Predicting genetic biodiversity in salamanders using geographic, climatic, and life history traits

Danielle J. Parsons [1], Abigail E. Green [2], Bryan C. Carstens [1] and Tara A. Pelletier [2]

Affiliations

[1] Museum of Biological Diversity and Department of Evolution, Ecology, and Organismal Biology; The Ohio State University; Columbus, OH USA 43210

[2] Department of Biology, Radford University, Radford, VA USA 24142

Abstract

The geographic distribution of genetic variation within a species reveals information about its evolutionary history, including responses to historical climate change and dispersal ability across various habitat types. We combine genetic data from salamander species with geographic, climatic, and life history data collected from open-source online repositories to develop a machine learning model designed to identify the traits that are most predictive of unrecognized genetic lineages. We find evidence of hidden diversity distributed throughout the clade Caudata that is largely the result of variation in climatic variables. We highlight some of the difficulties in using machine-learning models on open-source data that are often messy and potentially taxonomically and geographically biased.

Introduction

Documenting biodiversity is an important first step in understanding both ecological and evolutionary processes (Gadelha *et al.*, 2021), particularly the functional roles that act to connect processes functioning at both shallow and deep time scales (Guralnick & Hill, 2009). Notably, any such documentation of biodiversity implicitly assumes that the units (e.g., species) are comparable across different geographic regions. Given that a Linnean shortfall (i.e., the ratio of recognized to unrecognized species (Whittaker *et al.*, 2005)) exists in most clades and may be substantial across Eukaryota (Mora *et al.*, 2011), it is not clear that this assumption is reasonable. An alternative approach is to utilize evolutionary significant units (Moritz, 1994), or genetic lineages, in place of species in broad analyses of biodiversity (e.g., (Mable, 2019)). This may be particularly useful in clades with relatively high degrees of morphological and ecological conservatism. One such clade is Caudata (i.e., salamanders and newts), which exhibits high frequencies of cryptic species (e.g., (Jockusch *et al.*, 2012; Camp & Wooten, 2016; Bernardes *et al.*, 2020)).

Identifying genetic lineages in Caudata can have important conservation implications. For example, Mead *et al.* (2005) discovered a new species of western *Plethodon* salamander that was originally thought to be either *P. elongatus* or *P. stormi* (Mead *et al.*, 2005). All three of these species are listed on the IUCN Red List as either near threatened (*P. elongatus*), vulnerable (*P. asupak*), or endangered (*P. stormi*). More recently, Parra Olea *et al*. (2020) discovered five cryptic lineages in *Chiropterotriton* from Mexico, several of which are threatened due to their restricted ranges (Parra Olea *et al.*, 2020). Species with small ranges and/or limited dispersal capabilities can be harder to protect because their distributions often do not fall within protected areas (Nauman & Olson, 2008) and small ranges are often used as a factor in assigning conservation priorities (Hortal *et al.*, 2015). Therefore, it is important to identify these lineages, as they could easily go unnoticed and unprotected. Many other species of salamander that would have otherwise gone unnoticed and have been recognized using molecular data have small ranges and likely need protection (Steffen *et al.*, 2014; Nishikawa & Matsui, 2014; Min *et al.*, 2016; Kuchta *et al.*, 2018; Okamiya *et al.*, 2018). The presence of cryptic diversity has been recently highlighted as a key component of undescribed biodiversity that requires greater attention (Bickford *et al.*, 2007; Pfenninger & Schwenk, 2007).

Efforts to conserve undescribed genetic diversity can be facilitated using computational methods that identify genetic lineages representing potentially hidden diversity in need of further investigation. The use of data science techniques has allowed biodiversity studies to expand their geographic and taxonomic focus to explore broader patterns of evolution, which can be difficult to assess using traditional meta-analysis methods (Lyman & Edwards, 2022). Macrogenetics, a relatively new field that merges biodiversity data with genetic data (Blanchet *et al.*, 2017; Leigh *et al.*, 2021), has been used to explore how human impacts influence levels of intraspecific genetic diversity (Miraldo *et al.*, 2016; Millette *et al.*, 2020), to study past and future climate refugia (Carstens *et al.*, 2018; Baranzelli *et al.*, 2022), and to quantify latitudinal biodiversity gradients (Gratton *et al.*, 2017; Pelletier & Carstens, 2018; Barrow *et al.*, 2021; Fonseca *et al.* 2023). Macrogenetic methods, particularly in combination with predictive modeling, can be used to inform conservation policies by identifying species, taxonomic groups, or geographic areas in need of further investigation (Pelletier *et al.*, 2018; Raposo *et al.*, 2021). Recently, such analyses have expanded to taxonomic work.

Parsons *et al.* (2022) analyzed mitochondrial DNA sequences from over 4000 species of mammals, representing roughly 66% of currently described species, and found that mammal diversity is largely under-described using molecular species delimitation methods on publicly available barcode data. This is useful for several reasons. A comprehensive list of genetic lineages that may represent species now exists that can help focus taxonomic efforts. Parsons *et al.* (2022) also found that taxa with small bodies, and large geographic distributions with variation in precipitation and isothermality, were more likely to contain cryptic diversity. While some of this might seem obvious (morphological differences are harder to observe in small-bodied animals and these animals may be harder to find), it does allow researchers to document characteristics of species, higher taxonomic groups, or even geographic regions that contribute to diversification and therefore biodiversity patterns. When done in disparate taxonomic groups (e.g., vertebrates, invertebrates, plants, and fungi) and at different levels (e.g., Class, Order, Family) this furthers our understanding of core evolutionary processes.

A similar approach was taken in birds. Using a tree-based molecular species delimitation method, Smith *et al.* (2018) found that latitude explained variation in phylogeographic breaks, while other traits pertaining to habitat and life history explained very little. In this case, phylogeographic structure was higher in the tropics. Conversely, in other organisms, isolation-by-distance within species is often higher at higher latitudes (multiple taxonomic groups: Pelletier & Carstens, 2018; amphibians: Amador *et al.,* 2023). Further, genetic variation within amphibians was best explained by range size and elevation, rather than latitude, in the neotropics (Amador *et al.,* 2023), while latitude was an important predictor of genetic diversity in the nearctic (Barrow *et al*., 2021). This suggests that differences exist in how genetic variation is distributed within species depending on which taxonomic groups are being examined, and at what spatial scale.

In order to expand these approaches, we conducted a computational assessment of genetic lineages in roughly 100 salamander species using the *phylogatR* database (Pelletier *et al.*, 2022). *PhylogatR* aggregates DNA sequence data from both GenBank and BOLD into sequence alignments, providing associated GBIF occurrence records (i.e., GPS coordinates) for each sequence. There are over 700 described species of salamanders belonging to nine families (Bánki, O. *et al.*, 2022), most located in the northern hemisphere. While salamanders contain a wide variety of life history strategies and habitats, they are likely to have high levels of cryptic diversity due to their moisture requirements and similar body forms. However, their eco-evolutionary processes can vary from species to species and sometimes oppose our expectations (Pelletier *et al.*, 2011, 2015; Pelletier & Carstens, 2016; Jones and Weisrock 2018; Pyron *et al.,* 2020; Dufresnes *et al.*, 2021). We follow methods from Parsons *et al.* (2022) and use molecular species delimitation methods to estimate the number of genetic lineages present in previously collected data that is both openly available and easily tractable. We then use a predictive modeling approach to determine whether any variables pertaining to geography, the environment, or life history traits contribute to the presence of genetic lineages within species. We also discuss some of the difficulties in using open-source data that are often messy and potentially taxonomically and geographically biased.

Materials and Methods

Collection of genetic and geographic data

We downloaded all available data from the *phylogatR* database (https://phylogatr.org/) using the search term ‘Caudata’ on 2/4/22. The uncleaned data represented four families, 93 different species, and 14 loci with a total of 3768 DNA sequences. To begin cleaning the data, we calculated nucleotide diversity (pi) values for each locus in every species and found outliers by setting lower and upper bounds of 2.5% (0) and 97.5% (0.2193634) respectively. For each of the four outliers and two species with missing pi values, we opened the DNA sequence file in Mesquite v3.7 (Maddison & Maddison, 2021) and removed any extremely short or non-overlapping sequences (Data S1). Additionally, we discovered a typo for the species *Batrachuperus karlschmidti* causing there to be two different species folders for the same species. Both the sequence and occurrence files were merged for the species and the sequence files were realigned to correct the error. Two species complexes were present in the dataset, and these were kept named as downloaded: *Triturus cristatus x dobrogicus macrosomus* and *Ambystoma laterale jeffersonianum* complex.

Species alignments from the download for both the mitochondrial genes Cytochrome oxidase I (*COI*) and Cytochrome b (*cytb*) were merged for all salamander species and aligned using MAFFT v7.5 (Katoh & Standley, 2013) with the default settings and including the –adjustdirection command to account for reverse complement sequences. We visually inspected alignment files for both genes and removed all short sequences, which we classified as those missing 50% or more of the second half of the sequence. Twenty-one sequences were removed from the *COI* alignment and 99 were removed from the *cytb* alignment, leaving totals of 768 and 908 sequences for *COI* and *cytb*, respectively. The sequences for seven species were completely removed from further analysis due to their short length (missing 50% or more of the second half of the sequences). In total, eighty-three species remained with an average of approximately 20 sequences per nominal species (see Data S2 for a list of identifiers corresponding to the sequences used in this study).

Species delimitation

We used three methods of species delimitation to determine the number of genetic lineages present in our samples. The GMYC is a tree-based method that takes a phylogenetic tree as input and finds a point in the tree where branching changes from within to between species (Pons *et al.*, 2006). The ABGD (Puillandre *et al.*, 2012) and ASAP (Puillandre *et al.*, 2021) methods are distance-based delimitation methods that use pairwise genetic distances to establish the threshold between intra- and inter-species divergence. Because each method is based on a specific set of assumptions, it is best to use multiple methods and compare their results in order to achieve a more accurate delimitation (Carstens *et al.*, 2013). By looking for concordance across methods, we can increase our confidence in the identified lineage boundaries and minimize the potential impact of bias introduced by any single method. While we report delimitation results from the genes *COI* and *cytb* for all methods, we used a consensus of delimitation results (among methods and loci) for assessing the role of geography, the environment, and life history traits in predicting salamander genetic diversity.

To estimate a species tree for input into the GMYC, we used BEAST v2.5.1 (Bouckaert *et al.*, 2019). We used the default parameters except for conducting 100,000,000 million generations, sampling every 5,000, and setting the model of sequence evolution to GTR+I+G (Abadi *et al.*, 2019). The log files were checked by eye using Tracer v1.7.2 (Rambaut *et al.*, 2018). ESS values were all over 1000 for both *cytb* and *COI*. We removed 10% as burnin and retained the maximum clade credibility tree using TreeAnnotator. After checking that the tree was binary and ultrametric, we used the R package *splits* (Ezard *et al.*, 2009) to conduct GMYC analyses. In each case we used the single threshold model and all other default settings. We conducted both ABGD and ASAP delimitation analyses via their web portals (https://bioinfo.mnhn.fr/abi/public/abgd/abgdweb.html and https://bioinfo.mnhn.fr/abi/public/asap/asapweb.html, respectively) using the default parameter settings.

Predictor variables

A variety of predictor variables were collected, including geographic and environmental values derived from georeferenced locality data (see Data S3). In addition, three life history traits were available from AmphiBIO, a global database for amphibian ecological traits (Oliveira *et al.*, 2017), for most of the species in our study: reproductive strategy (direct developing, larval phase), habitat (terrestrial, fossorial, aquatic, or some combination of these), and body size (total length). To supplement this dataset and fill in any missing trait values, we used AmphibiaWeb (AmphibiaWeb, 2023) and other online sources (Data S4).

To extract species specific data related to its environmental distribution, we utilized 42 GIS data layers (see Data S4 for data layer details), including all 19 BIOCLIM layers from the CHELSA database (Karger *et al.*, 2017; Karger, Dirk Nikolaus *et al.*, 2021) at 1 km resolution, elevation (Aster global digital elevation model version 2, 2011), population density (Socioeconomic Data And Applications Center (SEDAC) Gridded Populations of the World (GPW), 2016), terrestrial habitat heterogeneity (Tuanmu & Jetz, 2015), gross domestic product (World Bank Development Economics Research Group (DECRG) Gross Domestic Product, 2010), global land cover classification (European Space Agency, 2009), global river classification (Ouellet Dallaire *et al.*, 2019), disaster risk (Peduzzi, 2019), anthropogenic biome (Ellis *et al.*, 2010), and various indicators of seasonal growth (Karger *et al.*, 2017; Karger, Dirk Nikolaus *et al.*, 2021). We utilized the R packages 'raster' (, 2016), 'rgdal' (, 2017), 'geosphere' (, 2016), and 'plyr' (Wickham, 2011) to extract species specific information from each layer using geographic occurrence records obtained from *phylogatR*. To represent the environmental variation within the occupied range of each species, we extracted the value of each environmental layer for each GPS coordinate associated with each species. We then took the mean and standard deviation for each environmental variable. To obtain species specific data related to geographic distribution we extracted the minimum, maximum, mean, and length of latitude and longitude from the GPS points of each species.

We used the R package 'mice' (Buuren & Groothuis-Oudshoorn, 2011) to impute trait values missing from our dataset (see Figure S1 for distribution of missing data and specific trait values imputed). The imputation method 'pmm' was used for all numeric variables and 'polyreg' was used for categorical variables (i.e., reproductive strategy and habitat). We ran the imputation 15 times (Figure S2) and then pooled the iterations to generate the final imputed values. The final database containing all trait values (both imputed and original) is available in Data S4.

Predictive modeling

We used the R package 'caret' (Kuhn, 2008) to generate a random forest classification model (Breiman, 2001) based on our previously generated database of predictor variables and a consensus of our species delimitation results. Two separate sets of consensus models were generated to assess the role of geography, environment, and life history traits on the presence of hidden diversity (Figure 1A). The first model (*all agree*) represents a strict consensus of delimitation results from species in which results from all methods of species delimitation agree (Figure 1B). Any species with conflicting delimitation results were excluded from analysis. The second model (*majority rules*) represents a majority rule consensus in which species are assigned to a response category based on relative support of delimitation results (Figure 1C). For each model, we used 70% of the data to train the model and the remaining 30% was set aside as a test set. Models were generated using 10-fold cross validation with five repeats to tune the parameter 'mtry', the number of variables randomly sampled at each split, and optimize the area under the receiver operating characteristic curve, ROC. After training, we extracted the variable importance measures mean decrease accuracy (MDA) and Gini impurity (Gini) from the final models. We then used the final models on the test set data to evaluate model performance. Model performance was evaluated across a variety of metrics including model accuracy, which reflects how well the predicted classifications agree with the observed classifications, and both positive and negative predictive value, which indicate the how the model performs on observations from each class. Additionally, we calculated the no information rate (NIR), the proportion of observations that fall into the majority class, and the p-value [Accuracy>NIR], to test for model significance. The top important predictor variables from our best model were compared using a Kruskal-Wallis test to determine if these variables are significantly different between species that do or do not contain hidden diversity.

Results

Genetic and geographic dataset

Our final dataset consisted of 1676 DNA barcoding sequences (Figure 2). Of these, 768 sequences were from the Cytochrome oxidase I gene (*COI*), and 908 sequences were from the Cytochrome b gene (*cytb*). These sequences were derived from 83 nominal species of salamanders, which were distributed among 26 distinct genera occurring across the globe. The dataset contained 13 species with sequences from the gene *cytb*. Comparatively, *COI* exhibited notably broader taxonomic coverage, with 77 nominal species represented. Out of the 83 species analyzed, only seven were shared between *COI* and *cytb*. Of the remaining 76 species, 70 were unique to *COI* and six were unique to *cytb*. To supplement the genetic data collected, a total of 1765 georeferenced occurrence records from *phylogatR* were utilized to collect a combination of geographic, environmental, and life history trait values for each nominal species present in the dataset.

Species delimitation and consensus assignment

Species delimitation results were generated by analyzing *COI* and *cytb* sequences from each nominal species under three different delimitation methods, ABGD, ASAP, and GMYC. We classified each nominal species as either containing genetic lineages or not containing genetic lineages based on the number of genetic groups predicted by each delimitation analysis. While taxonomic overlap between *COI* and *cytb* was narrow, delimitation results for species shared by both loci were mostly congruent with respect to species classification. Of the seven species with sequences from both genes, only two species produced conflicting results regarding the presence of genetic lineages within a specific taxon based on loci. Delimitation results across different methods showed slightly less agreement. Classifications resulting from the GMYC and ASAP methods were similar across species. These methods, on average, resulted in slightly fewer predicted species per nominal species than the ABGD method (see Figure 3 for predicted species numbers).

To account for this variation in our final predictive models, we generated two consensus classifications to evaluate concordance between delimitation results from different methods and loci. The results of our consensus models indicate that roughly 2/3rds of the nominal salamander species used in this analysis are likely to contain genetic lineages that may be unexplored diversity. The strictest of these classifications produced a consensus model (*all agree*) consisting of 51 total species, 41 of which were classified as containing hidden diversity and 10 of which were classified as not containing hidden diversity. The remaining consensus model (*majority rules*) consisted of 83 total species, of which 51 were classified as containing genetic lineages and 32 were not (Figure 3).

Predictive modeling

For our *majority rules* and *all agree* consensus classifications, we developed random forest classification models using all available predictor data. To assess potential correlation between variables in our dataset we used the R package 'corrplot' (Taiyun Wei & Viliam Simko, 2021) to generate a correlation matrix of our predictor variables (Figure S3). Due to the presence of strong correlations between several of the geographic and environmental variables in our dataset we performed multiple random forest models with progressive sets of correlated variables removed at different cutoff values (i.e., |correlation coefficient| > 0.75; 0.85; 0.9). The results of these random forest models are presented below (Table 1).

All random forest models were found to have high predictive accuracy, with the *majority rules* and *all agree* models achieving accuracies of 75-85% and 87-93%, respectively, in identifying nominal species likely to contain hidden diversity. Although these results may initially seem to suggest that all our models are able to make meaningful predictions, further examination of additional model evaluation metrics reveals potential overfitting and inflation of predictive power. For example, despite the high accuracy of the models, the 95% confidence intervals for these values are broad with an average length of nearly 40% for most of the models

(Tables 1 and 2). Additionally, the no information rates (NIRs), a measure of prediction significance based on the underlying dataset that needs to be exceeded in order for model results to be significant, are particularly high for the *all agree* consensus models, where the class frequencies are more skewed towards species predicted to harbor hidden diversity. The high NIR values combined with wide confidence intervals result in a p-value [Accuracy > NIR] greater than 0.05 in all models, except for the *majority rules* consensus using a correlation cutoff of 0.90. While all our models show high accuracy, when the additional model evaluation metrics are considered only one has strong predictive power. Therefore, we only used the *majority rules* consensus using a correlation cutoff of 0.90 for interpreting variable importance of our data.

Evaluation of variable importance

We extracted variable importance measurements from each predictive model using the variable importance metrics MDA and Gini. While there was some overlap of top predictors between different models (Figure 4; Figure S4), no specific predictors were consistently predicted to be of significantly higher importance than other predictors in the model. Instead, importance was split across numerous predictors that were found to be unstable between models. This instability supports previous indications that many of the predictive models are likely prone to overfitting. Despite the lack of a strong set of standout predictors across models, one pattern does emerge that is applicable to the species in our dataset. Of the top ten most important predictors in each model, approximately 85% are measurements of standard deviation (vs. measurements of mean values or life history traits) (Data S5). This is supported by further examination of our one model that was able to predict significantly better than random, the *majority rules* consensus with a correlation coefficient cutoff of 0.90, in which the top five most important predictors are measurements of standard deviation. Significance testing indicates that species identified as containing hidden genetic lineages often have ranges characterized by a larger variance in annual and seasonal precipitation, isothermality, and net primary productivity than species not identified as harboring hidden genetic lineages (Figure 5).

Discussion

When identifying genetic lineages or delimiting species, it is important to recognize that species concepts are complex and often differ based on various factors, such as geographic location, reproductive isolating mechanisms, genetic markers, and taxonomic practices. Therefore, it is essential to approach species delimitations with caution and to recognize that they represent a hypothesis or starting point rather than a definitive answer (Hillis, 2019). In addition, while mitochondrial data can be suitable for preliminary assessments of species diversity (Gostel & Kress, 2022), these assessments should be considered in tandem with other species information and relevant data when describing species boundaries. However, with recent advances in technology rapidly increasing the quantity of publicly accessible genetic and geographic datasets, these data offer a cost effective and efficient way to explore large-scale patterns and predictors of intraspecific genetic variation (e.g., Miraldo *et al.*, 2016; Pelletier & Carstens, 2018; Yiming *et al.*, 2021).

Our results suggest that there are genetic lineages that may warrant further investigation distributed within Caudata. Adequately documenting biodiversity, both at the species and population level, is a first step in understanding the eco-evolutionary processes generating this diversity. However, in most clades, the Linnean shortfall is likely to influence broad scale patterns detected using macrogenetic approaches (Hortal *et al.*, 2015), making it essential to consider how the taxonomic designations used to inform these approaches influence the patterns detected. This is particularly important when dealing with clades suspected of harboring high levels of cryptic diversity. For example, Miraldo *et al*. (2016) generated the first global map of genetic diversity within species of mammals and amphibians. One of their main conclusions was that amphibians displayed lower levels of genetic variation in areas with higher human impact. Similarly, in amphibians, several recent studies have found within species genetic diversity to be lower in temperate regions in species with smaller ranges and at higher elevations (Barrow *et al*., 2020; Amador 2023). The methods used to detect these patterns are based on current taxonomic knowledge, and as such, rely on the assumption that the species designations used are accurate. However, if species descriptions inaccurately reflect biological diversity, nominal species that contain cryptic species will display higher levels of genetic diversity, while not reflecting true within species variation, potentially skewing our interpretation of any patterns that result.

Evaluating support for identified genetic lineages

While our delimitation of genetic lineages are a starting point, or hypothesis generation step, for evaluating a species in nature where complex processes, such as hybrid zones, and adequate sampling must be considered (Hillis, 2019), we believe these computational approaches are useful for targeting species in further need of examination. We conducted a literature search to explore whether the nominal species in our dataset have been previously explored from a species delimitation approach. We used the online American Museum of Natural History taxonomic and nomenclatural database, Amphibian Species of the World (Darrel, 2024), to evaluate current taxonomic research in each nominal species of salamander predicted to contain hidden diversity in our consensus model. Species in which we were able to identify research-based support for the potential of undescribed diversity were recorded, along with the related articles in which the diversity was described as well as the type of data used (see Data S7). Nearly 70% of species the majority rules consensus suggests harbor hidden lineages contain results that also support the potential splitting of species into separate lineages. Out of these about 38% were explored using mt DNA only, 10% with nuclear DNA only, 35% using a combination of both nuclear and mt DNA and 17% using mt DNA, nuclear DNA and morphology. Just under 10% of the species display a complex history of hybridization, making delimitations difficult, a situation not uncommon in salamanders (Denton *et al*., 2018; Pyron *et al*., 2020). We were unable to find results for roughly 25% of our species data. We encountered 5 species in which the results of previous delimitation work was either unclear or considered highly contested (e.g., *Ichthyosaura alpestris*, *Batrachuperus karlschmidti*, *Batrachuperus taibaiensis,* and *Salamandrella schrenckii*). Taxonomy is dynamic field (Raposo *et al.,* 2020) and given our search, it can be difficult to use current open-source data relying solely on species names. However, the current literature largely supports the delimitation results found here and suggests a number of species in further need of investigation (see citations in Data S7, formal name changes, and an ability to update current open-source databases to reflect these changes). Additionally, even though there are limitations to using current open-source data that might not keep up to date with current taxonomy, we can still determine what factors might predict the presence of hard-to-find species.

Significant Predictors of Diversity

Significance testing of the most important predictors from our best model (*majority rules* consensus with a correlation coefficient cutoff of 0.90) indicates that the species which our analysis identified as containing hidden genetic lineages often have ranges characterized by a larger variance in annual and seasonal precipitation, isothermality, and net primary productivity when compared to species that were not identified as containing hidden genetic lineages by our analysis (Figure 5B). And while the order of the most important traits is unstable across different models, across all models most of the traits found to be important were measurements of standard deviation (vs. measurements of mean values or life history traits) (Data S5). This suggests that the presence of variation in climate, rather than any species-specific trait or characteristic is the most identifiable driving force of within species genetic diversity for salamanders at this scale. Species traits were not a predictor of intraspecific genetic diversity in amphibians (Barrow *et al*., 2021; Amador *et al.,* 2023) using a different measure of genetic variation within species (nucleotide diversity). Using similar methods, our results in salamanders differ from that found in mammals, where body size and range size were the most important predictors (Parsons *et al*., 2022).

These findings are somewhat consistent with other studies of salamander diversification. Reproductive mode (larval stages, direct development) and habitat (combinations of terrestrial, aquatic, arboreal) vary across species and have evolved multiple times but have not been found to directly correlate with speciation, though being a direct developer might increase diversification rates (Liedtke *et al*., 2022). Alternatively, in one species which has intraspecific variation in habit, *Salamandra salamandra*, terrestrial-breeding individuals exhibited greater geographic genetic differentiation (Lourenço *et al*., 2019). Not surprisingly, this species showed conflicting results in our delimitation analyses. In vertebrate clades, terrestrial organisms tend to have higher diversification rates than aquatic organisms (Wiens, 2015), but we did not have a large number of fully terrestrial species in our dataset, which might have limited our ability to detect this as an important predictor. Given that salamanders are relatively constrained in body form and ecological niches, variation in climatic variables seems like a reasonable explanation for species containing cryptic diversity. This follows the suggestion that change in climatic niche variables increases diversification rates in plethodontid salamanders (Kozak & Wiens, 2010). Diversification rates in frogs and salamanders have been shown to be higher near the tropics (Wiens 2007), so one might expect latitude to be an important predictor. However, latitude was not included in the list of predictor variables that were likely to be important (Figure 4).

Predictive modeling as a tool to address the Linnean shortfall

Recently, Parsons *et al*. (2022) used publicly available genetic barcoding data to develop a predictive framework to identify mammalian clades most likely to contain hidden species and determine specific trait complexes that indicate where hidden mammal diversity is likely to exist. We adopted a similar approach to evaluate genetic lineages in the clade Caudata, a group which differs from mammals in several key aspects, including species richness and sampling intensity. We focused on a lower taxonomic level so there are fewer recognized species of salamanders (<1000; 'AmphibiaWeb', 2023) compared to the mammal dataset, making the ability to produce robust predictive models more challenging. Additionally, there was a smaller proportion of available data for salamanders than mammals (~10% compared to 60% of described species). However, these smaller datasets might be more realistic in that they are more representative of the type of data most likely to be available for the taxonomic groups that are in greatest need of attention from taxonomists.

While the predictive models generated in this study actually have a higher overall accuracy than those used in Parsons *et al.* (2022) (see Table 3), relying on this metric alone to evaluate the performance of predictive models can be misleading (Provost *et al.*, 1998). For classification models, model accuracy depends on how well the predicted classifications match the observed classifications. While seemingly straightforward, accuracy does not account for other model characteristics that may be influencing model behavior, such as the class frequencies of the underlying dataset (Kuhn & Johnson, 2013). In cases where one class occurs at a much higher frequency than the other, a predictive model can attain a high accuracy by simply always predicting the higher class. Therefore, an important benchmark to consider when interpreting overall model accuracy is the frequency at which the majority class occurs, the no information rate (NIR). If a model's accuracy is not significantly higher than the NIR (i.e., p-value [Accuracy > NIR]), it can remain unclear whether the model is making meaningful decisions. In our models, the overall accuracy was found to be high, but the 95% confidence intervals for the accuracy values are very wide for most of the models. In addition, because the dataset is skewed towards species classified as containing hidden diversity, the p-value [Accuracy > NIR] was found to be significant in only one model. This is important to point out because even though there are large datasets available, choosing the right analytical tools can remain challenging depending on the use of the predictive models. Beyond analytical tools, it's also important to consider your dataset, and how the characteristics of your dataset are affecting the results you obtain. Considering the scale of not only the dataset, but also the analytical methods used and the pattern one is attempting to examine is especially important in meta-analyses, as different patterns emerge at different scales (Gurevitch *et al.*, 2018).

Conclusions

Here, we chose to utilize biodiversity data from *phylogatR* (i.e., genetic data for which directly associated specimen locality information is available) to avoid potential discrepancies between the distribution of the genetic and geographic data analyzed. By doing so we hoped to gain a more fine-grain understanding of how species genetic diversity is influenced by geographic and environmental factors (Leigh *et al.*, 2021). However, making this choice significantly decreased the amount of data available and led to a greatly reduced dataset. Our study included 1676 DNA barcoding sequences from the genes *COI* and *cytb* (768 and 908 sequences each, respectively). However, a 3/31/23 search of GenBank for salamander barcoding sequences from the genes *COI* and *cytb* returned a total of 17097 sequences (4468 and 12629 sequences each, respectively; see Data S6). Similarly, while we were able to obtain 1765 occurrence records tied to the genetic sequences used in this study, a GBIF search for geographic occurrences tied to salamander preserved specimens and material samples returned 675243 records (see Data S6). This study highlights the lack of genetic data with easily-associated geographic information.

The numerous benefits of making biological data more broadly available have been repeatedly demonstrated (Wüest *et al.*, 2020). And recent years have seen a significant increase in the amount of available specimen and biodiversity data. The utility of these data to address large scale patterns of biodiversity, such as those examined in this study, is enhanced by our ability to integrate and synthesize data across different data sources, types, and taxonomic groups (Heberling *et al.*, 2021). Our study highlights the importance of not just making these data available, but making them available in a way that is standardized and will facilitate integration and re-use for future generations to come (e.g., Colella *et al.*, 2021; Hardisty *et al.*, 2022).

Acknowledgments

We thank Radford University Office of Undergraduate Research for supporting AEG through the Accelerated Research Opportunities and Research Rookies programs. We thank members of the Carstens lab for their comments on a draft of this manuscript. Funding for DJP was provided by a Presidential Fellowship from The Ohio State University and by NSF-1910623 awarded to BCC.

Supporting Information

File S1. (Contains **Figure S1:** Distribution of missing data in the salamander trait database; **Figure S2:** Distribution of imputed trait data; **Figure S3:** Correlation matrix of predictor variables; **Figure S4:** Variable importance for predictive models; **Table S1:** Comparison of model accuracy confidence intervals between salamander and mammal predictive models; **Data S1.** Nucleotide diversity of Caudata sequences from *phylogatR*; **Data S2.** *PhylogatR* identification numbers for records analyzed; **Data S3.** Final dataset of response and predictor variables; **Data S4.** Variable specifics and source information; **Data S5.** Variable importance extended results; **Data S6.** Results of search for publicly available genetic and geographic salamander data; **Data S7.** Results of literature search for genetic lineages in recognized salamander species.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research was funded the National Science Foundation via DBI-1911293 to TAP and DBI-1910623 to BCC.

Data Availability Statement

Data are uploaded to Dryad (set to private for peer review) and reviewers can access it using the following link: https://datadryad.org/stash/share/4tXXsub0cPan4BqKe1KS6l5SRGwf69F5p2zOmv7QYes. When the data is made public, the final DOI will be https://doi.org/10.5061/dryad.m63xsj474. Code related to this manuscript, including data cleaning, imputation, predictive modeling, and significance testing has been deposited in GitHub (https://github.com/parsons463/HiddenSalamanders). All remaining data are available in the manuscript and/or supporting information.

References

Abadi S, Azouri D, Pupko T, Mayrose I. **2019**. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications* 10: 934.

Amador L, Arroyo-Torres I, Lisa N. Barrow LN. 2023. Machine learning and phylogenetic models identify predictors of genetic variation in Neotropical amphibians. *bioRxiv*.

AmphibiaWeb. **2023**.

Aster global digital elevation model version 2. **2011**.

Bánki, O., Roskov, Y., Döring, M., Ower, G., Vandepitte, L., Hobern, D., Remsen, D., Schalk, P., DeWalt, R. E., Keping, M., Miller, J., Orrell, T., Aalbu, R., Adlard, R., Adriaenssens, E. M., Aedo, C., Aescht, E., Akkari, N., Alfenas-Zerbini, P. 2022. Catalogue of Life Checklist (Y. Roskov, Ed.; Version 2022-05-20).

Baranzelli MC, Cosacov A, Sede SM, Nicola MV, Sérsic AN. **2022**. Anthropocene refugia in Patagonia: A macrogenetic approach to safeguarding the biodiversity of flowering plants. *Biological Conservation* 268: 109492.

Barrow LN, Fonseca EM, Thompson CEP, Carstens BC. **2021**. Predicting amphibian intraspecific genetic diversity with machine learning: Challenges and prospects for integrating traits, geography, and genetic data, *Molecular Ecology Resources* 21: 2718-2831.

Bernardes M, Le MD, Nguyen TQ, Pham CT, Pham AV, Nguyen TT, Rödder D, Bonkowski M, Ziegler T. **2020**. Integrative taxonomy reveals three new taxa within the Tylototriton asperrimus complex (Caudata, Salamandridae) from Vietnam. *ZooKeys* 935: 121–164.

Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I. **2007**. Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution* 22: 148–155.

Blanchet S, Prunier JG, De Kort H. **2017**. Time to Go Bigger: Emerging Patterns in Macrogenetics. *Trends in Genetics* 33: 579–580.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Wu CH, Xie D, Zhang C, Stadler T, Drummond AJ. **2019**. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis (M Pertea, Ed.). *PLOS Computational Biology* 15: e1006650.

Breiman L. **2001**. Random Forests. *Machine Learning* 45: 5–32.

Buuren S van, Groothuis-Oudshoorn K. **2011**. **mice** : Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software* 45.

Camp CD, Wooten JA. **2016**. Hidden in Plain Sight: Cryptic Diversity in the Plethodontidae. *Copeia* 104: 111–117.

Carstens BC, Morales AE, Field K, Pelletier TA. **2018**. A global analysis of bats using automated comparative phylogeography uncovers a surprising impact of Pleistocene glaciation. *Journal of Biogeography* 45: 1795–1805.

Carstens BC, Pelletier TA, Reid NM, Satler JD. **2013**. How to fail at species delimitation. *Molecular Ecology* 22: 4369–4383.

Colella JP, Stephens RB, Campbell ML, Kohli BA, Parsons DJ, Mclean BS. **2021**. The Open-Specimen Movement. *BioScience* 71: 405–414.

Darrel FR. 2024. Amphibian Species of the World: an Online Reference. Version 6.2 (December 2023). Electronic Database accessible at https://amphibiansoftheworld.amnh.org/index.php. *American Museum of Natural History,* New York, USA.

Denton RD, Morales AE, Gibbs HL. 2018. Genome-specific histories of divergence and introgression between an allopolyploid unisexual salamander lineage and two ancestral sexual species. *Evolution*. 72: 1689–1700.

Dufresnes C, Brelsford A, Jeffries DL, Mazepa G, Suchan T, Canestrelli D, Nicieza A, Fumagalli L, Dubey S, Martínez-Solano I, Litvinchuk SN, Vences M, Perrin N, Crochet PA. **2021**. Mass of genes rather than master genes underlie the genomic architecture of amphibian speciation. *Proceedings of the National Academy of Sciences* 118: e2103963118.

Ellis EC, Klein Goldewijk K, Siebert S, Lightman D, Ramankutty N. **2010**. Anthropogenic transformation of the biomes, 1700 to 2000: Anthropogenic transformation of the biomes. *Global Ecology and Biogeography*: no-no.

European Space Agency. **2009**. ESA GlobCover project.

Foster J. Provost, Tom Fawcett, Ron Kohavi. **1998**. The Case against Accuracy Estimation for Comparing Induction Algorithms. *Machine learning: proceedings of the fifteenth international conference, Madison, Wisconsin, July 24 - 27, 1998*. San Francisco, Calif: Morgan Kaufmann.

Fonseca EM, Pelletier TA, Decker SK, Parsons DJ, Carstens BC. 2023. Pleistocene glaciations caused the latitudinal gradient of within-species genetic diversity, *Evolution Letters* 7: 331–338.

Gadelha LMR, Siracusa PC, Dalcin EC, Silva LAE, Augusto DA, Krempser E, Affe HM, Costa RL, Mondelli ML, Meirelles PM, Thompson F, Chame M, Ziviani A, Siqueira MF. **2021**. A survey of biodiversity informatics: Concepts, practices, and challenges. *WIREs Data Mining and Knowledge Discovery* 11.

Geosphere: Spherical trigonometry. **2016**.

Gostel MR, Kress WJ. **2022**. The Expanding Role of DNA Barcodes: Indispensable Tools for Ecology, Evolution, and Conservation. *Diversity* 14: 213.

Gratton P, Marta S, Bocksberger G, Winter M, Keil P, Trucchi E, Kühl H. **2017**. Which Latitudinal Gradients for Genetic Diversity? *Trends in Ecology & Evolution* 32: 724–726.

Guralnick R, Hill A. **2009**. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics* 25: 421–428.

Gurevitch J, Koricheva J, Nakagawa S, Stewart G. **2018**. Meta-analysis and the science of research synthesis. *Nature* 555: 175–182.

Hardisty AR, Ellwood ER, Nelson G, Zimkus B, Buschbom J, Addink W, Rabeler RK, Bates J, Bentley A, Fortes JAB, Hansen S, Macklin JA, Mast AR, Miller JT, Monfils AK, Paul DL, Wallis E, Webster M. **2022**. Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet. *BioScience* 72: 978–987.

Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. **2021**. Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences* 118: e2018093118.

Hillis DM. **2019**. Species Delimitation in Herpetology. *Journal of Herpetology* 53: 3.

Hortal J, de Bello F, Diniz-Filho JAF, Lewinsohn TM, Lobo JM, Ladle RJ. **2015**. Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics* 46: 523–549.

Jockusch EL, Martínez-Solano I, Hansen RW, Wake DB. **2012**. Morphological and molecular diversification of slender salamanders (Caudata: Plethodontidae: Batrachoseps) in the southern Sierra Nevada of California with descriptions of two new species. *Zootaxa* 3190: 1.

Jones KS, Weisrock DW. 2018. Genomic data reject the hypothesis of sympatric ecological speciation in a clade of *Desmognathus* salamanders. *Evolution* 72: 2378–2393.

Karger DN, Conrad O, Böhner J, Kawohl T, Kreft H, Soria-Auza RW, Zimmermann NE, Linder HP, Kessler M. **2017**. Climatologies at high resolution for the earth’s land surface areas. *Scientific Data* 4: 170122.

Karger, Dirk Nikolaus, Conrad, Olaf, Böhner, Jürgen, Kawohl, Tobias, Kreft, Holger, Soria-Auza, Rodrigo Wilber, Zimmermann, Niklaus E., Linder, H. Peter, Kessler, Michael. **2021**. Climatologies at high resolution for the earth’s land surface areasCHELSA V2.1 (current). : 2.1 KB.

Katoh K, Standley DM. **2013**. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30: 772–780.

Kozak KH, Wiens JJ. **2010**. Accelerated rates of climatic-niche evolution underlie rapid species diversification: Niche evolution and rapid diversification. *Ecology Letters* 13: 1378–1389.

Kuchta SR, Brown AD, Highton R. **2018**. Disintegrating over space and time: Paraphyly and species delimitation in the Wehrle's Salamander complex. *Zoologica Scripta* 47: 285–299.

Kuhn M. **2008**. Building Predictive Models in *R* Using the **caret** Package. *Journal of Statistical Software* 28.

Kuhn M, Johnson K. **2013**. *Applied Predictive Modeling*. New York, NY: Springer New York.

Leigh DM, van Rees CB, Millette KL, Breed MF, Schmidt C, Bertola LD, Hand BK, Hunter ME, Jensen EL, Kershaw F, Liggins L, Luikart G, Manel S, Mergeay J, Miller JM, Segelbacher G, Hoban S, Paz-Vinas I. **2021**. Opportunities and challenges of macrogenetic studies. *Nature Reviews Genetics* 22: 791–807.

Liedtke HC, Wiens JJ, Gomez-Mestre I. **2022**. The evolution of reproductive modes and life cycles in amphibians. *Nature Communications* 13: 7039.

Lourenço A, Gonçalves J, Carvalho F, Wang IJ, Velo-Antón G. **2019**. Comparative landscape genetics reveals the evolution of viviparity reduces genetic connectivity in fire salamanders. *Molecular Ecology* 28: 4573–4591.

Lyman RA, Edwards CE. **2022**. Revisiting the comparative phylogeography of unglaciated eastern North America: 15 years of patterns and progress. *Ecology and Evolution* 12.

M. Keesey. PhyloPic.

Mable BK. **2019**. Conservation of adaptive potential and functional diversity: integrating old and new approaches. *Conservation Genetics* 20: 89–100.

Mead LS, Clayton DR, Nauman RS, Olson DH, Pfrender ME. **2005**. Newly discovered populations of salamanders from Siskiyou County California represent a species distinct from Plethodon stormi. *Herpetologica* 61: 158–177.

Mesquite: a modular system for evolutionary analysis. **2021**.

Millette KL, Fugère V, Debyser C, Greiner A, Chain FJJ, Gonzalez A. **2020**. No consistent effects of humans on animal genetic diversity worldwide (A Mooers, Ed.). *Ecology Letters* 23: 55–67.

Min MS, Baek HJ, Song JY, Chang MH, Poyarkov NAJr. **2016**. A new species of salamander of the genus Hynobius (Amphibia, Caudata, Hynobiidae) from South Korea. *Zootaxa* 4169: 475.

Miraldo A, Li S, Borregaard MK, Flórez-Rodríguez A, Gopalakrishnan S, Rizvanovic M, Wang Z, Rahbek C, Marske KA, Nogués-Bravo D. **2016**. An Anthropocene map of genetic diversity. *Science* 353: 1532–1535.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. **2011**. How Many Species Are There on Earth and in the Ocean? (GM Mace, Ed.). *PLoS Biology* 9: e1001127.

Moritz C. **1994**. Defining ‘Evolutionarily Significant Units’ for conservation. *Trends in Ecology & Evolution* 9: 373–375.

Mueller RL, Macey JR, Jaekel M, Wake DB, Boore JL. **2004**. Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. *Proceedings of the National Academy of Sciences* 101: 13820–13825.

Nauman RS, Olson DH. **2008**. Distribution and Conservation Of Plethodon Salamanders On Federal Lands In Siskiyou County, California. *Northwestern Naturalist* 89: 1.

Nishikawa K, Matsui M. **2014**. Three new species of the salamander genus Hynobius (Amphibia, Urodela, Hynobiidae) from Kyushu, Japan. *Zootaxa* 3852: 203.

Okamiya H, Sugawara H, Nagano M, Poyarkov NA. **2018**. An integrative taxonomic analysis reveals a new species of lotic *Hynobius* salamander from Japan. *PeerJ* 6: e5084.

Oliveira BF, São-Pedro VA, Santos-Barrera G, Penone C, Costa GC. **2017**. AmphiBIO, a global database for amphibian ecological traits. *Scientific Data* 4: 170123.

Ouellet Dallaire C, Lehner B, Sayre R, Thieme M. **2019**. A multidisciplinary framework to derive global river reach classifications at high spatial resolution. *Environmental Research Letters* 14: 024003.

Parra Olea G, Garcia-Castillo MG, Rovito SM, Maisano JA, Hanken J, Wake DB. **2020**. Descriptions of five new species of the salamander genus *Chiropterotriton* (Caudata: Plethodontidae) from eastern Mexico and the status of three currently recognized taxa. *PeerJ* 8: e8800.

Parsons DJ, Pelletier TA, Wieringa JG, Duckett DJ, Bryan C. Carstens. **2022**. Analysis of biodiversity data suggests that mammal species are hidden in predictable places. *Proceedings of the National Academy of Sciences* 119: e2103400119.

Peduzzi P. **2019**. The Disaster Risk, Global Change, and Sustainability Nexus. *Sustainability* 11: 957.

Pelletier TA, Carstens BC. **2016**. Comparing range evolution in two western *Plethodon* salamanders: glacial refugia, competition, ecological niches, and spatial sorting. *Journal of Biogeography* 43: 2237–2249.

Pelletier TA, Carstens BC. **2018**. Geographical range size and latitude predict population genetic structure in a global survey. *Biology Letters* 14: 20170566.

Pelletier TA, Carstens BC, Tank DC, Sullivan J, Espíndola A. **2018**. Predicting plant conservation priorities on a global scale. *Proceedings of the National Academy of Sciences* 115: 13027–13032.

Pelletier TA, Crisafulli C, Wagner S, Zellmer AJ, Carstens BC. **2015**. Historical Species Distribution Models Predict Species Limits in Western *Plethodon* Salamanders. *Systematic Biology* 64: 909–925.

Pelletier TA, Duffield DA, DeGrauw EA. **2011**. Rangewide Phylogeography of the Western Red-Backed Salamander (Plethodon vehiculum). *Northwestern Naturalist* 92: 200–210.

Pelletier TA, Parsons DJ, Decker SK, Crouch S, Franz E, Ohrstrom J, Carstens BC. **2022**. PhylogatR: Phylogeographic data aggregation and repurposing. *Molecular Ecology Resources* 22: 2830–2842.

Pfenninger M, Schwenk K. **2007**. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evolutionary Biology* 7: 121.

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP. **2006**. Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects (M Hedin, Ed.). *Systematic Biology* 55: 595–609.

Puillandre N, Brouillet S, Achaz G. **2021**. ASAP: assemble species by automatic partitioning. *Molecular Ecology Resources* 21: 609–620.

Puillandre N, Lambert A, Brouillet S, Achaz G. **2012**. ABGD, Automatic Barcode Gap Discovery for primary species delimitation: ABGD, AUTOMATIC BARCODE GAP DISCOVERY. *Molecular Ecology* 21: 1864–1877.

Pyron RA, O'Connell KA, Lemmon EM, Lemmon AR, Beamer DA. 2020. Phylogenomic data reveal reticulation and incongruence among mitochondrial candidate species in Dusky Salamanders (*Desmognathus*). *Molecular Phylogenetics and Evolution*. 146: 1055-7903.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. **2018**. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7 (E Susko, Ed.). *Systematic Biology* 67: 901–904.

Raposo MA, Kirwan GM, Lourenço ACC, Sobral G, Bockmann FA, Stopiglia R. **2021**. On the notions of taxonomic 'impediment', 'gap', 'inflation' and 'anarchy', and their effects on the field of conservation. *Systematics and Biodiversity* 19: 296–311.

Raposo MA, Kirwan GM, Lourenço ACC, Sobral G, Bockmann FA, Stopiglia R. 2021. On the notions of taxonomic 'impediment', 'gap', 'inflation' and 'anarchy', and their effects on the field of conservation. *Systematics and Biodiversity*. 19: 296-311.

raster: Geographic data analysis and modeling. **2016**.

rgdal: Bindings for the Geospatial Data Abstraction Library. **2017**.

Socioeconomic Data And Applications Center (SEDAC) Gridded Populations of the World (GPW). **2016**.

Steffen MA, Irwin KJ, Blair AL, Bonett RM. **2014**. Larval masquerade: a new species of paedomorphic salamander (Caudata: Plethodontidae: Eurycea) from the Ouachita Mountains of North America. *Zootaxa* 3786: 423.

T. Ezard, T. Fujisawa, T. G. Barraclough. **2009**. SPLITS: species' limits by threshold statistics.

Taiyun Wei, Viliam Simko. **2021**. R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.92).

Tuanmu MN, Jetz W. **2015**. A global, remote sensing-based characterization of terrestrial habitat heterogeneity for biodiversity and ecosystem modelling: Global habitat heterogeneity. *Global Ecology and Biogeography* 24: 1329–1339.

Whittaker RJ, Araújo MB, Jepson P, Ladle RJ, Watson JEM, Willis KJ. **2005**. Conservation Biogeography: assessment and prospect: Conservation Biogeography. *Diversity and Distributions* 11: 3–23.

Wickham H. **2011**. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software* 40.

Wiens JJ. **2015**. Explaining large-scale patterns of vertebrate diversity. *Biology Letters* 11: 20150506.

World Bank Development Economics Research Group (DECRG) Gross Domestic Product. **2010**.

Wüest RO, Zimmermann NE, Zurell D, Alexander JM, Fritz SA, Hof C, Kreft H, Normand S, Cabral JS, Szekely E, Thuiller W, Wikelski M, Karger DN. **2020**. Macroecology in the age of Big Data – Where to go from here? *Journal of Biogeography* 47: 1–12.

Yiming L, Siqi W, Chaoyuan C, Jiaqi Z, Supen W, Xianglei H, Xuan L, Xuejiao Y, Xianping L. **2021**. Latitudinal gradients in genetic diversity and natural selection at a highly adaptive gene in terrestrial mammals. *Ecography* 44: 206–218.

Table 1. Results of *majority rules* consensus predictive models. Model metrics for each random forest predictive model generated using the *majority rules* consensus classifications are shown.

Majority Rules Models	**Original**	**ǀCorrelationǀ > 0.75**	**ǀCorrelationǀ > 0.85**	**ǀCorrelationǀ > 0.90**
Accuracy	0.75	0.75	0.75	0.8333
Accuracy (95% CI)	(0.5329, 0.9023)	(0.5329, 0.9023)	(0.5329, 0.9023)	(0.6262, 0.9526)
No Information Rate	0.625	0.625	0.625	0.625
Pos Pred Value	0.7368a	0.8	0.7647	0.7895
Neg Pred Value	0.8	0.6667	0.7143	1
P-Value [Acc > NIR]	0.1453	0.1453	0.1453	0.02435

Table 2. Results of *all agree* consensus predictive models. Model metrics for each random forest predictive model generated using the *all agree* consensus classifications are shown.

All Agree Models	**Original**	**ǀCorrelationǀ > 0.75**	**ǀCorrelationǀ > 0.85**	**ǀCorrelationǀ > 0.90**
Accuracy	0.8667	0.9333	0.8667	0.8667
Accuracy (95% CI)	(0.5954, 0.9834)	(0.6805, 0.9983)	(0.5954, 0.9834)	(0.5954, 0.9834)
No Information Rate	0.8	0.8	0.8	0.8
Pos Pred Value	0.8571	0.9231	0.8571	0.8571
Neg Pred Value	1	1	1	1
P-Value [Acc > NIR]	0.398	0.1671	0.398	0.398

Table 3. Summary of results of mammal predictive models presented in Parsons *et al*. (Parsons *et al.*, 2022). Model metrics for each random forest predictive model generated using data from the class Mammalia are shown.

Mammal Models	**ABGD COI**	**ABGD cytb**	**GMYC COI**	**GMYC cytb**	**consensus**
Accuracy	0.737	0.68	0.6429	0.6517	0.781
Accuracy (95% CI)	(0.6802, 0.7885)	(0.6333, 0.7241)	(0.5821, 0.7004)	(0.6014, 0.6996)	(0.7273, 0.8285)
No Information Rate	0.7222	0.6235	0.6128	0.5488	0.6533
Pos Pred Value	0.56667	0.6304	0.17271	0.6624	2.85E-06
Neg Pred Value	0.75833	0.6937	0.5571	0.6345	0.807
P-Value [Acc > NIR]	0.32	0.008792	0.6735	3.00E-05	2.85E-06

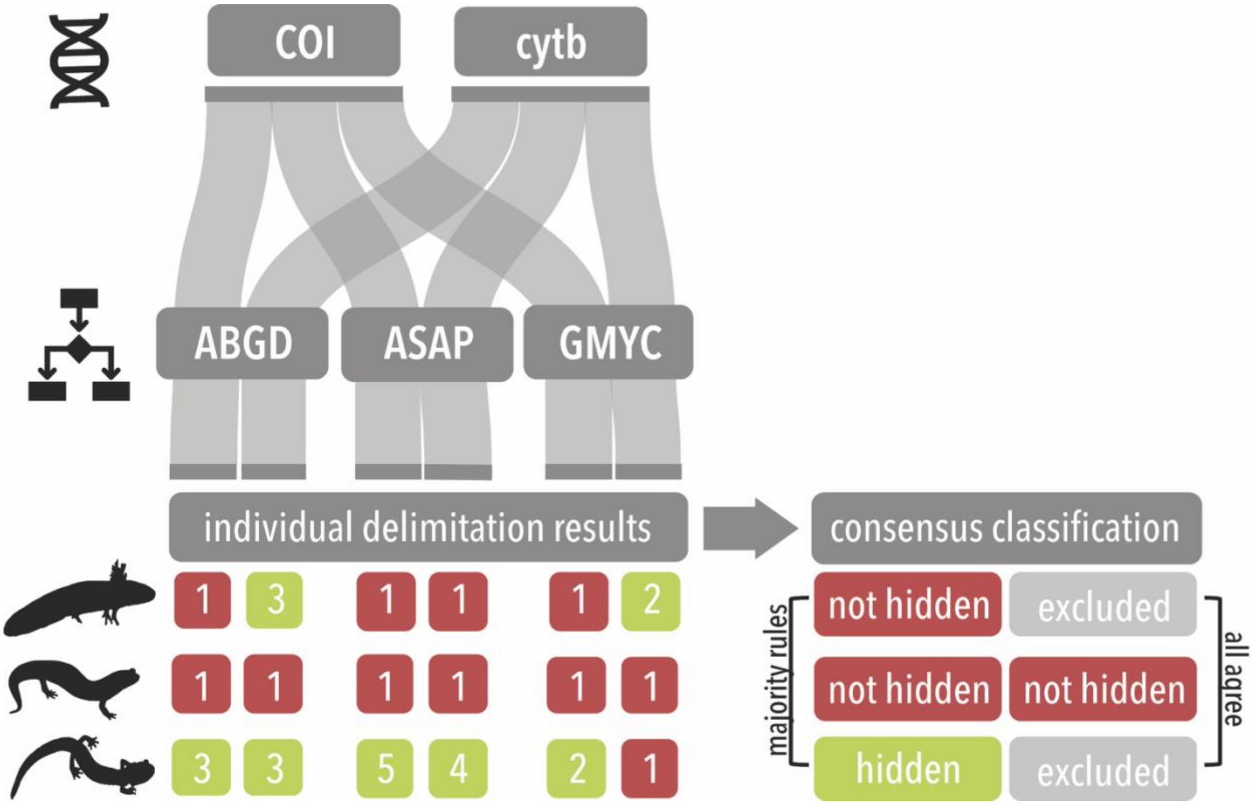


Figure 1. Consensus classification of species delimitation results. A, Flowchart describing the process of generating a consensus of delimitation results (among different methods and loci). B, C, Pipeline for classifying nominal species as either containing or not containing hidden diversity in each consensus analysis (*all agree* and *majority rules*, respectively).

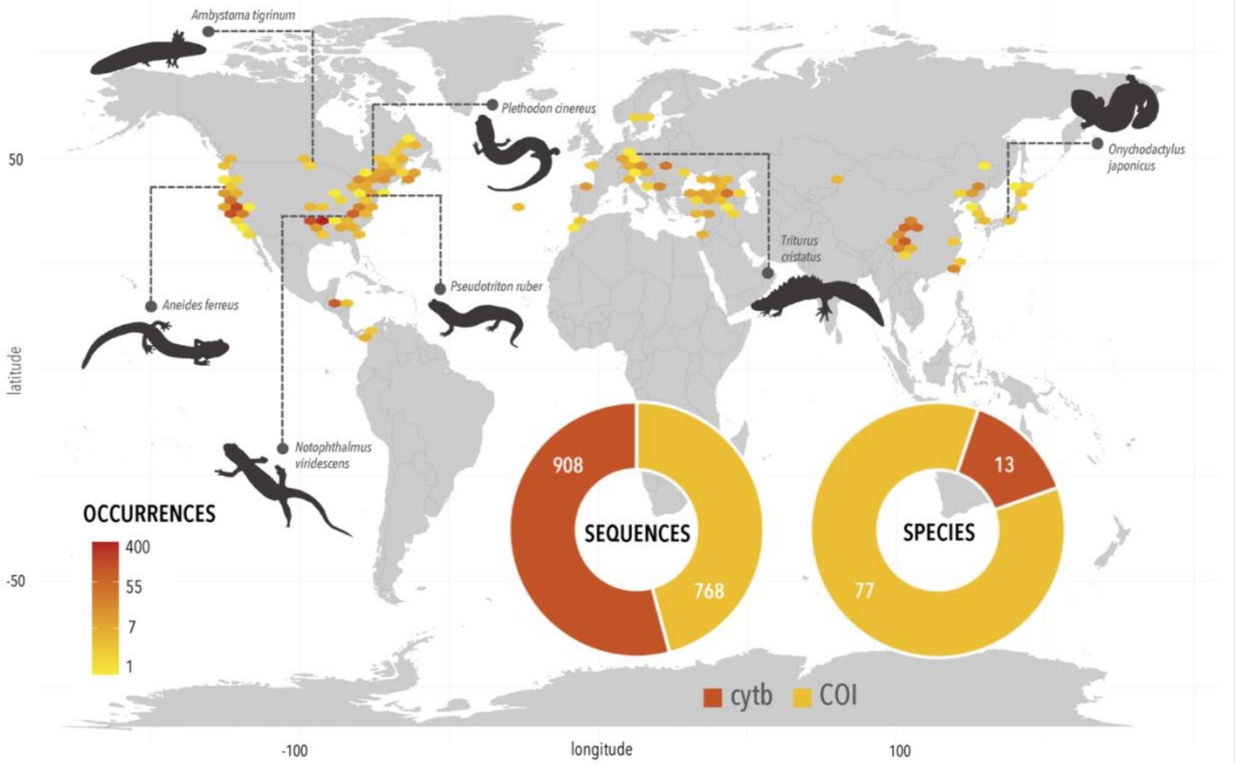
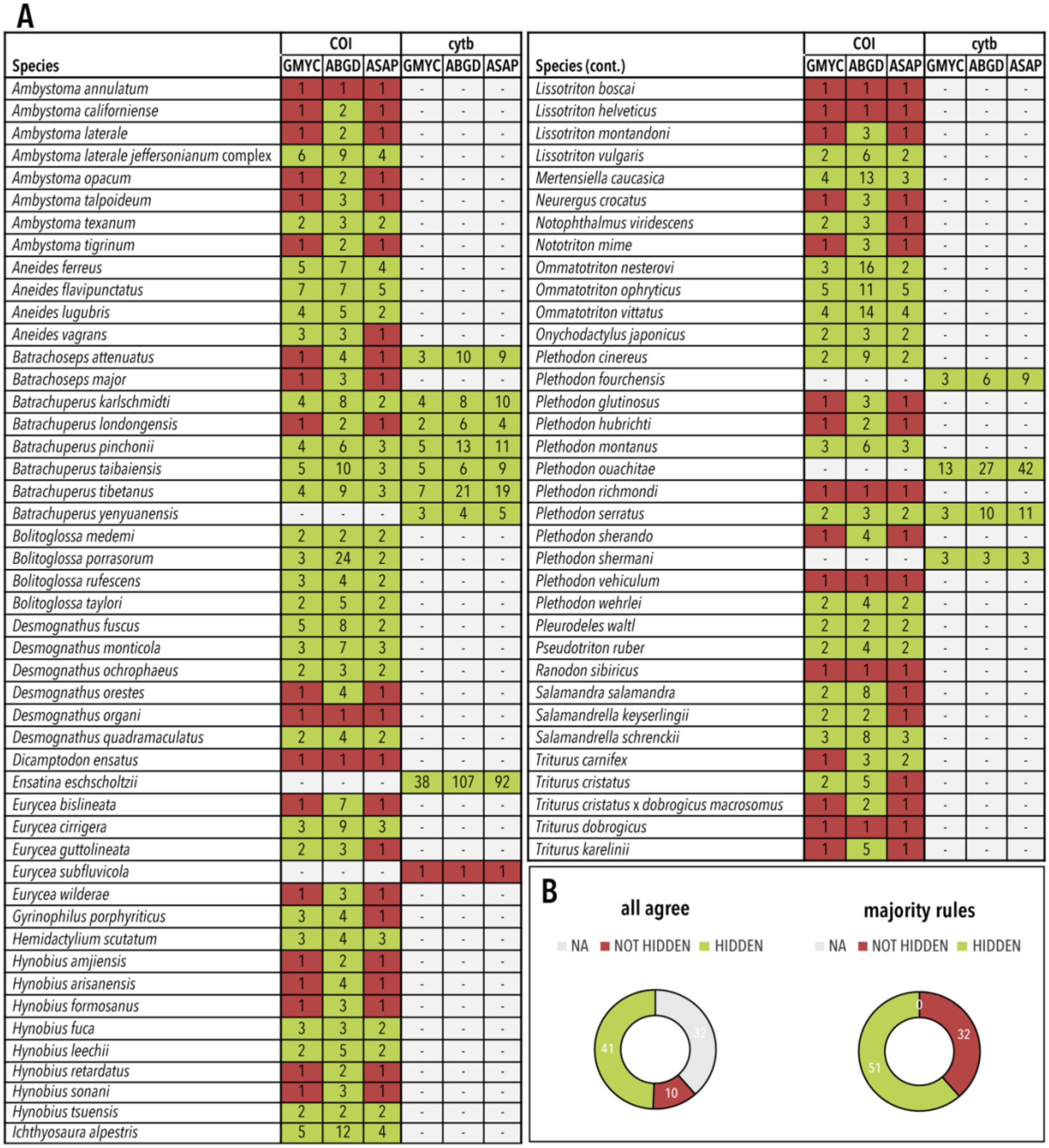


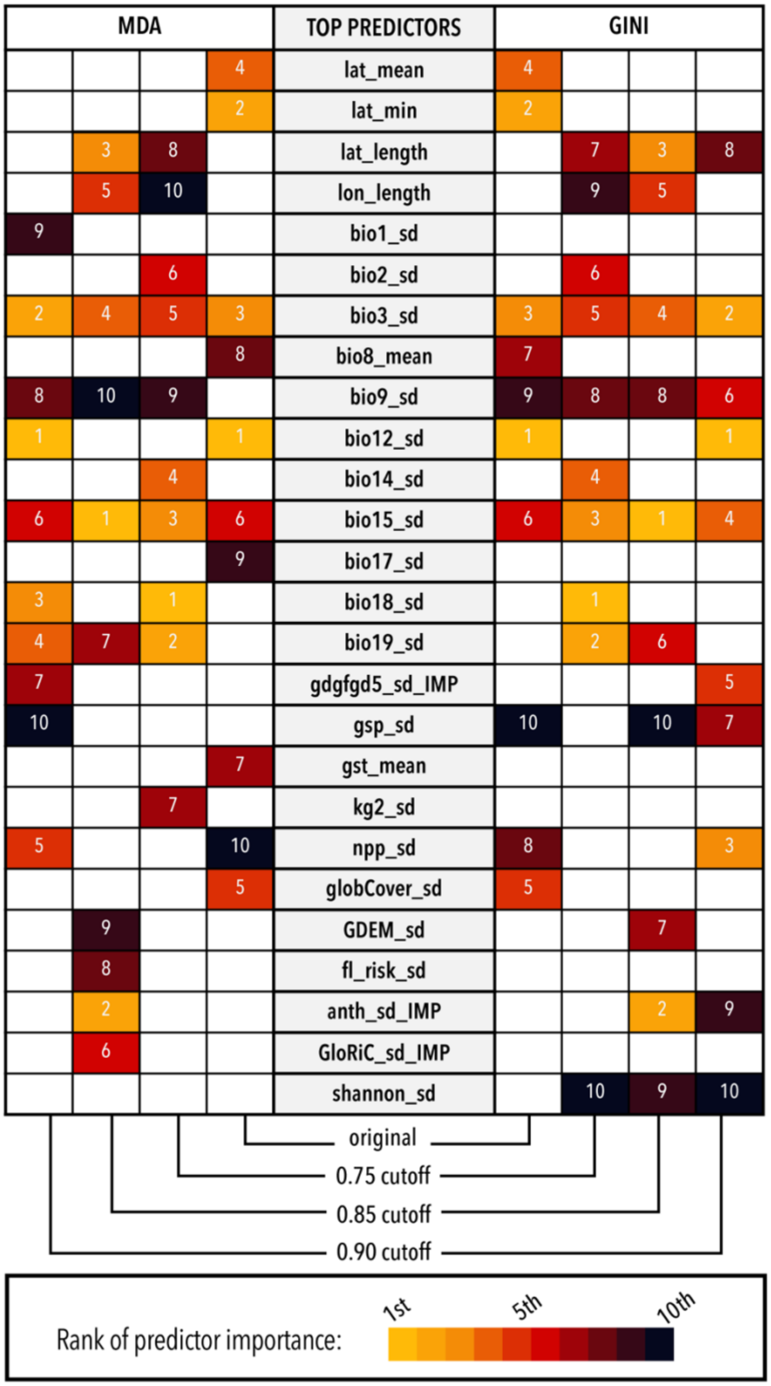
Figure 2. Geographic spread of salamander data. Map shows geographic distribution of salamander occurrences pulled from *phylogatR* (Pelletier *et al.*, 2022) and used in these analyses. Pie charts show the total number of *cytb* and *COI* sequences used (left) and the number of species represented by those *cytb* and *COI* sequences (right). Salamander figures in black were obtained from Phylopic (M. Keesey) and are licensed under public domain.


A
B
all agree
NA
NOT HIDDEN
HIDDEN
32
10
41
majority rules
NA
NOT HIDDEN
HIDDEN
0
32
51

Species	COI			cytb		
	GMYC	ABGD	ASAP	GMYC	ABGD	ASAP
Ambystoma annulatum	1	1	1	-	-	-
Ambystoma californiense	1	2	1	-	-	-
Ambystoma laterale	1	2	1	-	-	-
Ambystoma laterale jeffersonianum complex	6	9	4	-	-	-
Ambystoma opacum	1	2	1	-	-	-
Ambystoma talpoideum	1	3	1	-	-	-
Ambystoma texanum	2	3	2	-	-	-
Ambystoma tigrinum	1	2	1	-	-	-
Aneides ferreus	5	7	4	-	-	-
Aneides flavipunctatus	7	7	5	-	-	-
Aneides lugubris	4	5	2	-	-	-
Aneides vagrans	3	3	1	-	-	-
Batrachoseps attenuatus	1	4	1	3	10	9
Batrachoseps major	1	3	1	-	-	-
Batrachuperus karlschmidti	4	8	2	4	8	10
Batrachuperus londongensis	1	2	1	2	6	4
Batrachuperus pinchonii	4	6	3	5	13	11
Batrachuperus taibaiensis	5	10	3	5	6	9
Batrachuperus tibetanus	4	9	3	7	21	19
Batrachuperus yenyuanensis	-	-	-	3	4	5
Bolitoglossa medemi	2	2	2	-	-	-
Bolitoglossa porrasorum	3	24	2	-	-	-
Bolitoglossa rufescens	3	4	2	-	-	-
Bolitoglossa taylori	2	5	2	-	-	-
Desmognathus fuscus	5	8	2	-	-	-
Desmognathus monticola	3	7	3	-	-	-
Desmognathus ochrophaeus	2	3	2	-	-	-
Desmognathus orestes	1	4	1	-	-	-
Desmognathus organi	1	1	1	-	-	-
Desmognathus quadramaculatus	2	4	2	-	-	-
Dicamptodon ensatus	1	1	1	-	-	-
Ensatina eschscholtzii	-	-	-	38	107	92
Eurycea bislineata	1	7	1	-	-	-
Eurycea cirrigera	3	9	3	-	-	-
Eurycea guttolineata	2	3	1	-	-	-
Eurycea subfluvicola	-	-	-	1	1	1
Eurycea wilderae	1	3	1	-	-	-
Gyrinophilus porphyriticus	3	4	1	-	-	-
Hemidactylium scutatum	3	4	3	-	-	-
Hynobius amjiensis	1	2	1	-	-	-
Hynobius arisanensis	1	4	1	-	-	-
Hynobius formosanus	1	3	1	-	-	-
Hynobius fuca	3	3	2	-	-	-
Hynobius leechii	2	5	2	-	-	-
Hynobius retardatus	1	2	1	-	-	-
Hynobius sonani	1	3	1	-	-	-
Hynobius tsuensis	2	2	2	-	-	-
Ichthyosaura alpestris	5	12	4	-	-	-
Species (cont.)						
Lissotriton boscai	1	1	1	-	-	-
Lissotriton helveticus	1	1	1	-	-	-
Lissotriton montandoni	1	3	1	-	-	-
Lissotriton vulgaris	2	6	2	-	-	-
Mertensiella caucasica	4	13	3	-	-	-
Neurergus crocatus	1	3	1	-	-	-
Notophthalmus viridescens	2	3	1	-	-	-
Nototriton mime	1	3	1	-	-	-
Ommatotriton nesterovi	3	16	2	-	-	-
Ommatotriton ophryticus	5	11	5	-	-	-
Ommatotriton vittatus	4	14	4	-	-	-
Onychodactylus japonicus	2	3	2	-	-	-
Plethodon cinereus	2	9	2	-	-	-
Plethodon fourchensis	-	-	-	3	6	9
Plethodon glutinosus	1	3	1	-	-	-
Plethodon hubrichti	1	2	1	-	-	-
Plethodon montanus	3	6	3	-	-	-
Plethodon ouachitae	-	-	-	13	27	42
Plethodon richmondi	1	1	1	-	-	-
Plethodon serratus	2	3	2	3	10	11
Plethodon sherando	1	4	1	-	-	-
Plethodon shermani	-	-	-	3	3	3
Plethodon vehiculum	1	1	1	-	-	-
Plethodon wehrlei	2	4	2	-	-	-
Pleurodeles waltl	2	2	2	-	-	-
Pseudotriton ruber	2	4	2	-	-	-
Ranodon sibiricus	1	1	1	-	-	-
Salamandra salamandra	2	8	1	-	-	-
Salamandrella keyserlingii	2	2	1	-	-	-
Salamandrella schrenckii	3	8	3	-	-	-
Triturus carnifex	1	3	2	-	-	-
Triturus cristatus	2	5	1	-	-	-
Triturus cristatus x *dobrogicus macrosomus*	1	2	1	-	-	-
Triturus dobrogicus	1	1	1	-	-	-
Triturus karelinii	1	5	1	-	-	-

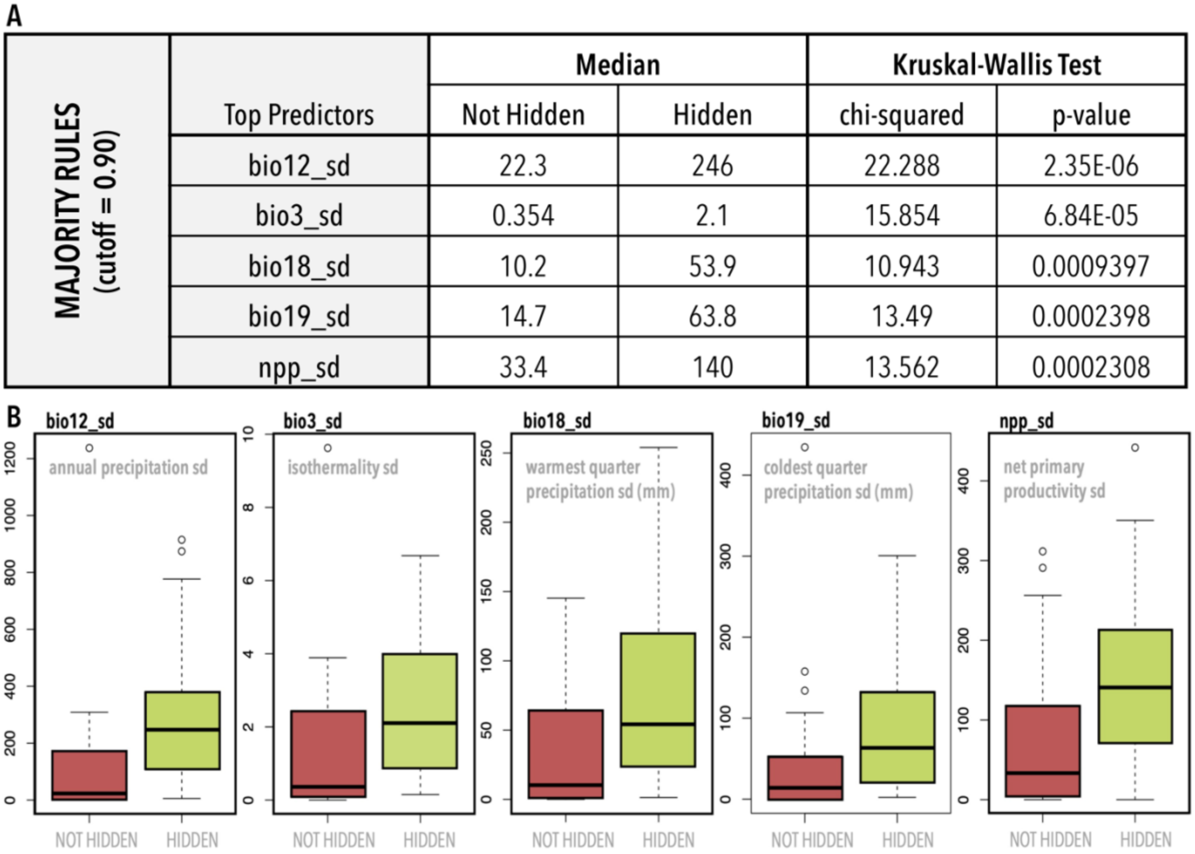
Figure 3. Species delimitation results. A, Graphs show the results of ABGD, ASAP, and GMYC species delimitation analyses of the genes *cytb* and *COI* for each nominal species. Numbers represent the predicted genetic lineages from each analysis. Results highlighted in red indicate no hidden genetic lineages were predicted (i.e., number of genetic lineages = 1). Results highlighted in green indicate hidden genetic lineages were predicted (i.e., number of genetic lineages > 1). Grey highlighting indicates that specific analysis was not performed due to a lack of data. B, Pie charts

display the number of nominal species classified as either containing or not containing hidden diversity in each consensus analysis (i.e., *all agree* and *majority rules*).



MDA				TOP PREDICTORS	GINI			
			4	lat_mean	4			
			2	lat_min	2			
	3	8		lat_length		7	3	8
	5	10		lon_length		9	5	
9				bio1_sd				
		6		bio2_sd		6		
2	4	5	3	bio3_sd	3	5	4	2
			8	bio8_mean	7			
8	10	9		bio9_sd	9	8	8	6
1			1	bio12_sd	1			1
		4		bio14_sd		4		
6	1	3	6	bio15_sd	6	3	1	4
			9	bio17_sd				
3		1		bio18_sd		1		
4	7	2		bio19_sd		2	6	
7				gdgfgd5_sd_IMP				5
10				gsp_sd	10		10	7
			7	gst_mean				
		7		kg2_sd				
5			10	npp_sd	8			3
			5	globCover_sd	5			
	9			GDEM_sd			7	
	8			fl_risk_sd				
	2			anth_sd_IMP			2	9
	6			GloRiC_sd_IMP				
				shannon_sd		10	9	10

Figure 4. Variable importance for predictive models generated using the *majority rules* consensus. Variables ranked among the top ten most important variables (based on MDA and Gini) from the predictive model generated at different correlation cut-offs are included.



MAJORITY RULES (cutoff = 0.90)	Top Predictors	Median		Kruskal-Wallis Test	
		Not Hidden	Hidden	chi-squared	p-value
	bio12_sd	22.3	246	22.288	2.35E-06
	bio3_sd	0.354	2.1	15.854	6.84E-05
	bio18_sd	10.2	53.9	10.943	0.0009397
	bio19_sd	14.7	63.8	13.49	0.0002398
	npp_sd	33.4	140	13.562	0.0002308

Figure 5. Difference in hidden vs not hidden trait values. A, Results of Kruskal-Wallis significance test on the top five most important predictors of the best model (*majority rules* – correlation cutoff 0.90). B, Corresponding boxplots for said predictors show a significant difference in the range of trait values between hidden and non-hidden genetic lineages.