# Predicting intraspecific diversity with machine learning: Challenges and prospects for integrating traits, geography, and genetic data

Lisa N. Barrow[1,2*] (https://orcid.org/0000-0001-7081-2432)

Emanuel Masiero da Fonseca[1] (https://orcid.org/0000-0002-2952-8816)

Coleen E. P. Thompson[1] (https://orcid.org/0000-0003-0591-7654)

Bryan C. Carstens[1] (https://orcid.org/0000-0002-1552-227X)

[1]Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 318 W. 12th Ave, Columbus, OH 43210

[2]Museum of Southwestern Biology and Department of Biology, University of New Mexico, Albuquerque NM 87131

[*]Corresponding author: lnbarrow@unm.edu

Running Title: AMPHIBIAN BIOGEOGRAPHY AND RANDOM FORESTS

## Abstract

The growing availability of genetic datasets, in combination with machine learning frameworks, offer great potential to answer long-standing questions in ecology and evolution. One such question has intrigued population geneticists, biogeographers, and conservation biologists: What determines intraspecific genetic diversity? This question is challenging to answer because many factors may influence genetic variation, including life history traits, historical influences, and geography, and the relative importance of these factors varies across taxonomic and geographic scales. Furthermore, interpreting the influence of numerous, potentially correlated variables is difficult with traditional statistical approaches. To address these challenges, we combined repurposed data with machine learning and investigated predictors of genetic diversity, focusing on Nearctic amphibians as a case study. We aggregated species traits, range characteristics, and >42,000 genetic sequences for 299 species using open-access scripts and various databases. After identifying important predictors of nucleotide diversity with random forest regression, we conducted follow-up analyses to examine the roles of phylogenetic history, geography, and demographic processes on intraspecific diversity. Although life history traits were not important predictors for this dataset, we found significant phylogenetic signal in genetic diversity within amphibians. We also found that salamander species at northern latitudes contain lower genetic diversity. Data repurposing and machine learning provide valuable tools for detecting patterns with relevance for conservation, but concerted efforts are needed to compile meaningful datasets with greater utility for understanding global biodiversity.



## Keywords

Caudata, data repurposing, latitude, random forests, nucleotide diversity, phylogenetic signal

## Introduction

Documenting patterns of species diversity and identifying the processes that produce these patterns have always been important goals of evolutionary biology (Darwin 1859; Wallace 1869). Modern researchers have access to troves of data from museum collections (e.g., iDigBio), genetic data repositories (e.g., NCBI GenBank), biodiversity portals (e.g., GBIF) and phenotypic databases (e.g., MorphoBank). While traditional statistical approaches can be used to explore some aspects of these data (e.g., a regression of body size against latitude), many factors are likely to interact with the correlated variables in any given species. Developing an understanding of the global processes that underlie such patterns requires the analysis of data from multiple species. Such an analysis can proceed with a variety of methods, encompassing two general approaches that attempt to solve this problem in different ways.

Meta-analysis and data repurposing represent two strategies for analyzing complex data from multiple species. Meta-analyses combine results from many studies to increase power and improve our understanding of key phenomena (e.g., Soltis *et al.* 2006; Field *et al.* 2009). Meta-analyses are able to draw from individual studies, each tailored to a focal taxon, but it is not easy to synthesize results across studies that differ in their aims, types of data, inference methods, and sampling design. It is more efficient to proceed via data repurposing (Sidlauskas *et al.* 2010), where data from multiple studies are reanalyzed in a common framework. Repurposed data can be effectively combined with machine learning techniques to investigate questions of interest (e.g., Pelletier & Carstens 2018). This strategy is consistent with broader goals in ecology and evolution to make data accessible and standardized (Carpenter *et al.* 2009; Peters 2010; Sidlauskas *et al.* 2010; Reichman *et al.* 2011; Parr *et al.* 2012; Pope *et al.* 2015; Soltis *et al.* 2016;

Blanchet *et al.* 2017). Furthermore, repurposing published data enhances their value and generates a greater return on investment in these data by funding agencies.

Of the many machine learning techniques available, we utilize the Random Forest approach introduced by Breiman (2001) because it enables an intuitive evaluation of the variables that influence the response variable. Random forest is a machine learning ensemble approach that uses multiple decision trees (a forest) to predict a user-defined response based on many potential predictor variables. Each individual decision tree consists of a subset of the data and a random ordering of variables at the nodes, making the correlation of predictor variables a non-issue. Individually, each tree is a weak predictor, but when many trees are combined, the resulting consensus prediction can be very effective. The importance of each variable is determined by examining the increase in prediction error after the removal of that variable while all others are left unchanged. A review of random forest classification and regression is provided by Liaw & Wiener (2002). Here, we apply this technique to address a long-standing question in evolutionary biology (Leffler *et al.* 2012): What factors determine genetic diversity within species?

*What determines intraspecific genetic diversity?*

Genetic variation within species is a fundamental aspect of diversity that has long intrigued population geneticists, biogeographers, and conservation biologists. Species with sufficient genetic diversity may have the ability to develop resistance to disease or adapt to environmental change (Frankham 1995; Jamieson & Allendorf 2012; Ekroth *et al.* 2019). In contrast, species with limited genetic diversity can experience detrimental effects due to inbreeding depression and may be at higher risk of extinction from catastrophic events such as disease outbreaks (Lande 1988; Hedrick & Kalinowski 2000; Frankham 2005). While the

87 importance of conserving intraspecific diversity is well-established (Moritz & Faith 1998; Moritz
88 2002; Paz-Vinas *et al.* 2018), it remains to be fully integrated as standard practice in global
89 conservation strategies (Laikre *et al.* 2020). In the past few decades, the availability of genetic
90 datasets has grown dramatically, and large amounts of genetic data are now available to
91 investigate diversity within species (Garrick *et al.* 2015). Predicting and understanding the
92 determinants of intraspecific genetic diversity is an increasingly attainable goal.
93 Many potential predictors of intraspecific genetic diversity have been implicated. Life
94 history traits, through their association with mutation rate and effective population size, are
95 expected to influence the accumulation and maintenance of genetic variation within species
96 (Ellegren & Galtier 2016). For example, Romiguier *et al.* (2014) investigated genome-wide
97 diversity of 76 animal species and found a correlation between genetic diversity and life history
98 strategy; long-lived species with few, large offspring and thus higher parental investment ("K-
99 strategists") had lower genetic diversity than species with many, small offspring ("r-strategists").
100 Chen *et al.* (2017) confirmed these results for animals and also found that longevity and mating
101 system explained genomic diversity in plants. Body size has been negatively correlated with
102 genetic diversity and structure in various taxonomic groups including mammals (Brüniche-Olsen
103 *et al.* 2018), bees (López-Uribe *et al.* 2019), and butterflies (Mackintosh *et al.* 2019), a
104 relationship that is presumed to correspond with limits to population abundance in larger-bodied
105 species (White *et al.* 2007). Ecological traits associated with dispersal propensity, including body
106 size, larval period, philopatry, or habitat specialization, could also influence whether or not
107 species exhibit genetic structure (Semlitsch 2008; Paz *et al.* 2015; Mims *et al.* 2016). However,
108 empirical support for associations between species traits and genetic diversity across taxonomic
109 groups has been mixed. Within butterflies, Mackintosh *et al.* (2019) found no relationship

between genetic diversity and longevity, propagule size, abundance, or host range. In Australian lizards, Grundler *et al.* (2019) found no relationship between genetic diversity and body size or habitat specialization, and Singhal *et al.* (2017) found the only significant predictor of intraspecific diversity to be the number of museum occurrence records, a possible proxy for abundance. These examples highlight the possibility that the determinants of genetic diversity may not be easily predictable and are likely to vary among taxonomic groups.

In addition to life history traits, geographic range characteristics can be useful metrics for predicting intraspecific genetic diversity. Range size is sometimes treated as a proxy for census population size, which should scale positively with genetic diversity, because species that occupy more area or that have historically been able to expand to new areas are expected to have higher abundances (Leffler *et al.* 2012; Singhal *et al.* 2017). Additionally, range size should be related to genetic structure because larger ranges typically encompass more habitat variation or physical barriers, leading to patterns of isolation by distance or environment, assuming the species are not capable dispersers (Wright 1943; Sexton *et al.* 2014). Incorporating geographic variables such as elevational, topographic, or climatic variation has illustrated the influence of these factors on population connectivity and phylogeographic structure (Wang 2012; Rodríguez *et al.* 2015). In a global study mapping amphibian genetic diversity, Miraldo *et al.* (2016) identified a latitudinal gradient of genetic diversity with higher diversity in the tropics. Similarly, at the intraspecific level, latitude has been associated with genetic structure and diversity for a broad range of taxa (Smith *et al.* 2017; Brüniche-Olsen *et al.* 2018; Pelletier & Carstens 2018). A plausible explanation is that the climatic and environmental instability experienced by taxa at higher latitudes during glacial cycles contributed to this pattern by reducing suitable habitat and population sizes (Hewitt 2000). The continued integration of intraspecific genetic variation,

133 species traits, and geographic variables is needed to better understand the drivers of genetic
134 diversity across different taxonomic groups and spatial scales.
135
136 *Nearctic amphibians as a case study*
137 Amphibians have long been considered the most imperiled class of vertebrates (Stuart *et*
138 *al.* 2004). More than 40% of data-sufficient amphibian species are considered threatened by the
139 International Union for Conservation of Nature (IUCN 2019), and considerable efforts have
140 focused on addressing the issue that a large proportion of species (~25%) are data deficient (Nori
141 *et al.* 2018). For example, in a global analysis of amphibians, González-del-Pliego *et al.* (2019)
142 found that body size and range size were correlated with threat status, estimated that half of data-
143 deficient species are threatened, and identified regions of the world (e.g., the Neotropics and
144 Southeast Asia) where a high number of data-deficient species are predicted to be threatened.
145 Using machine learning, Howard & Bickford (2014) predicted that >63% of data-deficient
146 species are threatened and identified similar geographic regions with high predicted extinction
147 risk. However, potentially useful information is lacking from these studies. Notably, intraspecific
148 genetic diversity and the factors that influence this measure are often absent in conservation
149 assessments, despite their relevance to understanding the long-term persistence and adaptive
150 capacity of species.
151 Our goal is to aggregate existing trait, geographic, and genetic data to investigate the
152 predictors of genetic diversity in amphibians. We focus on Nearctic amphibians as an initial test
153 case given the long history of study, tractable number of species, and presumed availability of
154 natural history and genetic information. This dataset includes all species native to Canada and the
155 U.S., a subset of which have ranges extending into Mexico. First, we compiled species' range

156 variables, natural history characteristics, and gene alignments using a series of open-access
157 databases and scripts. Second, we used a machine learning algorithm, random forest
158 classification, to classify species in terms of IUCN conservation status and identify important
159 predictors of IUCN status. Third, we used random forest regression to identify predictors of
160 intraspecific nucleotide diversity. Finally, we investigate the relevance of phylogenetic history
161 and demographic processes in contributing to current patterns of genetic diversity identified
162 across amphibian species.
163
164 **Materials and Methods**
165 *Species traits*
166 We compiled trait data for amphibian species focusing on traits related to life history and
167 ecology that may influence intraspecific genetic diversity. The following traits were initially
168 extracted from the AmphiBIO database (Oliveira *et al.* 2017): development mode (larval or
169 direct), maximum body size (mm), age at maturity (min and max years), longevity (max years),
170 clutch size (min and max number of eggs), and egg size (min and max in mm). We retrieved
171 additional body size information from the Peterson field guides to reptiles and amphibians of the
172 United States (Powell *et al.* 2016; Stebbins & McGinnis 2018). We used maximum snout-vent
173 length (SVL) for Anura (frogs), and maximum total length for Caudata (salamanders; note that
174 SVL was not always available for Caudata). The remaining traits of interest (neoteny, breeding
175 habitat, larval period, time to hatching, dispersal distance, and home range size) were compiled
176 primarily from species accounts in AmphibiaWeb <https://amphibiaweb.org>. Neoteny, or the
177 retention of juvenile characteristics as adults, was categorized as ‘yes’ (a species is always
178 neotenic), ‘no’ (neoteny has never been reported), or ‘some’ (neoteny has been reported in some

179 populations, but not all). Breeding habitat, including the habitat in which eggs and/or larvae
180 develop, was categorized as 'terrestrial', or 'aquatic', with aquatic species further categorized as
181 'permanent', 'ephemeral', or 'aquatic generalist'. Larval period was first recorded numerically
182 (min and max days) and then binned into categories: 'none' (direct developing), 'short' (0–90
183 days), 'mid' (91–365 days), or 'long' (>1 year). Time to hatching (min and max days), dispersal
184 distance (max recorded in meters), and home range size (max recorded in meters squared) were
185 recorded when available, but were unavailable for >40% of species and were thus excluded from
186 further analyses in the present study (Fig. S1).
187
188 *Range characteristics*
189 Species range maps for amphibians were downloaded as shapefiles from the IUCN Red
190 List of Threatened Species Version 6.1 (www.iucnredlist.org/) on 22 January 2019. Range maps
191 were manually edited in QGIS v. 2.18.2 as needed to exclude non-native ranges originally
192 included in IUCN range maps, and to generate range maps for species that have been described
193 recently from a portion of a former species' range (e.g., *Acris blanchardi* split from *Acris*
194 *crepitans*; Gamble *et al.* 2008). Altitude and bioclimatic variables were downloaded from
195 WorldClim v. 1.4 (Hijmans *et al.* 2005) at a spatial resolution of 2.5 minutes. We extracted the
196 following information from the list of focal amphibian species: range size (square km), latitude
197 (min, max, extent = max-min, midpoint = (max+min)/2), altitude (min, max, average, standard
198 deviation (SD)), slope (min, max, average, SD), and bioclimatic variables (min, max, average,
199 SD) using a custom script in R and the packages 'rgdal' (Bivand *et al.* 2019), 'raster' (Hijmans
200 2019a), and 'geosphere' (Hijmans 2019b). Analyses were conducted in R v. 3.6 (R Core Team
201 2019) and the scripts and datasets are available on Dryad (###).

*Genetic sequences*

Sequences were downloaded from GenBank in two ways and were processed using a series of R and Python scripts modified from Pelletier & Carstens (2018). First, we obtained georeferenced sequences for all Nearctic amphibian species by querying accession numbers linked to records downloaded from GBIF.org (28 November 2018). Second, we downloaded all GenBank sequences for the 19 focal amphibian families, regardless of whether they were georeferenced. For each dataset, we sorted sequences by species and gene and generated multiple sequence alignments for every species by gene combination using Mafft v. 7.402 (Katoh & Standley 2013). After summarizing the number of species and locus lengths for each species, we determined that cytochrome b (*cytb*) was the best-represented gene (Fig. S2). Alignments for *cytb* were then visually inspected in Geneious v. 8.1.9 (https://www.geneious.com) and edited to remove misaligned sequences prior to further analysis. Alignments for six additional anuran species were added from previously published datasets (Barrow *et al.* 2015, 2017) available on Dryad or as unannotated mitochondrial genomes on GenBank (Accession numbers MF198257–MF198403; note these were missed by our scripts because gene names are not defined).

For each species, we calculated nucleotide diversity using the nuc.div() function in the 'pegas' R package (Paradis 2010). To assess sensitivity of nucleotide diversity to sample size, or the number of individuals sequenced, we randomly subsampled sequences from each *cytb* dataset with replacement to generate 100 datasets for each sample size between two and 25 sequences. We calculated nucleotide diversity for each dataset and examined the variance in the estimates from the 100 replicate datasets for each sample size. Given the steep decline in variance observed as sample size increased above five (Fig. S3), we conducted analyses of genetic diversity with

species that included at least five sequences. Analyses were repeated using both the original value of nucleotide diversity for the full datasets and the median value of nucleotide diversity from the 100 datasets with five randomly sampled sequences.

*Random forest classification*

We aimed to determine whether conservation status of U.S. amphibian species can be predicted by range characteristics or species traits and which variables are most important. IUCN Red List categories for each species were obtained from AmphibiaWeb and include Least Concern (LC), Near Threatened (NT), Vulnerable (VU), Endangered (EN), Critically Endangered (CR), Extinct in the Wild (EW), Extinct (EX), or Data Deficient (DD). Given the relatively small number of species in each category, we excluded DD species and combined all categories other than LC into a single category ('nonLC'). We then applied random forest classification in the R package 'randomForest' (Liaw & Wiener 2002). To determine the importance of each variable, we examined the mean decrease in accuracy (MDA) after removing each variable from the predictive function. Downsampling was used to account for the unevenness of the response variable categories, subsampling the majority class ('LC') 100 times to match the sample size of the minority class ('nonLC'). We built each random forest with 2000 trees, averaged the MDA for each variable across iterations, and assessed model accuracy using classification error rates. After conducting analyses with all variables of interest, we built classifiers with reduced sets of predictors for simplicity of presentation and to potentially improve model accuracy.

246 *Random forest regression*

247 We used random forest regression to predict intraspecific nucleotide diversity based on
248 the range characteristics and species traits described above and two additional variables:
249 taxonomic family and the number of sequences for each species. Variable importance was
250 determined based on the percent increase in mean squared error (IncMSE) and model
251 performance was assessed as the percent variance explained by the models. We conducted
252 analyses for the dataset including species with at least five sequences and then conducted separate
253 analyses for Anura and Caudata because some species traits are not comparable or applicable to
254 both orders. For example, body size is measured differently between the two orders, only
255 salamanders exhibit neoteny, and very few frog species native to the region have direct
256 development. To examine the consistency of variable importance, we conducted 100 iterations
257 with 2000 trees each, then averaged the percent variance explained and IncMSE for each variable
258 across iterations. After conducting analyses with all variables of interest, we repeated analyses
259 with reduced sets of predictors as described above.

260

261 *Phylogenetic comparative methods*

262 Random forest analyses pointed to the importance of taxonomy for predicting nucleotide
263 diversity (see Results). To better understand these results, we downloaded phylogeny subsets
264 from VertLife.org for the U.S. amphibian species in our dataset. Phylogenetic relationships are
265 from Jetz and Pyron (2018) and are based on a time-calibrated posterior tree distribution for
266 nearly all extant amphibian species, generated using a combination of phylogenetic inference and
267 taxonomic assignment. We tested for phylogenetic signal in continuous life history traits and
268 nucleotide diversity using the phylosig() function and generated continuous trait maps using the

269 contMap() function in the R package 'phytools' (Revell 2012). Additional analyses and tree
270 pruning were conducted using the R package 'ape' (Paradis & Schliep 2019). To account for
271 phylogenetic relatedness between important continuous variables (see Results), we calculated
272 phylogenetically independent contrasts using the pic() function, which implements the method
273 described by Felsenstein (1985). We then assessed relationships between these continuous
274 variables using linear models.
275
276 *Species distribution models*
277 To test predictions about the potential influence of past demographic events on genetic
278 diversity, we modeled species distributions in the present and hindcasted these to the Last Glacial
279 Maximum (LGM). Occurrence records for each species were downloaded from GBIF.org (04
280 July 2019) using a polygon to encompass the ranges of all Nearctic amphibians. Custom scripts
281 (available from Dryad) were used to generate a file of occurrence localities for each species with
282 *cytb* data, removing any points that fell outside the IUCN range map (0.5 degree buffer width) for
283 that species, and thinning occurrence points to retain unique localities at a threshold of 10 km
284 using the R package 'spThin' (Aiello-Lammens *et al.* 2015). For species with a very large
285 number of occurrences, we randomly sampled 3,000 localities prior to thinning to save
286 computational time, and for species with small ranges and <25 localities, we did not thin datasets.
287 We estimated species distribution models (SDMs) for 138 species with *cytb* data.
288 We downloaded 19 bioclimatic layers from WorldClim v. 1.4 (Hijmans et al. 2005) for
289 current climate at a resolution of 30 seconds (~1 $km^2$) and for the LGM (CCSM4) at the highest
290 resolution available (2.5 min). We conducted SDM analyses in R using the packages 'biomod2'
291 (Thuiller *et al.* 2019), 'raster' (Hijmans 2019a), 'rgdal' (Bivand *et al.* 2019), 'rgeos' (Bivand &

Rundel 2019), and 'HDInterval' (Meredith & Kruschke 2018). Bioclimatic layers for each time period were stacked, clipped to the extent of North America (-170, -52, 12, 72), and masked by current land borders. We tested for correlations between current climate layers and removed highly correlated variables (>0.8), retaining five temperature (bio2, bio3, bio7, bio8, and bio10) and five precipitation variables (bio13, bio14, bio15, bio18, and bio19). Ensemble models were generated with Generalized Linear Models (GLM), Random Forest (RF), and Maxent (MAXENT.phillips) with five replicates, 80% of the data used for training and 20% for testing, and the 'ROC' evaluation metric. Models were hindcasted onto the LGM climate and were projected onto the current climate with the same resolution as the LGM (2.5 min).

For each time period, the five replicate models were averaged prior to calculating stability metrics. We converted models to binary presence/absence layers using two different thresholds, either 95% or 99%. The resulting models include the total area in which the model probability is equal to or greater than the probability threshold where 95% or 99% of the occurrence points are contained. For each threshold, we calculated the total suitable area for current and LGM models, determined the area of overlap between the two time periods, and calculated the overlap index (OI) metric described by Hijmans and Graham (2006). The OI describes the proportion of the current range that was also suitable during the LGM, with higher values indicating higher niche stability. We also determined whether the suitable area in the LGM was smaller (range contraction in the LGM) or larger (range expansion in the LGM) than the suitable area in the current model.

## Results

*Data summary*

315 We compiled data for 299 amphibian species native to the U.S. and Canada, 44 of which
316 have ranges extending into Mexico. The dataset consists of 197 salamanders in nine families and
317 102 frogs in 10 families. Complete trait data were available for body size, development, breeding
318 habitat, and neoteny, with various degrees of missing data for the remaining traits (Fig. S1).
319 Range characteristics derived from IUCN range maps were available for 268 species, with data
320 unavailable for recently described species (since 2008) or those with very small ranges. IUCN
321 conservation status has been assigned for 253 species, with 169 species ranked as "Least
322 Concern" and 84 species in one of the other risk categories. We assembled 4,759 georeferenced
323 GenBank sequences that were linked to GBIF records, but only 37 species (20 salamanders, 17
324 frogs) were represented by at least five sequences from at least two localities. Preliminary
325 analyses indicated there was no predictive power in models of genetic structure with this small
326 dataset. We therefore focused analyses on non-spatial genetic datasets and used metrics of overall
327 intraspecific genetic diversity. We assembled 42,067 GenBank sequences from 263 species,
328 which consisted primarily of mitochondrial genes (Fig. S2). The best represented gene was *cytb*,
329 with at least five sequences available for each of 147 species.
330
331 *Predictors of IUCN Conservation Risk*
332 For the initial model of IUCN status with 253 U.S. amphibian species, the out-of-bag
333 (OOB) estimate of error rate for the random forest classifier including all variables of interest was
334 17.8%, with a higher classification error for the 'nonLC' class (31.0%) compared to the 'LC'
335 class (11.2%). The OOB error rate for a model with a subset of predictors was 15.4% overall,
336 with 30.0% for 'nonLC' and 8.3% for 'LC'. Using downsampling, average error rates were
337 20.1% (overall) and error rates for the two classes were more similar: 19.8% (nonLC), and 20.5%

(LC). For the dataset with a subset of predictors, the average error rates were 19.6% (overall), 17.0% (nonLC), and 22.2% (LC).

The most important predictors of IUCN status were consistently total range size and latitudinal extent (Fig. 1a), regardless of which predictor set was used and whether or not downsampling was employed. Other important predictors were the number of GBIF occurrence records for a species and the standard deviation of bioclimatic variables including temperature seasonality (bio4), precipitation of the driest month (bio14), minimum temperature of the coldest month (bio6), and mean temperature of the driest quarter (bio9). Species traits were not predictive of IUCN status for this dataset. As expected given the criteria used for IUCN risk assessments, which incorporate aspects of population size and geographic range, the species ranked as Least Concern had larger ranges (Fig. 1b; $t = 8.14$, $p = 7.9 \times 10^{-14}$), broader latitudinal extents (Fig. 1c; $t = 11.43$, $p < 2.2 \times 10^{-16}$), more GBIF occurrence records (Fig. 1d; $t = 5.80$, $p = 2.9 \times 10^{-8}$), and more variation in bioclimatic variables (e.g., bio4: Fig. 1e; $t = 9.68$, $p < 2.2 \times 10^{-16}$) than non-LC species.

*Predictors of intraspecific genetic diversity*

Full models with all predictors included 137 amphibian species with *cytb* data. These models explained an average of 19.6% variance in original nucleotide diversity and 24.9% variance in median nucleotide diversity based on 100 sequence datasets with five randomly sampled sequences. The most important predictors were consistently taxonomic family and the number of sequences, indicating that nucleotide diversity is influenced by phylogenetic relatedness and sample size. Some bioclimatic variables, including the average precipitation of the wettest month (bio13), average precipitation of the warmest quarter (bio18), average

precipitation of the driest quarter (bio17), and the standard deviation of mean temperature of the driest quarter (bio9), were also ranked as important. Models with a subset of predictors explained an average of 25.6% variance in original nucleotide diversity and 30.1% variance in median nucleotide diversity. Taxonomic family, the number of sequences, and average precipitation of the wettest month (bio13) remained the most important predictors of nucleotide diversity, while species traits were not predictive of nucleotide diversity for this set of species (Fig. 2a). Given the consistent results with both the original and median nucleotide diversity metrics, we hereafter report only results for the latter.

When we considered each order separately, the Anura dataset included only 39 species with *cytb* data and there was no predictive power in these models (variance explained was negative). Full models for Caudata, however, included 98 species and explained an average of 20.5% of the variance in median nucleotide diversity. The most important predictors of nucleotide diversity were consistently the number of sequences and the minimum latitude of a species range. Models with a subset of predictors explained an average of 28.9% of the variance in median nucleotide diversity. The number of sequences and minimum latitude were the most important predictors (Fig. 2b). Models with fewer species (n=67) and additional trait information available explained an average of 28.3% variance in median nucleotide diversity. Minimum latitude remained the most important predictor and species traits were not important predictors of nucleotide diversity (Fig. 2c).

*Phylogenetic signal, minimum latitude, and climatic niche stability*

We found phylogenetic signal in species-wide nucleotide diversity (lambda = 0.152, p = 0.008). Phylogenetic signal is also present in several traits related to reproduction and ecology

384 including clutch size, larval period, and body size (Table S1), although these traits do not predict
385 nucleotide diversity (described above and further tested with phylogenetic generalized linear
386 mixed models; results not shown). The salamander families Sirenidae and Plethodontidae
387 included multiple species with high nucleotide diversity (Fig. 3), but within Caudata,
388 phylogenetic signal in nucleotide diversity was not significant (lambda = 0.105, p = 0.605).
389 Notably, there was no phylogenetic signal in the number of sequences sampled, suggesting that
390 the phylogenetic signal present in nucleotide diversity is not an artefact of sampling.
391 After accounting for phylogenetic relationships there was a negative correlation between
392 species-wide nucleotide diversity and minimum latitude in Caudata (Fig. 4). Species with more
393 northern ranges tended to have lower nucleotide diversity (98-species dataset: $R^2 = 0.057$, p =
394 0.0102; 67-species dataset: $R^2 = 0.115$, p = 0.003; Fig. S4). Results from SDMs did not,
395 however, support the hypothesis that species at higher latitudes were more heavily impacted by
396 glaciation at the LGM. At the 95% lowest presence threshold, we inferred LGM contraction of 76
397 Caudata species and LGM expansion for 20 species, but there was no difference in minimum
398 latitude of the two groups (Student's t-test: t = -1.59, p = 0.12). At the same threshold, 49
399 Caudata species had no overlap between the LGM and current models, indicating low niche
400 stability, while 47 species had at least some overlap, indicating predicted areas of stability. There
401 was no difference in minimum latitude of the two groups (t = 0.83, p = 0.41). Results were
402 similar at the 99% threshold, with no difference in minimum latitude between species inferred to
403 have undergone LGM contraction or expansion (t = -1.03, p = 0.31) and no difference in
404 minimum latitude between species that had no or some niche overlap between LGM and current
405 models (t = 0.07, p = 0.95).
406

## Discussion

Combining repurposed data with machine learning is a powerful strategy for addressing broad questions in evolutionary biology and conservation (Sidlauskas *et al.* 2010; Howard & Bickford 2014). Here, we apply these approaches to investigate a classic question: What factors determine intraspecific genetic variation? We highlight the prospects and challenges of combining data repurposing and machine learning to address questions that involve many complex data types and predictors. We then discuss our results from a case study that not only leads to insights in Nearctic amphibian biogeography, but also emphasizes the relevance of continued studies in this framework to inform biodiversity conservation.

### *Prospects and challenges for applying machine learning to repurposed data*

Data repurposing can meet some of the same goals as meta-analysis, such as establishing what we (as a field) understand about key patterns and pinpointing underlying mechanisms. Both types of investigation play an important role by potentially identifying connections among many variables. However, machine learning techniques make data repurposing far more appealing than meta-analysis as a strategy for extracting the most information out of publicly available data. While meta-analyses certainly provide important insights, genetic datasets collected and analyzed separately across studies, which may use disparate sample designs and report incongruent metrics, can limit the number of useable datasets and make interpretation difficult (Emel & Storfer 2012; López-Uribe *et al.* 2019). Once data are assembled, machine learning enables researchers to analyze very large datasets that incorporate many complex variables. By including a large suite of potential predictors, random forests can identify potential links among variables in an unbiased manner. This feature is useful for complex topics such as the determinants of genetic

diversity, where researchers might be biased by previous findings and focus only on a subset of variables. One caveat is that machine learning analysis is only as useful as the response variable used for a particular question. For example, while predictive models of binary questions (e.g., 'LC' versus 'nonLC' IUCN status) will likely be able to explain more variation than those that rely on regression techniques, not all questions can easily be reduced to a binary outcome (e.g., a continuous response such as nucleotide diversity). In this case, with a moderate number of species, we were able to predict IUCN status with fairly high accuracy (~20% error rate), while the best random forest regression models explained <30% of the variance in nucleotide diversity.

The application of machine learning techniques to repurposed data offers great potential for addressing key questions in ecology and evolutionary biology, but several challenges remain (Table 1). Our focus on assembling a curated set of species allowed us to identify several errors that would have been missed by automated pipelines. While useful for future work, this effort limited both the taxonomic and geographic scale of our study. For example, it seems plausible that the associations between life history traits and genetic diversity identified previously may only be apparent across broader taxonomic scales (e.g., Romiguier *et al.* 2014) or when more species can be included that span substantial trait variation (e.g., Brüniche-Olsen *et al.* 2018). Expanding the set of taxa from this study to a broader regional or global scale is a clear next step that will require addressing two substantial challenges. Compiling meaningful trait datasets for large numbers of species is time-consuming, and for many species, natural history information is not easily available. Solving this challenge will require collaborative efforts by the research community to standardize trait information, contribute to trait datasets, and make these datasets easily available (e.g., Oliveira *et al.* 2017). In the future, increasing efforts to digitize natural history collections and develop informatics tools will ultimately enable the incorporation of

453 individual and population-level trait variation into these types of studies (Guralnick *et al.* 2016;
454 Hedrick *et al.* 2020). The second major challenge to extending the scope of this study is the
455 limited amount of georeferenced genetic data available for most species.

456 Two main axes of sampling effort are important to interpreting intraspecific genetic
457 diversity: the number (and location) of individuals sampled and the number (and type) of loci
458 sampled. Researchers have typically had to make a trade-off between these two for logistical
459 reasons. Although the number of loci that can feasibly be sequenced is growing rapidly (Garrick
460 *et al.* 2015), the vast majority of available sequences are not directly linked to localities (Marques
461 *et al.* 2013). Increasing the amount of georeferenced genetic data available per species will enable
462 within-population variation to be disentangled from among-population variation and will foster a
463 better understanding of the mechanisms that influence intraspecific genetic diversity. Currently,
464 the genetic sequences available for most animal species are mitochondrial genes and it remains
465 unclear how representative mitochondrial DNA actually is for describing intraspecific genetic
466 diversity (Bazin *et al.* 2006; Mulligan *et al.* 2006; Galtier *et al.* 2009). Furthermore, it is likely
467 that available molecular markers do not accurately reflect adaptive variation (Mittell *et al.* 2015),
468 which is the component of variation most relevant to designing conservation strategies that
469 promote species persistence (Moritz 2002). Despite these issues, comparing genetic diversity
470 among species yields important insights, and new findings will be enhanced by continued efforts
471 to link existing genetic data to localities and to collect genome-scale data from a large number of
472 non-model species encompassing variation in geographic complexity and life history traits.
473

*Predictors of genetic diversity in Nearctic amphibians*

Amphibians play critical ecological roles as consumers (Whiles *et al.* 2006), prey (Zipkin *et al.* 2020), and indicators of aquatic and terrestrial environmental health (Vitt *et al.* 1990; Welsh & Ollivier 1998; Kerby *et al.* 2010). Given these roles, additional efforts to synthesize diverse datasets are warranted to inform conservation priorities. Range size and related variables are important predictors of conservation risk, but they are not predictive of genetic diversity in Nearctic amphibians. Rather, minimum latitude predicted genetic diversity within salamanders, with generally lower intraspecific genetic diversity for species at higher (i.e., more northern) latitudes. Intraspecific genetic diversity also appears to be phylogenetically conserved, a finding perhaps explained by conserved life history traits such as clutch size or body size that are associated with intraspecific genetic diversity (e.g., Romiguier *et al.* 2014; Paz *et al.* 2015). While life history traits are not predictors of genetic diversity for this dataset, taxonomic family was identified as important. This is likely driven by two salamander families, Sirenidae and Plethodontidae, which contain most species with the highest intraspecific diversity. Since more than half of U.S. amphibian species are plethodontid salamanders, the lack of a relationship between traits and genetic diversity might be explained by the limited trait variation (e.g., all are relatively small-bodied and have small clutch sizes) but considerable variation in genetic diversity within the group (Fig. 3).

The relevance of geography for understanding patterns of genetic diversity is well known (Wright 1943; Hewitt 2000). The negative association between latitude and intraspecific genetic diversity likely relates to population history since high intraspecific diversity at lower latitudes may be attributed to long-term stability (Miraldo *et al.* 2016) and low diversity at northern latitudes may be explained by population bottlenecks during glacial periods (Hewitt 2000).

Published phylogeographic studies on some of the species included in our study support this prediction, attributing lower genetic diversity at higher latitudes to postglacial range expansion (Highton & Webster 1976; Carstens *et al.* 2005; Radomski *et al.* 2020). However, the effect of past climate cycles on intraspecific genetic diversity is clearly dependent on refugial dynamics, as a species that is restricted into a single glacial refugium would exhibit different patterns than one where populations were isolated in separate refugia. One complicating factor that we did not address here is whether current species taxonomy is an appropriate unit of comparison, or whether some species include cryptic lineages that have not been formally described. Future studies with georeferenced localities and assignment of sequences to lineages to estimate within-population diversity should help clarify these issues further.

*Conclusions*

Identifying the predictors of intraspecific genetic diversity and the scale for which they are predictive are important goals for both evolutionary biology and applied conservation. We demonstrated the potential for combining repurposed data with machine learning techniques to investigate these issues in Nearctic amphibians. Our findings indicate that life history traits do not predict genetic diversity in this dataset, but future studies should incorporate additional species and trait variation on a global scale. Range size is an important criterion for assessing species conservation risk, but it does not predict genetic diversity within Nearctic amphibians. Instead, we found that minimum latitude was an important predictor of genetic diversity within salamanders, suggesting that this aspect of a species' range represents an important component to consider when assessing species risk. Northern latitude species that harbor low genetic diversity may be even more vulnerable to future climate change scenarios or disease outbreaks. Identifying

520 these potential risks by combining new and existing datasets can lead to proactive management
521 strategies that preserve remaining genetic diversity.
522
523 **Acknowledgements**
524 We thank Drew Duckett, Flavia Mol Lanna, Danielle Parsons, Megan Smith, Tara Pelletier,
525 Jamin Wieringa, and Tamaki Yuri for helpful discussions, and the many contributors to
526 AmphibiaWeb, GenBank, GBIF, and the IUCN Red List for archiving the data that made this
527 study possible. This work was supported by NSF DBI 1910623, The Ohio State University
528 (OSU) President's Postdoctoral Scholars Program, and the OSU Department of Evolution,
529 Ecology and Organismal Biology. EMF thanks the Coordenação de Aperfeiçoamento de Pessoal
530 de Nível Superior (CAPES) for his doctoral fellowship (process #88881.170016/2018).

## References


Aiello-Lammens ME, Boria RA, Radosavljevic A, Vilela B, Anderson RP (2015) spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, **38**, 541–545.

AmphibiaWeb (2018) <https://amphibiaweb.org>. *University of California, Berkeley, CA, USA. Accessed 18 Dec 2018.*

Barrow LN, Bigelow AT, Phillips CA, Lemmon EM (2015) Phylogeographic inference using Bayesian model comparison across a fragmented chorus frog species complex. *Molecular Ecology*, **24**, 4739–4758.

Barrow LN, Soto-Centeno JA, Warwick AR, Lemmon AR, Moriarty Lemmon E (2017) Evaluating hypotheses of expansion from refugia through comparative phylogeography of south-eastern Coastal Plain amphibians. *Journal of Biogeography*, **44**, 2692–2705.

Bazin E, Glémin S, Galtier N (2006) Mitochondrial Genetic Diversity in Animals. *Science*, **312**, 570–572.

Bivand R, Keitt T, Rowlingson B (2019) rgdal: Bindings for the "Geospatial" Data Abstraction Library. R package version 1.4-8. , https://cran.r-project.org/package=rgdal.

Bivand R, Rundel C (2019) rgeos: Interface to Geometry Engine - Open Source ('GEOS'). R package version 0.4-3. https://CRAN.R-project.org/package=rgeos.

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

Breiman L (2001) Random forests. *Machine Learning*, **45**, 5–32.

Brüniche-Olsen A, Kellner KF, Anderson CJ, DeWoody JA (2018) Runs of homozygosity have utility in mammalian conservation and evolutionary studies. *Conservation Genetics*, **19**, 1295–1307.

Carpenter SR, Armbrust EV, Arzberger PW *et al.* (2009) Accelerate Synthesis in Ecology and Environmental Sciences. *BioScience*, **59**, 699–701.

Carstens BC, Brunsfeld SJ, Demboski JR, Good JM, Sullivan J (2005) Investigating the evolutionary history of the Pacific Northwest mesic forest ecosystem: hypothesis testing within a comparative phylogeographic framework. *Evolution*, **59**, 1639–1652.

Chamberlain SA, Szöcs E (2013) taxize: taxonomic search and retrieval in R. *F1000Research*, **2**.

Chen J, Glémin S, Lascoux M (2017) Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular Biology and Evolution*, **34**, 1417–1428.

Darwin C (1859) *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. John Murray, London.

Ekroth AKE, Rafaluk-Mohr C, King KC (2019) Host genetic diversity limits parasite success beyond agricultural systems: A meta-analysis. *Proceedings of the Royal Society B: Biological Sciences*, **286**.

Ellegren H, Galtier N (2016) Determinants of genetic diversity. *Nature Reviews Genetics*, **17**, 422–433.

Emel SL, Storfer A (2012) A decade of amphibian population genetic studies: Synthesis and recommendations. *Conservation Genetics*, **13**, 1685–1689.

Field R, Hawkins BA, Cornell H V. *et al.* (2009) Spatial species-richness gradients across scales: A meta-analysis. *Journal of Biogeography*, **36**, 132–147.

Frankham R (1995) Effective population size/adult population size ratios in wildlife: A review.

*Genetics Research*, **66**, 95–107.
Frankham R (2005) Genetics and extinction. *Biological Conservation*, **126**, 131–140.
Galtier N, Nabholz B, GlÉmin S, Hurst GDD (2009) Mitochondrial DNA as a marker of molecular diversity: A reappraisal. *Molecular Ecology*, **18**, 4541–4550.
Gamble T, Berendzen PB, Bradley Shaffer H, Starkey DE, Simons AM (2008) Species limits and phylogeography of North American cricket frogs (Acris: Hylidae). *Molecular Phylogenetics and Evolution*, **48**, 112–125.
Garrick RC, Bonatelli IAS, Hyseni C *et al.* (2015) The evolution of phylogeographic datasets. *Molecular Ecology*, **24**, 1164–1171.
GBIF.org GBIF.org (28 November 2018) GBIF Occurrence Download https://doi.org/10.15468/dl.r6bshz.
GBIF.org GBIF.org (04 July 2019) GBIF Occurrence Download https://doi.org/10.15468/dl.rk2srg.
González-del-Pliego P, Freckleton RP, Edwards DP *et al.* (2019) Phylogenetic and Trait-Based Prediction of Extinction Risk for Data-Deficient Amphibians. *Current Biology*, **29**, 1557-1563.e3.
Grundler MR, Singhal S, Cowan MA, Rabosky DL (2019) Is genomic diversity a useful proxy for census population size? Evidence from a species-rich community of desert lizards. *Molecular Ecology*.
Guralnick RP, Zermoglio PF, Wieczorek J *et al.* (2016) The importance of digitized biocollections as a source of trait data and a new VertNet resource. *Database*, **2016**.
Hedrick BP, Heberling JM, Meineke EK *et al.* (2020) Digitization and the Future of Natural History Collections. *BioScience*, **XX**, 1–9.
Hedrick PW, Kalinowski ST (2000) Inbreeding depression in conservation biology. *Annual Review of Ecology and Systematics*, **31**, 139–162.
Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.
Highton R, Webster TP (1976) Geographic protein variation and divergence in populations of the salamander Plethodon cinereus. *Evolution*, **30**, 33–45.
Hijmans RJ (2019a) raster: Geographic Data Analysis and Modeling. R package version 3.0-7. https://CRAN.R-project.org/package=raster.
Hijmans RJ (2019b) geosphere: Spherical Trigonometry. R package version 1.5-10. https://CRAN.R project.org/package=geosphere.
Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.
Howard SD, Bickford DP (2014) Amphibians over the edge: Silent extinction risk of Data Deficient species. *Diversity and Distributions*, **20**, 837–846.
IUCN (2019) The IUCN Red List of Threatened Species. *Version 2019-3*, http://www.iucnredlist.org. Downloaded 6 Mar 2020.
Jamieson IG, Allendorf FW (2012) How does the 50/500 rule apply to MVPs? *Trends in Ecology and Evolution*, **27**, 578–584.
Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.
Kerby JL, Richards-Hrdlicka KL, Storfer A, Skelly DK (2010) An examination of amphibian sensitivity to environmental contaminants: Are amphibians poor canaries? *Ecology Letters*, **13**, 60–67.

621 Laikre L, Hoban S, Bruford MW *et al.* (2020) Post-2020 goals overlook genetic diversity.
622 *Science*, **367**, 1083–1085.
623 Lande R (1988) Genetics and demography in biological conservation. *Science*, **241**, 1455–1460.
624 Leffler EM, Bullaughey K, Matute DR *et al.* (2012) Revisiting an Old Riddle: What Determines
625 Genetic Diversity Levels within Species? *PLoS Biology*, **10**.
626 Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News*, **2**, 18–22.
627 López-Uribe MM, Jha S, Soro A (2019) A trait-based approach to predict population genetic
628 structure in bees. *Molecular Ecology*, **28**, 1919–1929.
629 Mackintosh A, Laetsch DR, Hayward A *et al.* (2019) The determinants of genetic diversity in
630 butterflies. *Nature Communications*, **10**, 1–9.
631 Marques AC, Maronna MM, Collins AG (2013) Putting GenBank Data on the Map. *Science*, **341**,
632 1341.
633 Meredith M, Kruschke J (2018) HDInterval: Highest (Posterior) Density Intervals. R package
634 version 0.2.0. https://CRAN.R-project.org/package=HDInterval.
635 Mims MC, Hauser L, Goldberg CS, Olden JD (2016) Genetic Differentiation , Isolation-by-
636 Distance , and Metapopulation Dynamics of the Arizona Treefrog ( Hyla wrightorum ) in an
637 Isolated Portion of Its Range. , 1–23.
638 Miraldo A, Li S, Borregaard MK *et al.* (2016) An Anthropocene map of genetic diversity.
639 *Science*, **353**, 1532–1535.
640 Mittell EA, Nakagawa S, Hadfield JD (2015) Are molecular markers useful predictors of adaptive
641 potential? *Ecology Letters*, **18**, 772–778.
642 Moritz C (2002) Strategies to protect biological diversity and the evolutionary processes that
643 sustain it. *Systematic Biology*, **51**, 238–254.
644 Moritz C, Faith DP (1998) Comparative phylogeography and the identification of genetically
645 divergent areas for conservation. *Molecular Ecology*, **7**, 419–429.
646 Mulligan CJ, Kitchen A, Miyamoto MM (2006) Comment on "Population size does not influence
647 mitochondrial genetic diversity in animals." *Science*, **314**, 1390.
648 Nori J, Villalobos F, Loyola R (2018) Global priority areas for amphibian research. *Journal of*
649 *Biogeography*, **45**, 2588–2594.
650 Oliveira BF, São-Pedro VA, Santos-Barrera G, Penone C, Costa GC (2017) AmphiBIO, a global
651 database for amphibian ecological traits. *Scientific Data*, **4**, 1–7.
652 Paradis E (2010) pegas: an R package for population genetics with an integrated-modular
653 approach. *Bioinformatics*, **26**, 419–420.
654 Paradis E, Schliep K (2019) Ape 5.0: An environment for modern phylogenetics and evolutionary
655 analyses in R. *Bioinformatics*, **35**, 526–528.
656 Parr CS, Guralnick R, Cellinese N, Page RDM (2012) Evolutionary informatics: Unifying
657 knowledge about the diversity of life. *Trends in Ecology and Evolution*, **27**, 94–103.
658 Paz-Vinas I, Loot G, Hermoso V *et al.* (2018) Systematic conservation planning for intraspecific
659 genetic diversity. *Proceedings of the Royal Society B: Biological Sciences*, **285**.
660 Paz A, Ibáñez R, Lips KR, Crawford AJ (2015) Testing the role of ecology and life history in
661 structuring genetic variation across a landscape: A trait-based phylogeographic approach.
662 *Molecular Ecology*, **24**, 3723–3737.
663 Pelletier TA, Carstens BC (2018) Geographical range size and latitude predict population genetic
664 structure in a global survey. *Biology Letters*, **14**.
665 Peters DPC (2010) Accessible ecology: Synthesis of the long, deep, and broad. *Trends in Ecology*

*and Evolution*, **25**, 592–601.
Pope LC, Liggins L, Keyse J, Carvalho SB, Riginos C (2015) Not the time or the place: The missing spatio-temporal link in publicly available genetic data. *Molecular Ecology*, **24**, 3802–3809.
Powell R, Conant R, Collins JT (2016) *Peterson Field Guide to Reptiles and Amphibians of Eastern and Central North America 4th Edition*. Houghton Mifflin Harcourt, Boston, MA.
R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Radomski TP, Hantak MM, Brown AD, Kuchta SR (2020) Multilocus phylogeography of eastern red-backed salamanders: cryptic appalachian diversity and postglacial range expansion. *Herpetologica*, **76**, 61–73.
Reichman OJ, Jones MB, Schildhauer MP (2011) Challenges and opportunities of open data in ecology. *Science*, **331**, 703–705.
Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223.
Rodríguez A, Börner M, Pabijan M *et al.* (2015) Genetic divergence in tropical anurans: deeper phylogeographic structure in forest specialists and in topographically complex regions. *Evolutionary Ecology*, **29**, 765–785.
Romiguier J, Gayral P, Ballenghien M *et al.* (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, **515**, 261–263.
Semlitsch RD (2008) Differentiating Migration and Dispersal Processes for Pond-Breeding Amphibians. *Journal of Wildlife Management*, **72**, 260–267.
Sexton JP, Hangartner SB, Hoffmann AA (2014) Genetic isolation by environment or distance: Which pattern of gene flow is most common? *Evolution*, **68**, 1–15.
Sidlauskas B, Ganapathy G, Hazkani-Covo E *et al.* (2010) linking big: The continuing promise of evolutionary synthesis. *Evolution*, **64**, 871–880.
Singhal S, Huang H, Title PO *et al.* (2017) Genetic diversity is largely unpredictable but scales with museum occurrences in a species-rich clade of Australian lizards. *Proceedings of the Royal Society B: Biological Sciences*, **284**.
Smith BT, Seeholzer GF, Harvey MG, Cuervo AM, Brumfield RT (2017) A latitudinal phylogeographic diversity gradient in birds. *PLoS Biology*, **15**, 1–25.
Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261–4293.
Soltis DE, Soltis PS, Soltis DE, Soltis PS, Soltis DE (2016) Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. *Plant Diversity*, **38**, 264–270.
Stebbins RC, McGinnis SM (2018) *Peterson Field Guide to Western Reptiles and Amphibians 4th Edition*. Houghton Mifflin Harcourt, Boston, MA.
Stuart SN, Chanson JS, Cox NA *et al.* (2004) Status and trends of amphibian declines and extinctions worldwide. *Science*, **306**, 1783–1786.
Thuiller W, Georges D, Engler R, Breiner F (2019) biomod2: Ensemble Platform for Species Distribution Modeling. R package version 3.3-7.1. https://CRAN.R-project.org/package=biomod2.
Vitt LJ, Caldwell JP, Wilbur HM, Smith DC (1990) Amphibians as harbingers of decay. *BioScience*, **40**, 418.
Wallace AR (1869) *The Malay Archipelago: The land of the orang-utan, and the bird of*

*paradise. A narrative of travel, with studies of man and nature*. Macmillan, London.
Wang IJ (2012) Environmental and topographic variables shape genetic structure and effective population sizes in the endangered Yosemite toad. *Diversity and Distributions*, **18**, 1033–1041.
Welsh HH, Ollivier LM (1998) Stream amphibians as indicators of ecosystem stress: A case study from California's redwoods. *Ecological Applications*, **8**, 1118–1132.
Whiles MR, Lips KL, Pringle CM *et al.* (2006) The effects of amphibian population declines on the structure and function of Neotropical stream ecosystems. *Frontiers in Ecology and the Environment*, **4**, 27–34.
White EP, Ernest SKM, Kerkhoff AJ, Enquist BJ (2007) Relationships between body size and abundance in ecology. *Trends in Ecology and Evolution*, **22**, 323–330.
Wright S (1943) Isolation by Distance. *Genetics*, **28**, 114–138.
Zipkin EF, DiRenzo G V., Ray JM, Rossman S, Lips KR (2020) Tropical snake diversity collapses after widespread amphibian loss. *Science*, **367**, 814–816.

## Data Accessibility

Scripts for data processing and analysis, DNA sequence alignments, and trait and climate datasets will be uploaded to Dryad doi:XXXX (Barrow *et al.* 2020).

## Author Contributions

LNB and BCC conceived the ideas; LNB, EMF, and CEPT gathered and analyzed the data; LNB and BCC wrote the paper with input from all authors.

## Tables and Figures

## Figure Captions

### Figure 1

Predictors of IUCN status for U.S. amphibians. a) Average variable importance (mean decrease in accuracy) for each predictor from 100 iterations is shown. The Least Concern (LC) class was downsampled to match the sample size of non-Least Concern (nonLC) species. b-e) Comparison of LC and nonLC species for the top four predictors. As expected based on the criteria used for IUCN status rankings, LC species have larger ranges and associated variables compared to nonLC species.

### Figure 2

Predictors of intraspecific *cytb* nucleotide diversity for a) 137 species of U.S. amphibians, b) 98 species of Caudata, and c) 67 species of Caudata with additional species traits. Average variable importance (increase in mean squared error) for each predictor from 100 iterations is shown.

### Figure 3

Phylogeny of U.S. amphibian species with *cytb* nucleotide diversity mapped as a continuous trait. Phylogenetic relationships are based on Jetz and Pyron (2018) and were downloaded from VertLife.org. The median nucleotide diversity from 100 datasets with five randomly-sampled sequences was mapped onto the tree using the contMap() function on phytools (Revell 2012).

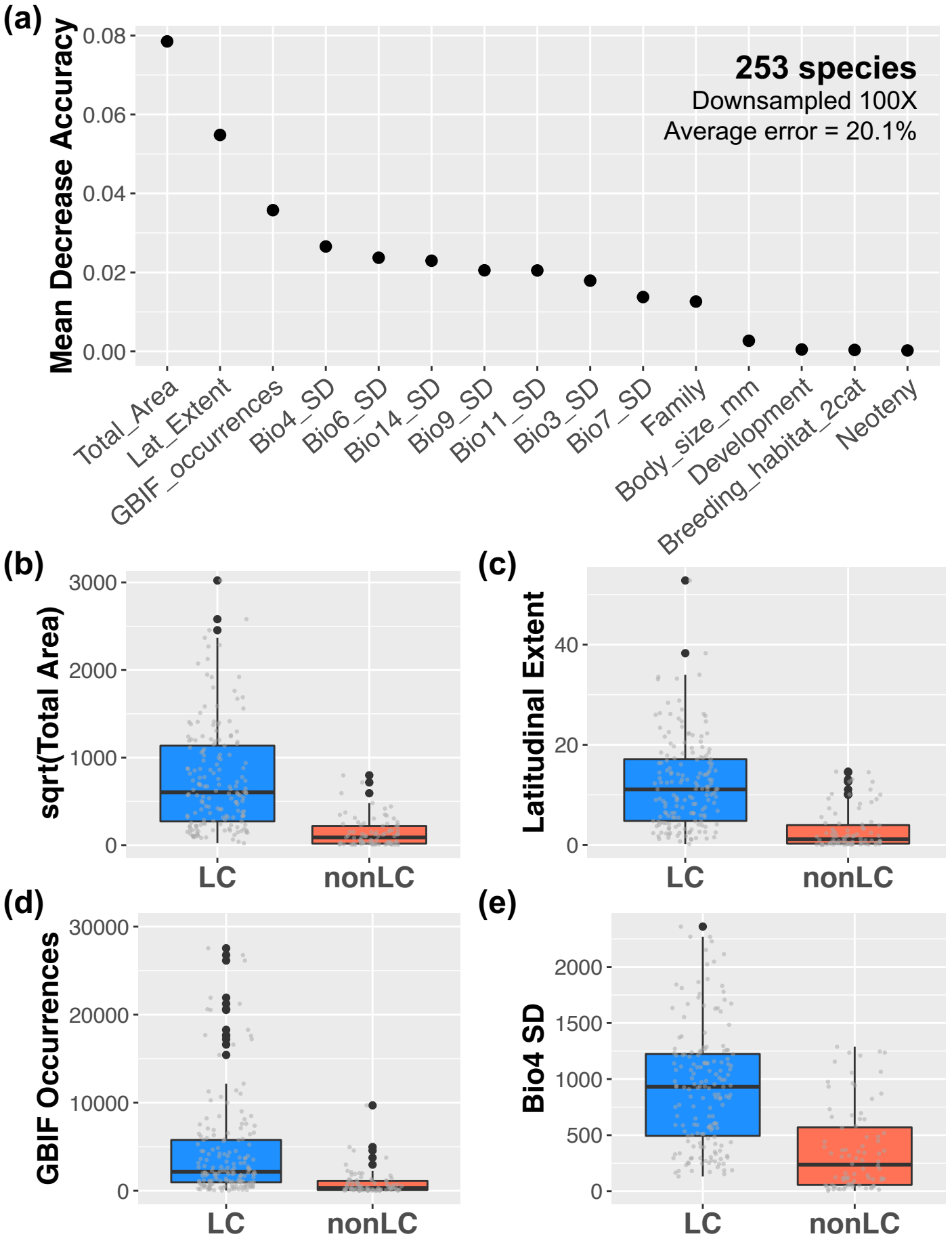
### Figure 4

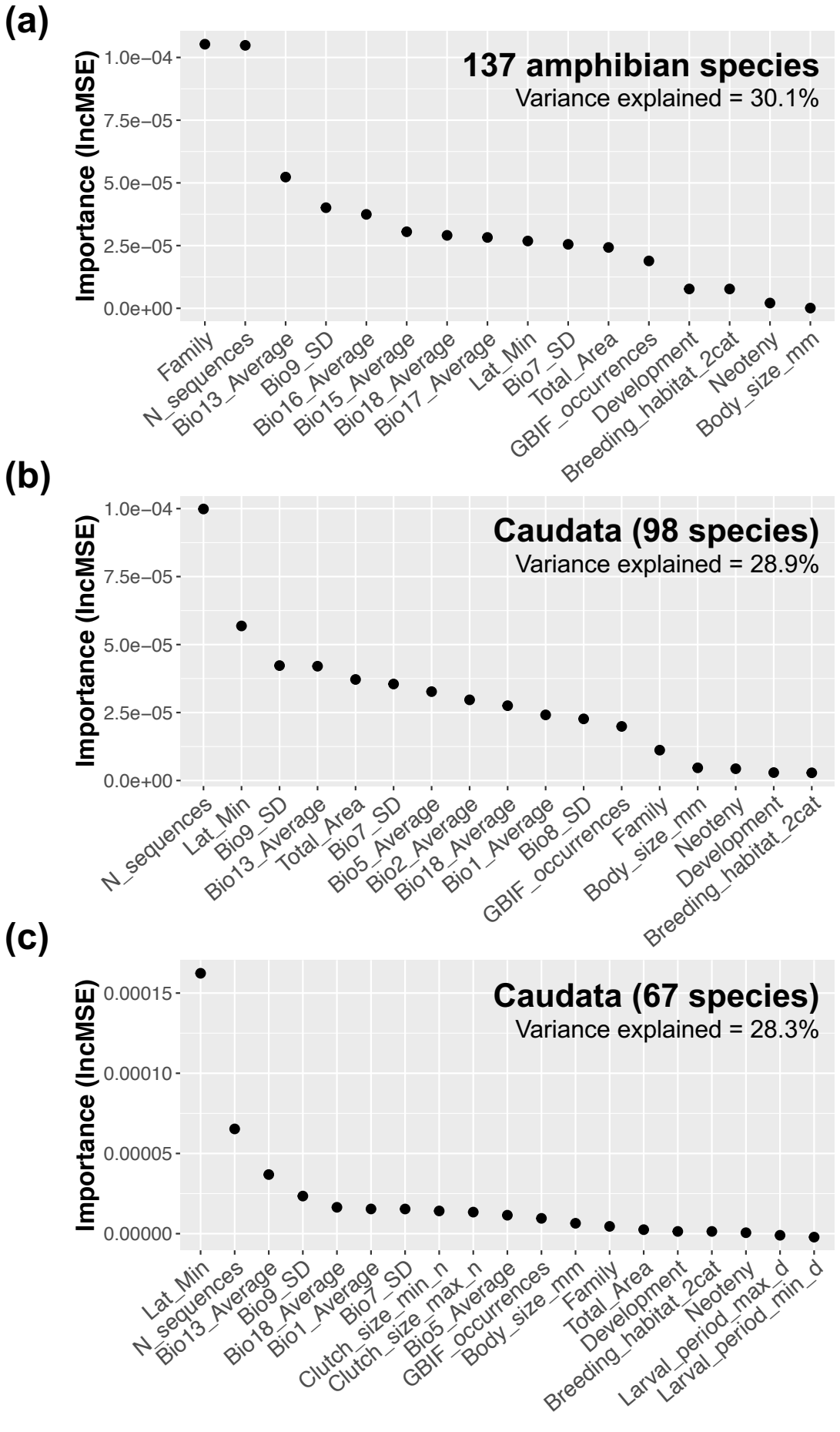
Minimum latitude and *cytb* nucleotide diversity for 98 Caudata species. a) Nucleotide diversity has a negative relationship with the minimum latitude of a species range. The median nucleotide diversity from 100 datasets with five randomly-sampled sequences is shown. b) Phylogenetic independent contrasts for the same metrics in a) using the pruned phylogeny from VertLife.org (Jetz and Pyron 2018).

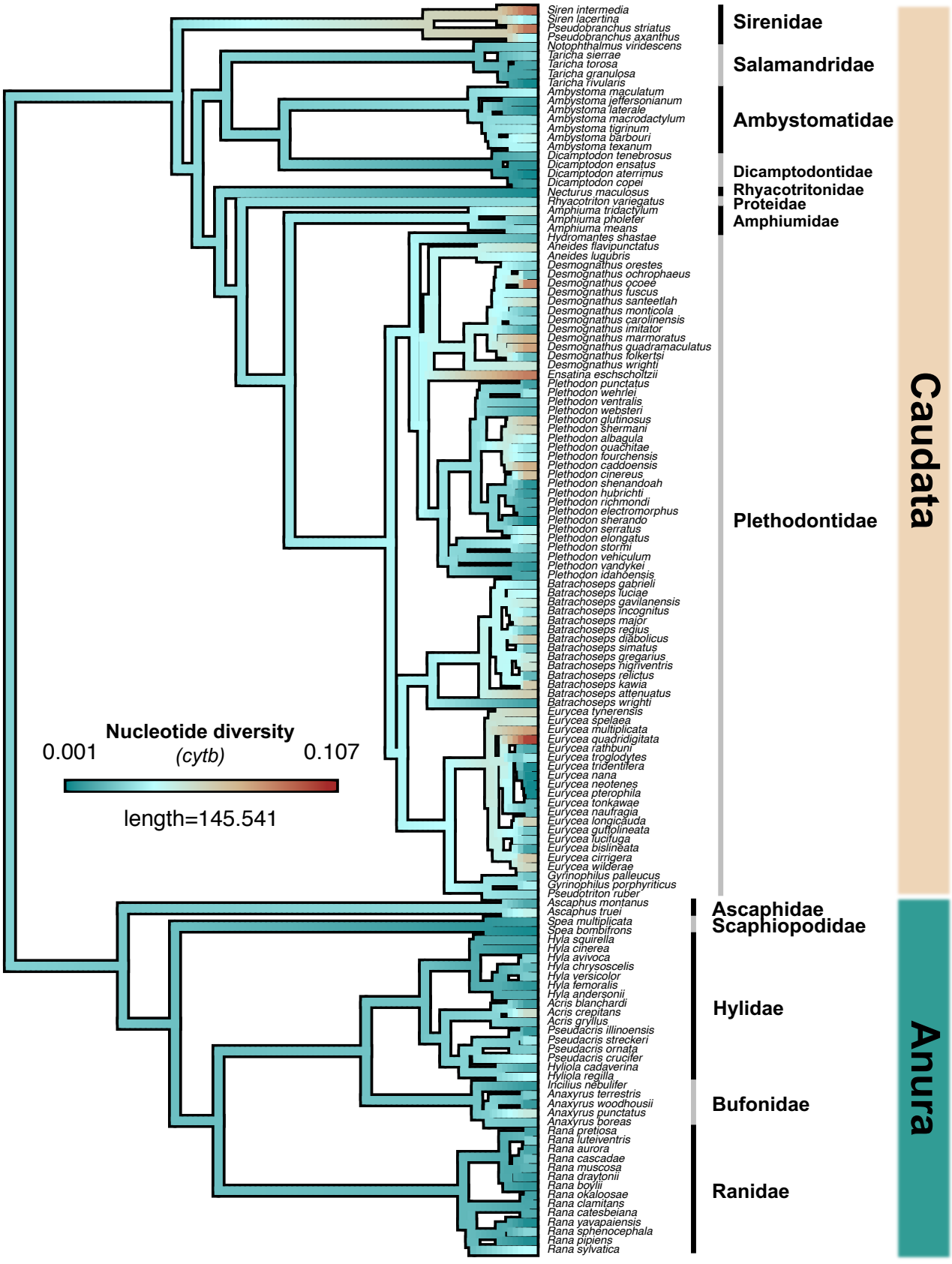
773 **Table 1**. Summary of challenges for repurposed data projects. Examples of challenges
774 encountered in the present study are provided with potential avenues for correction in the future.

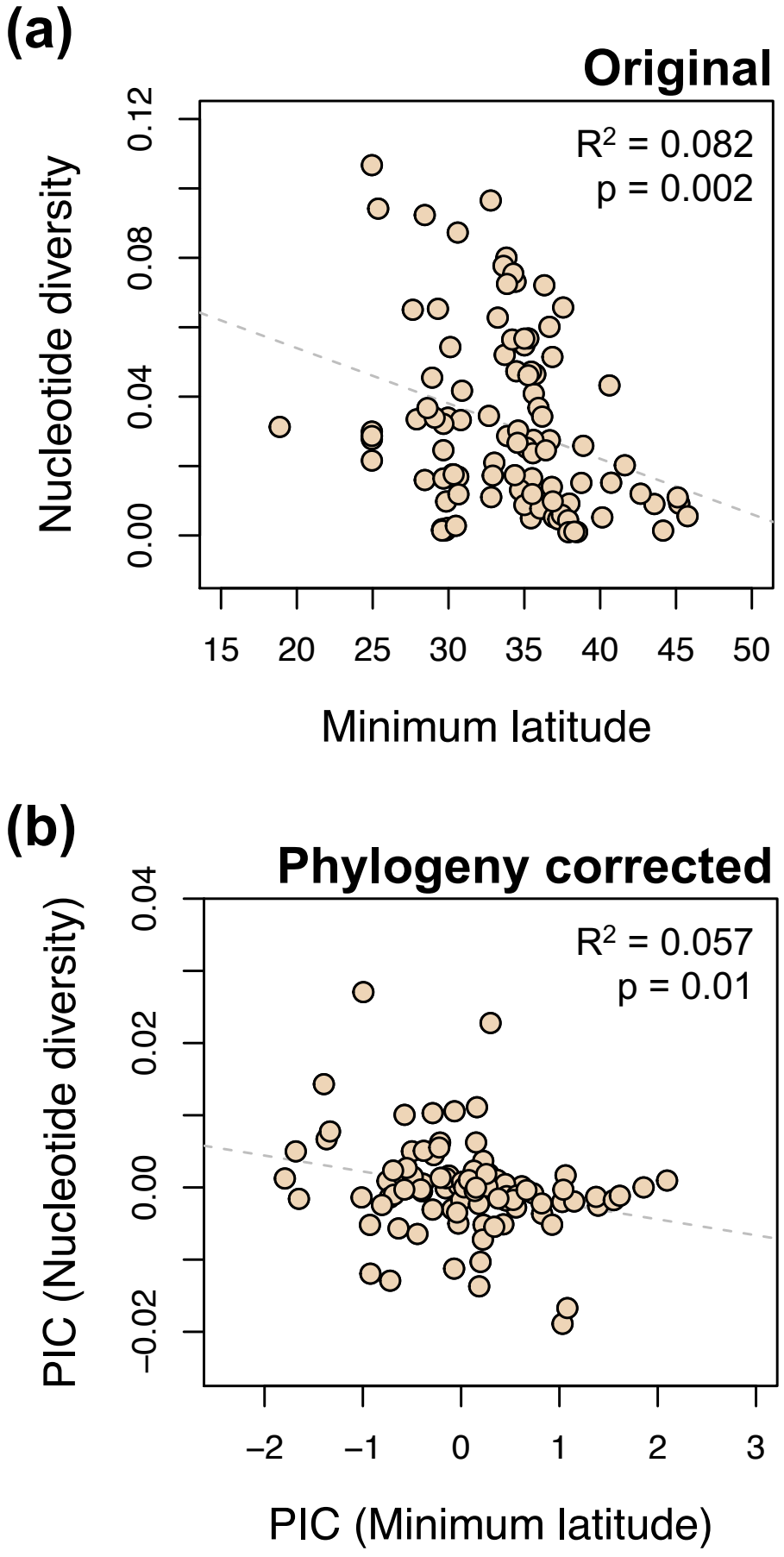
| | *Challenge* | *Example or Description* | *Potential for Correction* |
|---|---|---|---|
| | | | |
| **Trait Data** | | | |
| | Lack of natural history information | Difficult to measure traits such as dispersal, longevity, development time | • Promote and fund these efforts.<br>• Focus on traits that can be attained from natural history collections. |
| | Recorded differently across sources | Body size for salamanders either SVL or TL depending on source | • Develop standards within a taxonomic group.<br>• Carefully check your data. |
| | Time consuming, difficult to automate | Three of these authors spent weeks compiling data | • More collaborative efforts.<br>• Make trait tables open access and extendable by research community.<br>• Incorporate individual variation using museum specimens and open access databases. |
| | Deciding how to treat quantitative traits | Min/max attainable for the most species, but may not capture variation of interest | |
| | Deciding how to treat categorical traits | Some ambiguity; variation across ranges for habitat type | |
| | | | |
| **Range Data** | | | |
| | Taxonomy mismatches across databases | e.g., *Dryophytes*/*Hyla*, *Hyliola*/*Pseudacris* | • Check for synonyms manually.<br>• Use 'taxize' R package[1] |
| | Species without IUCN range maps | Newly described or split species, e.g., *Acris blanchardi* | • Revise manually.<br>• Use occurrence records. |
| | Introduced localities or errors in range maps or occurrence records | e.g., *Hyla cinerea* in Puerto Rico, *Rana pipiens* historical records throughout Mexico | • Check data; Revise range maps.<br>• Use script to remove records outside of range maps. |
| | | | |
| **Genetic Data** | | | |
| | Most sequences are not georeferenced | 4,759 GenBank accessions linked to GBIF localities of 42,067 assembled (<12%) | • Link sequences to georeferenced voucher specimens in databases.<br>• Include lat/lon when uploading. |
| | Different genes sequenced, most mitochondrial | Fig. S1; 217 species represented by at least one gene; only 147 with cytb | • Standardize loci collected.<br>• Identify comparable metrics.<br>• Shift to genome-scale datasets. |
| | Inconsistencies during GenBank upload | Different gene names (cytb, cytochrome b, Cyt-b, etc.) | • Use script to relabel names.<br>• Standardize names in databases. |
| | Data not uploaded to GenBank correctly (or at all) | "UNVERIFIED" sequences without genes annotated properly were missed | • Manually add data from other sources (e.g., Dryad).<br>• Correct GenBank records. |
| | Errors in automated alignments | Mis-aligned sequences identified, particularly when lengths differed | • Inspect alignments.<br>• Test different software and optimize settings. |
| | Small or uneven sample sizes across species | N individuals sampled was associated with diversity | • Check sensitivity to sampling.<br>• Collect more data! |

775 [1](Chamberlain & Szöcs 2013). SVL = snout-vent length; TL = total length.


(a)
Mean Decrease Accuracy
0.08
0.06
0.04
0.02
0.00
253 species
Downsampled 100X
Average error = 20.1%
Total_Area
Lat_Extent
GBIF_occurrences
Bio4_SD
Bio6_SD
Bio14_SD
Bio9_SD
Bio11_SD
Bio3_SD
Bio7_SD
Family
Body_size_mm
Development
Breeding_habitat_2cat
Neoteny
(b)
sqrt(Total Area)
3000
2000
1000
0
LC
nonLC
(c)
Latitudinal Extent
40
20
0
LC
nonLC
(d)
GBIF Occurrences
30000
20000
10000
0
LC
nonLC
(e)
Bio4 SD
2000
1500
1000
500
0
LC
nonLC


(a)
137 amphibian species
Variance explained = 30.1%
Importance (IncMSE)
1.0e−04
7.5e−05
5.0e−05
2.5e−05
0.0e+00
Family
N_sequences
Bio13_Average
Bio9_SD
Bio16_Average
Bio15_Average
Bio18_Average
Bio17_Average
Lat_Min
Bio7_SD
Total_Area
GBIF_occurrences
Development
Breeding_habitat_2cat
Neoteny
Body_size_mm
(b)
Caudata (98 species)
Variance explained = 28.9%
Importance (IncMSE)
1.0e−04
7.5e−05
5.0e−05
2.5e−05
0.0e+00
N_sequences
Lat_Min
Bio9_SD
Bio13_Average
Total_Area
Bio7_SD
Bio5_Average
Bio2_Average
Bio18_Average
Bio1_Average
Bio8_SD
GBIF_occurrences
Family
Body_size_mm
Neoteny
Development
Breeding_habitat_2cat
(c)
Caudata (67 species)
Variance explained = 28.3%
Importance (IncMSE)
0.00015
0.00010
0.00005
0.00000
Lat_Min
N_sequences
Bio13_Average
Bio9_SD
Bio18_Average
Bio1_Average
Bio7_SD
Clutch_size_min_n
Clutch_size_max_n
Bio5_Average
GBIF_occurrences
Body_size_mm
Family
Total_Area
Development
Breeding_habitat_2cat
Neoteny
Larval_period_max_d
Larval_period_min_d


Nucleotide diversity
0.001
(cytb)
0.107
length=145.541
Siren intermedia
Siren lacertina
Pseudobranchus striatus
Pseudobranchus axanthus
Notophthalmus viridescens
Taricha sierrae
Taricha torosa
Taricha granulosa
Taricha rivularis
Ambystoma maculatum
Ambystoma jeffersonianum
Ambystoma laterale
Ambystoma macrodactylum
Ambystoma tigrinum
Ambystoma barbouri
Ambystoma texanum
Dicamptodon tenebrosus
Dicamptodon ensatus
Dicamptodon aterrimus
Dicamptodon copei
Necturus maculosus
Rhyacotriton variegatus
Amphiuma tridactylum
Amphiuma pholeter
Amphiuma means
Hydromantes shastae
Aneides flavipunctatus
Aneides lugubris
Desmognathus orestes
Desmognathus ochrophaeus
Desmognathus ocoee
Desmognathus fuscus
Desmognathus santeetlah
Desmognathus monticola
Desmognathus carolinensis
Desmognathus imitator
Desmognathus marmoratus
Desmognathus quadramaculatus
Desmognathus folkertsi
Desmognathus wrighti
Ensatina eschscholtzii
Plethodon punctatus
Plethodon wehrlei
Plethodon ventralis
Plethodon websteri
Plethodon glutinosus
Plethodon shermani
Plethodon albagula
Plethodon quachitae
Plethodon fourchensis
Plethodon caddoensis
Plethodon cinereus
Plethodon shenandoah
Plethodon hubrichti
Plethodon richmondi
Plethodon electromorphus
Plethodon sherando
Plethodon serratus
Plethodon elongatus
Plethodon stormi
Plethodon vehiculum
Plethodon vandykei
Plethodon idahoensis
Batrachoseps gabrieli
Batrachoseps luciae
Batrachoseps gavilanensis
Batrachoseps incognitus
Batrachoseps major
Batrachoseps regius
Batrachoseps diabolicus
Batrachoseps simatus
Batrachoseps gregarius
Batrachoseps nigriventris
Batrachoseps relictus
Batrachoseps kawia
Batrachoseps attenuatus
Batrachoseps wrighti
Eurycea tynerensis
Eurycea spelaea
Eurycea multiplicata
Eurycea quadridigitata
Eurycea rathbuni
Eurycea troglodytes
Eurycea tridentifera
Eurycea nana
Eurycea neotenes
Eurycea pterophila
Eurycea tonkawae
Eurycea naufragia
Eurycea longicauda
Eurycea guttolineata
Eurycea lucifuga
Eurycea bislineata
Eurycea cirrigera
Eurycea wilderae
Gyrinophilus palleucus
Gyrinophilus porphyriticus
Pseudotriton ruber
Ascaphus montanus
Ascaphus truei
Spea multiplicata
Spea bombifrons
Hyla squirella
Hyla cinerea
Hyla avivoca
Hyla chrysoscelis
Hyla versicolor
Hyla femoralis
Hyla andersonii
Acris blanchardi
Acris crepitans
Acris gryllus
Pseudacris illinoensis
Pseudacris streckeri
Pseudacris ornata
Pseudacris crucifer
Hyliola cadaverina
Hyliola regilla
Incilius nebulifer
Anaxyrus terrestris
Anaxyrus woodhousii
Anaxyrus punctatus
Anaxyrus boreas
Rana pretiosa
Rana luteiventris
Rana aurora
Rana cascadae
Rana muscosa
Rana draytonii
Rana boylii
Rana okaloosae
Rana clamitans
Rana catesbeiana
Rana yavapaiensis
Rana sphenocephala
Rana pipiens
Rana sylvatica
Sirenidae
Salamandridae
Ambystomatidae
Dicamptodontidae
Rhyacotritonidae
Proteidae
Amphiumidae
Plethodontidae
Ascaphidae
Scaphiopodidae
Hylidae
Bufonidae
Ranidae
Caudata
Anura


(a)
Original
R² = 0.082
p = 0.002
Nucleotide diversity
0.00
0.04
0.08
0.12
15
20
25
30
35
40
45
50
Minimum latitude
(b)
Phylogeny corrected
R² = 0.057
p = 0.01
PIC (Nucleotide diversity)
−0.02
0.00
0.02
0.04
−2
−1
0
1
2
3
PIC (Minimum latitude)