitation leads to identification of more biologically relevant species*

Megan L. Smith[1,2] and Bryan C. Carstens[1]

[1]*Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio 43210*

[2]*E-mail: megansmth67@gmail.com*

Received March 28, 2019

Accepted October 12, 2019

Most approaches to species delimitation to date have considered divergence-only models. Although these models are appropriate for allopatric speciation, their failure to incorporate many of the population-level processes that drive speciation, such as gene flow (e.g., in sympatric speciation), places an unnecessary limit on our collective understanding of the processes that produce biodiversity. To consider these processes while inferring species boundaries, we introduce the R-package *delimitR* and apply it to identify species boundaries in the reticulate taildropper slug (*Prophysaon andersoni*). Results suggest that secondary contact is an important mechanism driving speciation in this system. By considering process, we both avoid erroneous inferences that can be made when population-level processes such as secondary contact drive speciation but only divergence is considered, and gain insight into the process of speciation in terrestrial slugs. Further, we apply *delimitR* to three published empirical datasets and find results corroborating previous findings. Finally, we evaluate the performance of *delimitR* using simulation studies, and find that error rates are near zero when comparing models that include lineage divergence and gene flow for three populations with a modest number of Single Nucleotide Polymorphisms (SNPs; 1500) and moderate divergence times (<100,000 generations). When we apply *delimitR* to a complex model set (i.e., including divergence, gene flow, and population size changes), error rates are moderate (~0.15; 10,000 SNPs), and, when present, misclassifications occur among highly similar models.

KEY WORDS: **Ecological speciation, machine learning, reinforcement, speciation, species delimitation.**

Historically, investigations that seek to identify species limits have been largely independent from those that explore the process of speciation. Due to recent advances in high-throughput sequencing techniques, evolutionary biologists can now collect tens of thousands of SNPs from any species complex at a reasonable cost. Such efforts have led to a rapid increase in the magnitude of phylogeographic datasets that may be informative at the level of the species boundary (Garrick et al. 2015). The field has seen a corresponding increase in available methods for species delimitation that use molecular data (e.g., Yang and Rannala 2010; Ence and Carstens 2011; Carstens et al. 2013; Leache et al. 2014b), with most relying on the multi-species coalescent model (MSC). Although the MSC is a powerful framework for estimating population sizes and divergence times while accounting for ancestral polymorphism and incomplete lineage sorting (Rannala and Yang 2003), it is limited to situations in which gene flow ceases immediately upon population divergence, corresponding to an allopatric mode of speciation. Allopatric speciation may or may not be common, but it is clearly not the only mechanism by which species arise in nature (Coyne and Orr 2004; Nosil 2012; Zachos 2016).

Processes other than lineage divergence play an important role in some modes of speciation. For example, gene flow during divergence is a hallmark of sympatric speciation (Coyne and Orr 2004), and has been implicated in a range of empirical systems, including Tennessee cave salamanders (Niemiller et al. 2008), flowering plants on Lord Howe Island (Papadopulos et al. 2011), *Heliconius* butterflies (Martin et al. 2013), and *Myotis* bats (Morales et al. 2016). Gene flow may also play an important role during the later stages of divergence via reinforcement, the process by

*This article corresponds to Peede, D., D'Agostino, E. and Ottenburghs, J. 2020. Digest: Species delimitation in the face of demographic processes. Evolution. https://doi.org/10.1111/evo.13919.

which gene flow among divergent populations results in hybrids with lower fitness and thereby increases positive assortative mating among members of the same lineage, leading eventually to reproductive isolation between co-occurring lineages (e.g., Hoskin et al. 2005; Kronforst et al. 2007). Other population-level evolutionary processes, including population size changes, are also important in some proposed models of speciation. For example, in founder-effect speciation (Mayr 1954; Gavrilets and Hastings 2017), a few individuals colonize a new area (generally via long-distance dispersal), and the new population is immediately isolated from the ancestral population (Templeton 2008; Gavrilets and Hastings 2017). Genetic drift and natural selection lead to a shift in adaptive peak, and speciation occurs (Mayr 1954). Ignoring these processes while considering only lineage divergence represents the major barrier to uniting systematic investigations into species limits and evolutionary investigations into the process of speciation, and this may prevent researchers from developing a conceptual understanding of the relative rates of allopatric versus other modes of speciation.

It is also likely that incorporating population-level processes into species delimitation may prevent errant findings. A key assumption of the MSC is that shared genetic polymorphism is a remnant of ancestral polymorphism and not due to gene flow. Indeed, simulation studies have shown that ignoring gene flow leads the MSC to overestimate population sizes and underestimate divergence times (e.g., Leaché et al. 2014a). For example, Bayesian Phylogenetics and Phylogeography (BPP), a popular implementation of the MSC for species delimitation (Yang and Rannala 2010), has been shown to delimit populations as species even when levels of gene flow between populations are moderate (Jackson et al. 2017a; Leaché et al. 2018), which may not be desirable under certain species concepts (e.g., the Biological Species Concept; Mayr 1942). It is clear that species delimitation efforts based on the MSC should proceed with caution, particularly when processes other than lineage divergence may have been important during speciation. However, few studies have considered other parameters when delimiting species (but see Camargo et al. 2012; Jackson et al. 2017a; Morales and Carstens 2018, which consider migration) primarily due to computational limitations.

Here, we introduce an approach that allows researchers to directly investigate the processes of speciation, ranging from allopatric speciation to founder-effect speciation to isolation with secondary contact. We implement this method in *delimitR*, an R-package that conducts demographic model selection under the coalescent. It builds upon an approach to demographic model selection (Smith et al. 2017) that represents SNP data using the site frequency spectrum (SFS) and uses machine learning to compare demographic models. Users can apply *delimitR* to their data using a default model set and user-specified priors where models differ not only in the presence or absence of population-level processes (i.e., migration and population size change) but also in the number of species or populations included. This default model set can also be amended or replaced entirely to tailor it to any focal system. This flexible framework enables researchers to design a model set based on prior knowledge of their focal taxa, and to compare different models of speciation (and models in which a speciation event does not occur). *delimitR* thus allows users to identify the process by which speciation occurred in their focal taxa while simultaneously inferring species limits. We describe *delimitR* below, and then use *delimitR* to conduct a preliminary investigation into speciation and species limits in the reticulate taildropper slug, *Prophysaon andersoni*. Finally, we apply *delimitR* to three published datasets to compare inferences made using this method to existing interpretations and evaluate its performance using simulations.

Materials and Methods

SPECIES DELIMITATION IN *delimitR*

delimitR modifies and expands an approach introduced by Smith et al. (2017) that uses the SFS and a random forest (RF) classifier to perform demographic model selection. Briefly, under the algorithm described by Smith et al. (2017), the user defines a set of models and specifies these models by hand in the coalescent simulator fastsimcoal26 (fsc26; Excoffier et al. 2013). Data are then simulated under each model in the form of a folded multidimensional SFS (mSFS) and summarized by making cells more coarse and creating a binned SFS. Finally, following Pudlo et al. (2015) an RF classifier is built from these simulated data, error rates are estimated, and the classifier is applied to the observed data to find the model producing data most similar to the observed data. More details are available in Smith et al. (2017).

This work expands on the approach of Smith et al. (2017) such that it can be used for species delimitation. First, we developed an R package (*delimitR*) to automate the simulation and summarization of data and demographic models. *delimitR* generates a default model set that includes divergence (or a lack thereof) and migration (in the form of secondary contact or divergence with gene flow). The user may modify this model space by implementing custom models in fsc26 and placing them in the working directory. Second, by summarizing the data using an SFS with a maximum dimensionality that matches the user-specified maximum number of populations, *delimitR* can evaluate models that differ in the number of lineages present. Although populations may be collapsed in the model (i.e., divergence time between sister species may be 0), the data are consistently summarized based on the maximum number of populations. When generated in this way, the expected SFS under each model differs in the number of SNPs shared between populations, but not in the number of bins used to summarize the data, which enables comparison across models that

differ in the number of lineages. Results from *delimitR* include two pieces of information relevant to species boundaries: (1) how many populations are present and (2) whether gene flow is occurring or has occurred between those populations. Most existing approaches to species delimitation consider only the first piece of information; this contributes to oversplitting that results from population genetic structure. The potential complication is that an interpretation of *delimitR* results requires an explicit definition of the operational species concept. For example, if a user infers a model with three populations, but with gene flow between the most recently diverged sister population pair, the researcher must decide whether these two populations can be considered species based on parameter estimates and external information.

Species delimitation in *delimitR* consists of three steps: (1) The default model set is generated, and the user may add models to this model set. (2) Under either the default or a user-specified model set, data are simulated in fsc26. (3) An RF classifier is constructed from the simulated data, error rates are calculated using out-of-bag (oob) error rates from simulated data, and the classifier is applied to the observed data, as in Smith et al. (2017). Details about generating the model set are given below, along with a brief description of Steps 2 and 3, but see Smith et al. (2017) for additional details.

Generating the model set

There are two approaches to defining a model set in *delimitR*: using the predefined model set generated by the program, or using a user-specified set of models. Under the predefined model set, *delimitR* considers models of divergence, divergence with gene flow, and divergence with secondary contact. To simplify the default model space, *delimitR* requires that the user provides a guide tree or a list of guide trees. The guide tree defines the relationships between putative species and is used to generate models with different numbers of lineages. For example, if the guide tree ((0,1),2) was provided, models with three species (0, 1, and 2), two species (0 + 1 and 2), and one species (0 + 1 + 2) would be considered (Fig. 1). Users may provide multiple guide trees or may include models outside of the default model set.

To generate the default model set, the user also provides a migration matrix, specifying which lineages can exchange genes. For example, given the guide tree above, the user could specify a migration matrix: $\begin{matrix} F & T & F \\ T & F & F \\ F & F & F \end{matrix}$ (Fig. 1). The above matrix specifies that gene flow can only occur between species 0 and species 1. The default model set considers symmetric migration, but asymmetric models can be specified by the user. The user must also specify whether they wish to include secondary-contact models and/or divergence-with-gene-flow models. In secondary-contact models, gene flow occurs between the species specified in the migration matrix but is limited to a recent time period (Fig. 1). Gene flow begins at time zero and ends at half of the time to the most recent coalescent event in the tree, although this timing can also be modified by users. In contrast, the divergence-with-gene-flow models implemented in *delimitR* allow gene flow between sister species that begins halfway between time zero and the coalescent event involving the two species and ends at the coalescent event between the two species (Fig. 1). If multiple migration parameters are considered in a divergence-with-gene-flow model, then gene flow will begin at the minimum of the two possible start times and end at the minimum of the two possible end times. Finally, the user must specify a maximum number of migration events to consider in any single model. For example, if the maximum number of migration events is set to two, no single model can include more than two migration edges. Additional examples are provided in Figures S1–S3. Finally, the user must provide information on divergence time priors, population size priors, migration rate priors, the number of samples per population, and the number of SNPs used to construct the SFS. Because we simulate unlinked SNPs and ignore invariable sites, *delimitR* does not require information on mutation rates or recombination rates which is rarely available in nonmodel systems. Given this information, *delimitR* will generate fsc26 input files for the default model set with the function setup_fsc2(). If the default model set is inadequate for a given system, the user needs to only generate the fsc26 input files and place them in the working directory. *delimitR* is thus applicable to any demographic scenarios that can be implemented in fsc26 and allows users to incorporate prior information about the system into their species-delimitation analysis.

Data simulation, model selection, and assessment of power

Following Smith et al. (2017), we use fsc26, a coalescent simulator, to simulate data under each model in the model set. Given a user-specified number of replicates *N*, *delimitR* will simulate *N* replicates under each model in the model set and output one mSFS per simulation using the function fastsimcoalsims(). We then use a binning strategy to further summarize the mSFS following Smith et al. (2017) and generate the binned SFS. We apply the same binning strategy to the simulated and observed data using the functions makeprior() and prepobserved(), respectively. We use the simulated data to construct an RF classifier, in which the bins of the binned SFS are the predictor variables and the model used to simulate the data is the response variable. To build the RF classifier, *delimitR* uses the R package "abcrf" (Pudlo et al. 2015). The RF classifier consists of a user-defined number *M* of decision trees. Each decision tree is constructed from a subset of the prior, and at each node in each decision tree a bin of the binned SFS is considered, and a binary decision rule is constructed based on the number of SNPs in the bin.

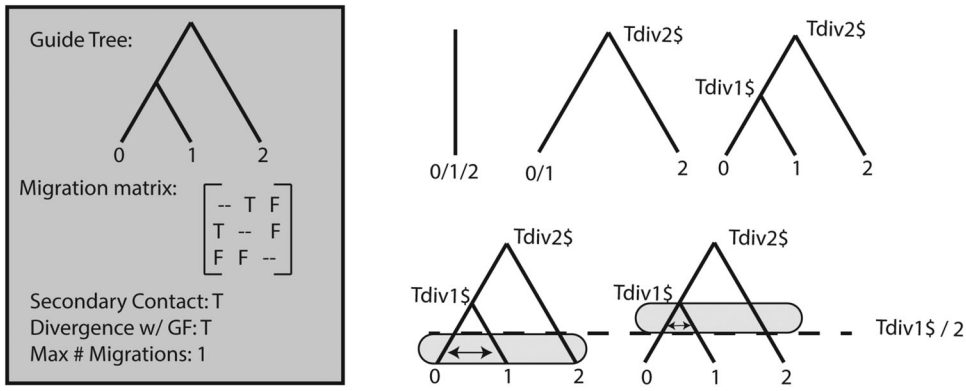


Figure 1. **An example of a default model set that can be considered in *delimitR*. GF is gene flow. In the gray box, we show the information the user would provide to specify the model set. On the right, we show the resulting models. Lines indicate divergent populations, and arrows indicate gene flow. The gray regions around migration arrows demonstrate the timing of migration. Divergence times are labeled with "Tdiv#$."**

delimitR uses oob error rates to assess the power of the RF classifier. Because only a portion of the prior is used for the construction of each decision tree, it can take an element of the prior (i.e., a dataset simulated under a known model), consider only decision trees constructed without reference to that element, and apply the RF classifier. When this classifier is applied to the datasets, we move down nodes until we reach the leaves of the trees, which are model indices in this case. Each decision tree votes for a model, and the model receiving the largest portion of votes is selected as the best model. We then calculate how often we choose an incorrect model. To construct the RF classifier and calculate oob error rates, *delimitR* uses the abcrf function from the R package "abcrf" (Pudlo et al. 2015), as implemented by the RF_build_abcrf() function. The predict.abcrf() function from the R package "abcrf" (Pudlo et al. 2015), as implemented by the RF_predict_abcrf() function, is then used to select the model that simulates data most similar to the observed data. Specifically, as described above, in our forest of decision trees, each node of each tree considers a particular bin in the SFS, and splits data based on the number of SNPs falling in that bin. This procedure is repeated until the tips of the trees are reached, before tallying the decision tree votes for any particular model. Finally, the posterior probability of the selected model is estimated by regressing against oob error rates following Pudlo et al. (2015).

SPECIATION AND SPECIES LIMITS IN TAILDROPPER SLUGS (GENUS *Prophysaon*)

To illustrate the application of *delimitR* to an empirical system, we used *delimitR* to understand speciation and estimate species limits in a system with uncertain species boundaries: *Prophysaon andersoni*, the reticulate taildropper slug. Previous work in this system has suggested the presence of multiple undescribed lineages (Smith et al. 2018), and phenotypic and ecological variation is evident across the range of this species (Burke 2013), but little is certain regarding species boundaries. SNP data from 88 individuals and one technical replicate (Fig. 2A; Table S1), were collected using GBS (Elshire et al. 2011) and assembled using ipyrad (Eaton and Overcast 2016). Data collection in *P. andersoni* is described in detail in the Supporting Information.

Population assignment

To provide a starting point for the number of putative species and the assignment of individuals to putative species, we applied Structure (Pritchard et al. 2000) to the *P. andersoni* data. Analyses were run for K values from 1 to 10 with 10 replicates per K value. The first 100,000 generations were discarded as burn-in, and 500,000 generations followed. We then used the command-line version of STRUCTURE HARVESTER (Earl 2012), along with visualizations of log-likelihood scores to determine the optimal values of K. Finally, we used CLUMMP to summarize and visualize results (Jakobsson and Rosenberg 2007). To estimate differentiation between populations using a traditional metric, we calculated F_{ST} between populations using the R package "PopGenome" (Pfeifer et al. 2014).

Species delimitation in delimitR

We used custom Python scripts (available on github) to construct a mSFS, with population assignments based on Structure results. We removed the technical replicate from the dataset before performing model selection. We required that all SNPs used in the construction of the mSFS be biallelic and sampled in at least 50% of the individuals in each population. For SNPs sampled in more than 50% of individuals in a population, we randomly down-sampled alleles, following Satler and Carstens (2017). We sampled only a single SNP from each locus, and we did not consider invariable sites. This allowed us to build an SFS from a matrix without missing data.

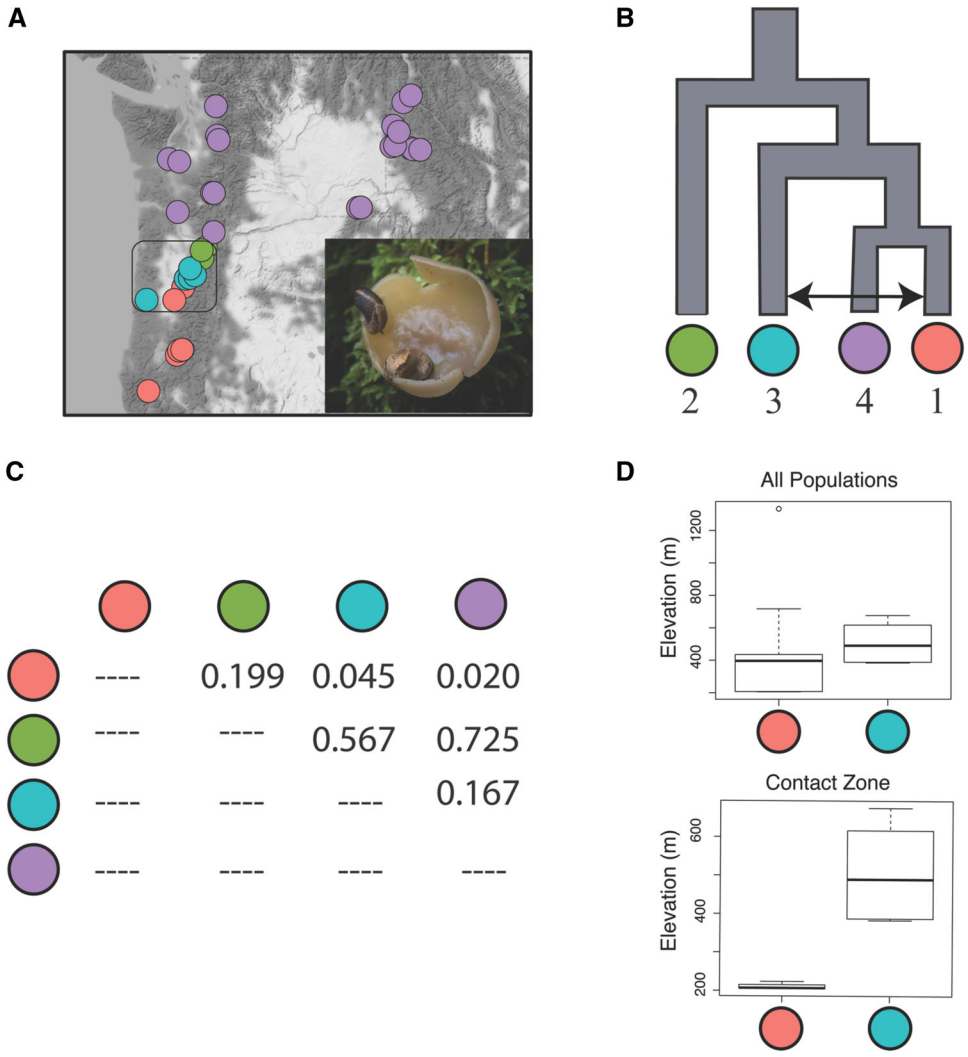


Figure 2. Results from *P. andersoni*. (A) Sampling map, with a black square surrounding samples considered to be "near the contact zone." Colors correspond to structure population assignments (colored based on population with highest posterior probability). Photo by M. Smith. (B) The best model, with the arrow representing secondary contact. (C) F_{ST} values between populations. (D) Elevational ranges for the two populations experiencing secondary contact both throughout their entire ranges and "near the contact zone."

Rather than using the default model set to analyze the *Prophysaon* data, we designed a model set in *delimitR* that considered all topologies for four populations, and all possible collapsed nodes (i.e., all topologies for 1, 2, 3, or 4 populations). We considered models that either lacked or included secondary contact after the Last Glacial Maximum for populations with evidence of admixture in Structure results, and we considered models that either lacked or included population expansion after the Last Glacial Maximum. When population expansion was included, it was assumed that all populations expanded, but rates of expansion could vary among populations. We only allowed a single migration event per model to further limit the model space. In total, our model set included 208 models. Our choice of parameters to include was informed by our prior knowledge of the system. Specifically, we considered population expansion after the Last Glacial Maximum because ecological modeling suggests that, during the Last Glacial Maximum, habitat for *P. andersoni* would likely have been much more fragmented than it is in the present (Smith et al. 2018). Likewise, we considered secondary contact after the Last Glacial Maximum to allow for the possibility that refugial populations came back into contact after expanding from isolated refugia. Priors were chosen to correspond with these climatic events hypothesized to have influenced diversity in *P. andersoni* (Pielou 2008) and a generation time of 1 year (COSEWIC 2006). We drew population sizes from a broad uniform (1000, 200,000) prior, and divergence times were drawn from a uniform (5000, 10,000,000) prior. Growth rates were drawn from a uniform (–0.001, –0.00035) prior, and population growth continued for 5000 generations.

Migration rates were drawn from a uniform (0.000005, 0.000025) prior, corresponding to 0.005–5 migrants per generation (Nm). Migration and population expansion ended 1000 to 20,000 generations ago. We simulated 10,000 mSFS under each model and then summarized the simulated and observed mSFS by binning (four classes per population). We built an RF classifier and applied this classifier to the observed data using functions in the R package *delimitR* as described above, and we used 1000 decision trees in the RF classifier due to the large number of models. We recorded the oob error rates, the selected model, and the approximated posterior probability of the selected model. We estimated parameters and their confidence intervals under the best model in fsc26 (Supporting Information). To assess the fit of the models to the data, we performed principal components analysis (PCA) on data simulated under the top 10 models and the empirical data using the prcomp function from the R package "stats" (R Core Team 2013) and plotted the first two axes.

To assess how including population-level parameters influenced the results of species delimitation, we conducted two additional species-delimitation analyses that considered only divergence models. First, we compared four divergence-only models in *delimitR.* All models that did not conflict with the topology from the full analysis were considered. As above, we simulated 10,000 datasets under each model. Priors on divergence times and population sizes were as above. We binned using four classes, and 1000 decision trees were used in the RF classifier. Second, we applied the widely used program BPP version 4.0.4 (Yang and Rannala 2010) to delimit species. Due to computational limitations of BPP, we were required to subsample 100 loci from our full dataset. However, we did use sequence data, rather than unlinked SNPs, in the BPP analyses. We created 10 down-sampled datasets and ran BPP on each of these. BPP analyses did not use a guide tree, and we used an inverse gamma prior (3, 0.004) on theta and an inverse gamma prior (3, 0.002) on tau, as these correspond to broadly uninformative priors (Flouri et al. 2018). To assess whether the priors were driving inferences, we repeated the analyses with an inverse gamma prior (3, 0.04) on theta and an inverse gamma prior (3, 0.02) on tau. We allowed mutation rates to vary across loci using the random-rates model of Burgess and Yang (2008), with rates drawn from a Dirichlet distribution D(5), and we constrained theta to be the same for all loci (heredity = 0). We discarded the first 2000 samples as burn-in followed by 20,000 samples.

ANALYSES OF PREVIOUSLY PUBLISHED EMPIRICAL DATASETS

To evaluate the performance of *delimitR* on a well-studied system and compare results to previous findings, we reanalyzed the data from Leaché et al. (2014b), which consist of five putative species of West African forest geckos (*Hemidactylus fasciatus*). Leaché and Fujita (2010) used divergence-only models in BPP to delimit species in this group, and, based on these results, three new species were named. In later work, SNP data were collected, and Leaché et al. (2014b) found that this dataset consisted of five species using Bayes factor delimitation without considering gene flow. We analyzed the same SNP data in *delimitR* and compared 13 models with a maximum of five populations that included gene flow between recently diverged sister species pairs (details in the Supporting Information).

Additionally, we applied *delimitR* to two datasets from Satler and Carstens (2017). These data are from two ecological associates of North American pitcher plants: a moth (*Exyra semicrocea*) and a spider (*Peucetia viridans*). *Exyra simicrocea* is an obligate inquiline commensal with the pitcher plant *Sarracenia alata*, whereas *P. viridans* is an opportunistic capture interrupter. Both datasets consist of two potential species for each nominal species: one east of the Mississippi River and another west of the Mississippi River, and previous results suggest high migration rates between populations east and west of the Mississippi for the spider *P. viridans* and low migration rates between populations east and west of the Mississippi for the moth *E. semicrocea.* We reanalyzed this data in *delimitR* considering panmictic models, divergence only models, divergence with gene flow models, and secondary contact models (details in the Supporting Information).

SIMULATION STUDIES

Finally, simulation studies were conducted to assess the accuracy of *delimitR*. First, we designed a simulation study that considered four scenarios in a three-population system (Fig. 3A). These scenarios included: (1) no population divergence; (2) divergence between two of the three populations; (3) divergence among all three populations; and (4) divergence among all three populations with secondary contact between the two most closely related populations. We conducted this analysis using both moderate (>50,000 generations ago) and recent (>5000 generations ago) divergence times, and refer to these studies as the moderate and recent simulation studies. We sampled 10 diploid individuals from each population (20 alleles). Population sizes were drawn from uniform (10,000, 100,000) priors. Divergence times between species 0 and 1 were drawn from a uniform (50,000, 100,000 generations) prior for the moderate-divergence-time study and from a uniform (5000, 10,000 generations) prior for the recent-divergence-time study. Divergence times between the ancestor of species 0 and 1 and species 2 were drawn from a uniform (1,000,000, 5,000,000 generations) prior for the moderate-divergence-time study and from a uniform (50,000, 100,000 generations) prior for the recent-divergence-time study. The migration rates were drawn from a uniform (0.000005, 0.00005) prior, corresponding to 0.05–5 Nm. We simulated 10,000 datasets under each of the four

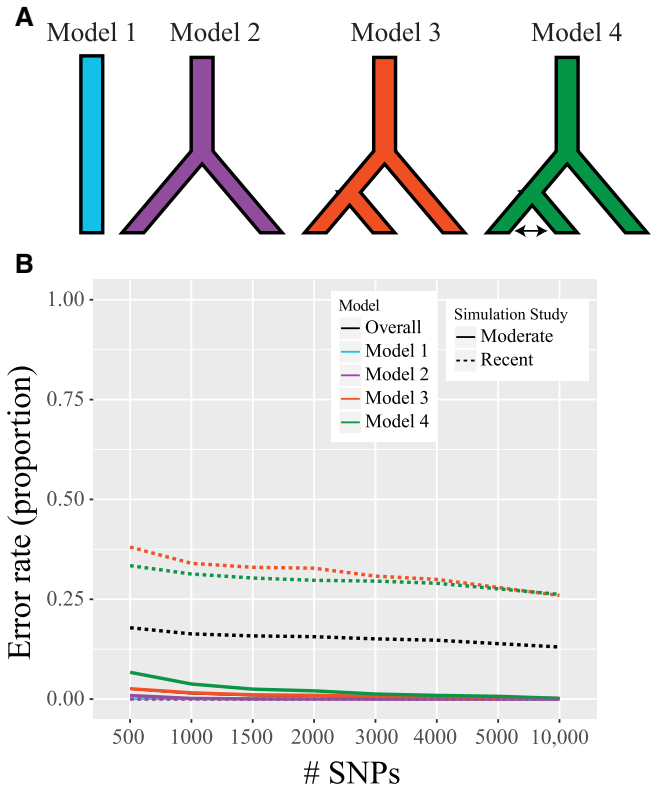


Figure 3. **Results of the first simulation study, which considered up to three species with the potential for secondary contact between the most recently diverged lineages. (A) The four models evaluated in the simulation study. (B) The results of the simulation study with moderate and recent divergence times. Oob error rates are reported as the proportion of simulations that were misclassified.**

models for both the moderate and recent divergence-time analyses. We evaluated the performance of the RF classifier using datasets containing 500, 1000, 1500, 2000, 3000, 4000, 5000, and 10,000 SNPs, and we used 10 classes per population to summarize the mSFS. We constructed an RF classifier with 500 decision trees and calculated oob error rates in *delimitR*. In addition to oob error rates, we used a cross-validation approach to assess the accuracy of our classifier. We simulated 1000 pseudo-observed datasets under each of the four models. We applied the RF classifier constructed above to each of these datasets using 500 decision trees and calculated how often the correct model was selected for each dataset.

To evaluate the power of *delimitR* to distinguish among more complex modes of speciation, we designed a simulation study based on a continent-island system. Island systems have long been of interest to evolutionary biologists, particularly with respect to speciation. With respect to island systems, founder-effect speciation as initially proposed by Mayr (1954) involves the founding of an island population by a small number of individuals from a continental source population. Subsequent genetic drift and a shift in selective regime lead to a shift in adaptive peaks, which ultimately results in a speciation event. Although founder-effect speciation is one potential driver of divergence in continent-island systems, it is not the only possible mechanism. For example, vicariance has been implicated as a driver of diversification in insular arthropods and shrews (Gillespie and Roderick 2002; Esselstyn et al. 2009). According to vicariance models, previously contiguous populations are isolated when island populations separate owing to geological events such as changing sea levels, and divergence occurs in isolation. Although geographic isolation may seem a key feature of speciation on islands, it has also been proposed that secondary contact could be an important driver of speciation in such scenarios, for example, in Darwin's finches (Grant et al. 1996). We considered six potential scenarios for an island-continental system: (1) no population divergence; (2) allopatric (vicariant) speciation; (3) divergence with secondary contact; (4) divergence with gene flow; (5) founder-effect speciation; and (6) founder effects with continuous gene flow (Fig. 4A, B). We sampled 10 diploid individuals from each population (20 alleles). Population sizes were drawn from a uniform (50,000, 100,000) prior for the island population and a uniform (75,000, 200,000) prior for the continental population. For models that included divergence, divergence times between the island and continental populations were drawn from a uniform (50,000, 100,000 generations) prior. For models with migration, the migration rate was drawn from a uniform prior that corresponded to 0.05–10 Nm. For models including a bottleneck, the proportion of the population that remained during the bottleneck was drawn from a uniform (0.001, 0.01) prior, which corresponds to 50–100 individuals, and the bottleneck lasted from 100 to 500 generations. We refer to this simulation study as the complex model set. Sample sizes, the number of SNPs, the number of decision trees, and the number of classes used to summarize the SFS matched those described in the first simulation study. As in the first simulation study, we used oob error rates and cross-validation (1000 datasets under each model) to assess the accuracy of the classifier. Accuracy and posterior probabilities were calculated as above.

In addition to the simulation studies described above, we performed additional simulation studies to assess (1) the ability of *delimitR* to distinguish amongst a model set that includes both secondary contact and divergence with gene flow; (2) the effects of changing the number of simulations or the number of decision trees on error rates; and (3) the effects of prior misspecification on error rates. These additional simulation studies are further described in the Supporting Information. All analyses were carried out on the Ohio Supercomputer (Ohio Supercomputer Center 1987).

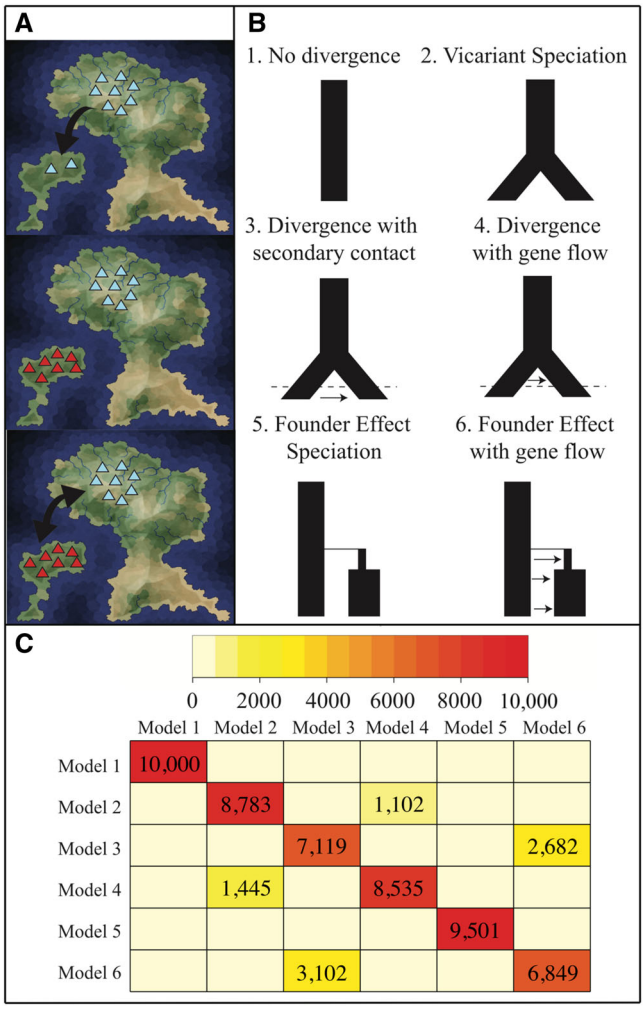


Figure 4. Results of the second simulation study. (A) In island-continent systems, there are several processes that may contribute to speciation, including (1) Founder effect speciation, in which a small number of individuals colonizes the island and speciates (top); (2) Vicariant speciation, in which a previously contiguous population diverges into two when island and mainland populations separate (middle); (3) secondary contact, in which a period of initial divergence is followed by a period of gene flow (and potentially reinforcement, or fusion; lower). (B) The six models evaluated in the simulation study. (C) The heatmap represents the oob error rates, in terms of the number of simulated datasets that were classified as belonging to a certain model. Each cell represents the number of simulations under the model (row) classified as belonging to each model in the model set (columns). Red along the diagonal indicates that most simulated datasets were correctly classified. For any cell with more than 1000 simulations, the number in the cell indicates the number of simulated datasets classified as belonging to that model. The map illustration was modified from the polygon map generator tool (redblobgames.com/maps/mapgen2).

Results

SPECIATION AND SPECIES LIMITS IN TAILDROPPER SLUGS (GENUS *Prophysaon*)

Data collection in P. andersoni

After filtering reads and aligning and assembling in ipyrad, we retained 18,625 loci. For Structure, we used 18,318 unlinked SNPs sampled in 50% of all samples. We used 8611 unlinked SNPs sampled in more than 50% of individuals in each population to construct the mSFS used in model selection.

Population assignment

Although the log likelihood continued to increase until K reached 6, increases were minimal after increasing K from 3 to 4 (Table S2). The delta K method supported a K of 2, but the delta K value for K = 3 was very close to that of K = 2. Given that individual assignment was stable, but populations tended to be collapsed as K moved from 4 to 2, and given that our downstream models allowed for populations to be collapsed, we used the results from K = 4 for the remaining analyses. Eight individuals from eight localities were assigned to Population 1, 20 individuals from four localities were assigned to Population 2, 17 individuals from eight localities were assigned to Population 3, and 43 individuals from 22 localities were assigned to Population 4. Population assignments are shown in Fig. 2A, and a table with admixture results is provided in Table S3. F_{ST} values ranged from 0.0197 between Population 1 and Population 4 to 0.7253 between Population 2 and Population 4 (Fig. 2C).

Species delimitation in delimitR

Oob error rates were low overall (oob error rates = 0.0528), but were >0.1 for nine models. In six of these nine cases, the model was most often misclassified as a model that matched the true model save for a difference (i.e., presence or absence) in population expansion. In three other cases, a model including secondary contact between sister species was most often misclassified as the same model without secondary contact. The best model was a four-population model with secondary contact between Population 1 and Population 3 (Fig. 2B). The model received 42 votes and had a posterior probability of 0.689. Ten models received more than 20 votes, and all the 10 included secondary contact. The top three models with >30 votes each were four population models that included secondary contact between Population 1 and Population 3 (Table 1). Parameter estimates and confidence intervals under the best model are reported in Table S4. In the PCA, the empirical data fell within the cloud of datasets simulated under the 10 best models (Fig. S4). Simulating and summarizing the data for all 208 models considered took 2385 h of CPU time. Building the RF classifier, selecting the best model, and calculating error rates and posterior probabilities took 280 h of CPU

Table 1. **The top 10 models for *P. andersoni*. The number of species in the model (# Species), the species between which secondary contact occurs in the model (Secondary contact), whether expansion occurs in the model (Expansion), the species tree topology (Topology), and the number of votes received when the RF classifier was applied to the observed data (# Votes).**

# Species	Secondary contact	Expansion	Topology	# Votes
4	Pop1 + Pop3	No	(((1,4),3),2)	42
4	Pop1 + Pop3	No	(((1,4),2),3)	36
4	Pop1 + Pop3	No	(((2,3),1),4)	31
3	Pop1/Pop4 + Pop3	No	((1/4,(2,3))	28
4	Pop1 + Pop3	No	(((1,3),2),4)	28
4	Pop1 + Pop3	No	((1,4),(2,3))	27
4	Pop1 + Pop4	No	(((2,3),1),4)	26
3	Pop1/Pop4 + Pop3	No	((1/4,2),3)	24
2	Pop1/Pop4 + Pop2/Pop3	Yes	(1/4,2/3)	23
3	Pop1/Pop4 + Pop3	No	((1/4,2),3)	22

time. However, due to parallelization the entire analysis took 68 h of actual time.

When approaches considering divergence-only models were used, oversplitting was common. BPP supported four species with a posterior probability of 1.0 in all replicates regardless of the prior settings. *delimitR* could easily distinguish among divergence only models (oob = 0.00045), and also supported a model with four species (posterior probability = 0.969).

ANALYSES OF PREVIOUSLY PUBLISHED EMPIRICAL DATASETS

When we analyzed the *Hemidactylus* gecko data delimited by Leaché et al. (2014b), results corroborated previous findings (Fig. S5). The best model was a five-population model with divergence only (posterior probability [pp] = 0.656), and most models receiving votes differed only in the presence or absence of migration parameters (Table S6). This result was expected given that migration rates as low as 0.05 Nm were considered, and error rates were highest between models that differed only in the presence or absence of migration parameters (Fig. S5). For the pitcher plant invertebrates from Satler and Carstens (2017), results corroborated previous findings in that migration models were strongly supported for both datasets. For *P. viridans*, the best model included secondary contact (pp = 0.944), whereas for *E. semicrocea* the best model included divergence with gene flow (pp = 0.798; Fig. S6 and Table S7). Additional information on these results is available in the Supporting Information.

SIMULATION STUDIES

In the simulation study with moderate divergence times (50,000–100,000 generations), overall error rates were low (0.0009–0.026) regardless of the number of SNPs used (Table 2; Fig. 3B). Oob error rates were zero for the model with no population divergence and highest (but still low; 0.0019–0.0672) for the model with three populations and secondary contact (Table 2; Fig. 3B), and error rates based on cross-validation were similarly low (Table S8). The moderate-divergence-time study with only 500 SNPs used 2 h and 13 min of CPU time, whereas the same study with 10,000 SNPs used 4 h and 3 min of CPU time. In the recent-divergence-time analyses, overall oob error rates were moderate (0.13–0.18) regardless of the number of SNPs used (Table 2; Fig. 3B). Oob error rates were near zero for the model with no population divergence and highest (0.26–0.38) for the three-population model with secondary contact and the three-population model (Table 2; Fig. 3B). Error rates based on cross-validation were similar (Table S9).

In the simulation study with the complex model set, overall oob error rates were moderate (0.15–0.32), but decreased when more SNPs were used (Table 3; Fig. 4C). Oob error rates were zero for the no-divergence model and highest (0.32, 10,000 SNPs) for the founder-effect model with gene flow (Table 3; Fig. 4C). In the cross-validation study, error rates were highest among models that included migration. In particular, both oob error rates and cross-validation error rates were high between the secondary-contact model and the founder-effect-with-gene-flow model (Fig. 4C; Table S10). This was not unexpected, because migration likely swamped the signal of the founder effect such that the patterns of genetic variation predicted by these models converged, making it difficult to distinguish between these scenarios. This study used between 27 min and 2 h and 11 min of CPU time, depending on the number of SNPs simulated.

In addition to these simulation studies, we evaluated the ability of *delimitR* to distinguish among models in a set including both divergence with gene flow and secondary contact. We found that error rates were low to moderate, and that misclassifications were most common between divergence-only models and divergence-with-gene-flow models (Fig. S7; Tables S11 and S12). We also verified that the number of simulations and the number of decision trees had minimal effects on error rates (Fig. S8). Finally, we investigated prior sensitivity and found that, when the prior was

Table 2. Error rates from the simulation study for each model and overall. Moderate divergence times were drawn from uniform (50,000, 100,000) priors. Recent divergence times were drawn from uniform (5000, 10,000) priors. All error rates in this table are given as proportions of simulations that were misclassified.

# SNPs	Divergence times	Overall	Model 1	Model 2	Model 3	Model 4
500	Moderate	0.026	0.000	0.009	0.026	0.067
1000	Moderate	0.014	0.000	0.002	0.016	0.038
1500	Moderate	0.009	0.000	0.001	0.011	0.025
2000	Moderate	0.008	0.000	0.000	0.010	0.021
3000	Moderate	0.005	0.000	0.000	0.007	0.013
4000	Moderate	0.004	0.000	0.000	0.005	0.009
5000	Moderate	0.003	0.000	0.000	0.004	0.007
10,000	Moderate	0.001	0.000	0.000	0.002	0.002
500	Recent	0.179	0.000	0.000	0.381	0.334
1000	Recent	0.163	0.000	0.000	0.340	0.313
1500	Recent	0.158	0.000	0.000	0.330	0.303
2000	Recent	0.156	0.000	0.000	0.328	0.298
3000	Recent	0.151	0.000	0.000	0.308	0.296
4000	Recent	0.147	0.000	0.000	0.300	0.290
5000	Recent	0.139	0.000	0.000	0.280	0.276
10,000	Recent	0.130	0.000	0.000	0.260	0.262

Table 3. Error rates from the island-continent simulation study for each model and overall. All error rates in this table are given as proportions of simulations that were misclassified.

# SNPs	Overall	M 1	M 2	M 3	M 4	M 5	M 6
500	0.318	0.000	0.412	0.497	0.315	0.178	0.507
1000	0.278	0.000	0.346	0.450	0.272	0.143	0.457
1500	0.260	0.000	0.315	0.436	0.253	0.130	0.423
2000	0.240	0.000	0.281	0.408	0.228	0.114	0.410
3000	0.217	0.000	0.242	0.377	0.206	0.095	0.385
4000	0.205	0.000	0.210	0.355	0.200	0.085	0.379
5000	0.192	0.000	0.190	0.333	0.185	0.079	0.363
10,000	0.154	0.000	0.122	0.288	0.147	0.050	0.315

violated, misclassifications happened in the expected direction (Fig. S9). For example, simulating data with shallower divergence times than considered in the prior tended to result in misclassifying divergence-only models as secondary-contact models, and simulating data with higher migration rates than considered in the prior tended to result in misclassifying divergence-with-secondary-contact models as models lacking divergence.

Discussion

UNDERSTANDING THE SPECIATION PROCESS

As illustrated by our results for *P. andersoni*, *delimitR* allows extraordinary flexibility by enabling users to focus on the process by which species may have formed as they conduct investigations into empirical systems. Notably, *delimitR* can be applied to evaluate demographic models consistent with a variety of modes of speciation under nearly any species concept. It requires researchers to use their expertise and familiarity with the focal system to identify reasonable priors on divergence times and migration rates and to decide which models should be included in the model set. This feature, rare among delimitation approaches (but see Jackson et al. 2017a), encourages explicit predictions that are based on developed hypotheses and requires researchers to be explicit about the species concept that they apply to their data, thereby increasing transparency and repeatability in species delimitation investigations by connecting the delimited species to the evolutionary processes by which they were formed. Inferences about the process of speciation that result from *delimitR* can form the basis for predictions that can then be tested using ecological and morphological data. Further, after model selection is performed in *delimitR*, researchers are able to estimate relevant parameters, such as migration rates and divergence times, using existing methods (e.g., fsc26). These estimates, along with the results from tests of process-based predictions using ecological

and morphological data, will enable researchers to distinguish between population-level and species-level differentiation, which may not be possible using divergence-only models and genetic data alone (Sukumaran and Knowles 2017; Jackson et al. 2017a). In general, process-based species delimitation such as that implemented in *delimitR* will both prevent erroneous inferences caused by ignoring population-level processes that drive speciation and allow researchers to infer how speciation happened in their focal system. This advance (*delimitR*) promotes biologically meaningful species delimitation.

SPECIES DELIMITATION AND SPECIATION IN TAILDROPPER SLUGS

Our results from the *Prophysaon* data suggest that secondary contact may have played an important role in speciation in taildropper slugs. This, along with parameter estimates (Table S4), suggests that *P. andersoni* survived in multiple refugia during the Last Glacial Maximum and that after the Last Glacial Maximum at least two of the four refugial populations came into contact and exchanged genes. Based on parameter estimates, these two lineages have exchanged genes at a rate equivalent to ~1.36 Nm (Table S4). Theoretically, this degree of secondary contact could lead either to lineage fusion or to reinforcement. If hybrids have similar fitness to the parental genotypes, fusion could occur. Alternatively, if hybrids have lower fitness than parentals, secondary contact could lead to reinforcement and, eventually, speciation. In *P. andersoni*, our ecological data suggest that these lineages are isolated by habitat. Specifically, near the contact zone (i.e., in Oregon), the two *P. andersoni* lineages experiencing secondary contact occupy distinct, nonoverlapping elevational ranges, whereas elevational ranges are overlapping and not statistically different across the entirety of the ranges of these two populations (Fig. 2D). The pattern of habitat isolation in *P. andersoni*, where differentiation is clear near the zone of contact and less clear in allopatric portions of the range, in combination with evidence of secondary contact and gene flow, suggests reinforcement as a driving force (Nosil 2012), but additional data from the contact zone would be valuable in further testing whether reinforcement has occurred. More accurate estimates of the timing of migration would help to understand contemporary levels of gene flow between these populations, and whole genome data analyzed with methods that consider linkage information could permit better estimates of this timing. Further, some mechanism of reproductive isolation is predicted if reinforcement has occurred. Terrestrial slugs locate mates by following the slime trails laid down by individuals during foraging, and work in other systems has demonstrated that slugs preferentially follow conspecific over heterospecific slime trails (reviewed in Ng et al. 2013). We consider slime trails an excellent candidate for an external reproductive isolating mechanism in this group, and future work will test for reproductive character displacement in this trait, which would lend further support to the hypothesis of reinforcement generated from this work.

Under a species concept invoking reproductive isolation, our results support up to four cryptic species within the nominal *P. andersoni*. Although migration is evident between two of these populations, ecological data suggest that reinforcement may be occurring. If true, these results demonstrate that reproductive isolation is present between these lineages, and they should continue to diverge. Given computational limits, our model space was limited to models with a single migration edge. We note that some of the 10 best models also included secondary contact between Population 1 and Population 4, and that some top models did not include divergence between these populations. Given geographic isolation between these two populations and recent divergence time estimates (Table S4), this suggests that divergence between these two populations has been very recent, which is not surprising given that much of the range of Population 4 would likely have been unsuitable for this species during the Last Glacial Maximum (Smith et al. 2018). Given the geographic isolation and separation by other, putatively reproductively isolated lineages, we expect that contemporary gene flow between these two lineages is low to absent, and that the lineages are diverging; however, additional ecological and morphological data evaluating differences between these lineages are necessary to determine their status. In terrestrial gastropods, both radular (or dental) morphology and reproductive morphology are often variable between species. Future work will describe radular morphology imaged using Scanning Electron Microcopy and reproductive morphology to identify phenotypic characters that can be used to distinguish between putative species of *P. andersoni*.

Had we relied on divergence-only models to estimate species limits in *Prophysaon andersoni*, our inferences would have been strikingly different. Results from BPP indicated that there were four lineages, but would not have suggested secondary contact as an important process in this group. It would have been considerably easier to apply existing methods (e.g., BPP) rather than developing and testing *delimitR*, but doing so would have prevented us from considering ecological speciation and reinforcement. Our *delimitR* results led to a reanalysis of our ecological data and to the interpretation of secondary contact and reinforcement as potentially having driven speciation between two putative species of *P. andersoni*.

MODEL SELECTION AND SPECIES DELIMITATION

delimitR is accurate across a wide range of parameter and model space. Generally, its performance improves with the number of SNPs and with increasing divergence times (Fig. 3), whereas recent divergent times tend to increase the difficulty of detecting migration (Fig. 3). Results show that *delimitR* struggles to identify the correct model when priors are misspecified, but that

misclassifications happen in the expected direction, highlighting that users should take care when defining priors and be cognizant of how misspecifications may affect results.

One inherent challenge to the application of any model-selection framework to empirical systems is determining which demographic models to include in the comparison set (Carstens et al. 2017). The number of models that could be compared is limited in many approaches, either due to analytical or practical considerations. An example of the former would include approaches that implement a demographic model that includes only a subset of the parameters considered by *delimitR* (e.g., Carstens et al. 2013), whereas an example of the latter would include methods that are more computationally intensive and as such are limited in the number of models that can be compared. Researchers are faced with difficult decisions when they perform model selection using such methods because they cannot include any conceivable model. Even simulation-based methods such as approximate Bayesian computation (Csilléry et al. 2012) and PHRAPL (Jackson et al. 2017b), where users can theoretically include any number of demographic models of custom design, see their accuracy decrease as the number of models that are included in the model set increases (Pelletier and Carstens 2014; Jackson et al. 2017a). When compared to a traditional approximate Bayesian computation approach, the RF approach used in *delimitR* has much lower error rates (Smith et al. 2017). Using *delimitR*, we were able to compare 208 models in a single model comparison step with low error rates. Future work is needed to understand the relationship between model space, the number of parameters, and accuracy in *delimitR*, but the results reported here suggest that *delimitR* may offer much more flexibility than previous methods by enabling researchers to compare large, complex model sets with low-to-moderate error rates.

MACHINE LEARNING, GENOMIC DATA, AND SPECIES DELIMITATION

Computational limits are the primary reason that previous approaches to species delimitation have not focused on the process of speciation. To circumvent this issue, *delimitR* uses a machine-learning algorithm (RF) for classification and the binned mSFS to summarize data. This combination allows us to compare a large set of models using large datasets in minutes of computational time, a task unmanageable using standard approaches to model comparison. The largest computational burden is the dataset simulation, and this burden can be eased with access to multiple processors. Furthermore, the RF approach used here automatically generates estimated error rates with effectively no additional computational expense, giving researchers a built-in approach to assessing statistical power, without conducting a full power analysis. This will encourage a nuanced interpretation of results when the power to choose among the models in the model set is low, and will prevent researchers from blindly comparing models among which they do not have enough data to distinguish. We have tested *delimitR* using datasets with moderate numbers of populations (up to five evaluated here), moderate-to-large numbers of SNPs (up to 10,000 evaluated here), and moderate numbers of individuals per population, and found that error rates were low and computing times were reasonable. Our implementation of *delimitR*, particularly when combined with parallel computing, enables complex model sets containing parameters that represent relevant evolutionary processes to be evaluated using moderate numbers of SNPs. Given the large numbers of SNP-based phylogeographic-scale datasets that are being collected, *delimitR* will enable researchers to better understand the diversity of evolutionary processes that create new species.

AUTHOR CONTRIBUTIONS

MLS and BCC designed the study. Funding and support was obtained by MLS and BCC. MLS collected data, wrote software, and performed analyses. MLS and BCC wrote and revised the manuscript.

ACKNOWLEDGMENTS

MLS was funded by a NSF GRFP (DGE-1343012), and this work was funded by NSF (DEB1457519). Ecological data collection was funded by the Society for the Study of Evolution and the Society for Systematic Biologists. The authors would like to thank members of the Carstens lab, members of D. Tank and J. Sullivan's labs, R. Garrick, and B. Stone for comments that improved the manuscript prior to publication, and the authors would like to thank P. Blischak for the name *delimitR*. The authors would like to thank the Ohio Super Computer for computing resources (allocation grant PAS1181-2). The authors would also like to thank M. Lucid of Idaho Fish and Game for donation of samples and Jesse Wallace for help with lab work.

DATA ARCHIVING

Raw reads are available on the NCBI SRA archive (PRJNA577570). Alignments are available on the Dryad Digital Repository (doi: 10.5061/dryad.2jm63xsjm). The R-package *delimitR*, as well as a full tutorial, is available on github (https://github.com/meganlsmith/delimitR).

LITERATURE CITED

Burgess, R., and Z. Yang. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol. Biol. Evol. 25:1979–1994.

Burke, T. E. 2013. Land snails and slugs of the pacific northwest. Oregon State Univ. Press, Corvallis, OR.

Camargo, A., M. Morando, L. J. Avila, and J. W. Sites. 2012. Species delimitation with ABC and other coalescent-based methods: a test of accuracy with simulations and an empirical example with lizards of the liolaemus darwinii complex (Squamata: Liolaemidae). Evolution 66:2834–2849.

Carstens, B. C., T. A. Pelletier, N. M. Reid, and J. D. Satler. 2013. How to fail at species delimitation. Mol. Ecol. 22:4369–4383.

Carstens, B. C., A. E. Morales, N. D. Jackson, and B. C. O'Meara. 2017. Objective choice of phylogeographic models. Mol. Phylogenet. Evol. 116:136–140.

COSEWIC. 2006. COSEWIC assessment and status report on the Blue-grey Taildropper slug Prophysaon coeruleum in Canada. Committee on the

Status of Endangered Wildlife in Canada. Ottawa, Canada. Available at http://www.sararegistry.gc.ca/virtual_sara/files/cosewic/sr_blue_grey_taildropper_e.pdf. Accessed January, 16 2014.

Coyne, J. A., and H. A. Orr. 2004. Speciation. Sinauer Associates, Inc., Sunderland, MA.

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

Earl, D. A. 2012. Structure harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4:359–361.

Eaton, D. A. R., and I. Overcast. 2016. ipyrad: interactive assembly and analysis of RADseq data sets. Available at http://ipyrad.readthedocs.io/. Accessed September 29, 2018.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379.

Ence, D. D., and B. C. Carstens. 2011. SpedeSTEM: a rapid and accurate method for species delimitation. Mol. Ecol. Resour. 11:473–480.

Esselstyn, J. A., R. M. Timm, and R. M. Brown. 2009. Do geological or climatic processes drive speciation in dynamic archipelagos? The tempo and mode of diversification in Southeast Asian shrews. Evol. Int. J. Org. Evol. 63:2595–2610.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9:e1003905.

Flouri, T., X. Jiao, B. Rannala, Z. Yang. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol. Biol. Evol. 35:2585–2593.

Garrick, R. C., I. a. S. Bonatelli, C. Hyseni, A. Morales, T. a. Pelletier, M. F. Perez, E. Rice, J. D. Satler, R. E. Symula, M. T. C. Thomé, et al. 2015. The evolution of phylogeographic datasets. Mol. Ecol. 24:1164–1171.

Gavrilets, S., and A. Hastings. 2017. Founder effect speciation: a theoretical reassessment. Am. Nat. 147:466–491.

Gillespie, R. G., and G. K. Roderick. 2002. Arthropods on islands: colonization, speciation, and conservation. Annu. Rev. Entomol. 47:595–632.

Grant, P. R., B. R. Grant, and J. C. Deutsch. 1996. Speciation and hybridization in island birds [and discussion]. Philos. Trans. R. Soc. B Biol. Sci. 351:765–772.

Hoskin, C. J., M. Higgie, K. R. McDonald, and C. Moritz. 2005. Reinforcement drives rapid allopatric speciation. Nature 437:1353–1356.

Jackson, N. D., B. C. Carstens, A. E. Morales, and B. C. O'Meara. 2017a. Species delimitation with gene flow. Syst. Biol. 66:799–812.

Jackson, N. D., A. E. Morales, B. C. Carstens, and B. C. O'Meara. 2017b. PHRAPL: phylogeographic inference using approximate likelihoods. Syst. Biol. 66:1045–1053.

Jakobsson, M., and N. A. Rosenberg. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806.

Kronforst, M. R., L. G. Young, and L. E. Gilbert. 2007. Reinforcement of mate preference among hybridizing Heliconius butterflies. J. Evol. Biol. 20:278–285.

Leaché and Fujita. 2010. Bayesian species delimitation in West African Forest geckos (*Hemidactylus fasciatus*). Proc Biol Sci. 277:3071–3077.

Leaché, A. D., R. B. Harris, B. Rannala, and Z. Yang. 2014a. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63:17–30.

Leaché, A. D., M. K. Fujita, V. N. Minin, and R. R. Bouckaert. 2014b. Species delimitation using genome-wide SNP data. Syst. Biol. 63:534–542.

Leaché, A. D., T. Zhu, B. Rannala, and Z. Yang. 2018. The spectre of too many species. Syst. Biol. 61:168–181.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Slazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, and C. D. Jiggins. 2013. Genome-wide evidence for speciation with gene flow in Heliconius Butterflies. Genome Res. 23:1817–1828.

Mayr, E. 1942. Systematics and the origin of species, from the viewpoint of a zoologist. Harvard Univ. Press, Cambridge, MA.

———. 1954. Change of genetic environment and evolution. Pp. 157–180 *in* J. Huxley, ed. Evolution as a process. George Allen & Unwin, Lond.

Morales, A. E. and B. C. Carstens. 2018. Evidence that *Myotis lucifugus* "subspecies" are five nonsister species, despite gene flow. Syst. Biol. 67:756–769.

Morales, A. E., N. D. Jackson, T. A. Dewey, B. C. O'Meara, and B. C. Carstens. 2016. Speciation with gene flow in North American Myotis bats. Syst. Biol. 66:440–452.

Ng, T. P. T., S. H. Saltin, M. S. Davies, K. Johannesson, R. Stafford, and G. A. Williams. 2013. Snails and their trails: the multiple functions of trail-following in gastropods. Biol. Rev. 88:683–700.

Niemiller, M. L., B. M. Fitzpatrick, and B. T. Miller. 2008. Recent divergence with gene flow in Tennessee cave salamanders (Plethodontidae: Gyrinophilus) inferred from gene genealogies. Mol. Ecol. 17:2258–2275.

Nosil, P. 2012. Ecological speciation. Oxford Univ. Press, Oxford, U.K.

Ohio Supercomputer Center. 1987. Ohio supercomputer. Ohio Supercomputer Center, Columbus, OH.

Papadopulos, A. S. T., W. J. Baker, D. Crayn, R. K. Butlin, R. G. Kynast, I. Hutton, and V. Savolainen. 2011. Speciation with gene flow on Lord Howe Island. Proc. Natl. Acad. Sci. 108:13188–13193.

Pelletier, T. A., and B. C. Carstens. 2014. Model choice for phylogeographic inference using a large set of models. Mol. Ecol. 23:3028–3043.

Pfeifer, B. et al. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31:1929–1936.

Pielou, E. C. 2008. After the ice age: the return of life to glaciated North America. University of Chicago Press, Chicago.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pudlo, P., J. M. Marin, A. Estoup, J. M. Cornuet, M. Gautier, and C. P. Robert. 2015. Reliable ABC model choice via random forests. Bioinformatics 32:859–866.

Rannala, B., and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.

R Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rundle, H. D., and P. Nosil. 2005. Ecological speciation. Ecol. Lett. 8:336–352.

Satler, J. D., and B. C. Carstens. 2017. Do ecological communities disperse across biogeographic barriers as a unit. Mol. Ecol. 26:3533–3545.

Schluter, D. 2000. The ecology of adaptive radiation. Oxford University Press, Oxford, U.K.

Smith, M. L., M. Ruffley, D. C. Tank, J. Sullivan, and B. C. Carstens. 2017. Demographic model selection using random forests and the site frequency spectrum. Mol. Ecol. 26:4562–4573.

Smith, M. L., M. Ruffley, A. M. Rankin, A. Espíndola, D. C. Tank, J. Sullivan, and B. C. Carstens. 2018. Testing for the presence of cryptic diversity in tail-dropper slugs (*Prophysaon*) using molecular data. Biol. J. Linn. Soc. 124:518–532.

Sukumaran, J and L. L. Knowles. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. 114:1607–1612.

Templeton, A. R. 2008. The reality and importance of founder speciation in evolution. BioEssays 30:470–479.

Yang, Z., and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. 107:9264–9269.

Zachos, F. E. 2016. Species concepts in biology. Vol. 801. Springer, Cham, Switzerland.

Associate Editor: Dr. David Weisrock
Handling Editor: Maria R. Servedio

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Sampling Localities, and population assignments from STRUCTURE.
Table S2. Results from K values 1-10 from Structure and StructureHarvester.
Table S3. Admixture results from STRUCTURE.
Table S4. Parameter estimates under the best model (Fig. 2) from fsc26 (Excoffier et al. 2013).
Table S5. Priors for Hemidactylus analysis. coal/bioko indicates the common ancestor of coal and bioko.
Table S6. Results from the Hemidactylus geckos. Model numbers correspond to those reported in Figure S5.
Table S7. Results from the two pitcher plant-associated invertebrates.
Table S8. Cross-validation results for simulation study #1 with moderate divergence times with 10,000 SNPs.
Table S9. Cross-validation results for simulation study #1 with recent divergence times with 10,000 SNPs.
Table S10. Cross-validation results for the island-continent simulation study with 10,000 SNPs.
Table S11. Oob error rates for the simulation test evaluating the ability of delimitR to distinguish among secondary contact and divergence with gene flow models.
Table S12. Cross-validation results for the simulation test evaluating the ability of delimitR to distinguish among secondary contact and divergence with gene flow models.
Figure S1. An example of a default model set that can be generated by delimitR.
Figure S2. An example of a default model set that can be generated by delimitR.
Figure S3. An example of a default model set that can be generated by delimitR.
Figure S4. PCA of the binned SFS for data simulated under the ten best models and the empirical data.
Figure S5. Results from the analysis of the Hemidactylus geckos dataset.
Figure S6. Results from the analysis of the two pitcher plant invertebrates.
Figure S7. The simulation test evaluating the ability of delimitR to distinguish among secondary contact and divergence with gene flow models.
Figure S8. The simulation test evaluating the effects of the number of simulations and the number of trees on error rates.
Figure S9. The cross-validation results for the simulation study evaluating the effects of prior misspecification.

Received: 1 February 2017 | Revised: 16 May 2017 | Accepted: 22 May 2017

DOI: 10.1111/mec.14223

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

Demographic model selection using random forests and the site frequency spectrum

Megan L. Smith[1] | Megan Ruffley[2,3] | Anahí Espíndola[2,3] | David C. Tank[2,3] | Jack Sullivan[2,3] | Bryan C. Carstens[1]

[1]Department of Evolution, Ecology & Organismal Biology, The Ohio State University, Columbus, OH, USA

[2]Department of Biological Sciences, University of Idaho, Moscow, ID, USA

[3]Biological Sciences, Institute for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, Moscow, ID, USA

Correspondence
Bryan C. Carstens, Department of Evolution, Ecology & Organismal Biology, The Ohio State University, Columbus, OH, USA.
Email: carstens.12@osu.edu

Funding information
Division of Environmental Biology, Grant/Award Number: DEB 14575199, DEB 1457726, DG-1343012; US National Science Foundation; NSF GRFP; Ohio State University

Abstract

Phylogeographic data sets have grown from tens to thousands of loci in recent years, but extant statistical methods do not take full advantage of these large data sets. For example, approximate Bayesian computation (ABC) is a commonly used method for the explicit comparison of alternate demographic histories, but it is limited by the "curse of dimensionality" and issues related to the simulation and summarization of data when applied to next-generation sequencing (NGS) data sets. We implement here several improvements to overcome these difficulties. We use a Random Forest (RF) classifier for model selection to circumvent the curse of dimensionality and apply a binned representation of the multidimensional site frequency spectrum (mSFS) to address issues related to the simulation and summarization of large SNP data sets. We evaluate the performance of these improvements using simulation and find low overall error rates (~7%). We then apply the approach to data from *Haplotrema vancouverense*, a land snail endemic to the Pacific Northwest of North America. Fifteen demographic models were compared, and our results support a model of recent dispersal from coastal to inland rainforests. Our results demonstrate that binning is an effective strategy for the construction of a mSFS and imply that the statistical power of RF when applied to demographic model selection is at least comparable to traditional ABC algorithms. Importantly, by combining these strategies, large sets of models with differing numbers of populations can be evaluated.

KEYWORDS
machine learning, model selection, phylogeography, RADseq

1 | INTRODUCTION

Since before the term "phylogeography" was coined (Avise et al., 1987), the discipline has developed in response to advances in data-acquisition technology (reviewed in Garrick, Bonatelli, & Hyseni, 2015). Recently, phylogeographic investigations have transformed from traditional studies using data from a handful of genetic loci to contemporary studies where hundreds or thousands of loci are collected (Garrick et al., 2015). With the proliferation of next-generation sequencing (NGS) data sets, researchers can now access genetic data to investigate complex patterns of divergence and diversification in nonmodel species. In recent years, the field has increasingly relied upon model-based methods (Nielsen & Beaumont, 2009). These methods are primarily of two classes: those that estimate parameters under a predefined model and those that compare a number of user-defined models. The former type of approach has expanded recently to methods that are applicable to NGS data sets. For example, sequential Markovian coalescent (SMC) approaches can estimate population size histories and divergence times using whole genomes (Terhorst, Kamm, & Song, 2016). However, such methods

require that researchers identify a model a priori, and are generally limited to relatively simple models that omit many potentially important parameters, due to computational constraints. For example, while Terhorst et al.'s SMC approach can estimate divergence times and population size changes, it does not incorporate gene flow between lineages. Instead, researchers may wish to compare models that include different parameters and determine which model best fits their data, and this has led to an increase in the use of approximate methods, due to the computational challenges of comparing such complex models. A particularly flexible method in this regard is approximate Bayesian computation (ABC; e.g., Beaumont, 2010), which has been used in a wide range of applications outside of population genomics and phylogeography, including ecology, epidemiology and systems biology (Beaumont, 2010).

ABC methods enable researchers to customize demographic models to their empirical system, and allows formalized model selection (Table 1). Under each prespecified model, parameters of interest, θ_i, are drawn from a prior distribution, $\pi(\theta)$, specified by the researcher (step 1). Data, x_i, are then simulated from the distribution of the data given the parameters, $p(x \mid \theta_i)$ (step 2), and a vector of summary statistics, S, is calculated from the simulated and empirical data (step 3). The efficiency of ABC is a result of the optimization. Simulations that exceed a user-defined threshold, ε, as measured by the distance function, $\rho(S(x_i), S(y))$, are rejected (step 4) such that the remaining θ_i constitute the posterior distribution. If data are simulated under multiple models, the proportions of simulations that each model contributes to the posterior distribution correspond to the posterior probabilities of the models under consideration (step 5). ABC was developed in the context of a handful of microsatellite loci (Pritchard, Seielstad, Perez-Lezaun, & Feldman, 1999), but in theory can be extended to any amount of data. In practice, however, extending it to large NGS data sets is difficult due to the "curse of dimensionality" (Blum, 2010). This term describes the situation that occurs as the vector of summary statistics grows large, as would be the case if data were summarized on a locus-by-locus basis for hundreds to thousands of loci, and simulation of data near the vector requires an increasingly large number of simulations, which leads to high error rates. Although ABC has been applied to large NGS data sets (e.g., Roux et al., 2010; Veeramah et al., 2015), these applications have typically required that researchers summarize thousands of loci using a small vector of summary statistics (e.g., in Roux et al., 2010; the average and standard deviation over loci for 11 summary statistics). Summarizing data from 1,000s of loci with dozens of summary statistics results in a substantial loss of the information content of the data and limits the number of models that researchers have statistical power to distinguish. While methods have been suggested to guide researchers in their choice of summary statistics (e.g., partial least-squares transformation; Wegmann, Leuenberger, & Excoffier, 2009), they still result in a large decrease in the information content of the data. Some recent studies have used the bins of the site frequency spectrum (SFS) as a summary statistic for ABC inference (e.g., Boitard, Rodriguez, Jay, Mona, & Austerlitz, 2016; Prates, Rivera, Rodrigues, & Carnaval, 2016; Stocks, Siol, Lascoux, & De Mita, 2014; Xue & Hickerson, 2015), but these approaches have not taken advantages of joint or multidimensional SFS (mSFS). Consideration of the mSFS is necessary to make inferences about multiple populations, but the dimensionality of the mSFS increases as the number of individuals and populations sampled increases such that the number of bins in the joint or multidimensional SFS becomes very large, and the "curse of dimensionality" becomes a limiting factor. One possible solution to the limitations of ABC that would allow researchers to avoid reducing their data to a small number of summary statistics is to follow Pudlo et al. (2015) in replacing the traditional rejection step (steps 4-5; Table 1) with a machine-learning approach such as Random Forests (RF) for model selection.

In the RF approach to phylogeographic model selection, the data simulation and summarization steps (Table 1, steps 1-3) remain unchanged from the traditional ABC algorithm. However, instead of using a rejection step that relies on a specified distance function between the observed and simulated data, model selection proceeds using a classification forest. This forest consists of hundreds of

TABLE 1 Comparison of the ABC and RF approaches to demographic model selection

Comparison of ABC and RF algorithms for model selection	
Both ABC and RF	
1. Draw parameters θ_i from the prior distribution $\pi(\theta)$.	
2. Simulate data x_i from the distribution of the data given the parameters $p(x \mid \theta_i)$.	
3. Summarize the data using some statistic $S(x_i)$.	
ABC	**RF**
4. Reject θ_i when some function $\rho(S(xi),S(y))$ measuring the distance between the simulated and observed data exceeds a user-defined threshold.	4. Train a RF classifier using $S(x_i)$ as predictor variables and the model under which the $S(x_i)$ were simulated as the response variable.
5. The retained θ_i approximate the posterior distribution and are used to approximate model posterior probabilities.	5. Apply classifier to the observed data set to choose the best model.
	6. Estimate the probability of misclassification for the observed data using oob error rates.

decision trees and is trained on the simulated data, with the summary statistics serving as the predictor variables and the generating model serving as the response variable. Once built, this classifier can be applied to the observed data. Decision trees will favour (i.e., vote for) a particular model, and the model receiving the most votes will be selected as the best model. Although this approach does not include the approximation of the posterior probability, in contrast to ABC approaches that utilize a rejection step, uncertainty in model selection can be estimated using the error rates of the constructed classifier. Both experimental (Hastie, Tibshirani, & Friedman, 2009) and theoretical (Biau, 2012; Scornet, Biau, & Vert, 2015) justifications of RF have been offered, with RF shown to be robust both to correlations between predictor variables (here, the summary statistics) and to the inclusion of a large number of noisy predictors. An additional advantage of the RF approach is the reduction in computational effort required for model selection, as >50-fold gains in computational efficiency have been reported (Pudlo et al., 2015).

Although the data simulation and summary statistic calculation steps (steps 2-3 in Table 1) of the ABC algorithm may be extended to NGS data sets from a first-principles argument, issues arise in the implementation. First, the simulation of data scales linearly with the number of loci and thus becomes computationally intensive when the data sets in question are large (Sousa & Hey, 2013). Additionally, calculating a set of traditional summary statistics for each locus for use as summary statistics is impractical given the large number of loci. Although it is possible to calculate certain traditional summary statistics directly from the SFS, rather than on a locus-by-locus basis, such a calculation results in the loss of much of the information content of the data (Sainudiin et al., 2011).

In response to these issues, we explore the use of the multidimensional site frequency spectrum (mSFS; the joint distribution of allele frequencies across three or more populations) for data simulation and summarization in the RF model selection algorithm. The mSFS is a useful summary of the SNP data sets that are frequently collected using NGS methods, and can be considered a complete summary of the data when all polymorphic sites are independent (i.e., unlinked) and biallelic (e.g., Gutenkunst, Hernandez, Williamson, & Bustamante, 2009). Furthermore, the mSFS is expected to reflect demographic events including expansion, divergence and migration (Gutenkunst et al., 2009), although inferences based on the SFS may be inaccurate when too few segregating sites are sampled (Terhorst & Song, 2015). To address this issue, we apply a binning approach to coarsen the mSFS. The use of the mSFS for data summary can also facilitate data simulation; for example, the coalescent simulation program fastsimcoal2 (FSC2) uses a continuous time approximation to calculate the mSFS from simulated SNP data (Excoffier, Dupanloup, Huerta-SÃnchez, Sousa, & Foll, 2013). Here, we propose an approach to phylogeographic model selection that combines the use of a RF classifier with the use of the mSFS to summarize NGS data. We apply this approach to evaluate demographic models in *Haplotrema vancouverense*, a land snail endemic to temperate rainforests of the Pacific Northwest of North America (PNW).

2 | MATERIALS AND METHODS

2.1 | Study system and models

2.1.1 | Study system

The PNW of North America can be divided into three distinct regions: the Cascades and Coastal Ranges in the west, the Northern Rocky Mountains in the east and the intervening Columbia Plateau (e.g., Figure 1; Brunsfeld, Sullivan, Soltis, & Soltis, 2000). The coastal and inland mountain ranges are characterized by mesic, temperate coniferous forests, but the intervening basin is characterized by a shrub–steppe ecosystem generated by the rain shadow of the Cascade Range that has developed since its orogeny in the early Pliocene. The Okanogan Highlands to the north and the Central Oregonian highlands to the south partially mitigate the ecological isolation of the inland and coastal forests, but the Columbia Plateau has nevertheless been a substantial barrier to dispersal for many of the taxa endemic to these temperate forests (e.g., Carstens, Brunsfeld, Demboski, Good, & Sullivan, 2005). In addition to being influenced by mountain formation, the distributions of taxa in the rainforests of the PNW have likely been impacted by climatic fluctuations throughout the Pleistocene (Pielou, 2008). Glaciers formed and retreated several times during these fluctuations, covering large portions of the northern parts of species' current ranges. Thus, species may have been entirely eliminated in the northern parts of their ranges or may have survived in small isolated glacial refugia.

Several biogeographic hypotheses have been proposed to explain the disjunct distribution of the PNW mesic forest endemics (reviewed in Brunsfeld et al., 2000). Here, we explore models that include from one to three glacial refugia (South Cascades, North Cascades and Clearwater River drainages). In one class of models, no refugia persisted in the inland region, and these models posit dispersal to the inland via either a southern or a northern route. In addition, to test whether or not there was population structure present, we evaluated models that included from one to four distinct populations (South Cascades, North Cascades, Clearwater River drainages and northern Idaho drainages). In total, we include 15 demographic models that differed in the number of populations, the number and location of refugia and the dispersal route (Figure 1; Fig. S1). We applied the approach proposed here to *Haplotrema vancouverense*, a land snail endemic to the PNW. No previous work has used genomic data to investigate the demographic history of this species. However, one study used environmental data to predict that *H. vancouverense* did not harbour cryptic diversity across the Columbia Basin (Espíndola et al., 2016).

2.2 | Specimen collection and data generation

Samples were collected for this study during the spring of 2015 and 2016, in addition to loans provided by the Idaho Fish and Game and museum collections (the Royal British Columbia Museum and the Florida Museum of Natural History). In total, we acquired 77 snails

from throughout the range of *H. vancouverense* (Figure 2; Table S1). This included 31 snails from 24 localities in the northern and southern Cascades and 46 snails from 18 localities in the Clearwater River and northern Idaho drainages. After collection, snails were preserved in 95% ethanol and DNA was extracted using Qiagen DNeasy Blood and Tissue Kits (Qiagen, Hilden, Germany) following the manufacturer's protocol. Prior to library preparation, DNA was quantified on a Qubit fluorometer (Life Technologies), and 200–300 nanograms of DNA was used for library preparation.

Library preparation followed the double-digest restriction-associated DNA (ddRAD) sequencing protocol developed in Peterson, Weber, Kay, Fisher, & Hoekstra, 2012, with modifications. DNA was digested using the restriction enzymes SbfII and MspI (New England Biolabs, USA), and adapters were ligated using T4 ligase (New England Biolabs). Ligated products were cleaned using magnetic beads in a PEG/NaCl buffer (Rohland & Reich, 2012). A subset of the ligation products was amplified and analysed by qPCR using the library quantification kit for Illumina libraries (KAPA Biosystems, USA) to ensure that no adapter had failed to ligate during the ligation step. All ligation products were quantified on the Qubit fluorometer (Life Technologies) and pooled across index groups in equimolar concentrations. 10–20 nanograms of this pool was used in each subsequent PCR. PCRs used the Phusion Master Mix (Thermo Fisher Scientific, USA) and were run for an initial step of 30 s at 98°C, followed by 16 cycles of 5 s at 98°C, 25 s at 60°C and 10 s at 72°C and a final extension for 5 min at 72°C. To minimize PCR bias, reactions were replicated seven times for each index group, and products were pooled within index groups. We analysed 4 μl of this pooled PCR product on a 1% agarose gel. A second clean-up using magnetic beads in a PEG/NaCl buffer (Rohland & Reich, 2012) was performed. Finally, PCR products were quantified on the Qubit fluorometer (Life Technologies) prior to selection for 300- to 600-bp fragments using the Blue Pippin (Sage Science, USA) following manufacturer's standard protocols. The remaining products were quantified using the Qubit fluorometer (Life Technologies) and the Bioanalyzer (Agilent Technologies, USA) before being pooled and sent for sequencing on an Illumina Hi-Seq at the Genomics Shared Resource Center at Ohio State University.

2.3 | Bioinformatics

Raw sequence reads were demultiplexed and processed using PYRAD (Eaton, 2014). Sites with a Phred quality score <20 were masked with Ns, and reads with more than four Ns were discarded. A minimum of ten reads was required for a locus to be called within an individual. Filtered reads were clustered using the program VSEARCH v.2.0.2 (https://github.com/torognes/vsearch) and aligned using MUSCLE v.3.8.31 (Edgar, 2004) under a clustering threshold of 85%. Consensus sequences with more than three heterozygous sites or more than two haplotypes for an individual were discarded, and loci represented in fewer than 60 per cent of individuals were discarded. Cut-sites and adapters were removed from sequences using the strict filtering in PYRAD.

To deal with missing data when constructing the mSFS, we applied a downsampling approach to maximize the number of SNPs included in the mSFS. A threshold of 50% was set in each population, meaning that only SNPs scored in at least half of the individuals in each population would be used in downstream analyses. For SNPs that exceeded this threshold, we randomly subsampled alleles. We repeated this downsampling approach ten times to create ten different mSFS to be used in downstream analyses in an attempt to account for rare alleles potentially missed during the downsampling procedure. Downsampling followed Thomé and Carstens (2016) and was performed using custom PYTHON scripts modified from scripts developed by J. Satler (https://github.com/jordansatler; modified version at https://github.com/meganlsmith). This approach was chosen over including only loci sampled across all individuals because such an approach would have limited the number of SNPs included in the

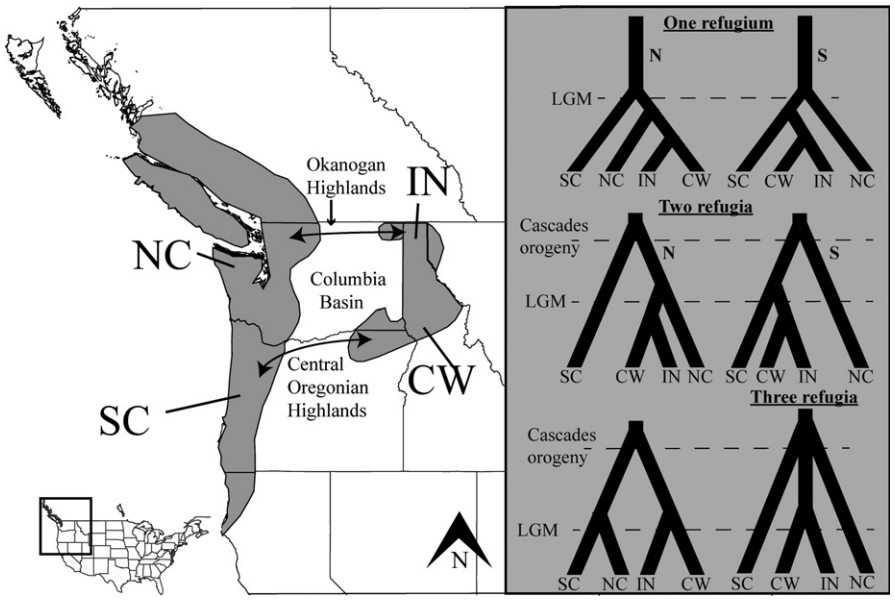


FIGURE 1 Map of the PNW illustrating the models tested in this study. NC, North Cascades; SC, South Cascades; IN, Northern Inland Drainages; CW, Clearwater drainages. The models tested included one to three refugia. When there were no inland refugia, dispersal could occur via either a northern or southern route. Additional models tested (Fig. S1) included from one to four populations. The heights of the bars indicate the time since colonization of the region τ_{col}, with taller bars indicating older populations. The shaded region on the map marks the distribution of *Haplotrema vancouverense*, reproduced from Burke (2013)

study and has been shown to bias parameter estimates due to the nonrandom sampling of genealogies (Huang & Knowles, 2014).

2.4 | Random forest model selection using the mSFS as a summary statistic

2.4.1 | Data simulation and summarization

The RF approach to model selection (Figure 3) follows the algorithm for RF model selection presented in Table 1. Parameters were drawn from prior distributions (Table S2) under each of the fifteen models considered (Figure 3; Step 1). mSFS were simulated in FSC2 (Excoffier et al., 2013) under each model, using a folded mSFS with a number of SNPs equivalent to the observed mSFS (Figure 3; Step 2). Monomorphic sites were not considered, and 10,000 replicate mSFS were simulated under each model in FSC2, leading to a total prior of 150,000 mSFS.

Given the number of populations included as well as the number of SNPs obtained by our sequencing protocol (see Results), use of all bins from the mSFS could result in limited coverage across the mSFS and thus to poor estimates of the mSFS; therefore, we used a custom Python script (https://github.com/meganlsmith) to coarsen the mSFS (Figure 3, Step 3). For example, for the "quartets" data set, SNPs were categorized based on which quartile they belonged to in each population, and all combinations of quartiles across populations were used as bins for a final data set consisting of 256 bins. In this example, the first bin would consist of SNPs occurring at a frequency <¼ in all four populations. We tested other binning strategies with the number of classes ranging from three to ten, enabling a joint exploration of the coarseness of the mSFS, the accuracy of model selection and the computational requirements of the classification procedure.

2.4.2 | Choosing the optimal binning strategy

To determine the optimal binning strategy, eight RF classifiers were constructed using the simulated data (i.e., Figure 3, Step 4), one at each level of mSFS coarseness considered here (i.e., 3–10 classes per population). Each classifier was constructed with 500 trees using the R package "ABCRF" (Pudlo et al., 2015), with the bins of the mSFS treated as the predictor variables and the generating model for each simulated data set treated as the response variable. At each node in each decision tree, the RF classifier considers a bin of the mSFS and constructs a binary decision rule based on the number of SNPs in the bin.

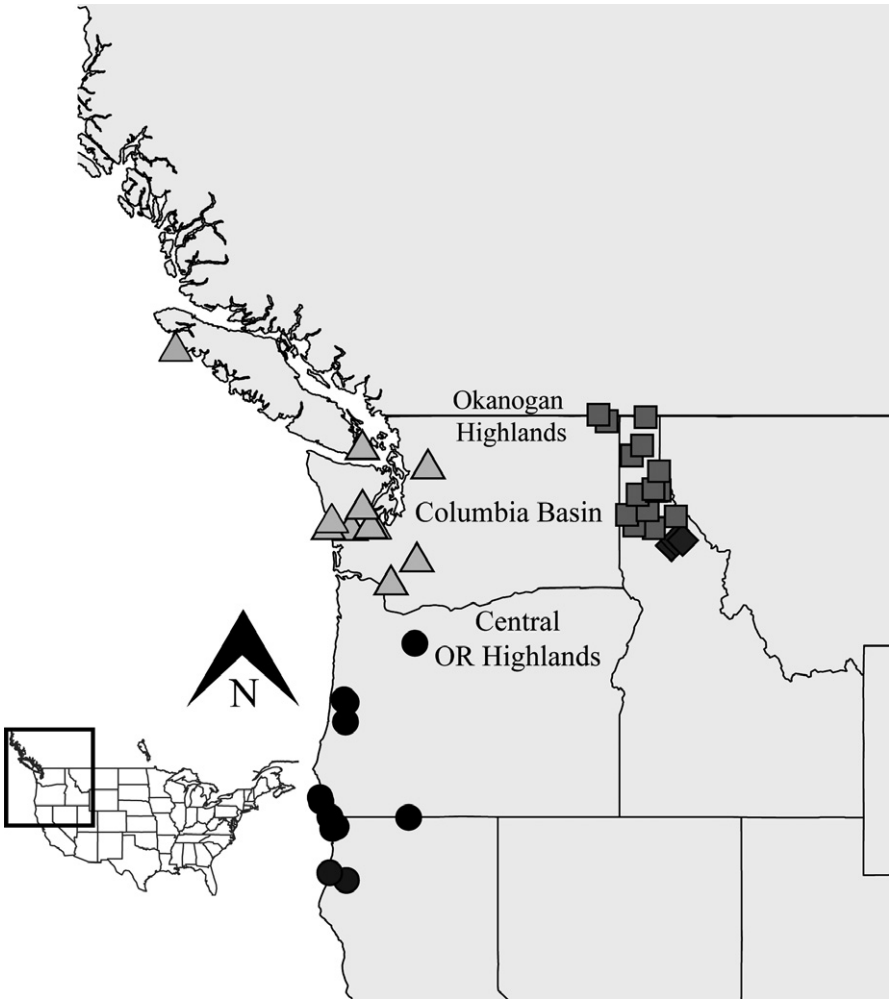


FIGURE 2 Collection localities for *H. vancouverense*. North Cascades = triangles; South Cascades = circles; Northern Inland Drainages = squares; Clearwater drainages = diamonds

When this classifier is applied to other data sets, it makes decisions at each node until it reaches a leaf of the decision tree, which in this instance is a model index. When a leaf is reached, the decision tree is said to "vote" for the model index assigned to that leaf. Each decision tree is constructed in reference to only a portion of the training data set, minimizing the correlation between decision trees. Prior to construction of the random forest, columns in which there was no variance in the entire prior (e.g., bins that contained no SNPs for any of the simulated data sets) were removed from the prior. These same columns were removed from the observed data set.

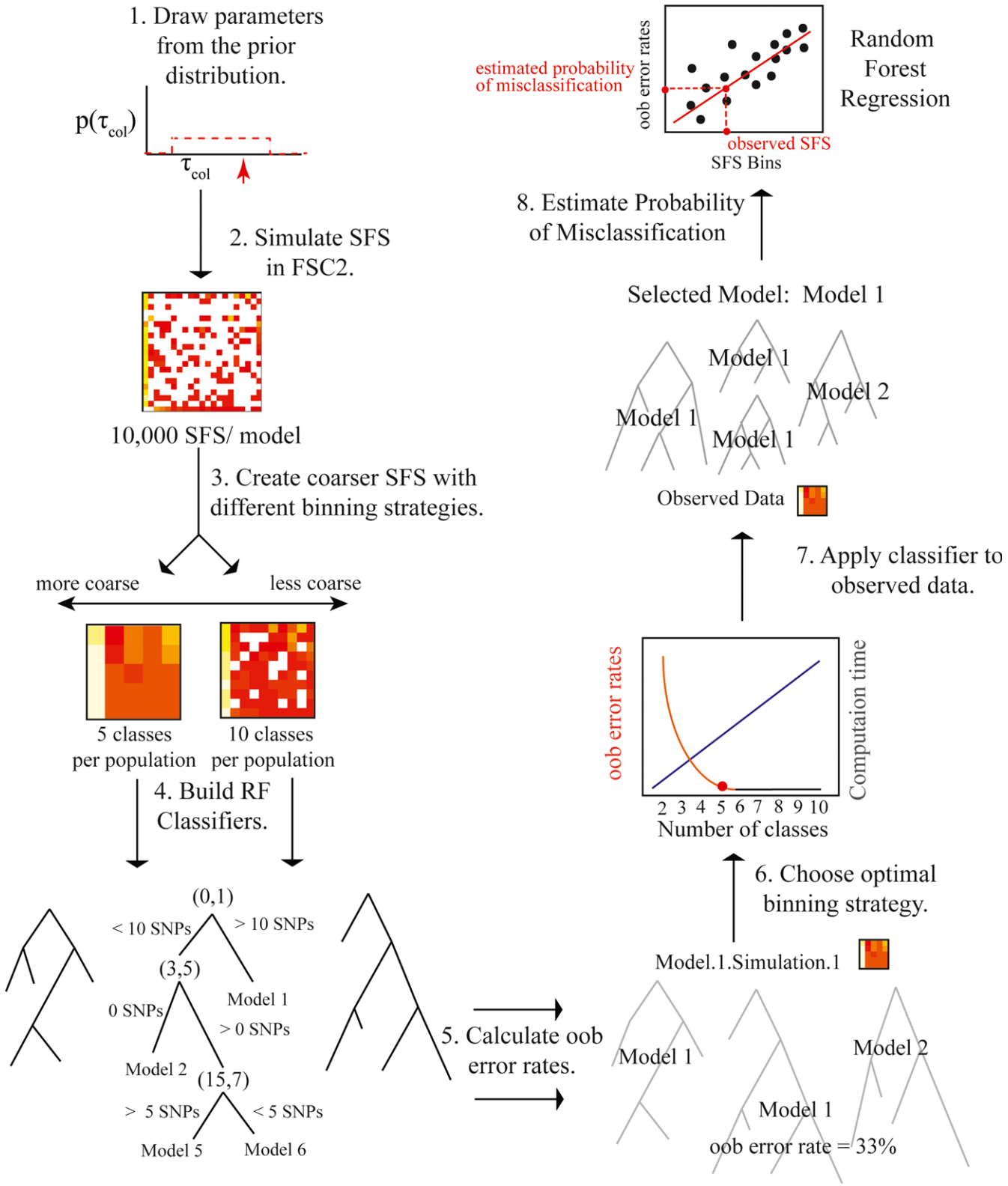


FIGURE 3 Flow chart illustrating the RF approach to model selection [Colour figure can be viewed at wileyonlinelibrary.com]

Because only a portion of the prior is used in the construction of each decision tree in RF classification, the error rate of the classifier can be assessed using the "out-of-the-bag" (oob) error rates (Figure 3; Step 5). Oob error rates are calculated by considering only decision trees constructed without reference to a particular element of the prior. For each simulated mSFS, we used a smaller classifier that consisted only of trees constructed without reference to the mSFS in question. We applied this classifier to the simulated mSFS and calculated the proportion of trees that voted for an incorrect model; this is the oob error rate for the simulated mSFS. To choose the optimal binning strategy (Figure 3, Step 6), we plotted the average misclassification rate and the computation effort required as a function of the binning strategy.

2.4.3 | Model selection and the misclassification rate

After the optimal binning strategy was determined, we applied the corresponding classifier to the observed data (Figure 3, Step 7). The "predict" function in the "abcrf" package was used to select the best model for the observed data, which was the model receiving the most votes (i.e., the model selected by the largest number of decision trees). One limitation of the RF approach is that the number of votes allocated to different models has no direct relationship to the posterior probabilities of the models and may be a poor measure of the probability of misclassification for the observed data. Following Pudlo et al. (2015), we estimated the probability of misclassification in a second step by regressing over the selection error in the prior to build a regression RF, in which the oob error rate is the response variable and the mSFS bins are the predictor variables. We then applied this RF to the observed data to estimate the probability of misclassification for the observed model (Figure 3; Step 8), again using the "predict" function in the R package "abcrf" (Pudlo et al., 2015). A Python script that simulates data, constructs a reference table, builds a classifier, selects the best model for the empirical data and calculates error rates and the probability of misclassification using FSC2 and the R package "abcrf" is available on github (https://github.com/meganlsmith).

To assess the power of the RF approach, we simulated 100 mSFS under each of the 15 models (Fig. S1), drawing priors from the same distributions used in model selection (Information on Prior Distributions; Table S2). We used custom python scripts (https://github.com/meganlsmith) to coarsen the simulated mSFS using five classes. We then applied the RF classifier built from the quintets prior to each of the simulated data sets using the "predict" function in the R package "abcrf" (Pudlo et al., 2015) and recorded which model was selected for each replicate.

2.5 | AIC-based model selection

To validate the results of our model selection using RF, we compared the above results to a commonly used information theoretic approach to phylogeographic model selection with NGS data sets (e.g., Carstens et al., 2013), where model selection in FSC2 followed the procedure suggested in Excoffier et al. (2013). FSC2 maximizes the composite likelihood of the observed data under an arbitrary number of models, and Akaike information theory can then be used to select among several tested models. The Brent algorithm implemented in FSC2 was used for parameter optimization, with parameter optimization replicated 100 times. For each replicate, 100,000 simulations were used for the calculation of the composite likelihood and 40 cycles of the Brent algorithm were used for parameter optimization. The maximum-likelihood estimates for the parameters were then fixed, and the likelihood was approximated for each model across 100 different replicates. The maximum likelihood across these 100 replicates for each model was used in model comparison. AIC scores were then calculated and converted to model weights as in Excoffier et al. (2013).

To assess the power of FSC2 to distinguish among the tested models, we used the same 100 simulated mSFS as in the RF power analysis. We used 100,000 simulations for the calculation of the composite likelihood, and 40 cycles of the Brent algorithm were used for parameter optimization. The likelihood was approximated for each model and used in model comparisons. AIC scores were calculated and converted to model weights as in Excoffier et al. (2013), and we recorded which model was selected for each replicate. Due to computational constraints, we did not perform the replication recommended for model selection in FSC2, as was done for the observed data. We also conducted a conventional ABC analysis (see Supporting Information).

3 | RESULTS

3.1 | Bioinformatics

After the filtering thresholds were applied, 1,943 loci were called in 77 individuals. This resulted in 1,716 unlinked biallelic SNPs and 5,996 total variable sites. When only unlinked SNPs were used, the downsampling approach resulted in data sets including SNPs from 12 alleles per locus from the Clearwater drainages, 14 alleles per locus from the North Cascades, 34 alleles per locus from the northern inland drainages and 17 alleles per locus from the South Cascades. These data sets included between 879 and 908 SNPs.

3.2 | RF model selection with the mSFS as a summary statistic

3.2.1 | Oob error rates and optimal binning strategy

Oob error rates decreased as the number of classes used to build the coarse mSFS increased, until the number of classes reached five (Figure 4). The error rate is no worse for five as opposed to a greater number of classes, and the computation effort increases considerably with larger numbers of classes (Figure 4). We therefore determined that five classes represented the optimal binning strategy for our data, and as such present results only from the "quintets" data set below. Using the "quintets" data set, the overall prior error

rate, calculated using oob error rates, was 6.59 per cent. Error rates varied across models (Fig. S2) and were highest between those models for which the only difference was whether dispersal occurred via a northern or a southern route (Table 2). Misclassification across models with different numbers of populations was less common, and data sets were never classified as belonging to a model having a different number or identity of refugia than the generating model (Table 2). Error rates appeared to plateau in relation to the number of trees used to construct the model, suggesting that more trees did not improve the predictive ability of the RF classifier (Fig. S3).

3.2.2 | Model selection with random forests and AIC-based model selection

In analyses of the "quintets" data set, RF selected the four-population model that included recent southern dispersal to the inland region (Figure 5: Model 1; Table 3). The next best model was similar, but with colonization of the inland region via a northern instead of a southern route (Figure 5: Model 2; Table 3). The probability of misclassification of the best model was estimated to be 0.3514 (corresponding to an approximated posterior probability of 0.6846). The best model did not change between data sets built with different binning strategies, but the probability of misclassification varied across data sets (Table 3). To account for variation in the downsampling procedure, ten downsampling replicates were analysed using five categories per population to bin the data; the best model did not change between data sets, but the misclassification probability of the best model varied across data sets (Table S3). This analysis (constructing the RF from the prior, calculating oob error rates and applying the RF classifier to the observed data) was run on six processors with 24GB RAM and used 78.9 min of CPU time. Under the likelihood-based approach, the best model was a four-population model of recent dispersal to the inland with colonization via a northern route (Figure 5: Model 2; Table 4). The next best model was the same, except that colonization of the inland occurred via a southern route (Figure 5: Model 1; Table 4). This analysis required more than 1,500 CPU hr, largely due to the replication required in calculating the composite likelihood.

3.2.3 | Power analyses in random forests and AIC-based model selection

In the power analysis for the RF approach, the overall error rate was 7.67 per cent (Table S4). The highest error rates were for models 1 and 2 (Fig. S1) at 22 and 42 per cent, respectively. The power analysis in RF used approximately ~285 CPU hr.

In the power analysis for the FSC2 approach, the overall error rate was 3.33 per cent (Table S4). The highest error rates were for models 1 and 2 at 10 and 15 percent, respectively. Although we were not able to perform a full power analysis (with fixed MLE parameter estimates and replicates to approximate the composite likelihood of each model) due to computational constraints, the partial power analysis used approximately 2,205 resource units (~22,205 CPU hr). Model selection results of the conventional ABC

TABLE 2 Summary of the probabilities of different types of classification errors

Misclassification probabilities	
Description of Misclassification	**Probability**
Misclassified as model with a different number of populations	1.44%
Misclassified as model with different number of refugia	0.00%
Misclassified as model with different dispersal route	4.95%

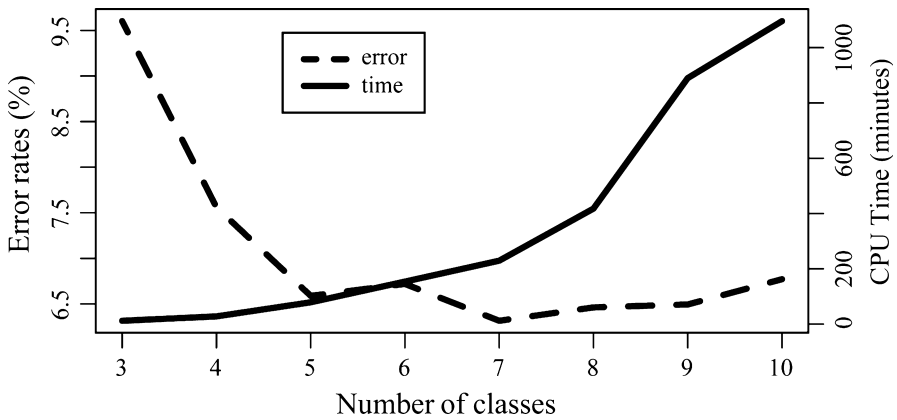


FIGURE 4 Error rates and computation time vs. the number of classes used to construct the mSFS. "Four classes" indicates that there were four categories of SNPs per population, for a total of 256 bins in a four-population multidimensional mSFS. All computations were performed on the Ohio Supercomputer, and CPU time indicates CPU time required to construct a Random Forest from the prior, estimate the oob error rates of the RF and apply this RF to the observed data. For up to six classes, computations were performed on six processors with 24GB of RAM. For seven and eight classes, computations were performed on twelve processors with 48GB of RAM. For nine and ten classes, computations were performed on a twelve processors with 192GB of RAM

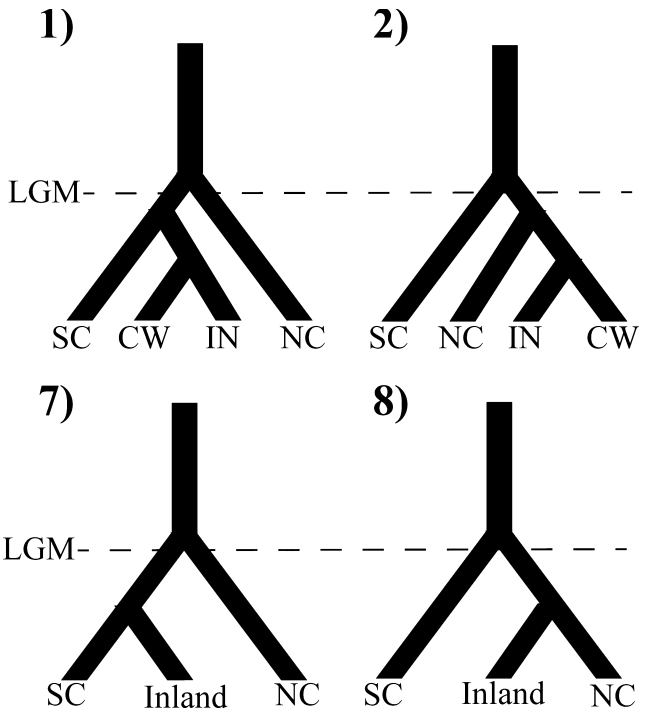


FIGURE 5 The four best models based on ABC and FSC2 results. All models include one refugium in the South Cascades. (1) and (2) include four populations, while (7) and (8) lump the two inland populations together. (1) and (7) posit a southern route of colonization of the inland rainforests, while (2) and (8) posit a northern route of colonization

analysis were similar, but our simulations suggested that the power to detect the best model was lower overall (Supporting Information).

4 | DISCUSSION

4.1 | Model selection using random forests

The combination of RF classification and the binning strategy for mSFS data appear to perform well in the context of phylogeographic model selection, and the use of the RF algorithm for model selection in place of a traditional ABC approach allowed us to circumvent many of the issues associated with using a traditional ABC approach on NGS data sets, with error rates much lower than those obtained when a classical ABC approach was applied to this data (ABC error rate = 30%, Supporting Information). The low error rate obtained in the RF approach to model selection (6.59%) can likely be attributed both to the more efficient approach to model selection and to the more complete summary of the data provided by the mSFS. Computational requirements (Figure 4) were much less than those of FSC2. In comparison with AIC-based methods, RF model selection is favourable in certain situations. Although AIC-based methods, such as FSC2, have proven powerful in certain contexts (Excoffier et al., 2013), the power of such analyses when applied to the smaller NGS data sets frequently collected using protocols such as ddRAD sequencing (Peterson et al., 2012) on nonmodel organisms has not been thoroughly evaluated in most studies using FSC2. Particularly when the number of bins in the mSFS greatly exceeds the number of SNPs, as is likely to occur as the number of populations increases, it may be inappropriate to use the full mSFS due to the reduction in the accuracy of parameter estimations (and thus of the likelihood calculation) that such data sets are expected to provide, as inferences based on SFS with small-to-moderate numbers of SNPs have been shown to be inaccurate (Terhorst & Song, 2015). Although FSC2 and the RF approach had similar power to distinguish among the models we tested, due to the computational requirements, it was difficult to assess the power of FSC2 given the data collected. The approach proposed here has the advantage of oob error rates, which enable an efficient evaluation of the power of the method given the collected data. Then, researchers can generate coarser mSFS according to the characteristics of their data and system.

TABLE 3 Model votes for the four best models and one minus the probability of misclassification of the selected model (an approximation of the posterior probability) for data sets with seven different levels of coarseness (3-10 categories for within population frequencies; 256-10,000 bins). Models 1, 2, 7 and 8 are illustrated in Figure 5

Results of ABC RF Model Selection					
# Categories	Model 1	Model 2	Model 7	Model 8	1-Pr(Misclassification)
3	234	121	80	45	0.6241
4	252	132	44	28	0.7292
5	212	109	72	52	0.6846
6	168	124	74	55	0.6933
7	170	100	56	43	0.6565
8	194	131	48	31	0.6546
9	134	129	61	59	0.6927
10	164	131	65	40	0.6364

While the RF approach has several advantages for model selection, joint estimation of parameters is not straightforward (but see Raynal et al., 2017). Additionally, as the monomorphic cell (the cell with counts of sites without variation) of the mSFS is not used in our approach, the timing of demographic events is relative rather than absolute. For cases when researchers prefer to test explicit a priori hypotheses based on geological data (e.g., Carstens et al., 2013), other approaches (including FSC2) should be preferred. In general, we suggest that parameter estimation using methods such as FSC2 using the model(s) selected following this approach as well as all available SNPs is likely the most effective strategy for non-model systems.

4.2 | Future directions

Model selection using RF has many potential advantages that future investigations should explore. Here, we highlight two such possibilities: (i) testing a large number of models and (ii) species delimitation. Using RF, we were able to test a moderate number ($N = 15$) of

TABLE 4 Results from the four best models, based on the results of the likelihood-based model selection in FSC2. Models differed in the number of populations and the route of dispersal

Model probabilities for the four best models							
Populations	**Refugia**	**Dispersal Route**	**K**	**LnLhood**	**AIC**	Δ_i	**wAIC**
4 (NC, SC, NID, CW)	SC	South	8	−7,351	14,718	3	0.173
4 (NC, SC, NID, CW)	SC	North	8	−7,349	14,715	0	0.827
3 (NC, SC, NID+CW)	SC	South	7	−7,705	15,423	708	0.000
3 (NC, SC, NID+CW)	SC	North	7	−7,712	15,438	723	0.000

NC, North Cascades; SC, South Cascades; NID, Northern Inland Drainages, CW, Clearwater.

demographic models without sacrificing our ability to distinguish between models. The error rate associated with model selection using traditional ABC algorithms appears to increase as the number of models increases, particularly when more than four models are included (Pelletier & Carstens, 2014). Our results suggest that it may be possible to compare a larger number of models using the RF model selection approach, allowing researchers to make fewer assumptions about the historical processes that may have influenced their focal organisms. The out-of-the-bag error rates generated in this approach allow researchers to assess whether they can distinguish among the models tested, given their data, and should thus prevent researchers from testing more models than they have the power to differentiate among. As with other approaches to demographic model selection, we were still limited in the number of models that we could compare, and our results can only highlight the best model among those tested. Although assessing model fit can help researchers understand how well their data fit a model, such an approach is not straightforward with the RF approach.

Additionally, we were able to compare models that included different numbers of populations with a low misclassification rate (1.44%). It has been challenging to use ABC in such cases because some of the most useful summary statistics are based on comparisons within and between populations (e.g., Hickerson, Dolman, & Moritz, 2006), and the summary statistic vectors used in such a comparison would necessarily have different dimensionalities. Here, we were able to circumvent this issue by calculating the mSFS as if there were four populations, regardless of the number of populations used to generate the SNP data. Our results suggest that the model selection approach implemented here could be used for population and potentially species delimitation. Additionally, although we focused on Random Forests here, other machine-learning algorithms have been used to infer demographic histories (e.g., Deep Learning; Sheehan & Song, 2016), and future work should investigate the use of these algorithms in conjunction with the binned mSFS.

4.3 | Empirical results

Results from both ABC (Table 3) and FSC2 (Table 4) suggest that *H. vancouverense* survived in one or more refugia in the south Cascades throughout the Pleistocene glacial cycles. A previous investigation using environmental and taxonomic data to make predictions concerning the evolutionary histories of organisms predicted that *H. vancouverense* colonized the inland after the Pleistocene (Espíndola et al., 2016), and the results presented here support this prediction. Following glaciation, *H. vancouverense* expanded its range north to the North Cascades and east to the Northern Rocky Mountains. Our results were incongruent across the ABC and FSC2 analyses in regard to whether *H. vancouverense* colonized the Northern Rockies via a northern route across the Okanogan highlands or via a southern route across the Central Oregonian highlands. In our RF analysis, these two models are misclassified at proportions of 0.19 (southern route classified as northern route) and 0.18 (northern route classified as southern route), based on out-of-the-bag error rates. In the power analysis for FSC2, these models were misclassified 10 and 15 percent of the time (Power Analysis in FSC2; Table S5), and in the power analysis for the RF approach, these two models were misclassified 22 and 42 per cent of the time. In combination with the ambiguity across methods, this suggests that we have limited power to distinguish between these two models, given the data collected here.

5 | CONCLUSION

Our results indicate that binning can be an effective strategy for the summarization of the mSFS. This comes at an important time, when SNP data sets from hundreds to thousands of SNPs are being collected from a variety of nonmodel species. Our work demonstrates that, using the binning strategy together with the RF strategy for model selection, researchers can make accurate phylogeographic inferences from NGS data sets that may be too small for accurate estimation of the true mSFS. Finally, we show that by allowing researchers to evaluate a larger number of models and to compare models with different numbers of populations, RF model selection could have important implications for the future of model-based approaches.

ACKNOWLEDGEMENTS

Funding was provided by the US National Science Foundation (DEB 1457726/14575199). MLS was supported by a NSF GRFP (DG-1343012) and a University Fellowship from The Ohio State University. We thank the Royal British Columbia Museum and the Florida Museum of Natural History for providing samples. We thank Michael Lucid of Idaho Fish and Game for donations of samples and the Ohio Supercomputer Center for computing resources (allocation grant PAS1181-1). We thank the Carstens laboratory for comments that improved this manuscript prior to publication. We would also like to thank Graham Stone and three anonymous reviewers for helpful comments during the review process.

DATA ACCESSIBILITY

Raw reads, the parameters used for data processing and a full SNP data set are available on Dryad (https://doi.org/10.5061/dryad.2j27b). Scripts developed as a part of the work presented here are available on github (https://github.com/meganlsmith).

AUTHOR CONTRIBUTIONS

M.L.S and B.C.C designed the study. Funding and support were obtained by B.C.C, D.C.T and J.S. M.L.S and M.R collected samples. M.L.S collected genomic data, performed genetic analyses and wrote the article. All authors edited the article and approved the final version of the article.

REFERENCES

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., ... Saunders, N. C. (1987). Intraspecific phylogeography: The mitochondrial dna bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, *18*, 489–522.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology and Systematics*, *41*, 379–406.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, *13*, 1063–1095.

Blum, M. G. B. (2010). Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association*, *105*, 491–1178.

Boitard, S., Rodriguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data - an approximate bayesian computation approach. *PLoS Genetics*, *12*, 1–36.

Brunsfeld, S. J., Sullivan, J., Soltis, D. E., & Soltis, P. S. (2000). Comparative phylogeography of north- western North America : A synthesis. *Special Publication-British Ecological Society*, *14*, 319–340.

Burke, T. E. (2013). *Land snails and slugs of the Pacific Northwest*. Corvallis, OR, USA: Oregon State University.

Carstens, B. C., Brennan, R. S., Chua, V., Bermingham, E., Lamb, T., Neigel, J. E., ... Saunders, N. C. (2013). Model selection as a tool for phylogeographic inference: An example from the willow Salix melanopsis. *Molecular Ecology*, *22*, 4014–4028.

Carstens, B. C., Brunsfeld, S. J., Demboski, J. R., Good, J. M., & 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.

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

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 113.

Espíndola, A., Ruffley, M., Smith, M. L., Carstens, B. C., Tank, D. C., & Sullivan, J. (2016). Identifying cryptic diversity with predictive phylogeography. *Proceedings of the Royal Society B: Biological Sciences*, *283*, 20161529.

Excoffier, L., Dupanloup, I., Huerta-SÃnchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*, e1003905.

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.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*, e1000695.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2nd edn. New York: Springer.

Hickerson, M. J., Dolman, G., & Moritz, C. (2006). Comparative phylogeographic summary statistics for testing simultaneous vicariance. *Molecular Ecology*, *15*, 209–223.

Huang, H., & Knowles, L. L. (2014). Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Systematic biology*, *65*(3), 357–365.

Nielsen, R., & Beaumont, M. A. (2009). Statistical inferences in phylogeography. *Molecular Ecology*, *18*, 1034–1047.

Pelletier, T. A., & Carstens, B. C. (2014). Model choice for phylogeographic inference using a large set of models. *Molecular Ecology*, *23*, 3028–3043.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, *7*, e37135.

Pielou, E. C. (2008). *After the ice age: The return of life to glaciated North America*. Chicago, IL, USA: University of Chicago Press.

Prates, I., Rivera, D., Rodrigues, M. T., & Carnaval, A. C. (2016). A mid-Pleistocene rainforest corridor enabled synchronous invasions of the Atlantic Forest by Amazonian anole lizards. *Molecular Ecology*, *25*, 5174–5186.

Pritchard, J., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, *16*, 1791–1798.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J., Gautier, M., & Robert, C. P. (2015). Reliable ABC model choice via random forests. *Bioinformatics*, *32*, 859–866.

Raynal, L., Marin, J., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2017). ABC random forests for Bayesian parameter inference. arXiv preprint, arXiv:1605.05537.

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing. *Genome Research*, *22*, 939–946.

Roux, C., Fraïsse, C., & Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2010). Shedding light on the grey zone of speciation along a continuum of genomic divergence. bioRxiv, 513–516.

Sainudiin, R., Thornton, K., Harlow, J., Booth, J., Stillman, M., Yoshida, R., ... Donnelly, P. (2011). Experiments with the site frequency spectrum. *Bulletin of Mathematical Biology*, *73*, 829–872.

Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, *43*, 1716–1741.

Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS Computational Biology*, *12*, 1–28.

Sousa, V., & Hey, J. (2013). Understanding the origin of species with genome-scale data: Modelling gene flow. *Nature Reviews Genetics*, *14*, 404–414.

Stocks, M., Siol, M., Lascoux, M., & De Mita, S. (2014). Amount of information needed for model choice in Approximate Bayesian Computation. *PLoS One*, *9*, 1–13.

Terhorst, J., Kamm, J. A., & Song, Y. S. (2016). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*, 303–309.

Terhorst, J., & Song, Y. S. (2015). Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences*, *112*, 7677–7682.

Thomé, M. T. C., & Carstens, B. C. (2016). Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. *Proceedings of the National Academy of Sciences*, *113*, 8010–8017.

Veeramah, K. R., Woerner, A. E., Johnstone, L., Gut, I., Gut, M., Marques-Bonet, T., ... Hammer, M. F. (2015). Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. *Genetics*, *200*, 295–308.

Wegmann, D., Leuenberger, C., & Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, *182*, 1207–1218.

Xue, A. T., & Hickerson, M. J. (2015). The aggregate site frequency spectrum for comparative population genomic inference. *Molecular Ecology*, *24*, 6223–6240.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

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

Biological Journal of the Linnean Society, 2018, **124**, 518–532. With 4 figures.

Testing for the presence of cryptic diversity in tail-dropper slugs (*Prophysaon*) using molecular data

MEGAN L. SMITH[1]*, MEGAN RUFFLEY[2,3], ANDREW M. RANKIN[2,3], ANAHÍ ESPÍNDOLA[2,3], DAVID C. TANK[2,3], JACK SULLIVAN[2,3] and BRYAN C. CARSTENS[1]

[1]*Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 318 W. 12th Avenue, 300 Aronoff Labs, Columbus, OH 43210-1293, USA*
[2]*Department of Biological Sciences, University of Idaho, 875 Perimeter Drive MS 3051, Moscow, ID 83844-3051, USA*
[3]*Institute for Bioinformatics and Evolutionary Studies (IBEST), Biological Sciences, University of Idaho, 875 Perimeter Drive MS 3051, Moscow, ID 83844-3051, USA*

Received 3 April 2018; revised 3 May 2018; accepted for publication 3 May 2018

The Pacific Northwest of North America contains two disjunct temperate rainforests, one in the Coastal and Cascades Ranges and another in the Northern Rocky Mountains. These rainforests harbour > 200 disjunct and endemic taxa, with coastal and inland populations separated by the Columbia Basin. For several taxa, molecular data have revealed cryptic diversity structured across the Columbia Basin. Here, we use information from previously studied taxa and a machine-learning framework to predict that tail-dropper slugs in the genus *Prophysaon* (*Prophysaon andersoni*, *Prophysaon coeruleum*, *Prophysaon dubium* and the *Prophysaon vanattae*/*Prophysaon humile* complex) should lack cryptic diversity. This prediction is supported by results from species distribution models (SDMs), which suggest that all taxa lacked suitable habitat in the inland rainforests during the Last Glacial Maximum. We collected *COI* data and tested these predictions using approximate Bayesian computation and found that models of recent dispersal between inland and coastal populations received strong support. Finally, we used posterior predictive simulations to show that the best model was a reasonable fit to the data for all taxa. Our study highlights the utility of predictive modelling in a comparative phylogeographical framework and illustrates how posterior assessments of model fit can improve confidence in model-based phylogeographical analysis.

ADDITIONAL KEYWORDS: approximate Bayesian computation – cryptic diversity – phylogeography – posterior predictive simulations.

INTRODUCTION

The Pacific Northwest of North America (PNW) supports two disjunct temperate rain forests, namely the Cascades and Coastal ranges in the west and the Northern Rocky Mountains in the east (Fig. 1). The inland and coastal rainforests were continuous before the orogeny of the Cascades range (2–5 Mya; Graham, 1999), when the elevation of the Cascades produced a rain shadow that led to the xerification of the Columbia Basin. This basin is now characterized by a shrub-steppe ecosystem and effectively separates inland and coastal rainforests by > 200 km of habitat that is unsuitable for rainforest endemic species. The isolation of these two rainforests is somewhat diminished to the south by the Central Oregon highlands and by the Okanogan highlands in the north, but the basin has still acted as a barrier to dispersal for several rainforest endemics (Carstens *et al.*, 2005). In addition to the orogeny of the Cascades in the Pliocene, Pleistocene climatic fluctuations have influenced the distributions of rainforest endemics in the PNW (Pielou, 2008). Specifically, glaciers intermittently covered large parts of species' contemporary ranges, and rainforest endemics may have been eliminated from northern parts of their ranges completely or may have survived in isolated refugia (Brunsfeld & Sullivan, 2005).

*Corresponding author. E-mail: megansmth67@gmail.com

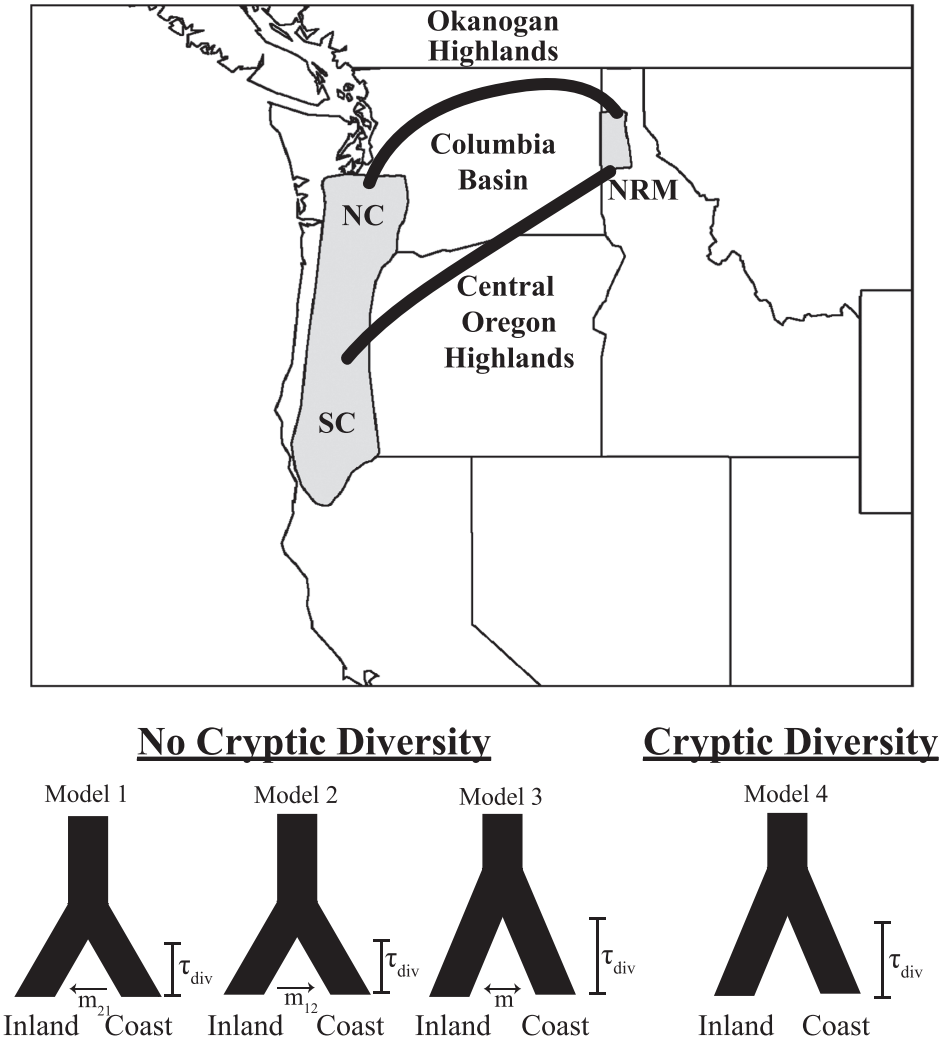


Figure 1. Phylogeographical models compared in the present study. Abbreviations: m_{12}, migration from population 1 into population 2; NC, North Cascades; NRM, Northern Rocky Mountains; SC, South Cascades; τ_{div}, divergence time. The grey areas on the map mark the distribution of *Prophysaon coeruleum*, adapted from Burke (2013). Black lines indicate northern and southern dispersal routes.

Several species with this disjunct distribution also have populations in the Blue and Wallowa Mountains of southeastern Washington and northeastern Oregon. Some studies have found that these populations are closely related to populations in the Northern Rockies (Nielson, Lohman & Sullivan, 2001), whereas in other species, such as the polydesmid millipede *Chonaphe armata* (Harger, 1872; Espíndola *et al.*, 2016) and *Prophysaon vanattae* (Pilsbry, 1948), populations in the Blue or Wallowa Mountains are phenotypically more similar to populations in the Cascades.

This compelling geological history has inspired a great deal of phylogeographical work, with several hypotheses proposed to explain the disjunct distributions of mesic forest endemics (reviewed by Brunsfeld *et al.*, 2001). The 'ancient vicariance' hypothesis posits that populations survived in both inland and coastal rainforests throughout the Pleistocene climatic fluctuations and that no gene flow has occurred between inland and coastal populations since the Pliocene. A variation on this hypothesis considered by Espíndola *et al.* (2016) posits pre-Pleistocene divergence between inland and coastal populations but allows for subsequent intermittent gene flow between inland and coastal populations along ephemeral habitat corridors at the margins of retreating glaciers. Another class of hypotheses reviewed by Brunsfeld *et al.* (2001), referred to as 'recent dispersal' hypotheses, posits that either inland or on the coast, no populations survived Pleistocene

climate fluctuations. These models predict post-Pleistocene divergence with subsequent dispersal either from coastal to inland or from inland to coastal populations. These models have received mixed support; data from several amphibian species suggest Pliocene divergence between inland and coastal populations and support the ancient vicariance model (Nielson *et al.*, 2001; Carstens *et al.*, 2004; Steele *et al.*, 2005), whereas data from dusky willows and robust lancetooth snails support a model of recent dispersal from coastal to inland rainforests (Carstens *et al.*, 2013; Smith *et al.*, 2017), and data from red alder support a more nuanced model of ancient vicariance, with intermittent gene flow in one or both directions (Ruffley *et al.*, 2018).

The ancient vicariance hypothesis predicts the presence of cryptic diversity structured across the Columbia Basin, owing to deep divergence between inland and coastal populations, and some species (e.g. the tailed frog, *Ascaphus montanus*; Nielson *et al.*, 2001) have been recognized after genetic data were collected to test this hypothesis. Recently, Espíndola *et al.* (2016) developed an approach to predict the presence or absence of such cryptic diversity that uses a machine-learning algorithm and genetic, taxonomic and environmental data from previously studied taxa to construct a classifier that attempts to predict the presence or absence of cryptic diversity in unsampled species. This method allows researchers to make predictions about which unstudied taxa are likely to harbour cryptic diversity, and may prove useful when limited resources force researchers to focus on only one or a few taxa. As part of the ongoing evaluation of this predictive framework for phylogeography, we apply random forest (RF) classification to terrestrial slugs endemic to the PNW and test the resulting predictions using species distribution models (SDMs) and genetic data.

MATERIAL AND METHODS

Study system and sampling

Slugs of the genus *Prophysaon* are endemic to the mesic forests of the PNW, with several species exhibiting the characteristic mesic forest disjunct distribution on either side of the Columbia Basin. Here, we focus on three disjunct species: *Prophysaon andersoni* (J.G. Cooper, 1872), *Prophysaon coeruleum* (Cockerell, 1890) and *Prophysaon dubium* (Cockerell, 1890). Additionally, we include two sister species: *Prophysaon vanattae* (Pilsbry, 1948) and *Prophysaon humile* (Cockerell, 1948). *Prophysaon vanattae* occurs only in the Cascades, whereas *P. humile* is endemic to the inland rainforest. *Prophysaon vanattae* and *P. humile* were classified as belonging to the subgenus *Mimetarion* (Pilsbry, 1948), along with *Prophysaon obscurum* (Cockerell, 1893) and *Prophysaon fasciatum* (Cockerell, 1890). *Prophysaon fasciatum* does not seem to be a distinct species (Pilsbry & Vanatta, 1898), and *P. obscurum* (Cockerell, 1893) is a narrow endemic found only south of the South Puget Sound and into western Washington and in the Columbia Gorge along the Washington–Oregon border (Burke, 2013). Molecular data indicate that *P. obscurum* is not distinct from *P. vanattae* (see Supporting Information: Gene Tree of this article; Wilke & Duncan, 2004), suggesting that it is appropriate to consider *P. vanattae* and *P. humile* as sister taxa that diverged across the Columbia Basin. Although this may seem to suggest deep divergence between the inland and coastal sister taxa, we chose to include these taxa in the present study because there has been no study of the genetic divergence between these two taxa, and it is unknown whether the phenotypic divergence used to delineate the two species corresponds to deep genetic divergence. We collected 223 samples from throughout the ranges of the focal taxa and other species of *Prophysaon* [*Prophysaon foliolatum* (Gould, 1851) and *P. obscurum*] from field collections and from museums (Fig. 2; Supporting Information, Table S1). Sites were visited during the autumn of 2016, and slugs were stored in 95% ethanol immediately after collection. Additional samples were requested from museum collections (Carnegie Museum of Natural History, Royal British Columbia Museum and the California Academy of Sciences).

Predicting cryptic diversity

We applied the approach developed by Espíndola *et al.* (2016) to predict whether the focal taxa harbour cryptic diversity structured across the Columbia Basin. Specifically, we trained an RF classification function on a set of taxa that had already been studied [the water vole *Microtus richardsoni* (J.E. Dekay, 1842), the tree *Salix melanopsis* (Nutt), the millipede *C. armata*, the frog *Ascaphus montanus*/*truei*, the salamander *Dicamptodon atterimus*/*copei* (Cope, 1867; Nussbaum, 1970) and the salamander *Plethodon idahoensis*/*vandykei* (Slater & Slipp, 1940; Van Denburgh, 1906)]. These reference taxa were classified as 'cryptic' or 'non-cryptic' based on genetic data provided by Espíndola *et al.* (2016), where 'cryptic' species had cryptic diversity structured across the Columbia Basin, and non-cryptic species did not. In addition to the data used by Espíndola *et al.* (2016) to train their predictive function, we included data from *Haplotrema vancouverense* (I. Lea, 1839), a snail endemic to the PNW for which genomic data have demonstrated a lack of cryptic diversity structured

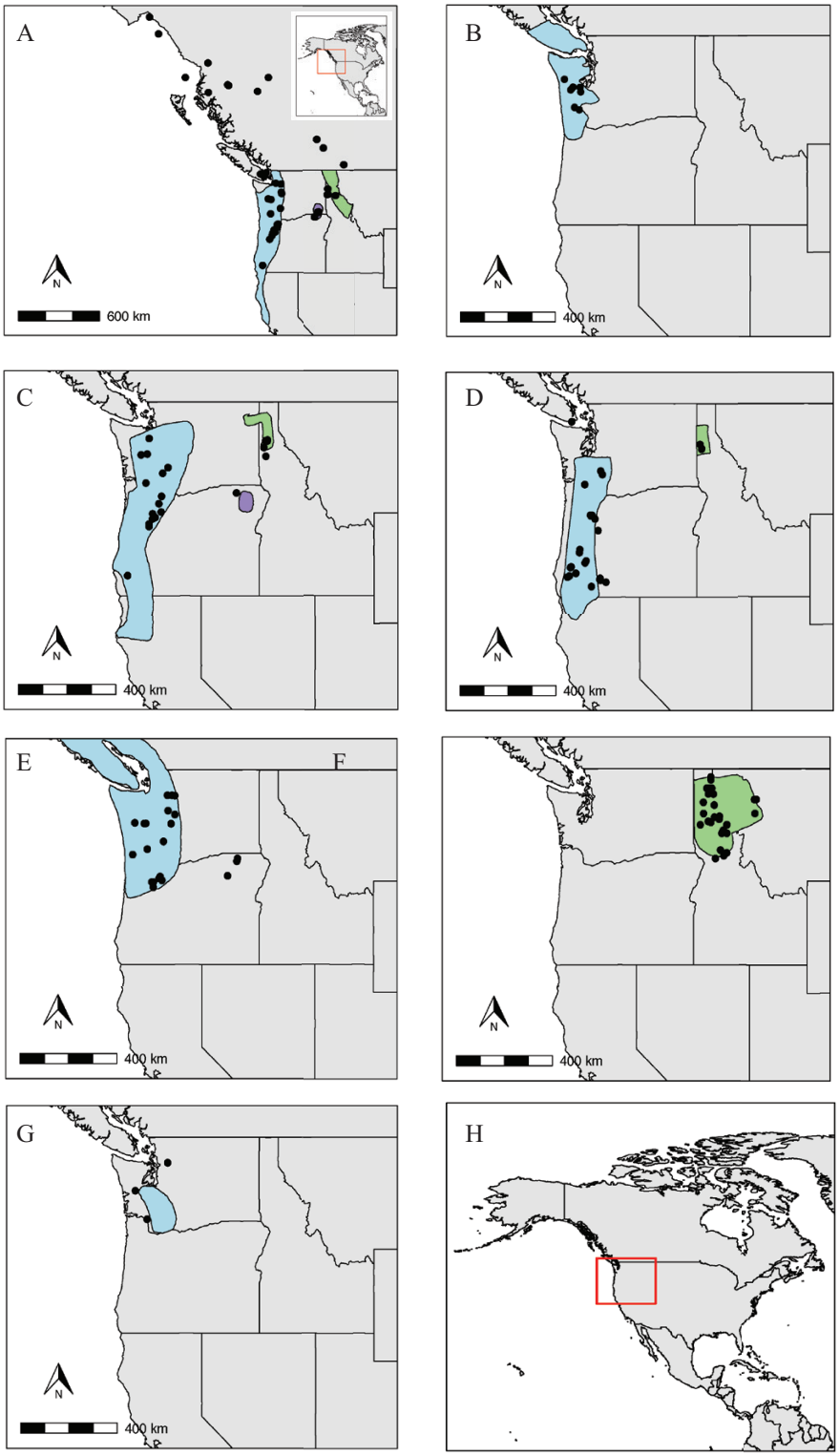


Figure 2. The distributions of all *Prophysaon* species considered in the present study, adapted from Burke (2013), with collection localities from the present study. Coastal ranges are shown in blue, inland ranges in green, and Blue and Wallowa ranges in purple. A, *Prophysaon andersoni*. B, *Prophysaon foliolatum*. C, *Prophysaon dubium*. D, *Prophysaon coeruleum*. E, *Prophysaon vanattae*. F, *Prophysaon humile*. G, *Prophysaon obscurum*. H, map showing context for maps B–G.

across the Columbia Basin (Smith *et al.*, 2017). We included this taxon because it is more closely related to *Prophysaon* than other species in the training set, and thus should improve the performance of the predictive function for the focal taxa. The occurrence data detailed by Espíndola *et al.* (2016) were also used here to extract eight environmental variables from the WorldClim database (Hijmans *et al.*, 2005): annual mean temperature, mean diurnal range, isothermality, maximum temperature of warmest month, temperature annual range, annual precipitation, precipitation seasonality and precipitation of driest quarter. In addition to environmental data, we used taxonomic information to train the classifier. Specifically, we classified each taxon as mollusc, arthropod, mammal, amphibian or plant. These broad taxonomic categories are intended to serve as a proxy for life-history characteristics that may influence traits such as dispersal ability. The dataset from Espíndola *et al.* (2016) combined with the data from *H. vancouverense* was used to train an RF classifier using the R package 'randomForest' (Liaw & Wiener, 2002). We used the same down-sampling strategy described by Espíndola *et al.* (2016) to account for differences in the number of cryptic and non-cryptic observations with 100 down-sampled replicates. We assessed the accuracy of these classifiers using cross-validation, where one taxon was omitted when building the RF classifier. The classifier was then applied to the omitted taxon, and the probability that the omitted taxon harboured or lacked cryptic diversity was calculated. We then applied the classifier to the four *Prophysaon* taxa with disjunct distributions and estimated the probability of each taxon harbouring or lacking cryptic diversity. Occurrence points for *Prophysaon* were from this study only, as the identification of occurrence data from data aggregators (e.g. the Global Biodiversity Information Facility, GBIF) could not be verified, and misidentifications are common in this group. Climate data were downloaded from WorldClim for these occurrence points for the eight bioclimatic variables listed above at a resolution of 30 arc s (Hijmans *et al.*, 2005). To evaluate the importance of the taxonomic predictors in driving the classifier, we also constructed and applied a classifier using no taxonomic variables and using a different set of classifications (Supporting Information: Random Forest Predictions) to evaluate the effects of including taxonomy in the classifier.

Species distribution models

In addition to the RF classifier, SDMs may help to predict whether taxa will harbour cryptic diversity structured across the Columbia Basin. Specifically in the PNW, where the presence or absence of cryptic diversity is largely driven by the persistence of habitat through the Pleistocene glacial cycles, hindcast SDMs can be a powerful tool for predicting the presence of cryptic diversity (Richards, Carstens & Knowles, 2007). If no suitable habitat is predicted in the Northern Rocky Mountains during the LGM (assuming accurate SDMs and palaeoclimate reconstructions), the focal taxa are likely to have colonized the inland after glaciation, and we would not expect cryptic diversity across the Columbia Basin. On the contrary, if suitable habitat persisted in the inland rainforests during the LGM, there may have been refugia present at that time, and deep genetic divergence might exist between inland and coastal populations.

We built SDMs using the ensemble method implemented in the R package 'biomod2' (Thuiller, Georges & Engler, 2014) and the ten modelling approaches available in the package (Supporting Information: Species Distribution Modeling-Ensemble Approach). For the SDMs, we used only samples collected from the present study, as misidentifications are common in this group. We removed museum specimens with incorrect registers (e.g. registers that fell in the Pacific Ocean) and duplicate samples. This data curation resulted in 42 occurrence points for *P. andersoni*, 29 for *P. coeruleum*, 23 for *P. dubium*, and 52 for the *P. vanattae*/*P. humile* sister species pair (Supporting Information, Table S2). Climate data were downloaded from the WorldClim database (Hijmans *et al.*, 2005). Current climate data were at a resolution of 30 arc s, and LGM data were at a resolution of 2.5′. We selected uncorrelated bioclimatic variables for each species ($r < 0.7$) and chose among highly correlated variables by prioritizing variables that we thought would be most important for the focal taxa. The selected variables are reported in the Supporting Information (Table S3). We then cropped these layers to the extent of the focal region, which was defined as −150 to −100° longitude and 35–65° latitude. We chose an extent broader than the range of the focal taxa because our aim was to hindcast the SDMs. Given that suitable habitat in the past might have occurred outside current ranges, we included areas not currently occupied by the focal taxa, which is the usual approach in phylogeographical hindcasting studies (e.g. Espíndola *et al.*, 2012; Gavin *et al.*, 2014).

We used several modelling methods (Supporting Information: Species Distribution Modeling-Ensemble Approach) and ran five replicates per model. We randomly sampled 10 000 pseudoabsences from the entire background area (defined by the extent of the focal region) using the 'random' strategy available in BioMod. We chose this strategy because results have been equivocal on which sampling strategy should be preferred, with performance varying greatly across

datasets and methods, but with random sampling tending to perform best across regression methods, and with a relatively small decrease in performance with random sampling compared with other sampling methods in classification and machine-learning techniques (Barbet-Massin *et al.*, 2012).

To build models in Maxent, we used a maximum of 1000 iterations to reach convergence and included linear, quadratic, product, threshold and hinge features. For parameters not specified above, we used the default BioMod parameterization. We used 80% of our data for training models and 20% for testing, and five replicates for each model to evaluate model performance. We performed three replicates to determine variable importance and evaluated models using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Models were rescaled, so that they could be combined later. We then built an ensemble model ignoring models with an ROC of < 0.85 and weighting all other models by ROC score. Finally, we forecast the ensemble model onto current and past (LGM) climate conditions. To evaluate the impacts of modelling choices on our SDM results, we also constructed SDMs using an alternative strategy and found that our results did not change substantially (Supporting Information: Species Distribution Modeling using Ecoregions and Maxent).

DNA ISOLATION, SEQUENCING AND ANALYSES

DNA was extracted using DNeasy Blood and Tissue kits (Qiagen), following the manufacturer's standard protocol. A 710 bp portion of the *COI* gene was amplified using the primer pair LCO1490 and HCO2198 (described by Folmer *et al.*, 1994). Forward and reverse reads were assembled in Geneious v6.1.7 (Kearse *et al.*, 2012) and edited by eye when necessary. All available sequences (46 additional sequences) for *Prophysaon* were downloaded from GenBank in August 2017. The *muscle* (Edgar, 2004) algorithm available in Geneious v.6.1.7 (Kearse *et al.*, 2012) was used to generate an alignment.

We used the AutoModel function in PAUP* v4.0a157 (Swofford, 2002) to select the best model of sequence evolution for the entire dataset. To select the best model of nucleotide substitution for our approximate Bayesian computation (ABC) analysis (see next subsection), we used the AutoModel function separately for datasets from each of the four disjunct taxa, excluding samples from the Blue and Wallowa Mountains (see next subsection). As a starting tree, we used a neighbor joining tree calculated from Jukes Cantor distances. We used the largest model set (11 substitution schemes), and considered models with and without gamma-distributed rate variation across sites and a proportion of invariable sites. We evaluated models using the small-sample-size corrected version of the Akaike information criterion (AICc), Bayesian information criterion (BIC) and decision theory (DT; Minin *et al.*, 2003).

To obtain posterior distributions of the parameters for the model of sequence evolution for use in the ABC analyses (see next subsection), we used MrBayes v.3.2.6 (Ronquist *et al.*, 2012). For each of the four disjunct taxa, MrBayes was run under the model of sequence evolution selected by PAUP if the model could be implemented. If the model selected using the methods above could not be implemented in MrBayes, we reran the AutoModel function in Paup* v4.0a157 (Swofford, 2002) using the reduced set of models (three substitution schemes). We conducted 400 independent runs, with 32 000 000 generations and four chains per run. For models that included gamma rate variation across sites, we used eight rate categories. We adjusted the temperature parameter to improve mixing, such that acceptance rates of initial runs were between 10 and 70%. We discarded 25% of runs as burn-in and combined the 100 independent runs for each of the focal taxa; the resulting posterior distributions of parameters were used in downstream analyses. We checked that the average deviation of split frequencies was < 0.01 for each run to assess convergence. Additionally, we estimated a maximum likelihood gene tree in GARLI v. 2.01 (Bazinet, Zwickl & Cummings, 2014), with the slugs *Hemphillia malonei* and *Zacoleus idahoensis* as outgroups (Supporting Information: Gene Tree).

DEMOGRAPHIC MODEL SELECTION USING APPROXIMATE BAYESIAN COMPUTATION

To test whether the focal taxa harboured cryptic diversity structured across the Columbia Basin, we first evaluated recent dispersal models (Fig. 1). The first two models consisted of post-Pleistocene divergence with dispersal and subsequent gene flow either from coastal to inland rainforests or from inland to coastal rainforests. These models correspond to a lack of suitable habitat in either the inland or the coastal rainforests during the Pleistocene glacial cycles and subsequent post-Pleistocene colonization. The third recent dispersal model included pre-Pleistocene divergence with subsequent gene flow in both directions, approximating a scenario in which there was ancient vicariance followed by secondary contact (Ruffley *et al.*, 2018). After comparing the recent dispersal models, we calculated the posterior probability of the best dispersal model and a model of pre-Pleistocene divergence with no subsequent gene flow (ancient vicariance). The models tested were parameterized such that population sizes could vary for each of the two populations. All divergence time and

population size parameters were drawn from uniform priors, which were adjusted after initial runs to ensure that simulated summary statistics were in the range of observed summary statistics for each taxon and region. Full information on the priors for each parameter and each taxon is available in the Supporting Information (Table S4). For the purpose of the ABC analysis, we did not consider the Blue and Wallowa populations, because we had too few samples from these regions to model these populations separately.

To compare these models within each of the four focal taxa, we simulated 100 000 datasets under each of the three migration models in ms (Hudson, 2002). We then used the program Seq-Gen v.1.3.4 (Rambaut & Grassly, 1997) to simulate sequence data from the gene trees simulated in ms. We simulated 574 bp, to match the observed data. Parameters for the model of sequence evolution were drawn from the posterior distribution of parameters from the MrBayes analyses (see previous subsection). We drew the scaling parameter from a uniform prior distribution (Supporting Information, Table S4). We then used a custom python script (available at https://github.com/meganlsmith) to calculate six summary statistics: π; the number of segregating sites (S); Watterson's θ (θ_W); π within each population; and the number of nucleotide differences between populations (*nucdiv*). We calculated π as $\pi = 2^{*}\sum_{i=2}^{N}\sum_{j=1}^{i-1} x_i x_j \pi_{ij}$, where x_i and x_j are the frequencies of the ith and jth sequences, π_{ij} is the number of nucleotide differences per nucleotide site between sequences i and j, and N is the number of unique sequences in the sample (Nei, 1979). Watterson's θ was calculated as the number of segregating sites divided by the $(N-1)$th harmonic number (Watterson, 1975). The number of nucleotide differences between subpopulations was simply the sum of the number of nucleotide differences between sequences in each of the two populations. We calculated the same summary statistics from the observed data using a python script.

We evaluated a range of rejection methods (simple rejection and logistic regression), tolerances (0.001, 0.005, 0.01 and 0.05) and summary statistic combinations using a custom R script and the R package 'abc' (Csilléry, François & Blum, 2012). We performed ten cross-validation replicates for each combination of rejection method, tolerance and summary statistic combination. We did not use more replicates owing to the large number of combinations being evaluated and the computational requirements of each replicate. Each method was evaluated based on the mean posterior probability of the correct model across the three migration models. However, using the logistic rejection method resulted in errors with many simulated and empirical datasets because there was too little variation in summary statistics in the selected region. Therefore, we considered only rejection methods for downstream analyses. The best three methods (based on the mean posterior probability of the best model across all three models) were evaluated further using 100 cross-validation replicates. When more than three methods had equivalent posterior probabilities across all models being compared, three methods were chosen at random, and 100 cross-validation replicates were run for these three methods. The best of these three methods was then selected based on the sum of the posterior probability of the correct model across all three models. We then used the winning method to calculate the posterior probabilities of each model for each taxon.

In a second step, we compared the recent dispersal model that had the highest posterior probability with the ancient vicariance model. The methods were the same as those in the first step of the ABC analysis. Cross-validation analyses were conducted in the same way as above, and the best model was selected based on the winning method.

Finally, we used posterior predictive distributions to assess the fit of the best model to the data directly. To generate the posterior predictive distribution, we simulated 100 datasets under each set of parameters from the posterior distribution of parameters under the best model. For each summary statistic used, we calculated the difference between the posterior predictive distribution of the summary statistic and the observed statistic, and evaluated whether the 95% highest density interval (HDI) included zero using the R package 'HDInterval' (Meredith & Kruschke, 2017). A 95% HDI that does not include zero indicates that the model is not a good fit to the data. In addition to the ABC analysis described here, we conducted an ABC analysis where we simulated only infinite sites data, rather than sequence data (Supporting Information: ABC with Infinite Sites Data).

RESULTS

Study system and sampling

We collected samples or downloaded data from seven described species of *Prophysaon*: *P. andersoni*, *P. coeruleum*, *P. dubium, P. foliolatum*, *P. humile*, *P. obscurum* and *P. vanattae* (Fig. 2; Supporting Information, Table S1). For *P. andersoni*, we collected or downloaded sequence data from 11 inland samples, 66 coastal samples, and four samples from the Blue and Wallowa mountains. For *P. coeruleum*, we collected or downloaded data from eight inland and 39 coastal

samples. For *P. dubium*, we collected or downloaded data from eight inland samples, 27 coastal samples, and one sample from the Blue and Wallowa mountains. For *P. vanattae* and *P. humile*, we collected or downloaded data from 38 inland samples, 33 coastal samples, and five samples from the Blue and Wallowa mountains.

Predicting cryptic diversity

The cross-validation analysis indicated that the RF classifier performed well for most species. For all species except *C. armata*, the mean posterior probability of the correct classification was > 0.90 (Fig. 3), whereas for *C. armata*, the mean posterior probability of a cryptic classification was 0.0125. This species was also identified as problematic by Espíndola *et al.* (2016), and these issues were attributed to difficulty in classifying *C. armata* as cryptic or non-cryptic using molecular data. All *Prophysaon* species groups were predicted to lack cryptic diversity, with a posterior probability > 0.98 (Fig. 3). However, when taxonomy was omitted or reclassified, these predictions changed. Specifically, without taxonomy, all focal taxa were predicted to harbour cryptic diversity, but with a low probability (Supporting Information, Fig. S1A), and when taxonomy was reclassified as vertebrate, invertebrate or plant, all taxa were predicted to harbour cryptic diversity with a high probability (Supporting Information, Fig. S1B).

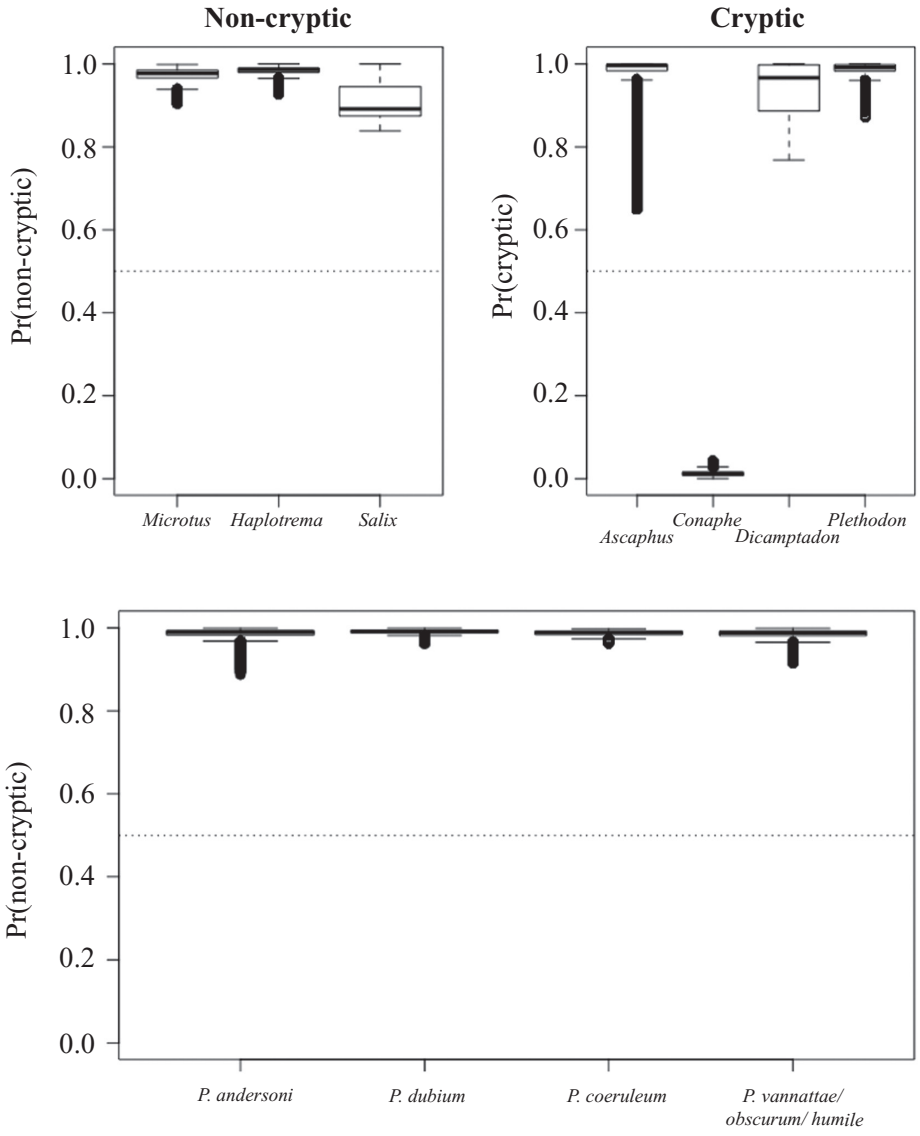


Figure 3. Results from the random forest analysis. Top: results from the cross-validation analysis. Bottom: classification results for *Prophysaon*.

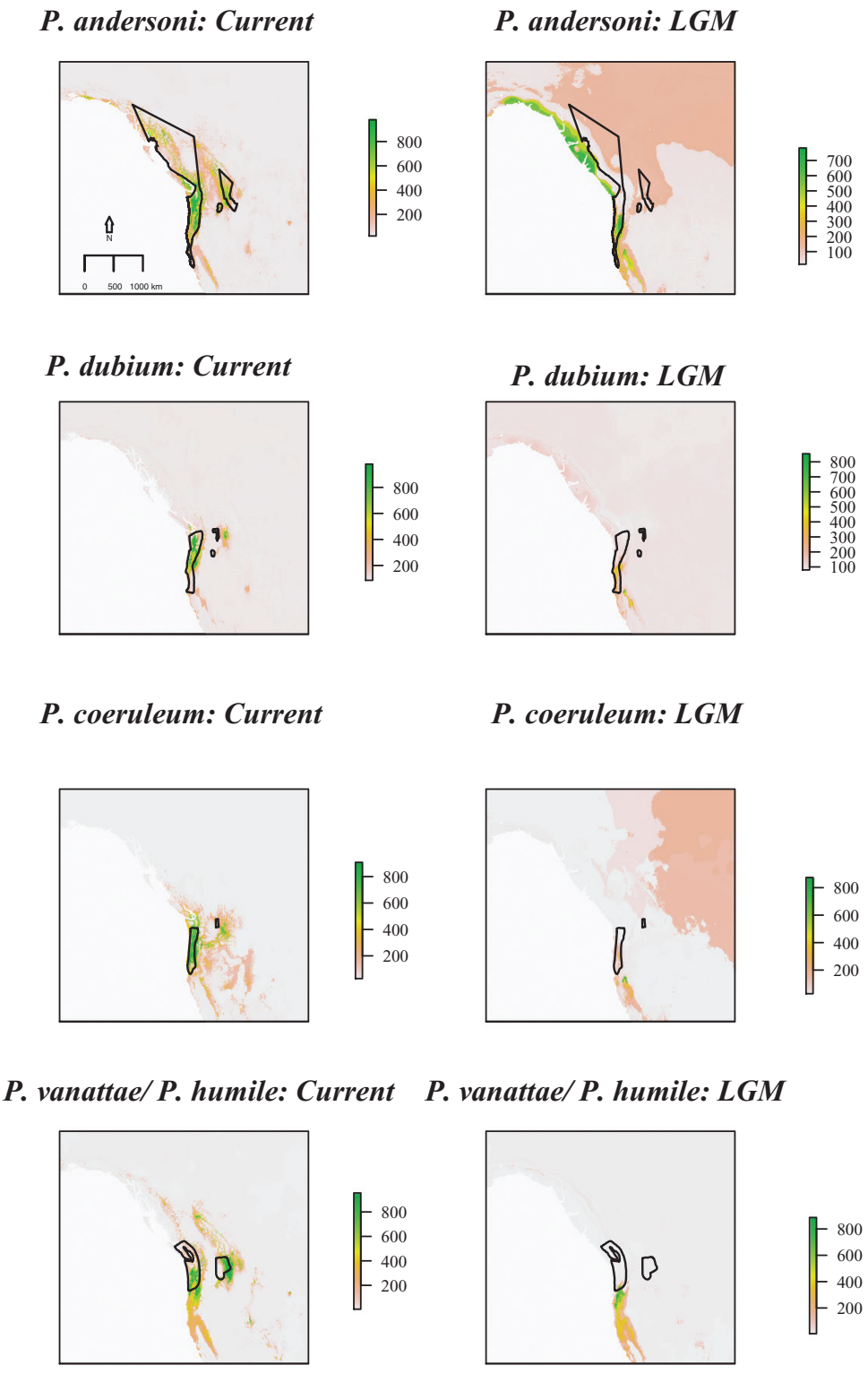


Figure 4. Species distribution models for the present and the Last Glacial Maximum (LGM). A, *Prophysaon andersoni*, current. B, *Prophysaon coeruleum*, current. C, *Prophysaon dubium*, current. D, *Prophysaon vanattae*/*Prophysaon humile*, current. E, *P. andersoni*, LGM. F, *P. coeruleum*, LGM. G, *P. dubium*, LGM. H, *P. vanattae*/*P.humile*, LGM. Current ranges, following Burke (2013), are outlined in black.

Species distribution models

The mean ROC AUC scores with and without models with an AUC < 0.85, respectively, were: 0.910 and 0.948 for *P. andersoni*, 0.925 and 0.975 for *P. dubium*, 0.923 and 0.963 for *P. coeruleum*, and 0.969 and 0.972 for *P. vanattae*/*P. humile*. Current distribution models indicated suitable habitat inland and on the coast (Fig. 4) and recovered the known range of the different taxa (Fig. 4). When SDMs were hindcast to conditions at the LGM, suitability values were extremely low in the inland portions of the species ranges (Fig. 4).

DNA isolation and sequencing

The final alignment included 574 bases of the mitochondrial gene *COI*, with no stop codons in the reading frame. There was an AT bias in base composition (A, 0.321; C, 0.112; G, 0.108; T, 0.459), comparable to that found in other invertebrate mitochondrial genes (Lin & Danforth, 2004; Liu *et al.*, 2012; Hilgers *et al.*, 2016). Maximum within-species maximum likelihood distances were as follows: *P. andersoni*, 0.447; *P. foliolatum*, 0.153; *P. dubium*, 0.147; *P. coeruleum*, 1.080; *P. vanattae*, 0.654; *P. humile*, 0.117; and *P. obscurum*, 0.083. The results of model selection in PAUP for the full dataset were concordant across AICc, BIC and DT methods (TPM2uf + I + Γ), and all analyses on the full dataset were conducted using this model. We also used PAUP to determine the best model for each disjunct taxon. For *P. andersoni*, *P. coeruleum* and *P. vanattae*/*P. humile*, the results supported an HKY + I + Γ model regardless of the criteria (i.e. AICc, BIC, DT), whereas a TPM3uf + I model was supported for *P. dubium*. However, as this model cannot be implemented in MrBayes, we reran model selection for *P. dubium* using a reduced model set that considered only the three substitution schemes that can be implemented in MrBayes and chose an HKY + I model. Parameter estimates from PAUP are reported in the Supporting Information (Table S5). For the MrBayes analyses there was no evidence of a lack of convergence, and all runs had a standard deviation of split frequencies < 0.01.

Demographic model selection using ABC

Recent dispersal models

The summary statistics for the observed data are reported in Table 1. Cross-validation analyses indicated moderate to low ability to distinguish among the three dispersal models. Across all species, the posterior probability of the correct model ranged from 0.361 to 0.763 in the cross-validation analysis (Table 2). For *P. andersoni*, *P. dubium* and the *P. vanattae*/*P. humile* sister species pair, model 1 had the highest posterior probability. For *P. coeruleum*, model 3 had the highest posterior probability (Table 3).

Recent dispersal vs. ancient vicariance

Cross-validation analyses indicated high power to distinguish between the recent dispersal and ancient vicariance models; the posterior probability of the correct model ranged from 0.973 to 0.999 (Table 4). For all species, the recent dispersal model received the highest posterior probability, and the posterior probability of the recent dispersal model was always one (Table 5). Across all species, the 95% highest density interval of the difference between the posterior predictive distribution and the observed data contained zero, meaning there was no indication of poor model fit. Plots of the posterior and posterior predictive distributions are available in the Supporting Information (Fig. S2). The results did not differ substantially when only infinite sites data, rather than sequence data, were simulated (Supporting Information: ABC with Infinite Sites Data).

DISCUSSION

Predicting cryptic diversity

The predictive model correctly predicted that all disjunct taxa included in the present study lacked cryptic diversity structured across the Columbia Basin. However, this finding might be driven by the relative lack of taxonomic breadth of the groups included in our model, because the only other mollusc included was *H. vancouverense*, which also lacks cryptic diversity. To gain a better understanding of the role of taxonomy in our predictions, we omitted taxonomy and found that we predicted that all *Prophysaon* species would harbour cryptic diversity with low to moderate posterior probabilities. We also ran the predictive model with coarsened taxonomic classification by recognizing only vertebrates, invertebrates and plants. In this case, there were two invertebrates in the model, the land snail *H. vancouverense* and the millipede *C. armata*, and *C. armata* was classified as cryptic. This again resulted in predictions that all species of *Prophysaon* would harbour cryptic diversity with high posterior probabilities. These results indicate that our RF predictions might be overly reliant on taxonomy, most probably because taxonomy serves as a proxy for important traits, including dispersal ability, that are

Table 1. Summary statistics calculated for the observed data for use in the approximate Bayesian computation (ABC) analyses

Species	π	S	Tajima's D	θ_H	H	π_{inland}	π_{coast}	*nucdiv*	θ_W
Prophysaon andersoni	0.0475	144	−0.452	10.0	0.772	0.00824	0.0509	23.2	29.3
Prophysaon coeruleum	0.0950	180	0.233	18.6	0.639	0	0.09819	55.1	40.8
Prophysaon dubium	0.0503	90	0.832	16.5	0.564	0.00859	0.0510	31.3	21.9
Prophysaon vanattae / *Prophysaon humile*	0.0860	172	−0.0692	15.5	0.437	0.04910	0.0801	62.7	35.6

Fay and Wu's H (H), Fay's θ_H statistic (θ_H), nucleotide divergence between inland and coastal populations (*nucdiv*), nucleotide diversity (ϖ), nucleotide diversity within the coastal population (π_{coast}), nucleotide diversity within the inland population (π_{inland}), number of segregating sites (S), and Watterson's theta (θ_W).

Table 2. Cross-validation results for the three recent dispersal models

Species	Method	Tolerance	Sumstats	pp(M1)	pp(M2)	pp(M3)
Prophysaon andersoni	Rejection	0.001	π (inland), π (coast), π_{12}	0.468	0.520	0.385
Prophysaon dubium	Rejection	0.001	π, S, π (inland), π (coast), π_{12}	0.465	0.464	0.361
Prophysaon vanattae / *Prophysaon humile*	Rejection	0.001	π, π (inland), π (coast), $\pi_{12,}$ θ_W	0.456	0.456	0.414
Prophysaon coeruleum	Rejection	0.001	π, S, π (inland), π_{12}	0.530	0.763	0.493

Models correspond to those shown in Figure 1, and the posterior probability (pp) of a model (M) is the mean posterior probability across the cross-validation replicates in which the data were simulated under that model.

Table 3. Approximate Bayesian computation results for the three recent dispersal models

Species	M1	M2	M3	BF
Prophysaon andersoni	0.697	0.034	0.269	2.60
Prophysaon dubium	0.540	0.173	0.287	1.88
Prophysaon vanattae / *Prophysaon humile*	0.580	0.290	0.130	2.00
Prophysaon coeruleum	0.287	0.300	0.413	1.38

The models (M) correspond to those shown in Figure 1, and the posterior probabilities are those calculated using the best method selected by cross-validation. The Bayes factors (BF) shown are those comparing the model having the highest posterior probability with the model having the second highest posterior probability.

not quantified directly. Continuing to add more species to the dataset, as this work has done, and generating data on the variables we are attempting to summarize with taxonomy will improve future classifications.

Given that *P. vanattae* and *P. humile* are currently described as different species, our finding of recent dispersal between these two is surprising. Although researchers commonly discuss the lack of phenotypic divergence when genetic divergence is present (cryptic diversity), the inverse is also a common phenomenon. Even within *Prophysaon*, subspecies designations based on morphology have been questioned in the light of molecular data (Wilke & Duncan, 2004). Within *P. vanattae*, there exists extensive phenotypic variation (Burke, 2013), which might exceed that between *P. vanattae* and *P. humile*. Given that taxonomic classifications drive how we group organisms, and thus our results in phylogeographic studies, this work highlights the need to revisit these classifications in light of phylogeographical evidence (such as that presented here) and genomic data. It is possible that our inference is driven by incorrectly treating all *P. vanattae* as a single species or by incorrectly splitting *P. vanattae* and *P. humile*, and future work should test these hypotheses using increased sampling, in terms of both individuals and loci. Furthermore, support for *P. vanattae* and *P. humile* as sister species is limited, and the gene tree estimated here does not support this relationship. It is possible that these species are not sister species, and future work including more loci should evaluate this relationship and reconsider the results reported here for these species. Careful species delimitation using multiple lines of evidence (e.g. morphological, ecological and genomic data) will be necessary to characterize these taxa accurately.

Assessing model fit

Given that the use of ABC for phylogeographical inference can tell us only which of the tested

Table 4. Cross-validation results for the step comparing the best recent dispersal model with the ancient vicariance model

Species	Recent dispersal model	Method	Tolerance	Sumstats	pp(recent dispersal)	pp(ancient vicariance)
Prophysaon andersoni	1	Rejection	0.001	π, S, π (inland), π (coast), θ_W	0.977	0.986
Prophysaon dubium	1	Rejection	0.001	S, π_{12}	0.977	0.981
Prophysaon vanattae/*Prophysaon humile*	1	Rejection	0.001	π, π (inland), π_{12}, θ_W	0.973	0.995
Prophysaon coeruleum	3	Rejection	0.001	π, S, π (coast)	0.984	0.999

Models correspond to those shown in Figure 1, and the posterior probability (pp) of a model is the mean posterior probability of the model across the cross-validation replicates in which the data were simulated under that model.

Table 5. Approximate Bayesian computation results for step comparing the best recent dispersal (RD) model with the ancient vicariance (AV) model

Species	RD model	p(RD)	p(AV)	BF
Prophysaon andersoni	1	1	0	Inf
Prophysaon dubium	1	1	0	Inf
Prophysaon vanattae/*Prophysaon humile*	1	1	0	Inf
Prophysaon coeruleum	3	1	0	Inf

The models correspond to those shown in Figure 1, and the posterior probabilities are those calculated using the best method selected by cross-validation. The Bayes factors (BF) shown are those comparing the model having the highest posterior probability with the model having the second highest posterior probability. Inf = infinite.

models is the best fit to the data, and not how well the models fit the data, it is essential to accompany model selection with tests of model fit. This need is particularly pronounced when analysing relatively limited datasets, such as the single mitochondrial locus analysed here. We addressed this issue with posterior predictive simulation tests of model fit, and we found no indication that the best model was a poor fit to the data for any of the focal taxa. This increases our confidence in the model selection results, because it suggests that not only was the selected model a better fit to the data than other models in the model set, but also the selected model was a reasonable fit to the data. This diminishes our concerns about how our choice of models might have influenced the results of model selection.

Conclusions

The work presented here represents a substantial expansion of the predictive framework described by Espíndola *et al.* (2016). It highlights the utility of this framework, while suggesting areas where it could be improved. The framework appears to make accurate predictions in all species analysed here, but these predictions are driven largely by taxonomy. Perhaps the most promising aspect of the predictive framework for phylogeography is its capacity to integrate phylogeographical research conducted at different points in time and in different empirical systems. An attribute of this integration is that the framework should increase in its utility and accuracy as more diverse taxa are incorporated. Our study has nearly doubled the number of species complexes (from seven to 12) that can be included in the predictive framework for the PNW temperate rainforest, and should improve the accuracy of this framework for future research via both the expansion of the taxon set and the careful exploration of the ABC methodology.

Indeed, such an iterative approach has been central to taxonomic research since its inception; hypotheses are formed, tested and then modified. The predictive framework developed by Espíndola *et al.* (2016), and used and expanded here, allows phylogeographical research to proceed in the same way. Information about previously studied species is used to construct hypotheses about what factors affect the phylogeographical histories of different taxa. In the present study these factors were environmental and taxonomic, but they could also include trait and life-history data. These hypotheses are then used to generate predictions about unstudied taxa. After these predictions are tested, we can modify our hypotheses by considering environmental, ecological

and other traits of the newly added taxa. This leads to a cyclical approach, in which we constantly modify and improve our hypotheses about how environmental and intrinsic factors drive species' responses to climatic and geological events, and provides a novel means to integrate phylogeographical datasets. This is no trivial accomplishment, given the scale of phylogeographical datasets that have been collected to date: a Web of Science search of the term phylogeograph* returned 15 911 hits (7 December 2017). Integrating these data into a common framework that allows researchers to develop and test hypotheses on a large scale has been a difficult task, and the predictive framework used here is a step towards that goal.

ACKNOWLEDGEMENTS

We would like to acknowledge Bill Leonard, Ben Stone and Tara Pelletier for assistance in the field. Funding was provided by the US National Science Foundation (DEB 1457726/14575199). M.L.S. was supported by an National Science Foundation Graduate Research Fellowship (DG-1343012). We thank the Royal British Columbia Museum, the Carnegie Museum and the California Academy of Sciences for providing samples. We also thank Michael Lucid of Idaho Fish and Game for donations of samples and the Ohio Supercomputer Center for computing resources (allocation grant PAS1181-1).

REFERENCES

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. 2012. Modelling species distributions to map the road towards carnivore conservation in the tropics. *Methods in Ecology and Evolution* **3:** 327–338.

Bazinet AL, Zwickl DJ, Cummings MP. 2014. A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0. *Systematic Biology* **63:** 812–818.

Brunsfeld SJ, Sullivan J. 2005. A multi-compartmented glacial refugium in the northern Rocky Mountains: evidence from the phylogeography of *Cardamine constancei* (Brassicaceae). *Conservation Genetics* **6:** 895–904.

Brunsfeld SJ, Sullivan J, Soltis DE, Soltis PS. 2001. Chapter 15 Comparative phylogeography of north-western North America: a synthesis. Special Publication-British Ecological Society **14**: 319–340.

Burke TE. 2013. *Land snails and slugs of the Pacific Northwest*. Corvallis, OR: Oregon State University Press.

Carstens BC, Brennan RS, Chua V, Duffie CV, Harvey MG, Koch RA, McMahan CD, Nelson BJ, Newman CE, Satler JD, Seeholzer G, Posbic K, Tank DC, Sullivan J. 2013. Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*. *Molecular Ecology* **22:** 4014–4028.

Carstens BC, Brunsfeld SJ, Demboski JR, Good JM, 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, Stevenson AL, Degenhardt JD, Sullivan J. 2004. Testing nested phylogenetic and phylogeographic hypotheses in the *Plethodon vandykei* species group. *Systematic Biology* **53:** 781–792.

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

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5:** 113.

Espíndola A, Pellissier L, Maiorano L, Hordijk W, Guisan A, Alvarez N. 2012. Predicting present and future intraspecific genetic structure through niche hindcasting across 24 millennia. *Ecology Letters* **15:** 649–657.

Espíndola A, Ruffley M, Smith ML, Carstens BC, Tank DC, Sullivan J. 2016. Identifying cryptic diversity with predictive phylogeography. *Proceedings of the Royal Society B: Biological Sciences* **283:** 20161529.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* **3:** 294–299.

Gavin DG, Fitzpatrick MC, Gugger PF, Heath KD, Rodríguez-Sánchez F, Dobrowski SZ, Hampe A, Hu FS, Ashcroft MB, Bartlein PJ, Blois JL, Carstens BC, Davis EB, de Lafontaine G, Edwards ME, Fernandez M, Henne PD, Herring EM, Holden ZA, Kong WS, Liu J, Magri D, Matzke NJ, McGlone MS, Saltré F, Stigall AL, Tsai YH, Williams JW. 2014. Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytologist* **204:** 37–54.

Graham A. 1999. *Late Cretaceous and Cenozoic history of North American vegetation: north of Mexico*. New York: Oxford University Press on Demand.

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

Hilgers L, Grau JH, Pfaender J, von Rintelen T. 2016. The complete mitochondrial genome of the viviparous freshwater snail *Tylomelania sarasinorum* (Caenogastropoda: Cerithioidea). *Mitochondrial DNA Part B: Resources* **1:** 330–331.

Hudson R. 2002. ms - a program for generating samples under neutral models. *Bioinformatics* **18**: 337–338.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28:** 1647–1649.

Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* **2:** 18–22.

Lin CP, Danforth BN. 2004. How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets. *Molecular Phylogenetics and Evolution* **30:** 686–702.

Liu GH, Wang SY, Huang WY, Zhao GH, Wei SJ, Song HQ, Xu MJ, Lin RQ, Zhou DH, Zhu XQ. 2012. The complete mitochondrial genome of *Galba pervia* (Gastropoda: Mollusca), an intermediate host snail of *Fasciola* spp. *PLoS ONE* **7**.

Meredith M, Kruschke J. 2017. HDInterval: highest (posterior) density intervals. R package version 0.1.3.

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

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76:** 5269–5273.

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.

Pielou EC. 2008. *After the ice age: the return of life to glaciated North America*. Chicago, IL: University of Chicago Press.

Pilsbry HA. 1948. *Land Mollusca of North America. Vol. 2, part 2*. Philadelphia, PA: The Academy of Natural Sciences of Philadelphia.

Pilsbry HA, Vanatta EG. 1898. Revision of the North American Slugs: Binneya, Hemphillia, Hesperarion, Prophysaon and Anadenulus. *Proceedings of the Academy of Natural Sciences of Philadelphi* **50:** 219–261.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* **13:** 235–238.

Richards CL, Carstens BC, Knowles LL. 2007. Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography* **34:** 1833–1845.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61:** 539–542.

Ruffley M, Smith ML, Espíndola A, Carstens BC, Sullivan J, Tank DC. 2018. Combining allele frequency and tree-based approaches improves phylogeographic inference from natural history collections. *Molecular Ecology* **27:** 1012–1024.

Smith ML, Ruffley M, Espíndola A, Tank DC, Sullivan J, Carstens BC. 2017. Demographic model selection using random forests and the site frequency spectrum. *Molecular Ecology* **26:** 4562–4573.

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.

Swofford DL. 2002. *PAUP* version 4.0 a157*. Sunderland, MA: Sinauer.

Thuiller W, Georges D, Engler R. 2014. *biomod2: Ensemble platform for species distribution modeling. R package version 3.1–64*. Available at: http://CRAN. R-project. org/package= biomod2

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7:** 256–276.

Wilke T, Duncan N. 2004. Phylogeographical patterns in the American Pacific Northwest: lessons from the arionid slug *Prophysaon coeruleum*. *Molecular Ecology* **13:** 2303–2315.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1. Sampling localities from this study, including samples from GenBank.
Table S2. Occurrence points used to build species distribution models (SDMs).
Table S3. Bioclimatic variables used in species distribution models.
Table S4. Priors for the approximate Bayesian computation (ABC) analysis.
Table S5. Parameters from PAUP* model selection results.
Table S6. Cross-validation results for the three recent dispersal models with infinite sites data.
Table S7. Approximate Bayesian computation (ABC) results for the three recent dispersal models with infinite sites data.
Table S8. Cross-validation results for the recent dispersal vs. ancient vicariance step with infinite sites data.
Table S9. Approximate Bayesian computation (ABC) results for the recent dispersal vs. ancient vicariance step with infinite sites data.
Figure S1. Results of predictions from the RF classifier when (A) no taxonomy and (B) a revised classification was used. The revised classification included three ranks: vertebrate, invertebrate and plant.
Figure S2. Posterior and posterior predictive distributions are shown compared with observed summary statistics when sequence data were simulated. Posterior distributions shown in blue, posterior predictive distributions are shown in red, and observed data are indicated by black dotted lines.
Figure S3. Species distribution models constructed in Maxent. The average across five replicates is shown both for the present and for the Last Glacial Maximum (LGM) for each species. A, *Prophysaon andersoni*, current. B,

P. andersoni, LGM. C, *Prophysaon dubium*, current. D, *P. dubium*, LGM. E, *Prophysaon coeruleum*, current. F, *P. coeruleum*, LGM. G, *Prophysaon vanattae* and *Prophysaon humile*, current. H, *P. vanattae* and *P. humile*, LGM.
Figure S4. Gene tree estimated in GARLIv2.0, with bootstrap support for major groups.
Figure S5. Majority rule consensus tree including all compatible groups. Bootstrap replicates were constructed in GARLIv2.0, and the consensus tree was computed in PAUP* v.4.0.

SHARED DATA

All sequences are available on GenBank (accession numbers: MH324506–MH324729). Scripts are available from github (https://github.com/meganlsmith/).

Title: The Frequency and Topology of Pseudoorthologs

Authors: Megan L. Smith[1] and Matthew W. Hahn[1]

[1]Department of Biology and Department of Computer Science, Indiana University, Bloomington, IN 47405, USA

Corresponding Author: Megan L. Smith; email: mls16@indiana.edu

ABSTRACT

Phylogenetics has long relied on the use of orthologs, or genes related through speciation events, to infer species relationships. However, identifying orthologs is difficult because gene duplication can obscure relationships among genes. Researchers have been particularly concerned with the insidious effects of pseudoorthologs—duplicated genes that are mistaken for orthologs because they are present in a single copy in each sampled species. Because gene tree topologies of pseudoorthologs may differ from the species tree topology, they have often been invoked as the cause of counterintuitive results in phylogenetics. Despite these perceived problems, no previous work has calculated the probabilities of pseudoortholog topologies, or has been able to circumscribe the regions of parameter space in which pseudoorthologs are most likely to occur. Here, we introduce a model for calculating the probabilities and branch lengths of orthologs and pseudoorthologs, including concordant and discordant pseudoortholog topologies, on a rooted three-taxon species tree. We show that the probability of orthologs is high relative to the probability of pseudoorthologs across reasonable regions of parameter space. Furthermore, the probabilities of the two discordant topologies are equal and never exceed that of the concordant topology, generally being much lower. We describe the species tree topologies most prone to generating pseudoorthologs, finding that they are likely to present problems to phylogenetic inference irrespective of the presence of pseudoorthologs. Overall, our results suggest that pseudoorthologs are unlikely to mislead inferences of species relationships under the biological scenarios considered here.

KEYWORDS: Paralogs, Phylogenetics, Orthologs, Birth-death model

Phylogenetics aims to reconstruct evolutionary relationships among species. Recent advances in sequencing technologies have drastically increased the amount of data available for phylogenetic inference (Scornavacca et al. 2020), which has led in turn to increased concern about how to assemble and filter large genomic and transcriptomic datasets. Central to most data-generating pipelines is the identification of orthologs, or genes related through speciation events, to the exclusion of paralogs, or genes related through duplication events (Fitch 1970). Because orthologous gene trees reflect only the species history, it has been argued that solely orthologs are appropriate for phylogenetic inference (e.g, Fernández et al. 2020; Kapli et al. 2020). Methods to extract orthologs from large datasets have therefore proliferated (reviewed in (Altenhoff et al. 2019a), e.g. (Ebersberger et al. 2009; Altenhoff et al. 2011, 2013; Dunn et al. 2013; Yang and Smith 2014)), but the task remains difficult, and pseudoorthologs (Koonin 2005) (or "hidden paralogs" (Doolittle and Brown 1994)), are thought to represent a particularly insidious problem. Pseudoorthologs are paralogs that are mistaken as orthologs because, due to patterns of differential duplication and loss, they are present in a single copy in each sampled species.

Pseudoortholog gene trees can differ from the species tree in their topology and branch lengths. Consider, for example, a scenario in which a duplication occurred in the ancestor of three species (A, B, and C), where species A and B are sister species (Figs. 1a,b). If one of the two copies is lost immediately, we can only sample genes with orthologous relationships (Fig. 1c). If one copy is retained in species A and species B, while the other is retained in species C, then we have a pseudoortholog that is topologically identical to the true ortholog, but which has a longer internal branch (Fig. 1d). Finally, if one copy is retained in species A (or B) and the

other is retained in species B (or A) and C, then we have a pseudoortholog with a topology that differs from the species tree topology (Figs. 1e,f). Because discordant pseudoorthologs are difficult to identify—and may introduce both branch length and topological heterogeneity—they are often invoked as the culprits behind counterintuitive results in phylogenetics.

Multiple studies have attempted to assess the influence of paralogs (including pseudoorthologs) on phylogenetic inference, though they have generally done so by comparing results filtered using different ortholog detection methods (Fernández et al. 2020), none of which are likely to remove pseudoorthologs. The results of these analyses have been mixed, with some studies finding substantial differences in inferred species trees (Altenhoff et al. 2019b; Siu-Ting et al. 2019; Cheon et al. 2020) and others finding minimal differences (Fernández et al. 2018; Kallal et al. 2018; Cheon et al. 2020). Furthermore, and in contrast to the long-held opinion that orthologs, not paralogs, should be used to infer species relationships, recent methodological developments explicitly allow for the inclusion of paralogs in phylogenetic inference (reviewed in Smith and Hahn 2021). In particular, quartet-based gene tree methods are robust to the inclusion of paralogs because the concordant topology is expected to be the most common topology, in the limit of a very large number of genes (Yan et al. 2021; Legried et al. 2020; Markin and Eulenstein 2020; Zhang et al. 2020). However, branch-length estimates, concordance factors, and measures of nodal support may still be impacted by paralog inclusion.

For researchers who wish to only use orthologs for phylogenetic inference, current practices for excluding putative pseudoorthologs can be particularly restrictive because excluding these genes from phylogenetic datasets is difficult. When extracting putative

orthologs, researchers typically rely on graph-based or tree-based approaches (Altenhoff et al. 2019a). Graph-based approaches are the most commonly used: they rely on the concept of reciprocal best hits (Li 2003) or length-normalized reciprocal best hits (Emms and Kelly 2015), followed by the application of clustering approaches to delineate orthologs. These methods assume that the two most closely related homologs between a pair of species should be orthologs. When pseudoorthologs are present they will be reciprocal best hits, despite not having an orthologous relationship, because true orthologs are absent. Tree-based approaches (e.g., Yang and Smith 2014) extract orthologs from the clusters identified using graph-based methods. Tree-based approaches are more computationally intensive, but in some cases may be able to identify and exclude pseudoorthologs by identifying excessively long branches. Even these approaches, though, will often fail to identify pseudoorthologs, particularly when duplication events were more recent. Another approach to removing putative pseudoorthologs relies on knowledge of a species tree, removing genes that show discordance between the gene and species tree for a set of pre-defined clades (Siu-Ting et al. 2019). This approach assumes that discordance between gene trees and the species tree with respect to *a priori* defined 'uncontestable' relationships is due to gene duplication and loss, although many other factors including incomplete lineage sorting and introgression may also lead to gene tree heterogeneity (Maddison 1997).

Thus, options for excluding pseudoorthologs from phylogenetic datasets are limited and likely ineffectual at removing pseudoorthologs, and those that do exist may lead to a drastic reduction in the amount of data available. For example, when using their approach based on the monophyly of predefined clades, Siu-Ting et al. (Siu-Ting et al. 2019) removed 637 of their 2656

putative orthologs. They found some differences between tree topologies and branch lengths inferred from these filtered and unfiltered datasets, but it is difficult to establish whether the removed genes were actually pseudoorthologs and how many pseudoorthologs remained after stringent filtering. A better understanding of when, and how stringently, we should filter our data to remove pseudoorthologs would clearly be helpful: it could prevent unnecessary filtering of informative genes from phylogenetic datasets and would guide researchers as to whether and when results should be interpreted with caution due to the potential presence of pseudoorthologs.

Despite long-standing concerns about the effects of pseudoorthologs on phylogenetic inference, no attempt has been made to calculate the probability of pseudoorthologs or to understand the regions of parameter space in which they may be most problematic. Here, we use a stochastic birth-death model to calculate the probabilities and branch lengths of orthologs and pseudoorthologs, including both concordant and discordant pseudoortholog topologies. In what follows, we first describe the model, and then explore regions of parameter space that are most likely to produce pseudoorthologs. We show that the probability of orthologs is high relative to the probability of pseudoorthologs across parameter space, and that the ratio of concordant to discordant topologies is even higher. Our results should reassure researchers concerned about the effects of pseudoorthologs on phylogenetic inference.

THE MODEL

Probabilities of Orthologs and Pseudoorthologs

To calculate the probabilities of orthologs and pseudoorthologs, we use a stochastic birth-death model (Bailey 1964). Previous work has applied birth-death models to gene trees with the aim of inferring orthology, reconciling gene and species trees, and accurately reconstructing gene trees (Arvestad et al. 2003, 2004; Rasmussen and Kellis 2011). Here, we evaluate a specific case by focusing on a rooted three-taxon species tree, considering scenarios that generate single-copy genes in order to estimate probabilities of orthologs and pseudoorthologs.

All calculations assume that there is one gene copy at the beginning of internal branch t_1 (Fig. 1a). When only a single duplication (and no loss) occurs on this branch, such that two copies exist at the most recent node, we treat each copy independently, generating two "daughter" gene trees (Fig. 1b). The independent evolution of each copy means that we can calculate probabilities of further gain and loss on all subsequent lineages, always beginning with a single copy at the base of the daughter gene trees. Since we always begin with a single copy, we use the following equations to calculate the probabilities of transitions along branches, where λ is the duplication rate and μ is the loss rate. The probability of starting with 1 copy and ending with n copies along a branch with length t can be calculated as (Bailey 1964):

$$p_n(t) = \begin{cases} (1-\alpha)(1-\beta)\beta^{n-1}, & n \geq 0 \\ \alpha, & n = 0 \end{cases}$$

where

$$\alpha = \frac{\mu(e^{(\lambda-\mu)t} - 1)}{\lambda e^{(\lambda-\mu)t} - \mu}, \beta = \frac{\lambda(e^{(\lambda-\mu)t} - 1)}{\lambda e^{(\lambda-\mu)t} - \mu}$$

When λ = μ, we use the following simplification (Bailey 1964):

$$p_n(t) = \begin{cases} \dfrac{(\lambda t)^{n-1}}{(1+\lambda t)^{n+1}}, & n \geq 0 \\ \dfrac{(\lambda t)}{(1+\lambda t)}, & n = 0 \end{cases}$$

Using the above equations, we can calculate the overall probabilities of different ortholog and pseudoortholog topologies. As an example, consider the concordant pseudoortholog in Figure 1d (see also Supporting Fig. S2a). We calculate the overall probability of this topology by multiplying: the probability of transitioning from 1 to 2 copies on branch t_1, the probability of transitioning from 1 to 0 copies on branch t_2 in one copy and 1 to 0 copies on branch t_3 in the other copy, and the probability of no changes on any of the other branches (online Appendix A). Note that there are two different arrangements that may lead to this outcome, depending on which daughter gene tree losses occur on. Similarly, we can calculate the probability of the type of ortholog shown in Figure 1a by calculating the total probability that there are no transitions on any branch (i.e. that the state is 1 at all nodes; Supporting Fig. S1a). In total, we consider six ortholog configurations (i.e. sets of events leading to orthologs; Supporting Fig. S1) that can each occur in from one to six different arrangements (depending on which exact copies are lost). We consider nine concordant pseudoortholog configurations (Supporting Fig. S2) and five configurations for each of the two discordant pseudoortholog topologies (Supporting Fig. S3), each of which can occur in one to six different arrangements. There are more configurations possible with more ancestral copies, but here we limit the number of copies at the end of branch t_1 to 3 (i.e. two duplications on branch t_1). Code to calculate the probabilities of orthologs, concordant pseudoorthologs, and discordant pseudoorthologs is given in online Appendix A.

Our model makes several assumptions. Most notably, we assume that there are no more than three copies at the end of branch t_1, and we avoid nested duplication scenarios. To evaluate whether these assumptions lead to accurate predictions, we compared the exact probabilities calculated according to the equations above to the proportions of each scenario observed in simulations in SimPhy (Mallo et al. 2016). We calculated the proportion of observed orthologs, concordant pseudoorthologs, and discordant pseudoorthologs by evaluating topology and branch lengths. If our assumptions are reasonable, we expect a 1:1 relationship between calculations and observations. We drew 100 sets of parameters from uniform priors (Supporting Table S1), and performed 10,000 simulations under each set of parameters in SimPhy. While our model does not exhaust all possible scenarios that could lead to pseudoorthologs, these simulations show that the scenarios we consider lead to accurate predictions of the numbers of orthologs, as well as the numbers of concordant and discordant pseudoorthologs (Supporting Fig. S4).

Expectations for Branch Lengths of Pseudoorthologs

In addition to differing topologically from the species tree, pseudoorthologs differ from the species tree in terms of branch lengths. For orthologs, the single internal branch of a rooted three-taxon tree is length t_2; this branch determines the phylogenetic signal within each gene tree, and is the focus here. For the simplest concordant pseudoortholog (Fig. 1d), the internal branch length is equal to the sum of t_2 and the time until the duplication event occurs in branch t_1 (v in Fig. 1), while for the simplest discordant pseudoortholog the internal branch length is equal to v (Figs. 1e,f). Furthermore, average internal branch lengths for the two discordant pseudoorthologs are always equal.

The value of v is the expected time back to the duplication event on t_1 conditional on a duplication event occurring on branch t_1. Because waiting times for events in the birth-death process are also exponentially distributed (Gernhard 2008), we can use a model similar to that for the multispecies coalescent (Mendes and Hahn 2018) to calculate times here. To find the expectation for v, we need only convert from the coalescent units used in Mendes and Hahn (Mendes and Hahn 2018) to duplication units, where one coalescent unit is equal to $\frac{1}{\lambda}$ here. These considerations lead to the following expectation for the time back to the duplication event:

$$E[v] = \frac{1}{\lambda} - \frac{t_1}{e^{t_1\lambda} - 1}$$

Although we have conditioned on a duplication event occurring in branch t_1, we have not conditioned on other events. For example, we have not conditioned on the absence of any subsequent duplication events or a subsequent loss on branch t_1. To evaluate whether the

assumptions made here led to reasonable predictions of the internal branch length, we again used simulations in SimPhy (Mallo et al. 2016). We drew 50 sets of parameters from uniform priors; all priors were the same as in Supporting Table S1 except μ and λ were drawn from U(0.004, 0.005) priors to ensure more pseudoorthologs. We performed 10,000 simulations under each set of parameters, and calculated the average internal branch lengths from simulations that produced either trees matching the concordant pseudoortholog shown in Figure 1d or trees matching the discordant pseudoortholog shown in Figure 1e. The expected branch lengths are a close match to simulated branch lengths (Supporting Fig. S5), and thus should provide accurate predictions of the internal branch lengths of pseudoorthologs.

The model presented here demonstrates that the expected internal branch length for concordant pseudoorthologs is always longer than the expected branch length for discordant pseudoorthologs, by the length of the internal branch t_2. Thus, even when pseudoorthologs are present, the total expected branch length supporting the concordant topology should exceed the expected branch length supporting the discordant topology. In other words, there is more phylogenetic signal in concordant trees than discordant ones. In addition, the internal branch supporting each of the two different discordant pseudoorthologs has the same expected length. This implies equal support for each of the two discordant topologies.

PROBABILITIES OF ORTHOLOGS AND PSEUDOORTHOLOGS

We used the model described above to explore how different parameters affected the probabilities of orthologs and pseudoorthologs, including both concordant and discordant

topologies. We begin by describing our results in terms of unconditional probabilities, which consider all scenarios resulting from our model, including those that do not produce one gene copy per species. Considering the rates of gene duplication (λ) and loss (μ), we found that higher rates of each decreased the overall probability of orthologs (Supporting Fig. S6a). The probability of orthologs decreases because there is a higher chance of duplication and loss events occurring: duplication creates additional copies, while loss means that no copy can be sampled from some species. A similar effect is generated by increasing all branch lengths.

By contrast, the probability of pseudoorthologs is maximized at intermediate values of λ and μ (Fig. 2a), because at least one duplication event and two loss events are required for pseudoorthologs (Figs. 1d-f). Values of these parameters that are too high decrease the probability of there being a single copy in each species; because pseudoorthologs require more losses than gains, slightly higher values of μ are possible. Similarly, the probability of pseudoorthologs is maximized at intermediate branch lengths of t_1 (Fig. 2b) and t_3 (Supplementary Fig. 6b), because at least one duplication is required on branch t_1 and at least one loss is required on branch t_3.

PROBABILITIES OF CONCORDANT AND DISCORDANT PSEUDOORTHOLOGS

To understand the relative probabilities of concordant and discordant pseudoortholog topologies, we examined many of the same rate and branch-length parameters. Increasing μ decreases the ratio of concordant pseudoorthologs to discordant pseudoorthologs, because

discordant pseudoorthologs require at least three losses while concordant pseudoorthologs can occur with only two losses (Supporting Fig. S6c). Increasing λ slightly increases the ratio of concordant pseudoorthologs to discordant pseudoorthologs (Supporting Fig. S6c), as there are more possible configurations leading to concordant pseudoorthologs than discordant pseudoorthologs when there is more than one duplication event (Supporting Figs. S2, S3). Changes to the lengths of branches t_2 and t_4/t_5 affect the relative probabilities of concordant and discordant pseudoorthologs: specifically, as branch t_2 gets longer and branches t_4/t_5 get shorter, concordant pseudoorthologs become more likely and discordant pseudoorthologs become less likely (Fig. 2c). This occurs because concordant pseudoorthologs can be generated either by a loss on t_2 or by losses on both t_4 and t_5 (Fig. 1d; Supporting Figs. S1a,b), while discordant pseudoorthologs require losses on branches t_4 and t_5 (Figs. 1e-f; Supporting Fig. S3). Both scenarios additionally require losses on t_3, and so the length of t_3 does not affect their relative frequencies. Note that results from the model presented here are also supported by results from simulations (Supporting Fig. S4).

We further explored the probabilities of all events conditional on a single copy being present in each species. These calculations directly address the chance that pseudoorthologs are mistaken for orthologs: the conditional probabilities represent the fraction of all single-copy genes that are orthologs, concordant pseudoorthologs, or discordant pseudoorthologs. We explored two general regions of parameter space, representing the range of values of λ and μ observed in empirical datasets: 0.002 and 0.005 per million years (Mendes et al. 2020). We considered a long length of branch t_3 (198.9 million years) across a range of lengths for branches

t_1, t_2, t_4, and t_5. The large value of t_3 mirrors a potentially difficult region of tree space (see next section), coupled with moderate and high rates of duplication and loss.

The conditional probability of orthologs given that a single copy is present in each species is very high when rates of duplication and loss are moderate (0.002, Fig. 3a, Supporting Fig. S7; minimum conditional probability of orthologs = 0.955), and is moderately high even when rates of duplication and loss approach the highest observed in empirical datasets (Fig. 3c; minimum conditional probability = 0.711). Furthermore, the ratio of concordant to discordant topologies is very high when duplication and loss rates are moderate (Fig. 3b; minimum = 76.7:1) and is still rather high even when rates of duplication and loss are high (Fig. 3d; minimum = 8.4:1). These ratios include both orthologs and concordant pseudoorthologs in the "concordant" category. Note again that we chose t_3 to mirror the most problematic regions of parameter space for these results; Supporting Figure S8 shows results for different values of t_3, confirming the impression that the scenario shown here in the main text is a worst-case scenario with regards to this branch length.

Notably, the probability of either of the two discordant pseudoorthologs can never exceed the probability of the concordant pseudoortholog, because it is always possible to generate a concordant pseudoortholog with the same number of duplication and loss events (and often fewer events). For example, a discordant pseudoortholog can be generated by a duplication on branch t_1, a loss on branches t_4 and t_3 in one copy, and a loss on branch t_5 in the other copy (Fig. 1e). A concordant pseudoortholog could also be generated by this pattern, as long as the losses on branches t_4 and t_5 occurred in the same copy, while the loss on branch t_3 occurred in the other copy (Supporting Fig. S2b). In reasonable regions of parameter space, the probability of

concordant pseudoorthologs is much higher than the probability of either discordant pseudoortholog because most concordant pseudoorthologs require one fewer loss event (i.e. the scenario shown in Fig. 1d; Supporting Figs. S2, S3). Moreover, the probabilities of the two discordant topologies are always equal, as these rely on the same events on the same branches, and only differ in terms of which branches are lost from which copy. Thus, one never expects either discordant topology to be significantly more frequent than the other. Finally, note again that in Figure 3 we are showing the probabilities of orthologs and pseudoorthologs conditional on sampling a single copy per species. However, the absolute probability of sampling a single copy per species at all is lowest in the regions of parameter space that maximize the probability of pseudoorthologs and discordant topologies (Supporting Fig. S9).

WORST-CASE SCENARIOS

In order to find the species tree topologies most prone to producing pseudoorthologs (especially discordant ones), we searched parameter space for such trees. Specifically, we searched for regions of parameter space that a) maximized the probability of pseudoorthologs conditional on a single copy per species, b) maximized the probability of discordant pseudoorthologs conditional on a single copy per species, and c) minimized the ratio of concordant topologies to discordant topologies (Supporting Table S2). We set bounds on all parameters (Supporting Table S3), and constrained the species tree to be ultrametric by setting t_5 equal to t_4 and requiring that $t_4 + t_2 = t_3$. We changed each parameter in turn, increasing or decreasing the value at random by a value chosen from a uniform prior distribution U(0.000001, 0.001) for μ and λ, and U(0.0001, 20) for branch lengths. For each optimization, we accepted

each change if it increased the probability (or decreased the ratio), and accepted the change one percent of the time if it decreased the probability (or increased the ratio). For each parameter we performed 100 optimization steps. We visited the parameters in the order: μ, λ, t_1, t_3, and t_2, and repeated the procedure ten times.

The maximum conditional probability of pseudoorthologs that we found was 0.285 (Supporting Table S2), and this value was only obtained with high values of λ and μ (Mendes et al. 2020). The highest conditional probability of either of the two discordant pseudoorthologs observed was 0.095, and again this involved high values of λ and μ (Supporting Table S2; Fig. 4). The minimum ratio of the concordant topology to either of the two discordant topologies was 8.5. This suggests that, even in the most problematic regions of parameter space, discordant pseudoorthologs will comprise fewer than 10% of all single-copy genes.

Equally importantly, our results show that the ratio of concordant to discordant topologies is lowest in a region of tree space in which discordance due to ILS is also likely to be a concern (Fig. 4)—when the internal branch of a three-species tree is very short. If we take units in millions of years and assume a species with a generation time of 29 years and an effective population size of 10,000, then the probability of either discordant topology for the worst-case species tree under the multispecies coalescent model is 0.225. By comparison, the conditional probability of either discordant topology under our model of duplication and loss in this same area of parameter space is 0.095, a value more than two times lower. Additionally, this region of parameter space involves very long branch lengths for t_1, t_3, and t_4/t_5 (nearly 200 million years)

329 and high rates of duplication and loss. In such species trees, pseudoorthologs are not likely to be
330 the biggest impediment to phylogenetic inference.
331
332 We also explored regions of parameter space that maximized the absolute probabilities of
333 pseudoorthologs and discordant pseudoorthologs, rather than the probabilities conditional on a
334 single copy per species (Supporting Table S4). The most notable difference was in the branch
335 lengths that maximized the probability of pseudoorthologs. When absolute probabilities are
336 considered, a long internal branch t_2 and short terminal branches t_4 and t_5 maximize this
337 probability because they maximize the probability of the concordant pseudoortholog (Supporting
338 Table S4). However, when conditional probabilities are considered, a shorter internal branch t_2
339 maximizes the probability of pseudoorthologs because, coupled with longer branches t_4 and t_5, a
340 shorter branch t_2 decreases the probability of orthologs.
341
342 **WHOLE GENOME DUPLICATIONS**
343
344 To evaluate the effects of whole genome duplication events (WGDs), we used
345 simulations in SimPhy (Mallo et al. 2016). Since WGDs cannot be specified in SimPhy, we
346 simulated two locus trees per replicate for the rooted three-taxon tree (and an outgroup); these
347 trees were treated as a pair of duplicates produced by WGD. These duplicates were identical in
348 terms of their branch lengths, but all subsequent duplication or loss events were independent
349 across the two copies. We simulated data under six conditions. We combined moderate (0.002)
350 and high (0.005) duplication and loss rates with three branch length conditions that are described
351 in Supporting Table S5. The final scenario was designed based on the worst-case results under

the original model (Fig. 4). We recorded the proportion of single-copy genes and the proportion of those genes that were orthologs, concordant pseudoorthologs, and discordant pseudoorthologs, and compared these values to the predicted probabilities under the model without WGDs described above. In general, WGDs lead to fewer single-copy genes (Supporting Table S5). Conditional on a single copy per species, WGDs lead to a lower proportion of orthologs, and higher proportions of concordant pseudoorthologs and each discordant pseudoortholog. Despite this, the proportion of genes with the concordant topology is always higher than the proportion of genes with either discordant topology, and the proportion of genes with either discordant topology never exceeds 0.15 even in the worst-case scenario (Supporting Table S5). This suggests that, although polyploidy offers unique challenges, the expectation that the concordant topology should always be the most frequent holds, at least in the scenario considered here.

EXTENDING THE MODEL TO LARGER TREES

Thus far we have considered the probability of orthologs and pseudoorthologs in a three-taxon species tree. While we might intuitively expect that the addition of more species would lower the relative probability of pseudoorthologs (because more losses would be required to mimic orthologs), we carried out additional analyses to evaluate slightly larger trees by adding a single extra taxon sister to Species A, Species C, or to internal branch t_1, assuming no discordance at the node uniting the new taxon and its sister species in the former two cases. These results support our prediction: adding taxa decreases the probability of all types of single-copy genes, including orthologs and concordant pseudoorthologs (Supporting Figs. S10a, S10b, S11, S12). However, adding leaves disproportionately decreases the probability of discordant

pseudoorthologs, particularly when branches are added as sister to Species A (Supporting Figs. S10c, S10d). This outcome occurs because discordant pseudoorthologs require losses on branches t_4 and t_5, and, if these losses do not occur before the split between the two sister branches including Species A, then the number of losses required increases by one. While the same is true when a branch is added sister to Species C, since branch t_3 is longer than branches t_4 and t_5, there is more time for the loss to occur prior to the added speciation event. Adding a new lineage to internal branch t_1 has similar effects (Supporting Text, Supporting Figs. S11, S12), with a decrease in the absolute probabilities of orthologs and pseudoorthologs; the concordant topology remains the most probable. Overall, these limited extra analyses indicate that results for a three-taxon tree represent a worst-case scenario for confusing pseudoorthologs with orthologs.

DISCUSSION

Pseudoorthologs have long been feared for their possible detrimental effects on species tree inference. Removing these genes is difficult, and methods have relied on removing long branches (e.g., Yang and Smith, 2014) or on the monophyly of other clades defined *a priori* (e.g., Siu-Ting et al. 2019), both of which may remove a substantial fraction of the data, and neither of which is guaranteed to remove all (or only) pseudoorthologs. Our results suggest that pseudoorthologs are unlikely to mislead phylogenetic inferences, given the assumptions of the model presented here. We find that pseudoorthologs are rare overall (Fig. 3a), and that pseudoorthologs with discordant topologies are expected to be much less common than genes with concordant topologies (Fig. 3b). Thus, our results suggest that regardless of the particular

method used to identify single-copy orthologs, discordant pseudoorthologs are unlikely to be mistakenly sampled; thus, they are unlikely to pose a challenge to phylogenetics.

Even in the most problematic regions of parameter space considered here, pseudoorthologs are unlikely to mislead topological inferences. First, orthologs are still substantially more likely than pseudoorthologs, and the concordant topology is still more than 8X as likely as either of the two discordant topologies. Second, topological heterogeneity is recognized as common across the tree of life due to many processes (Bravo et al. 2019), and in this particular region of parameter space we expect discordance to be high due to incomplete lineage sorting (Fig. 4). Since discordance is likely to be high irrespective of the presence of pseudoorthologs, the use of methods that are robust to discordance (e.g., ASTRAL, Mirarab et al. 2014; Zhang et al. 2018) should be of utmost importance. Moreover, our modelling results corroborate previous findings that quartet methods such as ASTRAL are statistically consistent under a model of gene duplication and loss (Legried et al. 2020; Markin and Eulenstein 2020). Such methods rely on the fact that, for a rooted three-taxon species tree, the concordant topology is always the most frequent, which is exactly what we find here among all single-copy genes. Additionally, the probabilities of the two discordant topologies are always equal, suggesting that methods for species tree inference (Chifman and Kubatko 2014) and tests of introgression (Huson et al. 2005; Vanderpool et al. 2020) based on symmetry in number of the two discordant topologies should not be misled by the inclusion of pseudoorthologs. Thus, even in the most problematic regions of parameter space, quartet-based methods should not be misled by the presence of pseudoorthologs. Finally, in these regions of parameter space—and when there are more than three species in a tree—single-copy genes shared across all species are particularly

rare (e.g. Supporting Fig. S9b). Thus, we expect that researchers would be unable to sample many single-copy genes in such cases, and should therefore consider explicitly including paralogs in their dataset to gain more phylogenetic markers (Smith and Hahn 2021).

In addition to changes in gene tree topologies, pseudoorthologs have different branch lengths than orthologs. Concordant pseudoorthologs are expected to have the longest internal branch lengths, with the expected branch length converging on $t_2 + 1/\lambda$ (the latter term representing the expected time to the duplication event) as the length of branch t_1 increases (Fig. 1d; (Mendes and Hahn 2018)). Discordant pseudoorthologs will have longer terminal branch lengths (Figs. 1e, f), but a shorter internal branch length. The expected internal branch length of discordant pseudoortholog will converge to $1/\lambda$ as the length of branch length t_1 increases, which may be either shorter or longer than the internal branch length t_2 of true orthologs. Since concordant pseudoorthologs are never expected to occur at a lower frequency than either of the discordant pseudoorthologs—and the internal branch is always longer by t_2—the total expected internal branch length supporting the true topology should always exceed that supporting either discordant topology. These results suggest that, with enough data, concatenation-based methods are also unlikely to be misled by pseudoorthologs. However, the expected branch lengths depend on more assumptions than do the calculated probabilities, as these calculations are conditioned only on a duplication occurring on branch t_1 and not on the presence of other necessary events. These calculations also depend upon the presence of only a single copy at the beginning of branch t_1, and relaxations of these assumptions may alter the expected branch lengths.

While the incidental inclusion of pseudoorthologs seems unlikely to affect inferences of species tree topology, many phylogenetic studies also aim to estimate concordance factors, nodal support, and branch lengths, and sometimes to test for the presence of introgression. Pseudoorthologs could lead to biased branch length estimates. Specifically, since internal branch lengths of concordant pseudoorthologs and external branch lengths of discordant pseudoorthologs are always expected to be longer than the corresponding branch lengths of orthologs, the presence of pseudoorthologs should lead to overestimates of branch lengths for methods that estimate branch lengths in substitutions per site. For methods that estimate branch lengths in coalescent units, pseudoorthologs should lead to underestimated branch lengths since they should introduce additional discordance relative to expectations under the multispecies coalescent model. For the same reason, the presence of pseudoorthologs may decrease measures of nodal support and concordance factors. However, the rarity of pseudoorthologs across most of parameter space (Fig. 3a) should minimize their effects on estimates of branch lengths, concordance factors, and nodal support values. Notably, the inclusion of pseudoorthologs should not affect inferences of introgression that rely on symmetries in minor site patterns or topologies (e.g., (Huson et al. 2005; Vanderpool et al. 2020)) since each of the two discordant topologies is equally likely.

We stress that the results presented here make a number of assumptions about the process of gene duplication and loss. Here we discuss some of these assumptions and the potential effects of their violations on the probabilities of observing orthologs and pseudoorthologs. Our model assumes a relatively simple process of gene duplication and loss, with constant rates through time and across lineages. Higher rates of gene duplication and loss during certain time intervals

could change the relative probabilities of orthologs, discordant pseudoorthologs, and concordant pseudoorthologs. For example, if rates of gene loss are higher immediately after gene duplication, then we would expect that the probability of completely losing one copy would be higher, and thus that the probability of orthologs would increase. Alternatively, if rates of gene loss were higher near the tips of the tree, then we might expect an increased probability of pseudoorthologs and an increased ratio of discordant to concordant topologies. However, it is difficult to construct a scenario in which either of these model violations leads to more overall discordant than concordant topologies.

Most of the results presented here have also assumed that there is a single gene copy at the beginning of branch t_1. However, an alternative scenario for pseudoorthologs is polyploidy, a special case of gene duplication and loss in which the entire genome is duplicated (Otto 2007). Polyploidy can lead to increased probabilities of pseudoorthologs conditional on sampling a single gene per species, though the probability of sampling single-copy genes will also be much lower in this scenario. When the taxa considered are autopolyploids, or polyploid taxa for which both sub-genomes come from the same species, then we need only condition on the polyploidy event having occurred at some point prior to branch t_1 for our model to apply. We used simulations to explore this scenario, finding that while WGDs decrease the overall probability of orthologs and increase the probability of pseudoorthologs conditional on sampling a single copy, the concordant topology is still always more likely than either discordant topology (Supporting Table S5). Notably, this is a relatively simple model of WGD, and more complex scenarios could include nested WGDs or biased gene retention across genome copies. Nested WGDs could lead both to increased probabilities of pseudoorthologs and to increased ratios of discordant to

concordant pseudoorthologs. However, it is still difficult to construct a scenario in which the probability of discordant pseudoorthologs would exceed that of concordant pseudorthologs, since either type of pseudoortholog can result from the same events on different copies. Even with biased gene retention, the only way to generate more discordant than concordant topologies is if gene retention varies across copies in a species-specific manner. In our simulations we only explored autopolyploidy. Allopolyploids are polyploid taxa in which each sub-genome comes from a different species (Otto 2007). In this case, there are two "concordant" trees, depending on which parental genome is considered. Paralogs from the different parental species are sometimes lost in a biased fashion (e.g., Chang et al. 2010). The main consequence of biased loss is that one set of orthologous species relationships will be retained over the other (Thomas et al. 2017). Of course, even more so in polyploids than in other taxa, excessive filtering to remove putative pseudoorthologs will decrease the amount of available data. Furthermore, the number of single-copy gene families will be limited in cases of polyploidy, and using multiple-copy gene families for phylogenetic inference is likely the ideal approach in this case (Smith and Hahn, 2021).

Based on the results presented here, pseudoorthologs are unlikely to be the frequent cause of problems in phylogenetic inference. How then can we explain previous results that claim to demonstrate the negative effects of pseudoorthologs on phylogenetic inference? Some studies have found differences in trees inferred from datasets filtered using different ortholog detection methods (Altenhoff et al. 2019b; Siu-Ting et al. 2019; Cheon et al. 2020). However, most ortholog detection methods are unlikely to remove pseudoorthologs, and thus comparisons of these methods are not informative with respect to the effects of pseudoorthologs. In cases where researchers specifically aim to exclude pseudoorthologs (e.g. Siu-Ting et al. 2019) and

511 inferences differ across filtered datasets, more stringent filtering may remove problematic
512 sequences other than pseudoorthologs—for example, alternative isoforms or error-prone
513 sequences. While such filtering may improve phylogenetic inference, this improvement cannot
514 be attributed to the removal of pseudoorthologs. Paralogs included in datasets of putative single-
515 copy orthologs may also not be true pseudoorthologs. In large genomic and transcriptomic
516 datasets, putative pseudoorthologs may instead be paralogs for which, for technical reasons,
517 different copies were assembled in each species (as appears to be the case in Brown and
518 Thomson 2017). In our study, we assume that all pseudoorthologs are a result of true biological
519 loss, rather than sampling artifacts; when random sampling occurs, the overall probability of
520 pseudoorthologs will be higher than observed here. However, even in the extreme case, where a
521 single paralog is sampled at random from each species, quartet-based methods appear to perform
522 well when enough data is available (Yan et al. 2021; Legried et al. 2020; Markin and Eulenstein
523 2020). With respect to the results presented here, randomly sampling species in the present
524 increases the probability of sampling pseudoorthologs. It may also change the proportions of
525 discordant and concordant topologies, since the probability of losing any particular copy due to
526 sampling will not depend on branch lengths. However, it remains true that the discordant
527 topology should never be more probable than the concordant topology if sampling is random.
528
529 Overall, our results suggest that pseudoorthologs are not likely to mislead phylogenetic
530 inference. Pseudoorthologs are rare across reasonable regions of parameter space, and even in
531 the most extreme scenarios considered, the concordant topology is always expected to be the
532 most common. These results should reassure researchers who are well-aware of the difficulties of
533 identifying and removing these genes from phylogenomic datasets, and should encourage

534 researchers to focus their filtering efforts elsewhere, for example on detecting and removing
535 assembly artefacts or on poorly aligned sequences.

FUNDING

This work was supported by a National Science Foundation postdoctoral fellowship to MLS (DBI-2009989) and an National Science Foundation grant to MWH (DEB-1936187).

ACKNOWLEDGMENTS

We thank Rafael Guerrero for helpful discussion, as well as three reviewers and Stephen Smith for constructive comments. This work was supported by a National Science Foundation postdoctoral fellowship to MLS (DBI-2009989) and an NSF grant to MWH (DEB-1936187).

DATA AVAILABILITY

All code for calculating the probabilities of orthologs and pseudoorthologs is reproduced and commented in online Appendix A.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository:
http://dx.doi.org/10.5061/dryad.[NNNN]

REFERENCES

Altenhoff A.M., Gil M., Gonnet G.H., Dessimoz C. 2013. Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS ONE. 8:e53786.

Altenhoff A.M., Glover N.M., Dessimoz C. 2019a. Inferring orthology and paralogy. In: Anisimova M., editor. Evolutionary Genomics: Statistical and Computational Methods. New York, NY: Springer. p. 149–175.

Altenhoff A.M., Levy J., Zarowiecki M., Tomiczek B., Warwick Vesztrocy A., Dalquen D.A., Müller S., Telford M.J., Glover N.M., Dylus D., Dessimoz C. 2019b. OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome Res. 29:1152–1163.

Altenhoff A.M., Schneider A., Gonnet G.H., Dessimoz C. 2011. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res. 39:D289–D294.

Arvestad L., Berglund A.-C., Lagergren J., Sennblad B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics. 19:i7–i15.

Arvestad L., Berglund A.-C., Lagergren J., Sennblad B. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. Proc. Eighth Annu. Int. Conf. Comput. Mol. Biol. - RECOMB 04.:326–335.

Bailey N.T.J. 1964. The elements of stochastic processes with applications to the natural sciences. New York, NY: John Wiley & Sons, Inc.

Bravo G.A., Antonelli A., Bacon C.D., Bartoszek K., Blom M.P.K., Huynh S., Jones G., Knowles L.L., Lamichhaney S., Marcussen T., Morlon H., Nakhleh L.K., Oxelman B., Pfeil B., Schliep A., Wahlberg N., Werneck F.P., Wiedenhoeft J., Willows-Munro S., Edwards S.V. 2019. Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. PeerJ. 7:e6399.

Brown J.M., Thomson R.C. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66:517–530.

Chang P.L., Dilkes B.P., McMahon M., Comai L., Nuzhdin S.V. 2010. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol. 11:R125.

Cheon S., Zhang J., Park C. 2020. Is phylotranscriptomics as reliable as phylogenomics? Mol. Biol. Evol. 37:3672–3683.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics. 30:3317–3324.

Doolittle W.F., Brown J.R. 1994. Tempo, mode, the progenote, and the universal root. Proc. Natl. Acad. Sci. 91:6721–6728.

588 Dunn C.W., Howison M., Zapata F. 2013. Agalma: an automated phylogenomics workflow.
589 BMC Bioinformatics. 14:330.

590 Ebersberger I., Strauss S., von Haeseler A. 2009. HaMStR: profile hidden Markov model based
591 search for orthologs in ESTs. BMC Evol. Biol. 9:157.

592 Emms D. M., Kelly S. 2015 OrthoFinder: phylogenetic orthology inference for comparative
593 genomics. Genome biology 20: 1-14.
594
595 Emms D.M., Kelly S. 2018. STAG: Species Tree Inference from All Genes. bioRxiv.:267914.

596 Fernández R., Gabaldon T., Dessimoz C. 2020. Orthology: definitions, prediction, and impact on
597 species phylogeny inference. In: Scornavacca C., Delsuc F., Galtier N., editors.
598 Phylogenetics in the Genomic Era. Open access book. p. 2.4:1-2.4:14.

599 Fernández R., Kallal R.J., Dimitrov D., Ballesteros J.A., Arnedo M.A., Giribet G., Hormiga G.
600 2018. Phylogenomics, diversification dynamics, and comparative transcriptomics across
601 the spider tree of life. Curr. Biol. 28:1489–1497.

602 Fitch W.M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19:99–113.

603 Gernhard T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–778.

604 Huson D.H., Klöpper T., Lockhart P.J., Steel M.A. 2005. Reconstruction of reticulate networks
605 from gene trees. In: Miyano S., Mesirov J., Kasif S., Istrail S., Pevzner P.A., Waterman
606 M., editors. Research in Computational Molecular Biology. Berlin, Heidelberg: Springer
607 Berlin Heidelberg. p. 233–249.

608 Kallal R.J., Fernández R., Giribet G., Hormiga G. 2018. A phylotranscriptomic backbone of the
609 orb-weaving spider family Araneidae (Arachnida, Araneae) supported by multiple
610 methodological approaches. Mol. Phylogenet. Evol. 126:129–140.

611 Kapli P., Yang Z., Telford M.J. 2020. Phylogenetic tree building in the genomic age. Nat. Rev.
612 Genet. 21:428–444.

613 Koonin E.V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39:309–
614 338.

615 Legried B., Molloy E.K., Warnow T., Roch S. 2020. Polynomial-time statistical estimation of
616 species trees under gene duplication and loss. J. Comput. Biol.:cmb.2020.0424.

617 Li L. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res.
618 13:2178–2189.

619 Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

620 Mallo D., de Oliveira Martins L., Posada D. 2016. SimPhy: phylogenomic simulation of gene,
621 locus, and species trees. Syst. Biol. 65:334–344.

Markin A., Eulenstein O. 2020. Quartet-Based inference methods are statistically consistent under the unified duplication-loss-coalescence model. arXiv.:2004.04299.

Mendes F.K., Hahn M.W. 2018. Why concatenation fails near the anomaly zone. Syst. Biol. 67:158–169.

Mendes F.K., Vanderpool D., Fulton B., Hahn M.W. 2020. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics.:btaa1022.

Mirarab S., Reaz R., Bayzid Md.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 30:i541–i548.

Otto S.P. 2007. The evolutionary consequences of polyploidy. Cell. 131:452–462.

Rasmussen M.D., Kellis M. 2011. A Bayesian approach for fast and accurate gene tree reconstruction. Mol. Biol. Evol. 28:273–290.

Scornavacca C., Delsuc F., Galtier N. 2020. Phylogenetics in the genomic era. Open access book available from https://hal.inria.fr/PGE/.

Siu-Ting K., Torres-Sánchez M., San Mauro D., Wilcockson D., Wilkinson M., Pisani D., O'Connell M.J., Creevey C.J. 2019. Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. Mol. Biol. Evol. 36:1344–1356.

Smith M.L., Hahn M.W. 2021. New approaches for inferring phylogenies in the presence of paralogs. Trends Genet. 37:174–187.

Thomas G.W.C., Ather S.H., Hahn M.W. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. Syst. Biol. 66:1007–1018.

Vanderpool D., Minh B.Q., Lanfear R., Hughes D., Murali S., Harris R.A., Raveendran M., Muzny D.M., Hibbins M.S., Williamson R.J., Gibbs R.A., Worley K.C., Rogers J., Hahn M.W. 2020. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. PLOS Biol. 18:e3000954.

Yan Z., Smith M.L. Du P., Hahn M.W., Nakhleh L. 2021. Species Tree Inference Methods Intended to Deal with Incomplete Lineage Sorting Are Robust to the Presence of Paralogs. Syst. Biol. syab056, https://doi.org/10.1093/sysbio/syab056

Yang Y., Smith S.A. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. Mol. Biol. Evol. 31:3081–3092.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153.

655 Zhang C., Scornavacca C., Molloy E.K., Mirarab S. 2020. ASTRAL-Pro: quartet-based species-
656 tree inference despite paralogy. Mol. Biol. Evol. 37:3292–3307.

657

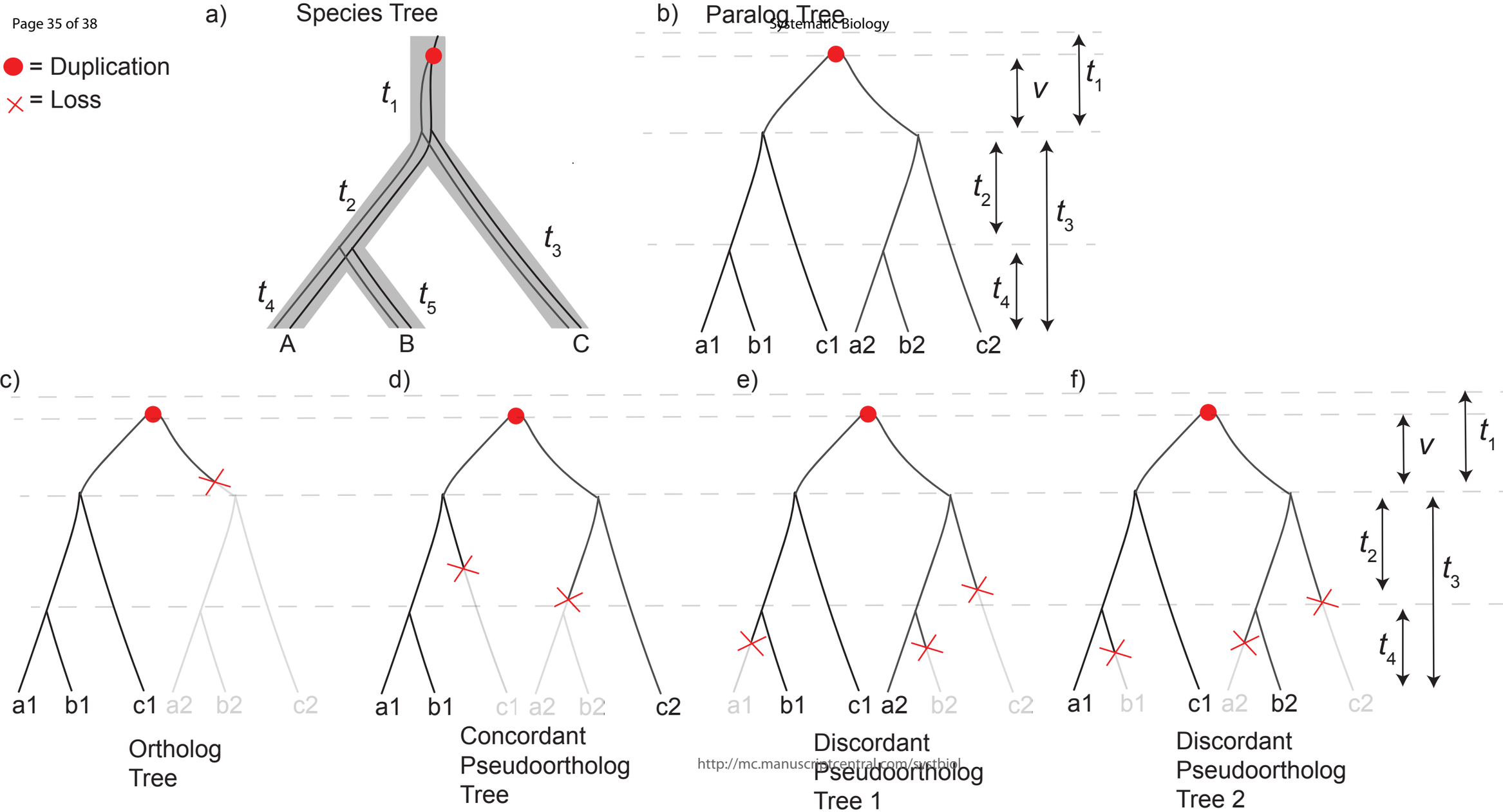
FIGURE CAPTIONS

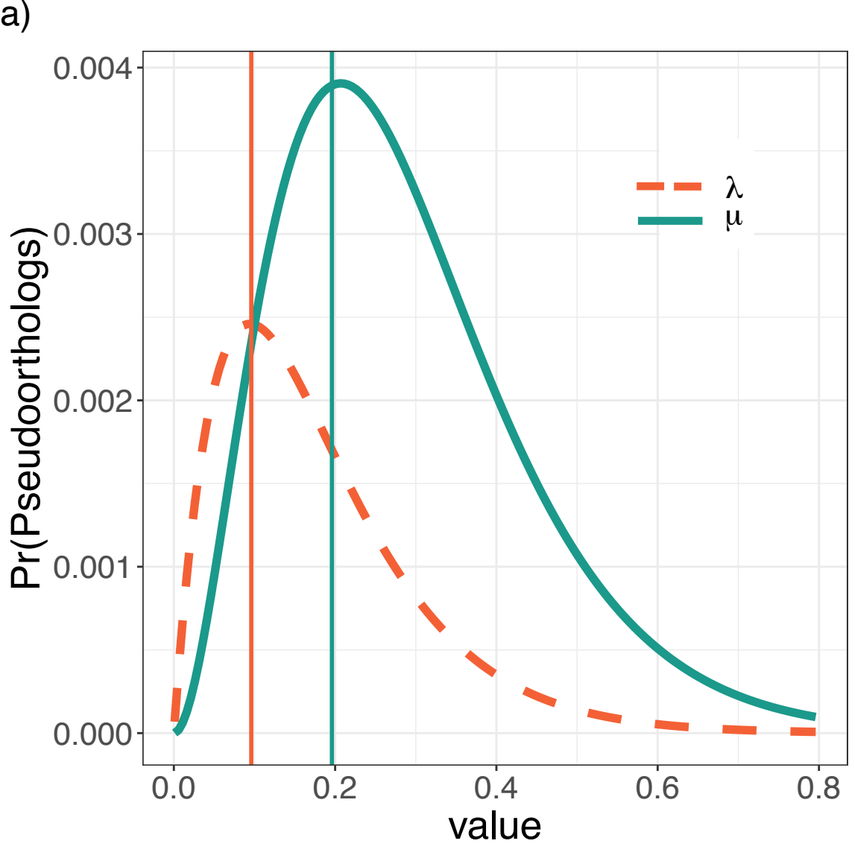
Figure 1. Orthologs and Pseudoorthologs. a) A rooted three-taxon species tree showing the relationships between Species A, B, and C is depicted by the grey outline. Within the tree, the duplication is indicated by a red dot, and the two daughter paralog trees are drawn. b) The full paralog tree depicting relationships between all gene copies. The lengths of all branches are shown on the right. c-f) The different relationships possible when a single copy in each species is present. Red X's indicate loss events. c) Orthologs require at least one loss, d) concordant pseudoorthologs require at least two losses, and e, f) discordant pseudoorthologs require at least three losses.

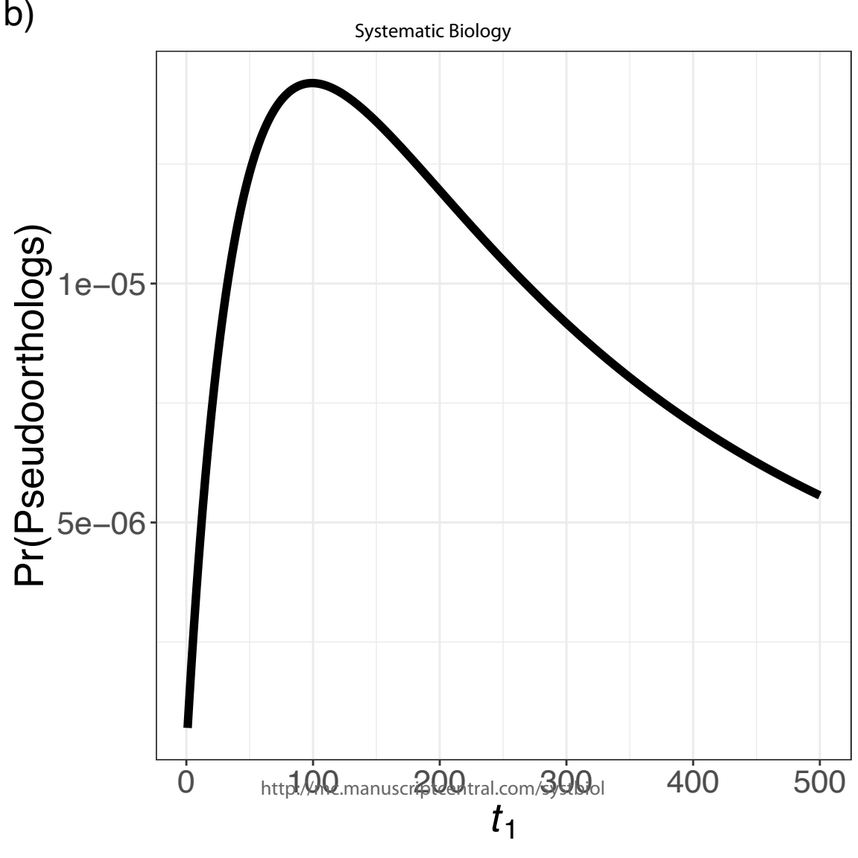
Figure 2. Effects of varying parameters on the probability of pseudoorthologs. a) The unconditional probability of pseudoorthologs is maximized at intermediate values of μ and λ. b) The unconditional probability of pseudoorthologs is maximized at intermediate values of branch length t_1. c) The relative unconditional probabilities of concordant and discordant pseudoorthologs depend on the ratio of branch lengths t_2 and t_4 (or t_5). As t_2 gets larger (x-axis becomes more positive) the unconditional probability of concordant pseudoorthologs increases and the unconditional probability of discordant pseudoorthologs decreases. The probability of one discordant pseudoortholog is shown, but the values are identical for the other discordant pseudoortholog.

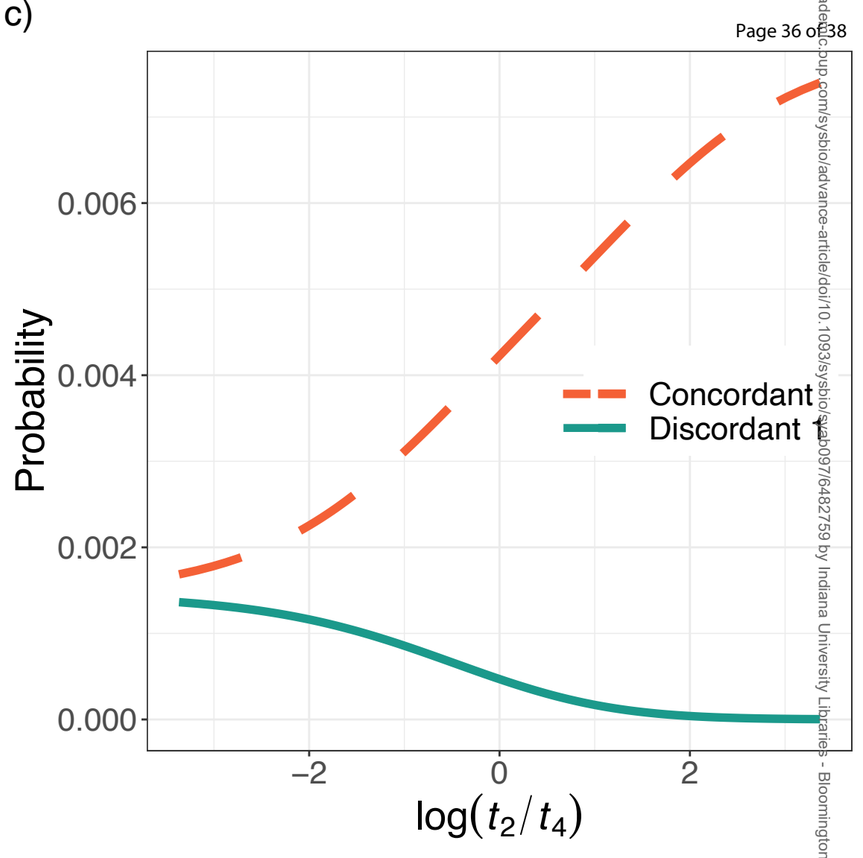
Figure 3. Probabilities of orthologs, pseudoorthologs, and discordance. Here, branch length t_3 = 198.9 mya. Branch length t_1 varies from 0.0001 to 200 mya, while branch length t_4 varies from

682 0.0001 to 198.8 mya; branch length t_2 is constrained such that the sum of t_2 and t_4 equals t_3. a)
683 The conditional probability of orthologs given that a single copy is present in each species, with
684 moderate rates of duplication and loss (λ=0.002 per my, μ=0.002 per my). b) The ratio of the
685 concordant topology (orthologs and concordant pseudoorthologs) to one discordant topology
686 given that a single copy is present in each species, with moderate rates of duplication and loss
687 (λ=0.002, μ=0.002). c) The conditional probability of orthologs given that a single copy is
688 present in each species with high rates of duplication and loss (λ=0.005, μ=0.005). d) The ratio
689 of the concordant topology (orthologs and concordant pseudoorthologs) to one discordant
690 topology given that a single copy is present in each species with high rates of duplication and
691 loss (λ=0.005, μ=0.005).
692
693 **Figure 4.** The species tree and parameters that minimize the ratio of the concordant topology to
694 the two discordant topologies. Discordant probabilities and ratios refer only to discordant
695 pseudoortholog 1, but the probabilities of the two discordant pseudoorthologs are equal.
696 Probabilities are conditional on sampling a single copy per species. To facilitate visualization,
697 the internal branch uniting Species A and B is not drawn to scale.









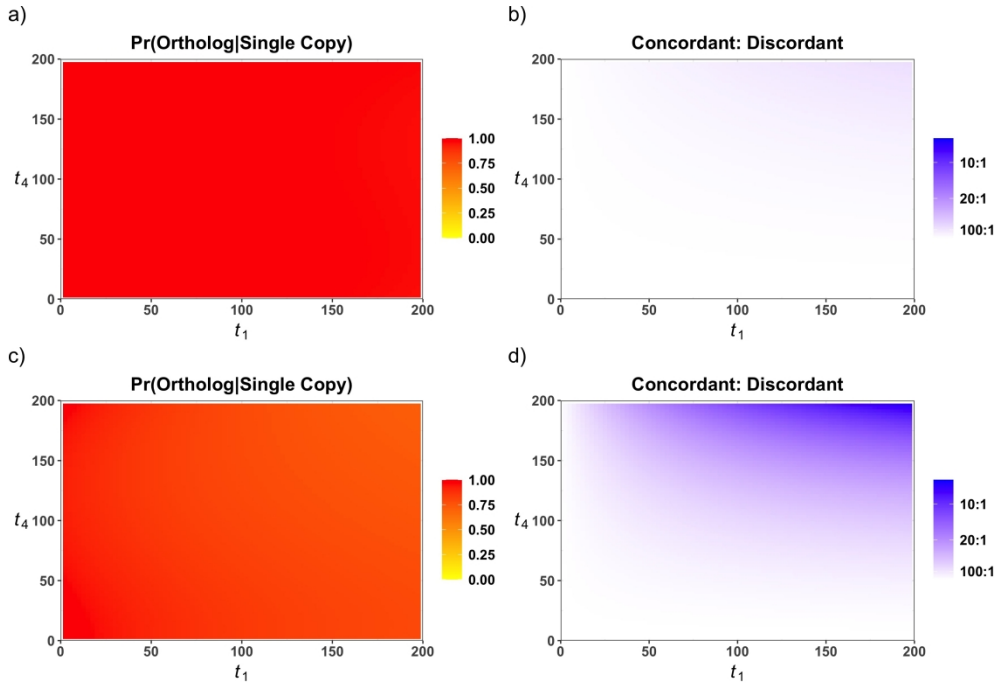
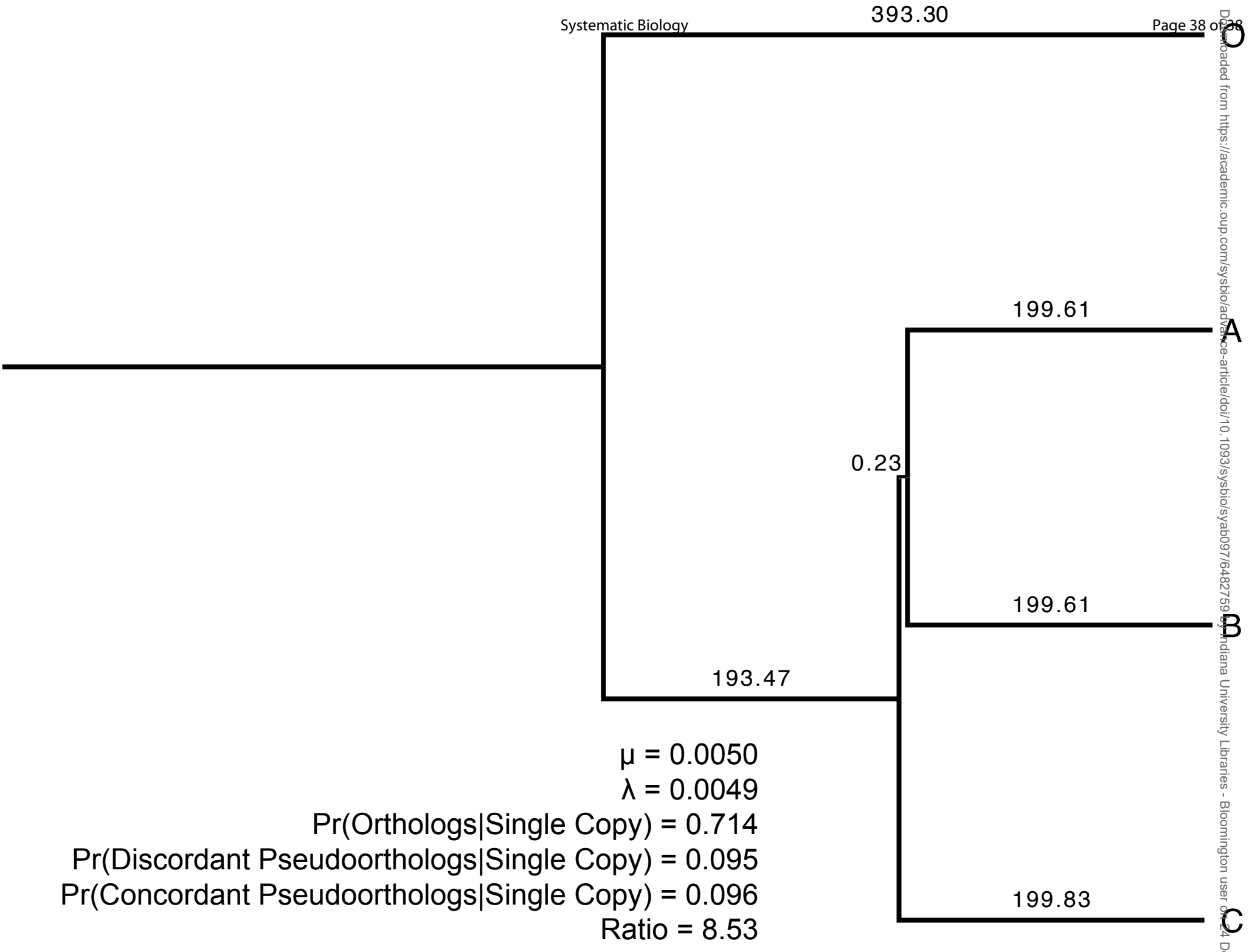


Figure 3. Probabilities of orthologs, pseudoorthologs, and discordance. Here, branch length t_3 = 198.9 mya. Branch length t_1 varies from 0.0001 to 200 mya, while branch length t_4 varies from 0.0001 to 198.8 mya; branch length t_2 is constrained such that the sum of t_2 and t_4 equals t_3. a) The conditional probability of orthologs given that a single copy is present in each species, with moderate rates of duplication and loss (λ=0.002 per my, μ=0.002 per my). b) The ratio of the concordant topology (orthologs and concordant pseudoorthologs) to one discordant topology given that a single copy is present in each species, with moderate rates of duplication and loss (λ=0.002, μ=0.002). c) The conditional probability of orthologs given that a single copy is present in each species with high rates of duplication and loss (λ=0.005, μ=0.005). d) The ratio of the concordant topology (orthologs and concordant pseudoorthologs) to one discordant topology given that a single copy is present in each species with high rates of duplication and loss (λ=0.005, μ=0.005).

1270x867mm (72 x 72 DPI)



Received: 1 February 2017 | Revised: 16 May 2017 | Accepted: 22 May 2017

DOI: 10.1111/mec.14223

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

Demographic model selection using random forests and the site frequency spectrum

Megan L. Smith[1] | Megan Ruffley[2,3] | Anahí Espíndola[2,3] | David C. Tank[2,3] | Jack Sullivan[2,3] | Bryan C. Carstens[1]

[1]Department of Evolution, Ecology & Organismal Biology, The Ohio State University, Columbus, OH, USA

[2]Department of Biological Sciences, University of Idaho, Moscow, ID, USA

[3]Biological Sciences, Institute for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, Moscow, ID, USA

Correspondence
Bryan C. Carstens, Department of Evolution, Ecology & Organismal Biology, The Ohio State University, Columbus, OH, USA.
Email: carstens.12@osu.edu

Funding information
Division of Environmental Biology, Grant/Award Number: DEB 14575199, DEB 1457726, DG-1343012; US National Science Foundation; NSF GRFP; Ohio State University

Abstract

Phylogeographic data sets have grown from tens to thousands of loci in recent years, but extant statistical methods do not take full advantage of these large data sets. For example, approximate Bayesian computation (ABC) is a commonly used method for the explicit comparison of alternate demographic histories, but it is limited by the "curse of dimensionality" and issues related to the simulation and summarization of data when applied to next-generation sequencing (NGS) data sets. We implement here several improvements to overcome these difficulties. We use a Random Forest (RF) classifier for model selection to circumvent the curse of dimensionality and apply a binned representation of the multidimensional site frequency spectrum (mSFS) to address issues related to the simulation and summarization of large SNP data sets. We evaluate the performance of these improvements using simulation and find low overall error rates (~7%). We then apply the approach to data from *Haplotrema vancouverense*, a land snail endemic to the Pacific Northwest of North America. Fifteen demographic models were compared, and our results support a model of recent dispersal from coastal to inland rainforests. Our results demonstrate that binning is an effective strategy for the construction of a mSFS and imply that the statistical power of RF when applied to demographic model selection is at least comparable to traditional ABC algorithms. Importantly, by combining these strategies, large sets of models with differing numbers of populations can be evaluated.

KEYWORDS
machine learning, model selection, phylogeography, RADseq

1 | INTRODUCTION

Since before the term "phylogeography" was coined (Avise et al., 1987), the discipline has developed in response to advances in data-acquisition technology (reviewed in Garrick, Bonatelli, & Hyseni, 2015). Recently, phylogeographic investigations have transformed from traditional studies using data from a handful of genetic loci to contemporary studies where hundreds or thousands of loci are collected (Garrick et al., 2015). With the proliferation of next-generation sequencing (NGS) data sets, researchers can now access genetic data to investigate complex patterns of divergence and diversification in nonmodel species. In recent years, the field has increasingly relied upon model-based methods (Nielsen & Beaumont, 2009). These methods are primarily of two classes: those that estimate parameters under a predefined model and those that compare a number of user-defined models. The former type of approach has expanded recently to methods that are applicable to NGS data sets. For example, sequential Markovian coalescent (SMC) approaches can estimate population size histories and divergence times using whole genomes (Terhorst, Kamm, & Song, 2016). However, such methods

require that researchers identify a model a priori, and are generally limited to relatively simple models that omit many potentially important parameters, due to computational constraints. For example, while Terhorst et al.'s SMC approach can estimate divergence times and population size changes, it does not incorporate gene flow between lineages. Instead, researchers may wish to compare models that include different parameters and determine which model best fits their data, and this has led to an increase in the use of approximate methods, due to the computational challenges of comparing such complex models. A particularly flexible method in this regard is approximate Bayesian computation (ABC; e.g., Beaumont, 2010), which has been used in a wide range of applications outside of population genomics and phylogeography, including ecology, epidemiology and systems biology (Beaumont, 2010).

ABC methods enable researchers to customize demographic models to their empirical system, and allows formalized model selection (Table 1). Under each prespecified model, parameters of interest, θ_i, are drawn from a prior distribution, $\pi(\theta)$, specified by the researcher (step 1). Data, x_i, are then simulated from the distribution of the data given the parameters, $p(x \mid \theta_i)$ (step 2), and a vector of summary statistics, S, is calculated from the simulated and empirical data (step 3). The efficiency of ABC is a result of the optimization. Simulations that exceed a user-defined threshold, ε, as measured by the distance function, $\rho(S(x_i), S(y))$, are rejected (step 4) such that the remaining θ_i constitute the posterior distribution. If data are simulated under multiple models, the proportions of simulations that each model contributes to the posterior distribution correspond to the posterior probabilities of the models under consideration (step 5). ABC was developed in the context of a handful of microsatellite loci (Pritchard, Seielstad, Perez-Lezaun, & Feldman, 1999), but in theory can be extended to any amount of data. In practice, however, extending it to large NGS data sets is difficult due to the "curse of dimensionality" (Blum, 2010). This term describes the situation that occurs as the vector of summary statistics grows large, as would be the case if data were summarized on a locus-by-locus basis for hundreds to thousands of loci, and simulation of data near the vector requires an increasingly large number of simulations, which leads to high error rates. Although ABC has been applied to large NGS data sets (e.g., Roux et al., 2010; Veeramah et al., 2015), these applications have typically required that researchers summarize thousands of loci using a small vector of summary statistics (e.g., in Roux et al., 2010; the average and standard deviation over loci for 11 summary statistics). Summarizing data from 1,000s of loci with dozens of summary statistics results in a substantial loss of the information content of the data and limits the number of models that researchers have statistical power to distinguish. While methods have been suggested to guide researchers in their choice of summary statistics (e.g., partial least-squares transformation; Wegmann, Leuenberger, & Excoffier, 2009), they still result in a large decrease in the information content of the data. Some recent studies have used the bins of the site frequency spectrum (SFS) as a summary statistic for ABC inference (e.g., Boitard, Rodriguez, Jay, Mona, & Austerlitz, 2016; Prates, Rivera, Rodrigues, & Carnaval, 2016; Stocks, Siol, Lascoux, & De Mita, 2014; Xue & Hickerson, 2015), but these approaches have not taken advantages of joint or multidimensional SFS (mSFS). Consideration of the mSFS is necessary to make inferences about multiple populations, but the dimensionality of the mSFS increases as the number of individuals and populations sampled increases such that the number of bins in the joint or multidimensional SFS becomes very large, and the "curse of dimensionality" becomes a limiting factor. One possible solution to the limitations of ABC that would allow researchers to avoid reducing their data to a small number of summary statistics is to follow Pudlo et al. (2015) in replacing the traditional rejection step (steps 4-5; Table 1) with a machine-learning approach such as Random Forests (RF) for model selection.

In the RF approach to phylogeographic model selection, the data simulation and summarization steps (Table 1, steps 1-3) remain unchanged from the traditional ABC algorithm. However, instead of using a rejection step that relies on a specified distance function between the observed and simulated data, model selection proceeds using a classification forest. This forest consists of hundreds of

TABLE 1 Comparison of the ABC and RF approaches to demographic model selection

Comparison of ABC and RF algorithms for model selection	
Both ABC and RF	
1. Draw parameters θ_i from the prior distribution $\pi(\theta)$.	
2. Simulate data x_i from the distribution of the data given the parameters $p(x \mid \theta_i)$.	
3. Summarize the data using some statistic $S(x_i)$.	
ABC	**RF**
4. Reject θ_i when some function $\rho(S(xi),S(y))$ measuring the distance between the simulated and observed data exceeds a user-defined threshold.	4. Train a RF classifier using $S(x_i)$ as predictor variables and the model under which the $S(x_i)$ were simulated as the response variable.
5. The retained θ_i approximate the posterior distribution and are used to approximate model posterior probabilities.	5. Apply classifier to the observed data set to choose the best model.
	6. Estimate the probability of misclassification for the observed data using oob error rates.

decision trees and is trained on the simulated data, with the summary statistics serving as the predictor variables and the generating model serving as the response variable. Once built, this classifier can be applied to the observed data. Decision trees will favour (i.e., vote for) a particular model, and the model receiving the most votes will be selected as the best model. Although this approach does not include the approximation of the posterior probability, in contrast to ABC approaches that utilize a rejection step, uncertainty in model selection can be estimated using the error rates of the constructed classifier. Both experimental (Hastie, Tibshirani, & Friedman, 2009) and theoretical (Biau, 2012; Scornet, Biau, & Vert, 2015) justifications of RF have been offered, with RF shown to be robust both to correlations between predictor variables (here, the summary statistics) and to the inclusion of a large number of noisy predictors. An additional advantage of the RF approach is the reduction in computational effort required for model selection, as >50-fold gains in computational efficiency have been reported (Pudlo et al., 2015).

Although the data simulation and summary statistic calculation steps (steps 2-3 in Table 1) of the ABC algorithm may be extended to NGS data sets from a first-principles argument, issues arise in the implementation. First, the simulation of data scales linearly with the number of loci and thus becomes computationally intensive when the data sets in question are large (Sousa & Hey, 2013). Additionally, calculating a set of traditional summary statistics for each locus for use as summary statistics is impractical given the large number of loci. Although it is possible to calculate certain traditional summary statistics directly from the SFS, rather than on a locus-by-locus basis, such a calculation results in the loss of much of the information content of the data (Sainudiin et al., 2011).

In response to these issues, we explore the use of the multidimensional site frequency spectrum (mSFS; the joint distribution of allele frequencies across three or more populations) for data simulation and summarization in the RF model selection algorithm. The mSFS is a useful summary of the SNP data sets that are frequently collected using NGS methods, and can be considered a complete summary of the data when all polymorphic sites are independent (i.e., unlinked) and biallelic (e.g., Gutenkunst, Hernandez, Williamson, & Bustamante, 2009). Furthermore, the mSFS is expected to reflect demographic events including expansion, divergence and migration (Gutenkunst et al., 2009), although inferences based on the SFS may be inaccurate when too few segregating sites are sampled (Terhorst & Song, 2015). To address this issue, we apply a binning approach to coarsen the mSFS. The use of the mSFS for data summary can also facilitate data simulation; for example, the coalescent simulation program fastsimcoal2 (FSC2) uses a continuous time approximation to calculate the mSFS from simulated SNP data (Excoffier, Dupanloup, Huerta-SÃnchez, Sousa, & Foll, 2013). Here, we propose an approach to phylogeographic model selection that combines the use of a RF classifier with the use of the mSFS to summarize NGS data. We apply this approach to evaluate demographic models in *Haplotrema vancouverense*, a land snail endemic to temperate rainforests of the Pacific Northwest of North America (PNW).

2 | MATERIALS AND METHODS

2.1 | Study system and models

2.1.1 | Study system

The PNW of North America can be divided into three distinct regions: the Cascades and Coastal Ranges in the west, the Northern Rocky Mountains in the east and the intervening Columbia Plateau (e.g., Figure 1; Brunsfeld, Sullivan, Soltis, & Soltis, 2000). The coastal and inland mountain ranges are characterized by mesic, temperate coniferous forests, but the intervening basin is characterized by a shrub–steppe ecosystem generated by the rain shadow of the Cascade Range that has developed since its orogeny in the early Pliocene. The Okanogan Highlands to the north and the Central Oregonian highlands to the south partially mitigate the ecological isolation of the inland and coastal forests, but the Columbia Plateau has nevertheless been a substantial barrier to dispersal for many of the taxa endemic to these temperate forests (e.g., Carstens, Brunsfeld, Demboski, Good, & Sullivan, 2005). In addition to being influenced by mountain formation, the distributions of taxa in the rainforests of the PNW have likely been impacted by climatic fluctuations throughout the Pleistocene (Pielou, 2008). Glaciers formed and retreated several times during these fluctuations, covering large portions of the northern parts of species' current ranges. Thus, species may have been entirely eliminated in the northern parts of their ranges or may have survived in small isolated glacial refugia.

Several biogeographic hypotheses have been proposed to explain the disjunct distribution of the PNW mesic forest endemics (reviewed in Brunsfeld et al., 2000). Here, we explore models that include from one to three glacial refugia (South Cascades, North Cascades and Clearwater River drainages). In one class of models, no refugia persisted in the inland region, and these models posit dispersal to the inland via either a southern or a northern route. In addition, to test whether or not there was population structure present, we evaluated models that included from one to four distinct populations (South Cascades, North Cascades, Clearwater River drainages and northern Idaho drainages). In total, we include 15 demographic models that differed in the number of populations, the number and location of refugia and the dispersal route (Figure 1; Fig. S1). We applied the approach proposed here to *Haplotrema vancouverense*, a land snail endemic to the PNW. No previous work has used genomic data to investigate the demographic history of this species. However, one study used environmental data to predict that *H. vancouverense* did not harbour cryptic diversity across the Columbia Basin (Espíndola et al., 2016).

2.2 | Specimen collection and data generation

Samples were collected for this study during the spring of 2015 and 2016, in addition to loans provided by the Idaho Fish and Game and museum collections (the Royal British Columbia Museum and the Florida Museum of Natural History). In total, we acquired 77 snails

from throughout the range of *H. vancouverense* (Figure 2; Table S1). This included 31 snails from 24 localities in the northern and southern Cascades and 46 snails from 18 localities in the Clearwater River and northern Idaho drainages. After collection, snails were preserved in 95% ethanol and DNA was extracted using Qiagen DNeasy Blood and Tissue Kits (Qiagen, Hilden, Germany) following the manufacturer's protocol. Prior to library preparation, DNA was quantified on a Qubit fluorometer (Life Technologies), and 200–300 nanograms of DNA was used for library preparation.

Library preparation followed the double-digest restriction-associated DNA (ddRAD) sequencing protocol developed in Peterson, Weber, Kay, Fisher, & Hoekstra, 2012, with modifications. DNA was digested using the restriction enzymes SbfII and MspI (New England Biolabs, USA), and adapters were ligated using T4 ligase (New England Biolabs). Ligated products were cleaned using magnetic beads in a PEG/NaCl buffer (Rohland & Reich, 2012). A subset of the ligation products was amplified and analysed by qPCR using the library quantification kit for Illumina libraries (KAPA Biosystems, USA) to ensure that no adapter had failed to ligate during the ligation step. All ligation products were quantified on the Qubit fluorometer (Life Technologies) and pooled across index groups in equimolar concentrations. 10–20 nanograms of this pool was used in each subsequent PCR. PCRs used the Phusion Master Mix (Thermo Fisher Scientific, USA) and were run for an initial step of 30 s at 98°C, followed by 16 cycles of 5 s at 98°C, 25 s at 60°C and 10 s at 72°C and a final extension for 5 min at 72°C. To minimize PCR bias, reactions were replicated seven times for each index group, and products were pooled within index groups. We analysed 4 μl of this pooled PCR product on a 1% agarose gel. A second clean-up using magnetic beads in a PEG/NaCl buffer (Rohland & Reich, 2012) was performed. Finally, PCR products were quantified on the Qubit fluorometer (Life Technologies) prior to selection for 300- to 600-bp fragments using the Blue Pippin (Sage Science, USA) following manufacturer's standard protocols. The remaining products were quantified using the Qubit fluorometer (Life Technologies) and the Bioanalyzer (Agilent Technologies, USA) before being pooled and sent for sequencing on an Illumina Hi-Seq at the Genomics Shared Resource Center at Ohio State University.

2.3 | Bioinformatics

Raw sequence reads were demultiplexed and processed using PYRAD (Eaton, 2014). Sites with a Phred quality score <20 were masked with Ns, and reads with more than four Ns were discarded. A minimum of ten reads was required for a locus to be called within an individual. Filtered reads were clustered using the program VSEARCH v.2.0.2 (https://github.com/torognes/vsearch) and aligned using MUSCLE v.3.8.31 (Edgar, 2004) under a clustering threshold of 85%. Consensus sequences with more than three heterozygous sites or more than two haplotypes for an individual were discarded, and loci represented in fewer than 60 per cent of individuals were discarded. Cut-sites and adapters were removed from sequences using the strict filtering in PYRAD.

To deal with missing data when constructing the mSFS, we applied a downsampling approach to maximize the number of SNPs included in the mSFS. A threshold of 50% was set in each population, meaning that only SNPs scored in at least half of the individuals in each population would be used in downstream analyses. For SNPs that exceeded this threshold, we randomly subsampled alleles. We repeated this downsampling approach ten times to create ten different mSFS to be used in downstream analyses in an attempt to account for rare alleles potentially missed during the downsampling procedure. Downsampling followed Thomé and Carstens (2016) and was performed using custom PYTHON scripts modified from scripts developed by J. Satler (https://github.com/jordansatler; modified version at https://github.com/meganlsmith). This approach was chosen over including only loci sampled across all individuals because such an approach would have limited the number of SNPs included in the

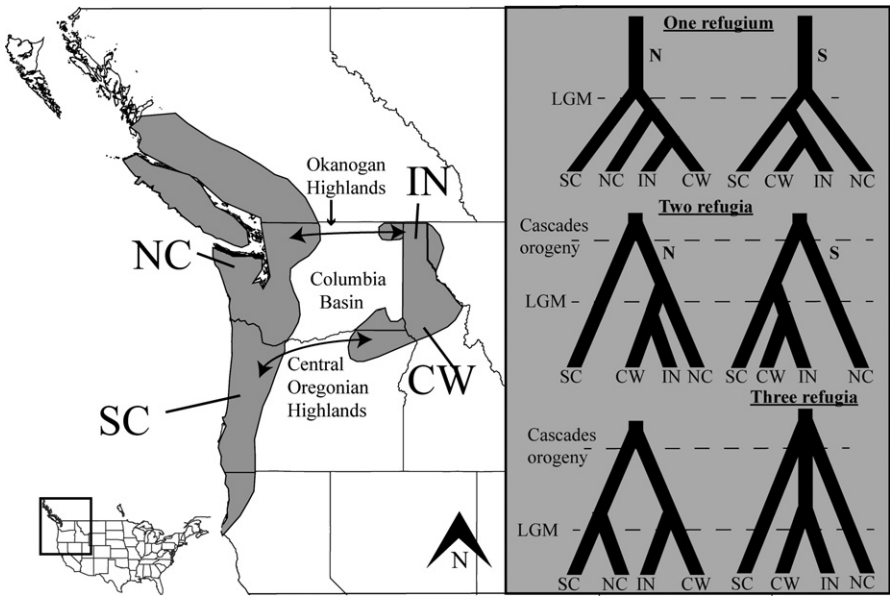


FIGURE 1 Map of the PNW illustrating the models tested in this study. NC, North Cascades; SC, South Cascades; IN, Northern Inland Drainages; CW, Clearwater drainages. The models tested included one to three refugia. When there were no inland refugia, dispersal could occur via either a northern or southern route. Additional models tested (Fig. S1) included from one to four populations. The heights of the bars indicate the time since colonization of the region τ_{col}, with taller bars indicating older populations. The shaded region on the map marks the distribution of *Haplotrema vancouverense*, reproduced from Burke (2013)

study and has been shown to bias parameter estimates due to the nonrandom sampling of genealogies (Huang & Knowles, 2014).

2.4 | Random forest model selection using the mSFS as a summary statistic

2.4.1 | Data simulation and summarization

The RF approach to model selection (Figure 3) follows the algorithm for RF model selection presented in Table 1. Parameters were drawn from prior distributions (Table S2) under each of the fifteen models considered (Figure 3; Step 1). mSFS were simulated in FSC2 (Excoffier et al., 2013) under each model, using a folded mSFS with a number of SNPs equivalent to the observed mSFS (Figure 3; Step 2). Monomorphic sites were not considered, and 10,000 replicate mSFS were simulated under each model in FSC2, leading to a total prior of 150,000 mSFS.

Given the number of populations included as well as the number of SNPs obtained by our sequencing protocol (see Results), use of all bins from the mSFS could result in limited coverage across the mSFS and thus to poor estimates of the mSFS; therefore, we used a custom Python script (https://github.com/meganlsmith) to coarsen the mSFS (Figure 3, Step 3). For example, for the "quartets" data set, SNPs were categorized based on which quartile they belonged to in each population, and all combinations of quartiles across populations were used as bins for a final data set consisting of 256 bins. In this example, the first bin would consist of SNPs occurring at a frequency <¼ in all four populations. We tested other binning strategies with the number of classes ranging from three to ten, enabling a joint exploration of the coarseness of the mSFS, the accuracy of model selection and the computational requirements of the classification procedure.

2.4.2 | Choosing the optimal binning strategy

To determine the optimal binning strategy, eight RF classifiers were constructed using the simulated data (i.e., Figure 3, Step 4), one at each level of mSFS coarseness considered here (i.e., 3–10 classes per population). Each classifier was constructed with 500 trees using the R package "ABCRF" (Pudlo et al., 2015), with the bins of the mSFS treated as the predictor variables and the generating model for each simulated data set treated as the response variable. At each node in each decision tree, the RF classifier considers a bin of the mSFS and constructs a binary decision rule based on the number of SNPs in the bin.

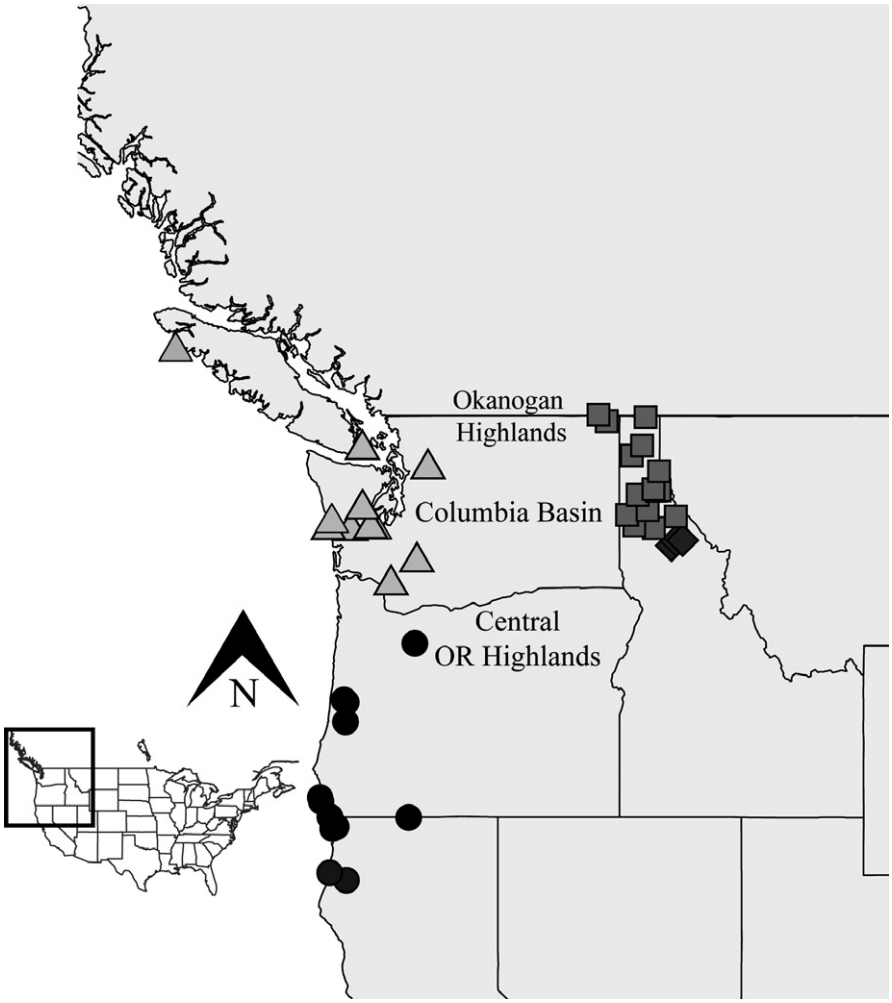


FIGURE 2 Collection localities for *H. vancouverense*. North Cascades = triangles; South Cascades = circles; Northern Inland Drainages = squares; Clearwater drainages = diamonds

When this classifier is applied to other data sets, it makes decisions at each node until it reaches a leaf of the decision tree, which in this instance is a model index. When a leaf is reached, the decision tree is said to "vote" for the model index assigned to that leaf. Each decision tree is constructed in reference to only a portion of the training data set, minimizing the correlation between decision trees. Prior to construction of the random forest, columns in which there was no variance in the entire prior (e.g., bins that contained no SNPs for any of the simulated data sets) were removed from the prior. These same columns were removed from the observed data set.

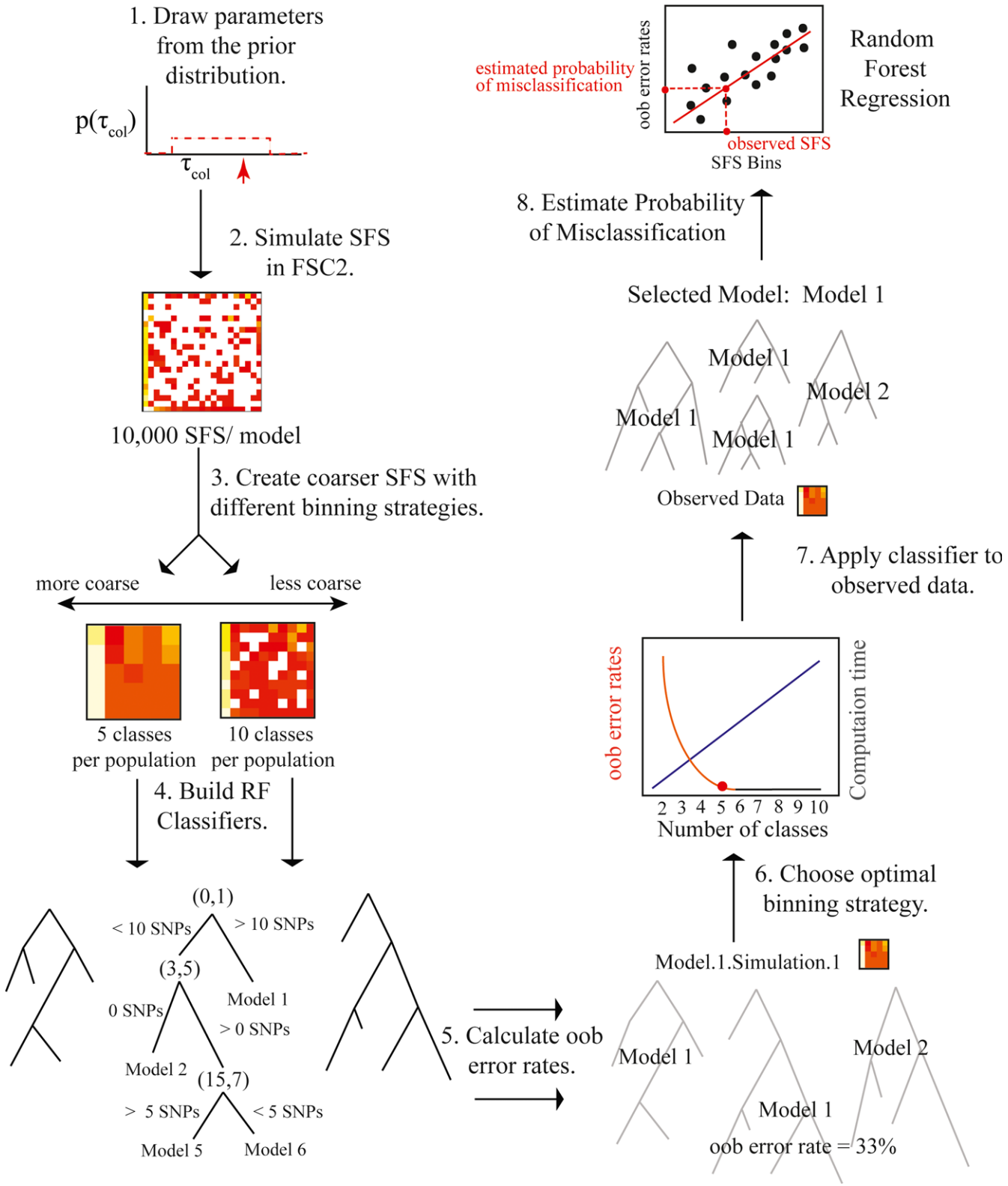


FIGURE 3 Flow chart illustrating the RF approach to model selection [Colour figure can be viewed at wileyonlinelibrary.com]

Because only a portion of the prior is used in the construction of each decision tree in RF classification, the error rate of the classifier can be assessed using the "out-of-the-bag" (oob) error rates (Figure 3; Step 5). Oob error rates are calculated by considering only decision trees constructed without reference to a particular element of the prior. For each simulated mSFS, we used a smaller classifier that consisted only of trees constructed without reference to the mSFS in question. We applied this classifier to the simulated mSFS and calculated the proportion of trees that voted for an incorrect model; this is the oob error rate for the simulated mSFS. To choose the optimal binning strategy (Figure 3, Step 6), we plotted the average misclassification rate and the computation effort required as a function of the binning strategy.

2.4.3 | Model selection and the misclassification rate

After the optimal binning strategy was determined, we applied the corresponding classifier to the observed data (Figure 3, Step 7). The "predict" function in the "abcrf" package was used to select the best model for the observed data, which was the model receiving the most votes (i.e., the model selected by the largest number of decision trees). One limitation of the RF approach is that the number of votes allocated to different models has no direct relationship to the posterior probabilities of the models and may be a poor measure of the probability of misclassification for the observed data. Following Pudlo et al. (2015), we estimated the probability of misclassification in a second step by regressing over the selection error in the prior to build a regression RF, in which the oob error rate is the response variable and the mSFS bins are the predictor variables. We then applied this RF to the observed data to estimate the probability of misclassification for the observed model (Figure 3; Step 8), again using the "predict" function in the R package "abcrf" (Pudlo et al., 2015). A Python script that simulates data, constructs a reference table, builds a classifier, selects the best model for the empirical data and calculates error rates and the probability of misclassification using FSC2 and the R package "abcrf" is available on github (https://github.com/meganlsmith).

To assess the power of the RF approach, we simulated 100 mSFS under each of the 15 models (Fig. S1), drawing priors from the same distributions used in model selection (Information on Prior Distributions; Table S2). We used custom python scripts (https://github.com/meganlsmith) to coarsen the simulated mSFS using five classes. We then applied the RF classifier built from the quintets prior to each of the simulated data sets using the "predict" function in the R package "abcrf" (Pudlo et al., 2015) and recorded which model was selected for each replicate.

2.5 | AIC-based model selection

To validate the results of our model selection using RF, we compared the above results to a commonly used information theoretic approach to phylogeographic model selection with NGS data sets (e.g., Carstens et al., 2013), where model selection in FSC2 followed the procedure suggested in Excoffier et al. (2013). FSC2 maximizes the composite likelihood of the observed data under an arbitrary number of models, and Akaike information theory can then be used to select among several tested models. The Brent algorithm implemented in FSC2 was used for parameter optimization, with parameter optimization replicated 100 times. For each replicate, 100,000 simulations were used for the calculation of the composite likelihood and 40 cycles of the Brent algorithm were used for parameter optimization. The maximum-likelihood estimates for the parameters were then fixed, and the likelihood was approximated for each model across 100 different replicates. The maximum likelihood across these 100 replicates for each model was used in model comparison. AIC scores were then calculated and converted to model weights as in Excoffier et al. (2013).

To assess the power of FSC2 to distinguish among the tested models, we used the same 100 simulated mSFS as in the RF power analysis. We used 100,000 simulations for the calculation of the composite likelihood, and 40 cycles of the Brent algorithm were used for parameter optimization. The likelihood was approximated for each model and used in model comparisons. AIC scores were calculated and converted to model weights as in Excoffier et al. (2013), and we recorded which model was selected for each replicate. Due to computational constraints, we did not perform the replication recommended for model selection in FSC2, as was done for the observed data. We also conducted a conventional ABC analysis (see Supporting Information).

3 | RESULTS

3.1 | Bioinformatics

After the filtering thresholds were applied, 1,943 loci were called in 77 individuals. This resulted in 1,716 unlinked biallelic SNPs and 5,996 total variable sites. When only unlinked SNPs were used, the downsampling approach resulted in data sets including SNPs from 12 alleles per locus from the Clearwater drainages, 14 alleles per locus from the North Cascades, 34 alleles per locus from the northern inland drainages and 17 alleles per locus from the South Cascades. These data sets included between 879 and 908 SNPs.

3.2 | RF model selection with the mSFS as a summary statistic

3.2.1 | Oob error rates and optimal binning strategy

Oob error rates decreased as the number of classes used to build the coarse mSFS increased, until the number of classes reached five (Figure 4). The error rate is no worse for five as opposed to a greater number of classes, and the computation effort increases considerably with larger numbers of classes (Figure 4). We therefore determined that five classes represented the optimal binning strategy for our data, and as such present results only from the "quintets" data set below. Using the "quintets" data set, the overall prior error

rate, calculated using oob error rates, was 6.59 per cent. Error rates varied across models (Fig. S2) and were highest between those models for which the only difference was whether dispersal occurred via a northern or a southern route (Table 2). Misclassification across models with different numbers of populations was less common, and data sets were never classified as belonging to a model having a different number or identity of refugia than the generating model (Table 2). Error rates appeared to plateau in relation to the number of trees used to construct the model, suggesting that more trees did not improve the predictive ability of the RF classifier (Fig. S3).

3.2.2 | Model selection with random forests and AIC-based model selection

In analyses of the "quintets" data set, RF selected the four-population model that included recent southern dispersal to the inland region (Figure 5: Model 1; Table 3). The next best model was similar, but with colonization of the inland region via a northern instead of a southern route (Figure 5: Model 2; Table 3). The probability of misclassification of the best model was estimated to be 0.3514 (corresponding to an approximated posterior probability of 0.6846). The best model did not change between data sets built with different binning strategies, but the probability of misclassification varied across data sets (Table 3). To account for variation in the downsampling procedure, ten downsampling replicates were analysed using five categories per population to bin the data; the best model did not change between data sets, but the misclassification probability of the best model varied across data sets (Table S3). This analysis (constructing the RF from the prior, calculating oob error rates and applying the RF classifier to the observed data) was run on six processors with 24GB RAM and used 78.9 min of CPU time. Under the likelihood-based approach, the best model was a four-population model of recent dispersal to the inland with colonization via a northern route (Figure 5: Model 2; Table 4). The next best model was the same, except that colonization of the inland occurred via a southern route (Figure 5: Model 1; Table 4). This analysis required more than 1,500 CPU hr, largely due to the replication required in calculating the composite likelihood.

3.2.3 | Power analyses in random forests and AIC-based model selection

In the power analysis for the RF approach, the overall error rate was 7.67 per cent (Table S4). The highest error rates were for models 1 and 2 (Fig. S1) at 22 and 42 per cent, respectively. The power analysis in RF used approximately ~285 CPU hr.

In the power analysis for the FSC2 approach, the overall error rate was 3.33 per cent (Table S4). The highest error rates were for models 1 and 2 at 10 and 15 percent, respectively. Although we were not able to perform a full power analysis (with fixed MLE parameter estimates and replicates to approximate the composite likelihood of each model) due to computational constraints, the partial power analysis used approximately 2,205 resource units (~22,205 CPU hr). Model selection results of the conventional ABC

TABLE 2 Summary of the probabilities of different types of classification errors

Misclassification probabilities	
Description of Misclassification	**Probability**
Misclassified as model with a different number of populations	1.44%
Misclassified as model with different number of refugia	0.00%
Misclassified as model with different dispersal route	4.95%

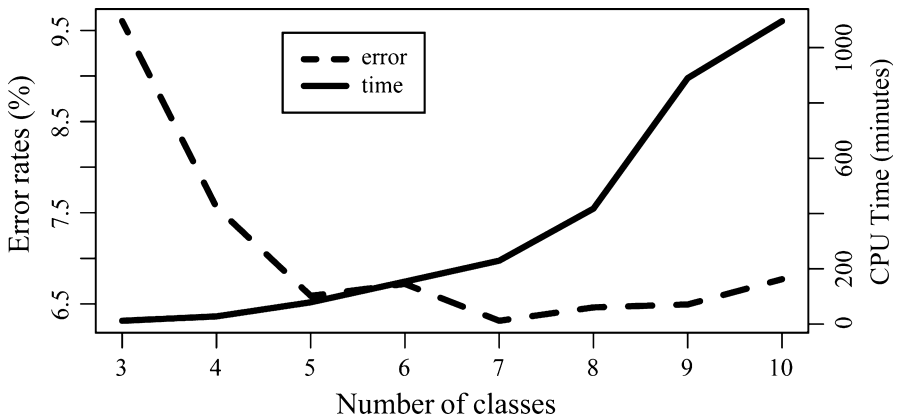


FIGURE 4 Error rates and computation time vs. the number of classes used to construct the mSFS. "Four classes" indicates that there were four categories of SNPs per population, for a total of 256 bins in a four-population multidimensional mSFS. All computations were performed on the Ohio Supercomputer, and CPU time indicates CPU time required to construct a Random Forest from the prior, estimate the oob error rates of the RF and apply this RF to the observed data. For up to six classes, computations were performed on six processors with 24GB of RAM. For seven and eight classes, computations were performed on twelve processors with 48GB of RAM. For nine and ten classes, computations were performed on a twelve processors with 192GB of RAM

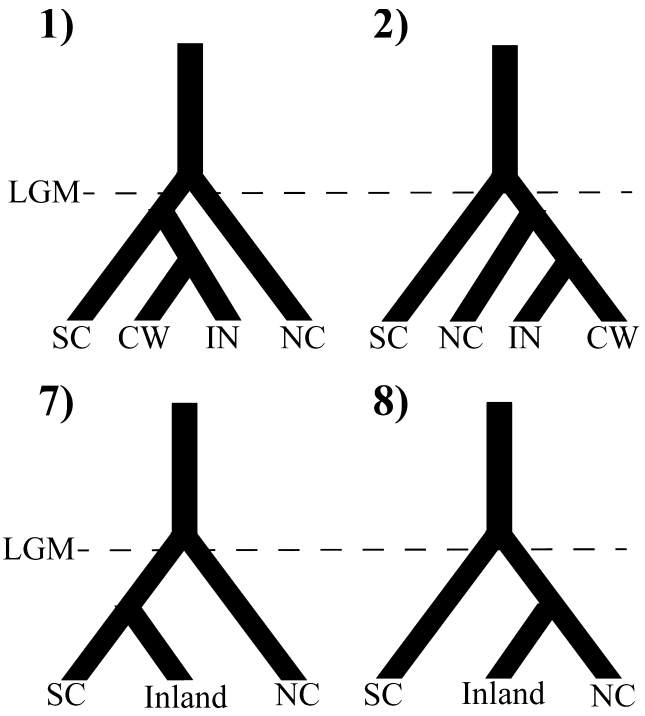


FIGURE 5 The four best models based on ABC and FSC2 results. All models include one refugium in the South Cascades. (1) and (2) include four populations, while (7) and (8) lump the two inland populations together. (1) and (7) posit a southern route of colonization of the inland rainforests, while (2) and (8) posit a northern route of colonization

analysis were similar, but our simulations suggested that the power to detect the best model was lower overall (Supporting Information).

4 | DISCUSSION

4.1 | Model selection using random forests

The combination of RF classification and the binning strategy for mSFS data appear to perform well in the context of phylogeographic model selection, and the use of the RF algorithm for model selection in place of a traditional ABC approach allowed us to circumvent many of the issues associated with using a traditional ABC approach on NGS data sets, with error rates much lower than those obtained when a classical ABC approach was applied to this data (ABC error rate = 30%, Supporting Information). The low error rate obtained in the RF approach to model selection (6.59%) can likely be attributed both to the more efficient approach to model selection and to the more complete summary of the data provided by the mSFS. Computational requirements (Figure 4) were much less than those of FSC2. In comparison with AIC-based methods, RF model selection is favourable in certain situations. Although AIC-based methods, such as FSC2, have proven powerful in certain contexts (Excoffier et al., 2013), the power of such analyses when applied to the smaller NGS data sets frequently collected using protocols such as ddRAD sequencing (Peterson et al., 2012) on nonmodel organisms has not been thoroughly evaluated in most studies using FSC2. Particularly when the number of bins in the mSFS greatly exceeds the number of SNPs, as is likely to occur as the number of populations increases, it may be inappropriate to use the full mSFS due to the reduction in the accuracy of parameter estimations (and thus of the likelihood calculation) that such data sets are expected to provide, as inferences based on SFS with small-to-moderate numbers of SNPs have been shown to be inaccurate (Terhorst & Song, 2015). Although FSC2 and the RF approach had similar power to distinguish among the models we tested, due to the computational requirements, it was difficult to assess the power of FSC2 given the data collected. The approach proposed here has the advantage of oob error rates, which enable an efficient evaluation of the power of the method given the collected data. Then, researchers can generate coarser mSFS according to the characteristics of their data and system.

TABLE 3 Model votes for the four best models and one minus the probability of misclassification of the selected model (an approximation of the posterior probability) for data sets with seven different levels of coarseness (3-10 categories for within population frequencies; 256-10,000 bins). Models 1, 2, 7 and 8 are illustrated in Figure 5

Results of ABC RF Model Selection					
# Categories	Model 1	Model 2	Model 7	Model 8	1-Pr(Misclassification)
3	234	121	80	45	0.6241
4	252	132	44	28	0.7292
5	212	109	72	52	0.6846
6	168	124	74	55	0.6933
7	170	100	56	43	0.6565
8	194	131	48	31	0.6546
9	134	129	61	59	0.6927
10	164	131	65	40	0.6364

While the RF approach has several advantages for model selection, joint estimation of parameters is not straightforward (but see Raynal et al., 2017). Additionally, as the monomorphic cell (the cell with counts of sites without variation) of the mSFS is not used in our approach, the timing of demographic events is relative rather than absolute. For cases when researchers prefer to test explicit a priori hypotheses based on geological data (e.g., Carstens et al., 2013), other approaches (including FSC2) should be preferred. In general, we suggest that parameter estimation using methods such as FSC2 using the model(s) selected following this approach as well as all available SNPs is likely the most effective strategy for non-model systems.

4.2 | Future directions

Model selection using RF has many potential advantages that future investigations should explore. Here, we highlight two such possibilities: (i) testing a large number of models and (ii) species delimitation. Using RF, we were able to test a moderate number ($N = 15$) of

TABLE 4 Results from the four best models, based on the results of the likelihood-based model selection in FSC2. Models differed in the number of populations and the route of dispersal

Model probabilities for the four best models							
Populations	**Refugia**	**Dispersal Route**	**K**	**LnLhood**	**AIC**	Δ_i	**wAIC**
4 (NC, SC, NID, CW)	SC	South	8	−7,351	14,718	3	0.173
4 (NC, SC, NID, CW)	SC	North	8	−7,349	14,715	0	0.827
3 (NC, SC, NID+CW)	SC	South	7	−7,705	15,423	708	0.000
3 (NC, SC, NID+CW)	SC	North	7	−7,712	15,438	723	0.000

NC, North Cascades; SC, South Cascades; NID, Northern Inland Drainages, CW, Clearwater.

demographic models without sacrificing our ability to distinguish between models. The error rate associated with model selection using traditional ABC algorithms appears to increase as the number of models increases, particularly when more than four models are included (Pelletier & Carstens, 2014). Our results suggest that it may be possible to compare a larger number of models using the RF model selection approach, allowing researchers to make fewer assumptions about the historical processes that may have influenced their focal organisms. The out-of-the-bag error rates generated in this approach allow researchers to assess whether they can distinguish among the models tested, given their data, and should thus prevent researchers from testing more models than they have the power to differentiate among. As with other approaches to demographic model selection, we were still limited in the number of models that we could compare, and our results can only highlight the best model among those tested. Although assessing model fit can help researchers understand how well their data fit a model, such an approach is not straightforward with the RF approach.

Additionally, we were able to compare models that included different numbers of populations with a low misclassification rate (1.44%). It has been challenging to use ABC in such cases because some of the most useful summary statistics are based on comparisons within and between populations (e.g., Hickerson, Dolman, & Moritz, 2006), and the summary statistic vectors used in such a comparison would necessarily have different dimensionalities. Here, we were able to circumvent this issue by calculating the mSFS as if there were four populations, regardless of the number of populations used to generate the SNP data. Our results suggest that the model selection approach implemented here could be used for population and potentially species delimitation. Additionally, although we focused on Random Forests here, other machine-learning algorithms have been used to infer demographic histories (e.g., Deep Learning; Sheehan & Song, 2016), and future work should investigate the use of these algorithms in conjunction with the binned mSFS.

4.3 | Empirical results

Results from both ABC (Table 3) and FSC2 (Table 4) suggest that *H. vancouverense* survived in one or more refugia in the south Cascades throughout the Pleistocene glacial cycles. A previous investigation using environmental and taxonomic data to make predictions concerning the evolutionary histories of organisms predicted that *H. vancouverense* colonized the inland after the Pleistocene (Espíndola et al., 2016), and the results presented here support this prediction. Following glaciation, *H. vancouverense* expanded its range north to the North Cascades and east to the Northern Rocky Mountains. Our results were incongruent across the ABC and FSC2 analyses in regard to whether *H. vancouverense* colonized the Northern Rockies via a northern route across the Okanogan highlands or via a southern route across the Central Oregonian highlands. In our RF analysis, these two models are misclassified at proportions of 0.19 (southern route classified as northern route) and 0.18 (northern route classified as southern route), based on out-of-the-bag error rates. In the power analysis for FSC2, these models were misclassified 10 and 15 percent of the time (Power Analysis in FSC2; Table S5), and in the power analysis for the RF approach, these two models were misclassified 22 and 42 per cent of the time. In combination with the ambiguity across methods, this suggests that we have limited power to distinguish between these two models, given the data collected here.

5 | CONCLUSION

Our results indicate that binning can be an effective strategy for the summarization of the mSFS. This comes at an important time, when SNP data sets from hundreds to thousands of SNPs are being collected from a variety of nonmodel species. Our work demonstrates that, using the binning strategy together with the RF strategy for model selection, researchers can make accurate phylogeographic inferences from NGS data sets that may be too small for accurate estimation of the true mSFS. Finally, we show that by allowing researchers to evaluate a larger number of models and to compare models with different numbers of populations, RF model selection could have important implications for the future of model-based approaches.

ACKNOWLEDGEMENTS

Funding was provided by the US National Science Foundation (DEB 1457726/14575199). MLS was supported by a NSF GRFP (DG-1343012) and a University Fellowship from The Ohio State University. We thank the Royal British Columbia Museum and the Florida Museum of Natural History for providing samples. We thank Michael Lucid of Idaho Fish and Game for donations of samples and the Ohio Supercomputer Center for computing resources (allocation grant PAS1181-1). We thank the Carstens laboratory for comments that improved this manuscript prior to publication. We would also like to thank Graham Stone and three anonymous reviewers for helpful comments during the review process.

DATA ACCESSIBILITY

Raw reads, the parameters used for data processing and a full SNP data set are available on Dryad (https://doi.org/10.5061/dryad.2j27b). Scripts developed as a part of the work presented here are available on github (https://github.com/meganlsmith).

AUTHOR CONTRIBUTIONS

M.L.S and B.C.C designed the study. Funding and support were obtained by B.C.C, D.C.T and J.S. M.L.S and M.R collected samples. M.L.S collected genomic data, performed genetic analyses and wrote the article. All authors edited the article and approved the final version of the article.

REFERENCES

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., ... Saunders, N. C. (1987). Intraspecific phylogeography: The mitochondrial dna bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, *18*, 489–522.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology and Systematics*, *41*, 379–406.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, *13*, 1063–1095.

Blum, M. G. B. (2010). Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association*, *105*, 491–1178.

Boitard, S., Rodriguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data - an approximate bayesian computation approach. *PLoS Genetics*, *12*, 1–36.

Brunsfeld, S. J., Sullivan, J., Soltis, D. E., & Soltis, P. S. (2000). Comparative phylogeography of north- western North America : A synthesis. *Special Publication-British Ecological Society*, *14*, 319–340.

Burke, T. E. (2013). *Land snails and slugs of the Pacific Northwest*. Corvallis, OR, USA: Oregon State University.

Carstens, B. C., Brennan, R. S., Chua, V., Bermingham, E., Lamb, T., Neigel, J. E., ... Saunders, N. C. (2013). Model selection as a tool for phylogeographic inference: An example from the willow Salix melanopsis. *Molecular Ecology*, *22*, 4014–4028.

Carstens, B. C., Brunsfeld, S. J., Demboski, J. R., Good, J. M., & 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.

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

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 113.

Espíndola, A., Ruffley, M., Smith, M. L., Carstens, B. C., Tank, D. C., & Sullivan, J. (2016). Identifying cryptic diversity with predictive phylogeography. *Proceedings of the Royal Society B: Biological Sciences*, *283*, 20161529.

Excoffier, L., Dupanloup, I., Huerta-SÃnchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*, e1003905.

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.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*, e1000695.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2nd edn. New York: Springer.

Hickerson, M. J., Dolman, G., & Moritz, C. (2006). Comparative phylogeographic summary statistics for testing simultaneous vicariance. *Molecular Ecology*, *15*, 209–223.

Huang, H., & Knowles, L. L. (2014). Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Systematic biology*, *65*(3), 357–365.

Nielsen, R., & Beaumont, M. A. (2009). Statistical inferences in phylogeography. *Molecular Ecology*, *18*, 1034–1047.

Pelletier, T. A., & Carstens, B. C. (2014). Model choice for phylogeographic inference using a large set of models. *Molecular Ecology*, *23*, 3028–3043.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, *7*, e37135.

Pielou, E. C. (2008). *After the ice age: The return of life to glaciated North America*. Chicago, IL, USA: University of Chicago Press.

Prates, I., Rivera, D., Rodrigues, M. T., & Carnaval, A. C. (2016). A mid-Pleistocene rainforest corridor enabled synchronous invasions of the Atlantic Forest by Amazonian anole lizards. *Molecular Ecology*, *25*, 5174–5186.

Pritchard, J., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, *16*, 1791–1798.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J., Gautier, M., & Robert, C. P. (2015). Reliable ABC model choice via random forests. *Bioinformatics*, *32*, 859–866.

Raynal, L., Marin, J., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2017). ABC random forests for Bayesian parameter inference. arXiv preprint, arXiv:1605.05537.

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing. *Genome Research*, *22*, 939–946.

Roux, C., Fraïsse, C., & Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2010). Shedding light on the grey zone of speciation along a continuum of genomic divergence. bioRxiv, 513–516.

Sainudiin, R., Thornton, K., Harlow, J., Booth, J., Stillman, M., Yoshida, R., ... Donnelly, P. (2011). Experiments with the site frequency spectrum. *Bulletin of Mathematical Biology*, *73*, 829–872.

Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, *43*, 1716–1741.

Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS Computational Biology*, *12*, 1–28.

Sousa, V., & Hey, J. (2013). Understanding the origin of species with genome-scale data: Modelling gene flow. *Nature Reviews Genetics*, *14*, 404–414.

Stocks, M., Siol, M., Lascoux, M., & De Mita, S. (2014). Amount of information needed for model choice in Approximate Bayesian Computation. *PLoS One*, *9*, 1–13.

Terhorst, J., Kamm, J. A., & Song, Y. S. (2016). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*, 303–309.

Terhorst, J., & Song, Y. S. (2015). Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences*, *112*, 7677–7682.

Thomé, M. T. C., & Carstens, B. C. (2016). Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. *Proceedings of the National Academy of Sciences*, *113*, 8010–8017.

Veeramah, K. R., Woerner, A. E., Johnstone, L., Gut, I., Gut, M., Marques-Bonet, T., ... Hammer, M. F. (2015). Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. *Genetics*, *200*, 295–308.

Wegmann, D., Leuenberger, C., & Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, *182*, 1207–1218.

Xue, A. T., & Hickerson, M. J. (2015). The aggregate site frequency spectrum for comparative population genomic inference. *Molecular Ecology*, *24*, 6223–6240.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

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