doi:10.1111/evo.1387

SSE SOCIETY for the STUDY of EVOLUTION

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.

EVOLUTION INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

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.

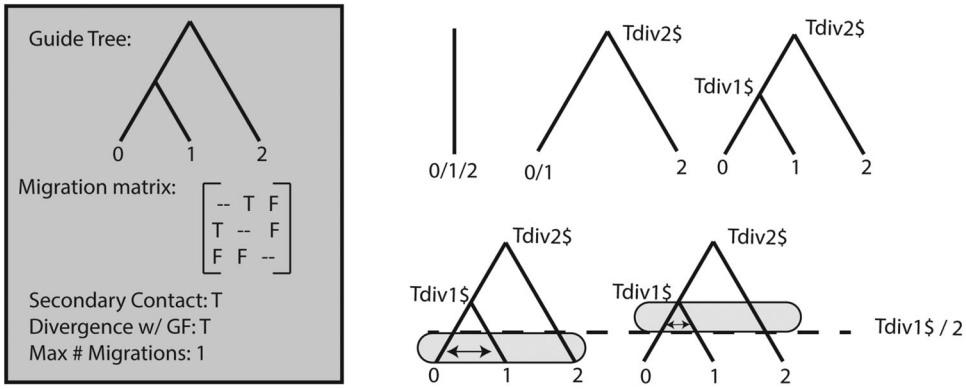

Figure 1. **An example of a default model set that can be considered in *delimitR*. GF is gene flow. In the gray box, we show the information the user would provide to specify the model set. On the right, we show the resulting models. Lines indicate divergent populations, and arrows indicate gene flow. 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.

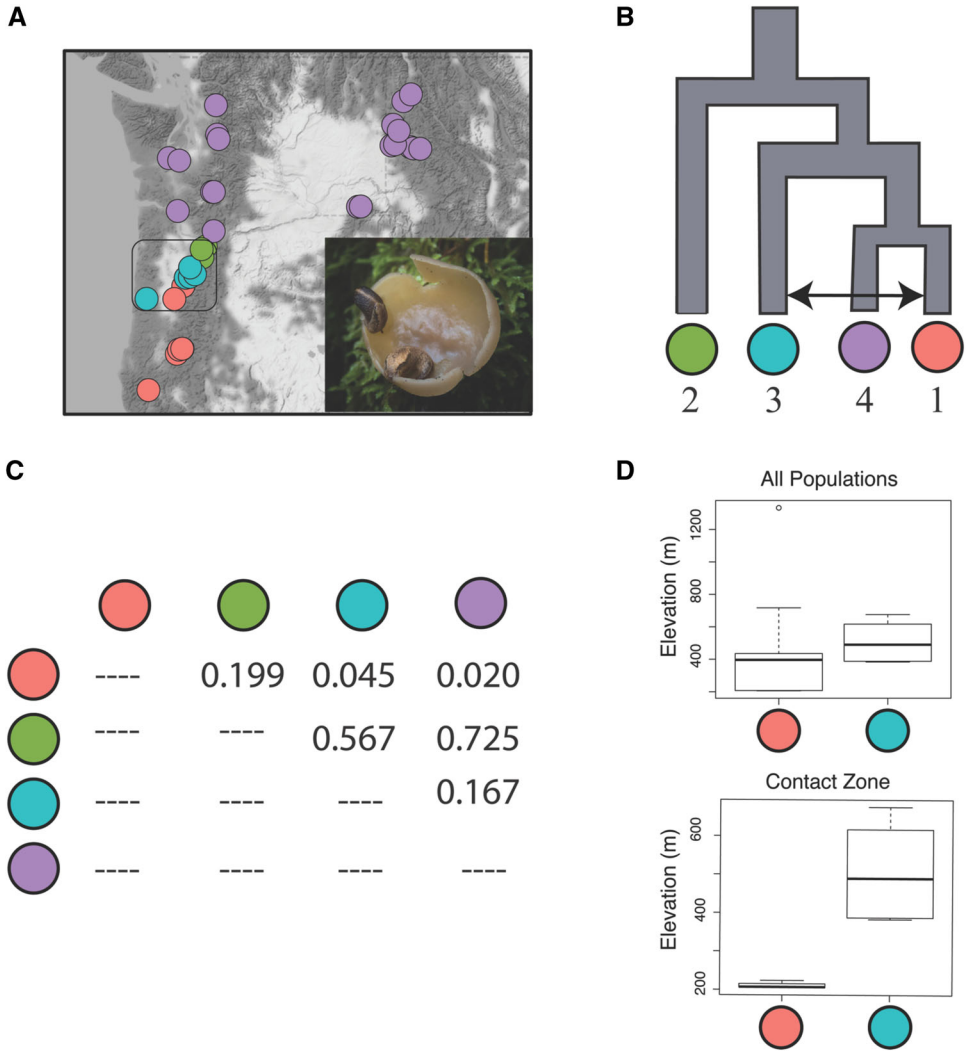

Figure 2. Results from *P. andersoni*. (A) Sampling map, with a black square surrounding samples considered to be "near the contact zone." Colors correspond to structure population assignments (colored based on population with highest posterior probability). Photo by M. Smith. (B) The best model, with 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 simicrocea*) 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

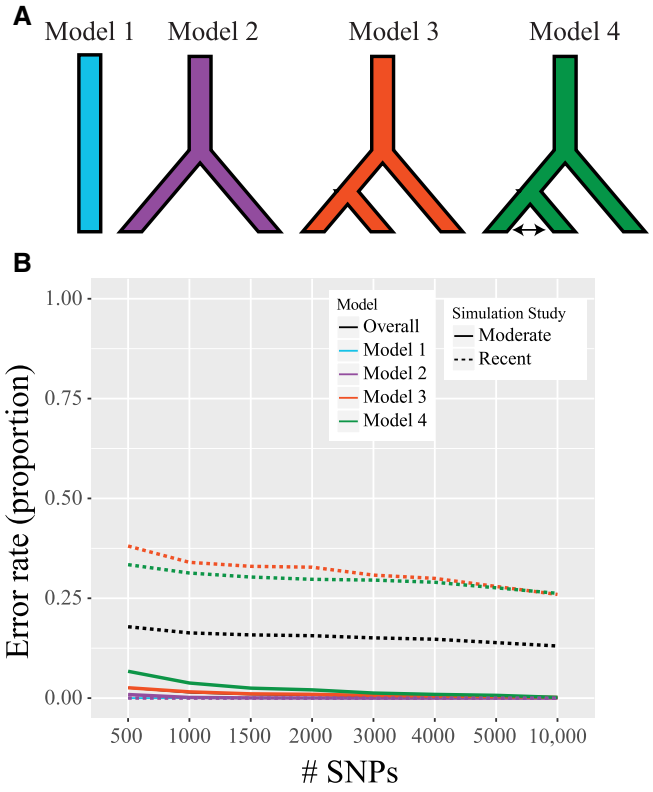

Figure 3. **Results of the first simulation study, which considered up to three species with the potential for secondary contact between the most recently diverged lineages. (A) The four models evaluated in the simulation study. (B) The results of the simulation study with moderate and recent divergence 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).

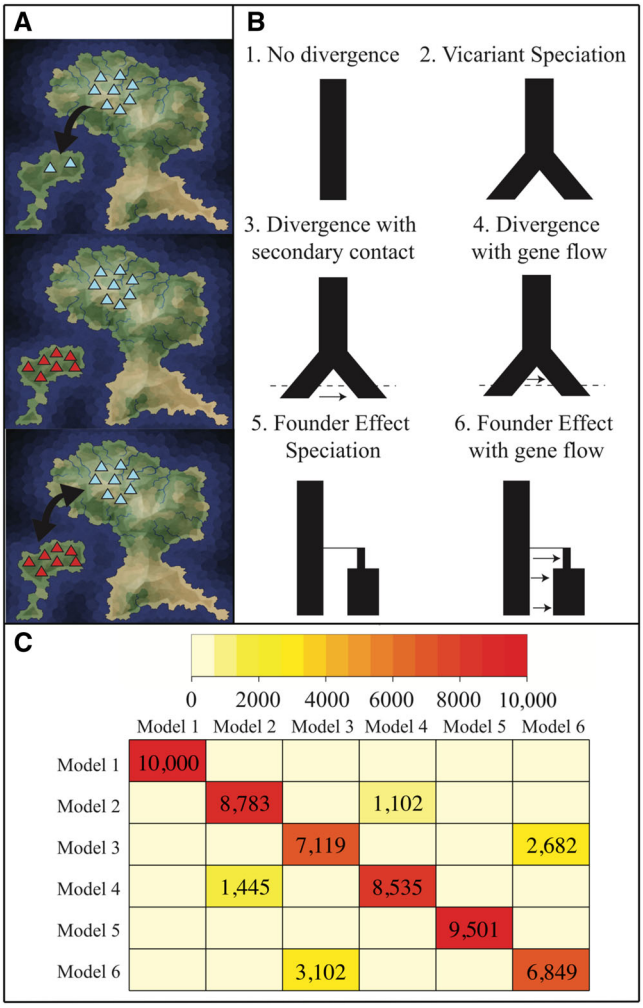

Figure 4. Results of the second simulation study. (A) In island-continent systems, there are several processes that may contribute to speciation, including (1) Founder effect speciation, in which a small number of individuals colonizes the island and speciates (top); (2) Vicariant speciation, in which 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 (red-blobgames.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.