Molecular Phylogenetics and Evolution 198 (2024) 108116

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Artificial intelligence enables unified analysis of historical and landscape influences on genetic diversity

Emanuel M. Fonseca [1], Bryan C. Carstens [*]

Museum of Biological Diversity & Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 1315 Kinnear Rd., Columbus OH 43212, USA

ARTICLE INFO

Keywords:
Demographic change
Isolation by distance
Landscape genetics
Machine learning
Phylogeography
Convolutional neural networks

ABSTRACT

While genetic variation in any species is potentially shaped by a range of processes, phylogeography and landscape genetics are largely concerned with inferring how environmental conditions and landscape features impact neutral intraspecific diversity. However, even as both disciplines have come to utilize SNP data over the last decades, analytical approaches have remained for the most part focused on either broad-scale inferences of historical processes (phylogeography) or on more localized inferences about environmental and/or landscape features (landscape genetics). Here we demonstrate that an artificial intelligence model-based analytical framework can consider both deeper historical factors and landscape-level processes in an integrated analysis. We implement this framework using data collected from two Brazilian anurans, the Brazilian sibilator frog (*Leptodactylus troglodytes*) and granular toad (*Rhinella granulosa*). Our results indicate that historical demographic processes shape most the genetic variation in the sibulator frog, while landscape processes primarily influence variation in the granular toad. The machine learning framework used here allows both historical and landscape processes to be considered equally, rather than requiring researchers to make an *a priori* decision about which factors are important.

1. Introduction

Inferring the processes that influence genetic variation in a spatial context is a key aim of phylogeography and landscape genetics (Avise et al., 2016; Manel et al., 2003). While both disciplines assay genetic variation from empirical systems, phylogeography gained prominence as researchers used Sanger methods to sequence organellar genes (e.g., Avise et al., 1987), while landscape genetics was more likely to utilize allelic markers (e.g., Allentoft et al. 2009). Early landscape genetic investigations utilized highly variable markers such as microsatellites to analyze the landscape processes that influence genetic variation across recent time scales (e.g., Manel et al., 2003; Holderegger & Wagner, 2008). Conversely, phylogeographic investigations continued to utilize sequence data because the variation in these data accumulates over longer time scales (Carstens et al., 2013) leading to inferences from deeper time periods (e.g., Carnaval et al., 2009; Peterman et al., 2014; Smith et al., 2014). Despite their shared goal, the disciplines retained separate identities due in part to extrinsic factors associated with temporal and spatial scales implied by various analytical analyses and in part to the assumption that different molecular markers were more informative at these different scales (Rissler, 2016). However, there is considerable overlap between these disciplines, particularly among biologists who are primarily motivated to learn as much as possible about their chosen focal species.

Phylogeography and landscape genetics have benefited from recent technical advances in throughput sequencing (Garrick et al., 2015; Storfer et al., 2018), SNP sequencing protocols (e.g., Peterson et al., 2012), and user-friendly bioinformatic processing software (e.g., ipyrad, Eaton & Overcast, 2020) have advanced to the point where researchers can collect genomic data consisting of thousands of SNPs from nearly any empirical system. These technologies have expanded the capacity of biologists to test hypotheses across a broad range of spatial and temporal scales (e.g., Myers et al., 2019; Vasconcellos et al., 2019; Wieringa et al., 2020). While any researcher can investigate both the landscape processes that influence genetic diversity and the deeper evolutionary history of their focal system using the same data, it remains the case that individual analytical methods are more appropriate to apply at either the landscape or phylogeographic scale. For researchers who collect

* Corresponding author.
E-mail address: carstens.1@osu.edu (B.C. Carstens).
[1] ORCID: 0000-0002-2952-8816.

https://doi.org/10.1016/j.ympev.2024.108116
Received 5 July 2023; Received in revised form 4 April 2024; Accepted 4 June 2024
Available online 12 June 2024

genomic datasets from natural systems, this wide range of potential methods presents a challenge because they are required to choose among a wide range of population genetic, landscape genetic, and phylogeographic methods for data analysis with little guidance as to which of these is likely to be most suitable.

Several factors which are likely to be unknown for any focal system combine to determine what types of data analyses may be optimal in a particular system. For example, the appropriate unit of analysis (i.e., lineage, population, sampling locality) may be unclear, which could prevent researchers from analyzing their data with phylogeographic methods that often require samples to be partitioned into populations or lineages before analysis. Researchers who are interested in modeling the demographic history of a species (i.e., refugial structure, population bottlenecks, or expansions) often perform a clustering analysis to assign samples to populations that are then used in the demographic modeling (e.g., Leaché & Fujita, 2010; Fonseca et al., 2018). However, since most clustering algorithms do not consider continuous processes such as isolation by distance (IBD, Wright, 1943; but see Bradburd et al., 2018), erroneous population delimitation and subsequent bias in downstream analyses can result if continuous processes influence genetic variation (Bradburd et al., 2018; Frantz et al., 2009). Likewise, some landscape genetic analyses assumes that genetic variation is maintained as a balance between migration and drift and if historical demographic processes such as demographic expansion have occurred the results could be biased (Wang, 2010; Bohonak & Vandergast, 2011; Epps & Keyghobadi, 2015). Finally, species boundaries may be unknown in the focal system, and this taxonomic uncertainty my lead to inflated estimates of some parameters when samples from multiple species are analyzed under a population genetic framework (e.g., Carstens and Dewey, 2010).

An integration of discrete historical factors and continuous landscape processes under a common analytical framework would enable researchers to identify the factors that are important in shaping genetic diversity in the focal species. Such a framework could conceivably take one of several forms, including conducting landscape genetic and phylogeographic analyses independently and developing a statistical framework for synthesizing these results or by conducting a single analysis that attempted to account for both deeper historical and landscape factors. An integrative analysis could include the development of a full likelihood framework, but this is likely a daunting task due to model complexity (Beaumont et al., 2002). Simulation-based methods have been proposed (e.g., Currat et al., 2004; Landguth & Cushman, 2010) that simulate genetic data in a spatial context, and exploring a framework developed around this option is clearly worthwhile. However, applying such a framework to empirical systems where life history parameters such as generation length, dispersal capacity, and mutational rates across the genome are poorly characterized will be challenging due to the impact that such parameters will have on simulation of genetic diversity. Integrating simulations within a supervised machine learning (SML) framework could enable the development of a more streamlined and efficient model. This might make it possible to design a less complex framework that is capable of identifying which factors (i.e., landscape, deeper historical) have most influenced the pattern of genetic diversity in the focal species (Schrider and Kern, 2018).

SML methods are a variety of artificial intelligence that seek to train a predictive model using a pre-classified dataset. In this application, a model would be trained using datasets where only historical factors or only landscape factors were evident. Since both factors are likely to be present, pretraining can also occur with data where both factors are present. Here, we provide an example of how discrete historical and continuous landscape processes can be accommodated under a framework that includes both types of processes. Our framework adopts a simulation-based approach that employs convolutional neural networks (CNNs), an artificial neural network (ANN) technique which mimics the biological neural network of human brain by artificially replicating the connection among neurons. While CNNs were developed for image and video classification, they can be applied to genetic data by changing how genetic data are represented in the analysis (Flagel et al., 2019). Rather than using summaries of the genetic data such as a range of statistics (e.g., Pritchard et al., 1999) or allele frequency spectra (e.g., Gutenkunst et al., 2009), a CNN approach incorporates an image of a DNA sequence alignment that has been processed to retain the salient features of the genetic diversity (e.g., Flagel et al., 2019). The advantage to this approach is that *a priori* decisions about the number of populations, the distribution of samples among populations, and the type of statistics that could be applied to summarize the genetic variation are not required (Fonseca et al., 2021). CNNs do not preclude the use of summary statistics, and other researchers have incorporated summary statistics into CNN frameworks (e.g., Blischak et al., 2021). In summary, an approach that (i) designs putative demographic and/or spatial models, (ii) simulates datasets under these models, and (iii) compares processed alignments of these simulated data using CNNs may have the potential to help researchers identify the types of factors that have influenced genetic diversity in their focal system. In order to explore this idea, we applied a CNN framework to two datasets that were previously collected from the sibilator frog *Leptodactylus troglodytes* and the granular toad *Rhinella granulosa*, two broadly co-distributed Neotropical anurans with different ecologies and evolutionary histories (Fig. 1).

2. Material and methods

We use data collected from two Neotropical anurans that are broadly distributed in arid vegetations in northeastern Brazil to illustrate how demographic and/or spatial processes can be accommodated under the same framework. These species were chosen because each have been the subject of a recent investigation that focused on identifying the best model of historical demography (e.g., Thomé et al., 2021a,b), which we assumed for the purpose of this exploration. Information about the data collected from each species are present in Table 1.

2.1. Neotropical anurans as a case study

The species considered here differ in their natural histories. The Brazilian sibilator frog (*Leptodactylus troglodytes*, Anura: Leptodactylidae) reproduces in underground chambers, where it produces a foam nest that allows the eggs to develop in humidity before the tadpoles hatch and complete their development in seasonal water bodies (Kokubum et al., 2009). This reproduction mode allows the Brazilian sibilator frog to reproduce continuously throughout the wet season. In contrast, the granular toad (*Rhinella granulosa*, Anura: Bufonidae) reproduces in bouts of explosive breeding (Wells, 2007) that are tightly associated with heavy rainfall events (Narvaes & Rodrigues, 2009) and have a reproductive physiology typical of desert anurans (Madelaire & Gomes, 2016). The granular toad is more dependent on water bodies than the sibilator frog, as both its eggs and larvae are fully aquatic.

The distribution of both species mostly overlaps, as both occur throughout the entirety of the xeric Caatinga biome (Fig. 1). The Brazilian sibilator frog also spans the northernmost part of the Cerrado savannah to the west of the Caatinga, whereas the granular toad occupies the northern Atlantic Forest biome east of the Caatinga. The phylogeographic patterns of the two species exhibit genetic structure across the landscape (Thomé et al., 2021a,b). The evolutionary history of the Brazilian sibilator frog appears to have been influenced by historical demographic events, such as population expansions and gene flow among discrete populations, while genetic diversity in the granular toad appears to be influenced largely by landscape factors, notably a pattern of isolation by distance across the landscape in which genetic similarity decays as a function of increased geographic distance (Thomé et al., 2021a,b). Other differences in their respective evolutionary histories include population limits that coincide with the São Francisco river for the sibilator frog, whereas for the granular toad population structure is best explained by a regional asynchrony of the wet season,

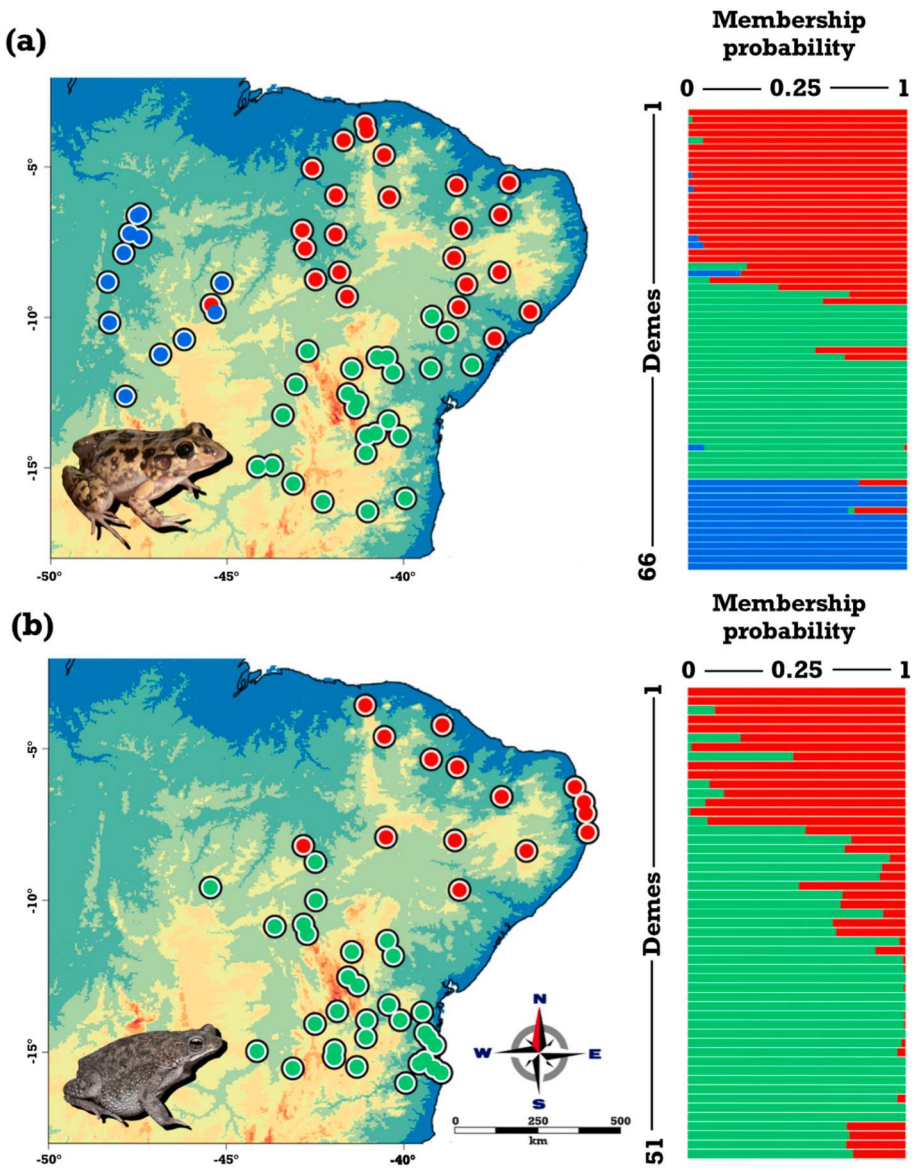

Fig. 1. Map showing the geographic distribution of sampled localities for (a) the Brazilian sibilator frog (*Leptodactylus troglodytes*) and (b) granular toad (*Rhinella granulosa*). Red, green, and blue circles represent north population (from Caatinga), south population (from Caatinga), and Cerrado population, respectively. Bar plots represent the membership probability for each species according to STRUCTURE and BAPS analyses. Map background depicts elevation variation in the study region, where elevation increases from blue (lower elevation) to red (higher elevation). *T* represents the divergence time and *m* the migration among the biological units. Photos: Ricardo Marques.

Table 1
Summary of sampling and data collected in two species investigated here. Shown for each species are the number of samples (n), the number of inferred populations (k), the number of sampling localities, the number of SNPs collected, the model selected from a phylogeographic model selection analysis, and the demographic inference made by the original research. Information is drawn from Thomé et al., 2021a,b.

species	n	k	localities	SNPs	selected model	demographic inference
Leptodactylus troglodytes	159	3	59	15,080	divergence, demographic size change	Pleistocene divergence with subsequent demographic expansion (Caatinga) or contraction (Cerrado)
Rhinella granulosa	80	2	51	7688	divergence, gene flow	isolation by environment via precipitation gradient

which structures populations due to differences in breeding times. Furthermore, environmental differences clearly influence the genetic structure in the sibilator frog, which shows a divergent population in the Cerrado, whereas in the granular toad population structure does not follow the sharp gradient between the wet Atlantic Forest and the dry Caatinga biome. The uniqueness in the evolutionary history and ecology of these species, which are both anurans and broadly co-distributed, led us to choose them for this exploration of how CNN could be applied to a demographic modeling that incorporates historical demographic and landscape processes.

2.2. Modeling demographic and landscape processes

To evaluate whether deep time processes, landscape processes, or

both shape patterns of genetic variation in these species, we built three evolutionary models that were designed to represent three extreme contrasts in the potential factors that could influence genetic diversity in a given species. The first of these was a historical demographic model that lacked any role for landscape processes, whereas the second was a model that only included landscape process. In this case, and isolation by distance (IBD) model was chosen because it represents process that common (Pelletier & Carstens, 2018), easy to simulate, and likely to be correlated with factors such as environmental distance in a terrestrial vertebrate. A third model that incorporates features of both the historical demographic and landscape models was also designed.

For the historical demographic models (Fig. 2a), we used the same genetic structure, evolutionary relationships, and demographic history recovered by Thomé et al., (2021a) and Thomé et al., (2021b). Both papers conducted phylogeographic model selection and chose a single model with strong support using information theoretic statistics. Thomé et al. found three and two genetic populations for *L. troglodytes* and *R. granulosa* (Fig. 1), respectively, with signatures of demographic expansions for two genetics clusters in *L. troglodytes* (red and green populations in Fig. 1a). We also included gene flow between all populations. For the sibilitator frog, we set the divergence time as minimum and maximum values as follow: T1) 500,000–1,000,000 generations before the present, and T2) 100,000–250,000 generations before the present. For the granular toad, we set the divergence time between 250,000–750,000 generations before the present. Migration rates were also set using a uniform distribution (from 1.0 x 10^{-6} to 1.0 x 10^{-7}). Finally, expansion time was set to have occurred between 10,000 and 50,000 generations before the present. All of these values were derived from confidence intervals of the point estimates that were made under the selected demographic model in the original Thomé et al. papers.

For the IBD model (Fig. 2b), we modeled a panmictic population that split into *n* demes (each deme is a geographic locality in the empirical datasets) 10,000–50,000 generations before the present. Population size on each was sampled from a uniform distribution (ranging from 50 to 200 haploid individuals) using deme sizes that were informed by observations from the field. In the model, each deme is connected among them as a function of geographic distance (i.e., demes further apart experience less gene flow than demes that are closer). Since IBD is a model in which the landscape is homogeneous with respect to dispersal, we create a landscape raster where all grid cells had the same value. Because of the uncertainty on taxa dispersal capacity across the landscape, each grid cell had an associated resistance prior ranging from 2 to 5 (reflecting low to high dispersal capacity). Then, we calculated the resistance distance among all points using the least-cost path function implemented in the R package "gdistance" (van Etten, 2017). Next, we converted the resistance matrix to a migration matrix by raising the resistance distance matrix to a second power and calculating its inverse. Thus, higher resistance values were associated to lower migration rate and vice-versa. We used such a function because we expect that gene flow decreases exponentially as geographic distance increases. The migration matrix calculated from the landscape resistance was used in the simulations as a proxy of genetic connectivity.

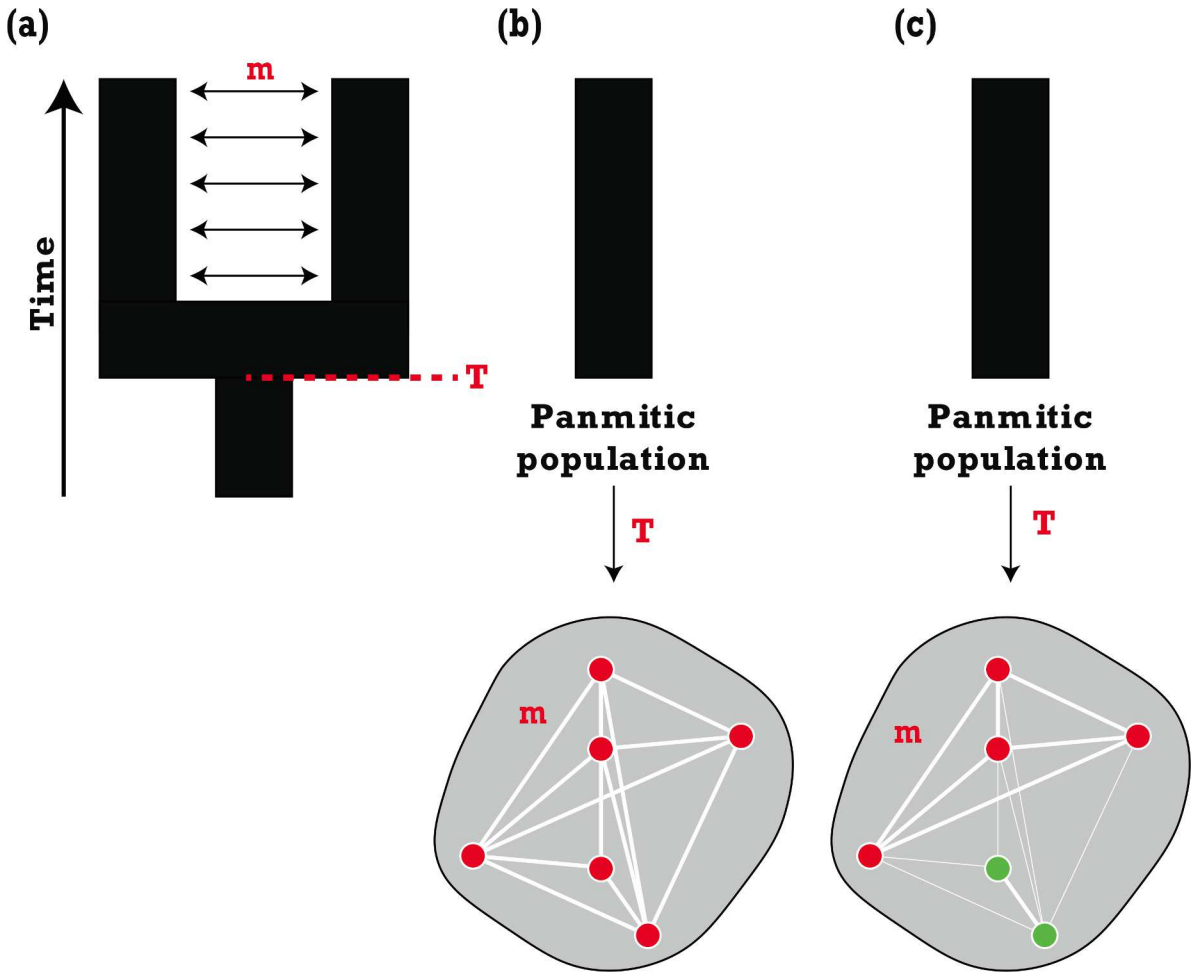

Fig. 2. Models tested for the Brazilian sibilator frog (*Leptodactylus troglodytes*) and the granular toad (*Rhinella granulosa*). (a) historical demographic model, (b) isolation by distance model, and (c) model with features of the historical demographic isolation by distance models. Note that models (a) and (c) incorporate changes in effective population size (if present; see methods).

For the third model (Fig. 2c), we included features of the historical demographic and IBD models. We first created an IBD model as described previously and allowed the inclusion of the evolutionary relationship and demographic change in each locality (if present). To account for the genetic structure, we multiplied the resultant migration matrix by a uniform prior ranging from 0.01 to 0.05. This prior was only applied if two demes come from different genetic clusters. To account for the possibility of demographic change in each deme (if present), we sampled an ancestral population size from a uniform distribution that ranged from 3 to 40 haploid individuals, a number derived by distributing the total Ne across the number of demes. Finally, we account for uncertainty in cluster membership probability by averaging the individual probability of all individuals present in a deme. Then, we assigned each deme to a population based on the average probability. For example, if a deme has a 90 % of chance of being from population A and 10 % from population B, this will represent the probability of this deme being part of population A or B on each simulation.

We used fastsimcoal2 (Excoffier et al., 2013) to simulate 2,500 datasets under each customized evolutionary model, which matched the empirical dataset in terms of the number of SNPs, localities, and individuals per locality. Customs R scripts were written to sort each SNP based on major allele frequency (higher to a lower frequency) after removing SNPs with the minimum allele frequency lesser than 5 %. Then, a second round of sorting was performed, from the point there was no more SNP within species 1, to sort SNPs within the second species. Within each species, individuals were randomly sorted. Next, we converted the alignment of each simulation and dataset into a biallelic matrix, with *n* rows and *k* columns, corresponding to the number of samples and SNPs, respectively. This matrix (i.e., the '1' and '0' of the biallelic SNPs) was converted to a black and white image with each SNP corresponding to a pixel in the image. Finally, rows (representing individuals) were organized assuming always the same spatial configuration. For the demographic model, we sampled individuals on each genetic cluster and assigned it randomly to a spatial locality that correspond to its genetic cluster.

2.3. Model selection using machine learning

We used a convolutional neural network (CNN) to calculate the relative probability of the different models given the empirical data for each species. We implemented a two-dimensional CNN architecture as follows: a two-dimensional convolution layer (kernel = 3 x 1) followed by a two-dimensional maximum pooling layer (kernel = 3 x 1). Then, the CNNs were flattened from the pooling layer and connected to an artificial neural network of 40 neurons and connected to the output layer with three neurons, each represent a different evolutionary model. For all layers, we used rectified linear unit activation functions (ReLU), except for the last one where we used a softmax function. This function is a generalization of the logistic function and used for multiclass prediction. We compiled the CNN using the Adam optimization procedure (Kingma & Ba, 2015), a categorical cross-entropy loss function, and a mini-batch size of 100. We ran the CNN for 10 epochs. The CNN was trained using 80 % of the simulated datasets and used the remaining 20 % to evaluate model accuracy. Lastly, we used the trained model to predict the model that likely generated the empirical dataset. We built all CNNs with the Keras python library (https://keras.io).

SML approaches such as the CNN implemented here enable researchers to analyze genetic data directly from a DNA alignment (e.g., Flagel et al., 2019; Fonseca et al., 2021; Perez et al., 2021; Torada et al., 2019) rather than after dividing the samples into populations or lineages and then estimating some metric from these groups. While this shift may appear superficial, it transforms data analysis in three ways. (*i*) Avoiding summary statistics maximizes the retention of the information contained within genetic data. (*ii*) The analysis of DNA alignments sidesteps the potentially difficult question of choosing which summary statistics are best suited for a particular system (e.g., Prangle et al., 2014). (*iii*) Avoiding summary statistics allows models that include both deep population divergence and continuous processes such as genetic IBD to be analyzed in the same model selection framework, which greatly increases the power of the resulting inference by increasing the range of potential processes considered by the analysis (e.g., Pelletier & Carstens, 2014).

3. Results

Our simulation approach couple with machine learning showed that for the sibilator frog the historical demographic scenario (model 1) was the best supported model with probability of 100 % (Table 2). Meanwhile for the granular toad, we recovered the isolation by distance model as the best model (model 2), with over 99 % of the total model probability (Table 2). Both trained CNNs model had a high accuracy when predicting the test set labels, reaching an overall accuracy of 99.2 % and 97.3 % for the sibilator frog and granular frog, respectively (Fig. 3). Our results are consistent with the findings reported by Thomé et al., (2021a) and Thomé et al., (2021b) in the sense that they show that the two species have distinct evolutionary histories that are likely due to differences in their life history and ecologies. These ecological differences likely impacted how each species responded to historical processes such as Pleistocene climatic fluctuations. Interestingly, the sibilator frog showed signals of recent population expansion in two of its three populations and we recovered the demographic model as the best support for this species. Recent investigations have shown that Caatinga community have responded synchronous to past Pleistocene climatic fluctuations (Bonatelli et al., 2021; Gehara et al., 2017). Thus, Quaternary climatic cycles likely promoted cycles of contraction followed expansion in the range of this species. Conversely, an isolation by distance model best described the evolutionary history of the granular toad, which showed no evidence of population size change over time.

The model selection framework proposed here is computationally efficient. For example, under the historical demographic model each simulation requires an average of 7 s to conduct (~2 s to conduct the simulation in fastsimcoal2 and ~ 5 s to process the image). Models with a spatial component were more computationally demanding, taking on average ~ 15 s to create the migration matrix, ~40 s to run the simulation, and ~ 5 s to process the image, but still reasonable. Once the data are simulated and processed, the actual analysis of the simulated dataset using the CNN is relatively fast. Each epoch of training required approximately 3 min, so the total CPU time required here was 30 min for the 10-epoch analysis. In total, the complete analysis would require fewer than 100 h on a modest computer (i.e., we generated these reference values using a 2018 Mac mini with a 1.6 GHz Intel Core i5 and 8 GB RAM). The analysis is amenable to parallelization, for example the data simulation could easily be conducted on a cluster computer.

4. Discussion

Evolutionary genetics is challenging because every species has a

Table 2
Probabilities of model comparisons obtained from three diversification scenarios tested using convolutional neural networks (CNN) for the Brazilian sibilator frog (*Leptodactylus troglodytes*) and the granular toad (*Rhinella granulosa*). The overall accuracies of the trained model were 99.2% and 97.3%for *L. troglodytes* and *R. granulosa*, respectively.

Model	*Leptodactylus troglodytes*	*Rhinella granulosa*
	Model probability	Model probability
Historical demographic model	1.0	0.01
Isolation by distance model	0.0	0.99
Historical demographic + isolation by distance model	0.0	0.0

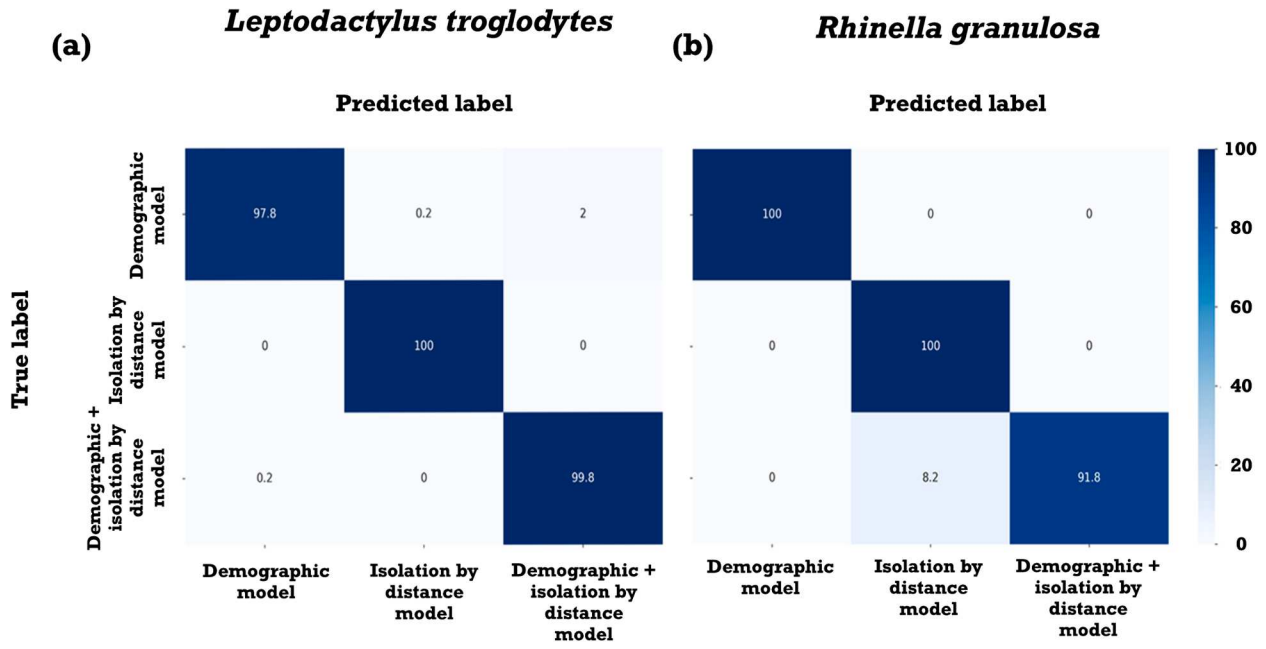

Fig. 3. Confusion matrix measuring the accuracy of the CNN on the training dataset. Numbers represent percentages, which were calculated based on 500 images for each model.

unique history and thus a unique response to their environment and the landscape that they inhabit. For example, even though the focal taxa are broadly co-distributed and have evolved under similar environmental and geological processes, they exhibit idiosyncratic evolutionary histories. Research is made much more interesting by this reality, because the unique interplay among factors that include life history, habitat specificity, and ecological niche engender questions that can be explored and hypotheses that can be tested. For example, the focal taxa differ in their in their reproductive strategies; the Brazilian sibilator frog reproduces continuously throughout the wet season, the granular toad is an explosive breeder with reproduction restricted to a short period annually. While the granular toad can be found in a broader range of habitat, the sibilator frog is also widespread. Genetic variation in the granular toad is best explained by landscape processes such as isolation by distance (Thomé et al., 2021b). In contrast, intraspecific genetic variation in the sibilator frog is largely shaped by historical divergence between the populations living in major biomes and sporadic gene flow that may be linked to climatic events (Thomé et al., 2021a). Notably, as data analysis was ongoing for the granular toad and the sibilator frog, it was conducted without a clear indication of whether landscape or historical processes were most important to the focal taxon. As a result, both publications discussed both landscape and historical factors as potential forces that influence genetic diversity (Thomé et al., 2021a,b).

Discordant genetic patterns among taxonomically-similar species have highlighted the importance of taxon-specific traits in phylogeography and landscape genetics (Papadopoulou & Knowles, 2016; Zamudio et al., 2016), and trait-based approaches have helped to elucidate the nature of discordant genetic patterns (e.g., Papadopoulou & Knowles, 2016; Sullivan et al., 2019). Recently, Bonatelli et al. (2021) demonstrated that habitat preference is one the strongest predictors of demographic responses across many taxa in the dry diagonal region (a region encompassing our study area). For this reason, it may be that the ephemeral nature of the granular toad reproduction is responsible for mitigating the effects of historical climate on survival. In contrast, in the sibilator frog, Pleistocene climatic fluctuations likely led to demographic collapses and to periods of small population sizes because of the continuously reproductive behavior of this species and its preference for a drier habitat which was affected by Pleistocene climate dynamics.

Knowledge about ecology, evolution, distribution, and population dynamics remains obscure for many species (Hortal et al., 2015). This uncertainty has important implications for data analysis. While researchers have available dozens of different programs and frameworks to help them to address their scientific questions, the choice of analytical method is ideally informed by information about the life history, distribution, and population dynamics of the focal species. When such information is lacking, there is a higher chance that inappropriate analytical methods will be used. Phylogeography and landscape genetics originally relied on different kinds of genetic data. For example, phylogeography traditionally incorporated sequence data, inferred historical processes based on gene genealogies, and has more recently incorporated summary statistics such as site frequency spectrum, nucleotide diversity, Tajima's D, and F_{ST}, that are calculated from discrete units (e.g., Garrick et al., 2020; Hickerson et al., 2006). In contrast landscape genetic investigations were traditionally based on allelic data such as microsatellites, and these data were analyzed in large part based on metrics calculated among demes or individuals (Waits & Storfer, 2015). More recently, researchers in both disciplines have begun to collect SNP data because these data are economical to generate in non-model systems. While the incorporation of these data has been a boon for both disciplines, they blur the distinction between them. One remaining difference is how genetic variation is partitioned, for example in lineages, populations, demes, or individuals. In contrast to methods such as Approximate Bayesian Computation (ABC), where choices made by the researcher about how to summarize genetic variation can influence the statistical power of the analysis (e.g., Beaumont, 2002; 2010), the use of alignment images here enhances the capacity to infer discrete and continuous genetic processes.

As phylogeography and landscape genetics have become more reliant on model-based methods over the last decades, the choice of an analytical model is often based on researcher beliefs regarding the processes that might have shaped the genetic variation (Zamudio et al., 2016). Since all models are inherently mis-specified to some extent given the complexity of the natural world (i.e., genetic drift, selection, hybridization, biological interactions, etc.), potential bias or erroneous inference can result if researchers choose an inappropriate analytical model (Koopman & Carstens, 2010). This is particularly problematic when researchers ignore either historical or landscape-level analyses when designing their investigations. In such cases, there is likely

confirmation bias (Nickerson, 1998) in the resulting inferences about the empirical system (Carstens et al., 2009). Our findings showed that two species, largely distributed over the same geographic region and likely evolving under similar climatic conditions, have different evolutionary histories. The inclusion of demographic and landscape genetics model under the same framework demonstrates the importance of including both discrete and continuous processes for a more reliable biological inference.

5. Conclusions

While thousands of investigations have been published that would be recognized as *phylogeography* or *landscape genetics*, it has been less common to explicitly consider landscape and deeper historical processes in the same investigation. Conducting an integrated analysis is possible when the genetic data can be analyzed in a framework that allows for both demographic history and landscape processes. Our work is a first step towards such a framework, but it can be improved by increasing both the number and variety of models. The incorporation of processes such as natural selection may be a logical expansion of the approach described here, as would information about habitat suitability and landscape resistance to dispersal. SML coupled with simulation-based exploration of models that incorporate both historical and landscape factors represent an important direction for molecular ecology because it can potentially unite multiple processes that accumulate over evolutionary time scales to processes acting in the present. It joins a growing number of research investigations that incorporate artificial intelligence approaches. For example, Thom et al. (2021) used artificial neural network and a stepping-stone model to demonstrate that populations in tropical mountains in the Brazilian Atlantic Forest have high rates of gene flow, and Pless et al. (2021) implemented a random forest analysis to map landscape connectivity in an invasive mosquito in North America. Burbrink et al. (2020) used an artificial neural network to infer how landscape and environmental features predicted the genetic structure of North American rat snakes and demonstrated that their model could accurately predict genetic distance. Kittlein et al. (2022) trained a CNN to predict local FST and mean allelic richness in a landscape genetic investigation. Like our work, the flexibility of the artificial intelligence methods utilized by these studies allowed researchers to ask (and answer) questions that were tailor-made to their focal systems

CRediT authorship contribution statement

Emanuel M. Fonseca: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Bryan C. Carstens:** .

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Please see GitHub (https://github.com/emanuelmfonseca/Demographic_LandscapeGenetics_Frogs) for all code used in the data analysis for this article.

Acknowledgments

We thank the Ohio Supercomputer Center (PAA0202) and the National Science Foundation for supporting this work (DEB-1831319). EMF thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for his doctoral fellowship (#88881.170016/2018). We thank M. Tereza Thomé for providing the Brazilian anuran SNP data analyzed here and for discussions about the ecological and evolutionary difference between species.

References

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., Saunders, N.C., 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Ann. Rev. Ecol. Syst. 18, 489–522.

Avise, J.C., Bowen, B.W., Ayala, F.J., 2016. In the light of evolution X: Comparative phylogeography. Proc. Nat. Acad. Sci. USA 113, 7957–7961. https://doi.org/10.1073/pnas.1604338113.

Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. Genetics 162, 2025–2035. https://doi.org/10.1111/j.1937-2817.2010.tb01236.x.

Blischak P. D., Barker M. S., Gutenkunst R.N. (2021). Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. Mol. Ecol. Resour. 21, 2676–2688. doi: 10.1111/1755-0998.13355. Epub 2021 Mar 8. PMID: 33682305; PMCID: PMC8675098.

Bohonak, A.J., Vandergast, A.G., 2011. The value of DNA sequence data for studying landscape genetics. Mol. Ecol. 20, 2477–2479. https://doi.org/10.1111/j.1365-294X.2011.05122.x.

Bradburd, G.S., Coop, G.M., Ralph, P.L., 2018. Inferring continuous and discrete population genetic structure across space. Genetics 210, 33–52. https://doi.org/10.1534/genetics.118.301333.

Carnaval, A.C., Hickerson, M.J., Haddad, C.F.B., Rodrigues, M.T., Moritz, C., 2009. Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. Science 323 (5915), 785–789. https://doi.org/10.1126/science.1166955.

Carstens, B.C., Stoute, H.N., Reid, N.M., 2009. An information-theoretical approach to phylogeography. Mol. Ecol. 18, 4270–4282. https://doi.org/10.1111/j.1365-294X.2009.04327.x.

Carstens, B.C., Brennan, R.S., Chua, V., Duffie, C.V., Harvey, M.G., Koch, R.A., Sullivan, J., 2013. Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*. Mol. Ecol. 22, 4014–4028. https://doi.org/10.1111/mec.12347.

Carstens, B.C., Dewey, T.A., 2010. Species delimitation using a combined coalescent and information-theoretic approach: an example from North American Myotis bats. Syst. Biol. 59, 400–414. https://doi.org/10.1093/sysbio/syq024.

Currat, M., Ray, N., Excoffier, L., 2004. SPLATCHE: A program to simulate genetic diversity taking into account environmental heterogeneity. Mol. Ecol. Notes 4, 139–142. https://doi.org/10.1046/j.1471-8286.2003.00582.x.

da S. Bonatelli, I.A., Gehara, M., Carstens, B.C., Colli, G.R., Moraes, E.M., 2021. Comparative and predictive phylogeography in the South American diagonal of open formations: Unraveling the biological and environmental influences on multitaxon demography. Mol. Ecol. 2020, 1–12. https://doi.org/10.1111/mec.16210.

Eaton, D.A.R., Overcast, I., 2020. ipyrad: interactive assembly and analysis of RADseq datasets. Bioinformatics 36, 2592–2594. https://doi.org/10.1093/bioinformatics/btz966.

Epps, C.W., Keyghobadi, N., 2015. Landscape genetics in a changing world: disentangling historical and contemporary influences and inferring change. Mol. Ecol. 24, 6021–6040. https://doi.org/10.1111/mec.13454.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., Foll, M., 2013. Robust demographic inference from genomic and SNP data. PLoS Genetics 9. https://doi.org/10.1371/journal.pgen.1003905.

Flagel, L., Brandvain, Y., Schrider, D.R., 2019. The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol. Biol. Evol. 36, 220–238. https://doi.org/10.1093/molbev/msy224.

Fonseca, E.M., Gehara, M., Werneck, F.P., Lanna, F.M., Colli, G.R., Sites, J.W., Rodrigues, M.T., Garda, A.A., 2018. Diversification with gene flow and niche divergence in a lizard species along the South American “diagonal of open formations”. J. Biogeog. 45 (7), 1688–1700. https://doi.org/10.1111/jbi.13356.

Frantz, A.C., Cellina, S., Krier, A., Schley, L., Burke, T., 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: Clusters or isolation by distance? J. App. Ecol. 46, 493–505. https://doi.org/10.1111/j.1365-2664.2008.01606.x.

Garrick, R.C., Bonatelli, I.A.S., Hyseni, C., Morales, A., Pelletier, T.A., Perez, M.F., Rice, E., Satler, J.D., Symula, R.E., Thomé, M.T.C., Carstens, B.C., 2015. The evolution of phylogeographic data sets. Mol. Ecol. 24, 1164–1171. https://doi.org/10.1111/mec.13108.

Garrick, R.C., Hyseni, C., Arantes, Í.C., 2020. Efficient summary statistics for detecting lineage fusion from phylogeographic datasets. J. Biogeog. 47, 2129–2140. https://doi.org/10.1111/jbi.13932.

Gehara, M., Garda, A.A., Werneck, F.P., Oliveira, E.F., da Fonseca, E.M., Camurugi, F., de M. Magalhães, F., Lanna, F.M., Sites, J.W., Marques, R., Silveira-Filho, R., São Pedro, V.A., Colli, G.R., Costa, G.C., Burbrink, F.T., 2017. Estimating synchronous demographic changes across populations using hABC and its application for a herpetological community from northeastern Brazil. Mol. Ecol. 26, 4756–4771. https://doi.org/10.1111/mec.14239.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5 (10), e1000695.

Hickerson, M.J., Dolman, G., Moritz, C., 2006. Comparative phylogeographic summary statistics for testing simultaneous vicariance. Mol. Ecol. 15, 209–223. https://doi.org/10.1111/j.1365-294X.2005.02718.x.

Holderegger, R., Wagner, H.H., 2008. Landscape Genetics. BioScience 58, 199–207. https://doi.org/10.1641/B580306.
Hortal, J., De Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M., Ladle, R.J., 2015. Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. Ann. Rev. Ecol. Evol. Syst. 46, 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400.
Kingma, D.P., Ba, J.L., 2015. Adam: a method for stochastic optimization. ArXiv Preprint ArXiv 1412, 6980.
Kittlein, M.J., Mora, M.S., Mapelli, F.J., Austrich, A., Gaggiotti, O.E., 2022. Deep learning and satellite imagery predict genetic diversity and differentiation. Meth. Ecol. Evol. 13, 711–721. https://doi.org/10.1111/2041-210X.13775.
Kokubum, M.N.C., Maciel, N.M., Matsushita, R.H., de Queiróz-Júnior, A.T., Sebben, A., 2009. Reproductive biology of the Brazilian sibilator frog *Leptodactylus troglodytes*. Herp. J. 19, 119–126.
Koopman, M.M., Carstens, B.C., 2010. Conservation genetic inferences in the carnivorous pitcher plant *Sarracenia alata* (Sarraceniaceae). Cons. Genet. 11, 2027–2038. https://doi.org/10.1007/s10592-010-0095-7.
Landguth, E.L., Cushman, S.A., 2010. Cdpop: A spatially explicit cost distance population genetics program. Mol. Ecol. Res. 10, 156–161. https://doi.org/10.1111/j.1755-0998.2009.02719.x.
Leaché, A.D., Fujita, M.K., 2010. Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). Proc. Roy. Soc. b: Biol. Sci. 277, 3071–3077. https://doi.org/10.1098/rspb.2010.0662.
Madelaire, C.B., Gomes, F.R., 2016. Breeding under unpredictable conditions: Annual variation in gonadal maturation, energetic reserves and plasma levels of androgens and corticosterone in anurans from the Brazilian semi-arid. Gen. Comp. Endocrin. 228, 9–16. https://doi.org/10.1016/j.ygcen.2016.01.011.
Manel, S., Schwartz, M.K., Luikart, G., Taberlet, P., 2003. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol. Evol. 18, 189–197. https://doi.org/10.1016/S0169-5347(03)00008-9.
Myers, E.A., Xue, A.T., Gehara, M., Cox, C.L., Davis Rabosky, A.R., Lemos-Espinal, J., Martínez-Gómez, J.E., Burbrink, F.T., 2019. Environmental heterogeneity and not vicariant biogeographic barriers generate community-wide population structure in desert-adapted snakes. Mol. Ecol. 28, 4535–4548. https://doi.org/10.1111/mec.15182.
Narvaes, P., Rodrigues, M.T., 2009. Taxonomic revision of *Rhinella granulosa* species group (Amphibia, Anura, Bufonidae), with a description of a new species. Arquivos Zool. 40, 1. 10.11606/issn.2176-7793.v40i1p1-73.
Nickerson, R.S., 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Rev. Gen. Psych. 2, 175–220.
Papadopoulou, A., Knowles, L.L., 2016. Toward a paradigm shift in comparative phylogeography driven by trait-based hypotheses. Proc. Nat. Acad. Sci. 113, 8018–8024. https://doi.org/10.1073/pnas.1601069113.
Pelletier, T.A., Carstens, B.C., 2014. Model choice for phylogeographic inference using a large set of models. Mol. Ecol. 23, 3028–3043. https://doi.org/10.1111/mec.12722.
Pelletier, T.A., Carstens, B.C., 2018. Geographical range size and latitude predict population genetic structure in a global survey. Biol. Let. 14 https://doi.org/10.1098/rsbl.2017.0566.
Perez, M.F., Bonatelli, I.A.S., Romeiro-Brito, M., Franco, F.F., Taylor, N.P., Zappi, D.C., Moraes, E.M., 2021. Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system. Mol. Ecol. Res. 22, 1016–1028. https://doi.org/10.1111/1755-0998.13534.
Peterman, W.E., Connette, G.M., Semlitsch, R.D., Eggert, L.S., 2014. Ecological resistance surfaces predict fine-scale genetic differentiation in a terrestrial woodland salamander. Mol. Ecol. 23, 2402–2413. https://doi.org/10.1111/mec.12747.
Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7, e37135.
Pless, E., Saarman, N.P., Powell, J.R., Caccone, A., Amatulli, G., 2021. A machine-learning approach to map landscape connectivity in Aedes aegypti with genetic and environmental data. Proc. Nat. Acad. Sci. 118 https://doi.org/10.1073/pnas.2003201118 e2003201118.
Prangle, D., Fearnhead, P., Cox, M.P., Biggs, P.J., French, N.P., 2014. Semi-automatic selection of summary statistics for ABC model choice. Stat. App. Genet. Molec. Biol. 13, 67–82.
Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W., 1999. Population growth of human Y chromosomes: A study of y chromosome microsatellites. Mol. Biol. Evol. 16, 1791–1798. https://doi.org/10.1093/oxfordjournals.molbev.a026091.
Rissler, L.J., 2016. Union of phylogeography and landscape genetics. Proc. Nat. Acad. Sci. USA 113, 8079–8086. https://doi.org/10.1073/pnas.1601073113.
Schrider, D.R., Kern, A.D., 2018. Supervised machine learning for population genetics: a new paradigm. Trends Genet. 34, 301–312. https://doi.org/10.1016/j.tig.2017.12.005.
Smith, B.T., McCormack, J.E., Cuervo, A.M., Hickerson, M.J., Aleixo, A., Cadena, C.D., Pérez-Emán, J., Burney, C.W., Xie, X., Harvey, M.G., Faircloth, B.C., Glenn, T.C., Derryberry, E.P., Prejean, J., Fields, S., Brumfield, R.T., 2014. The drivers of tropical speciation. Nature 515, 406–409. https://doi.org/10.1038/nature13687.
Storfer, A., Patton, A., Fraik, A.K., 2018. Navigating the interface between landscape genetics and landscape genomics. Front. Genet. 9, 68. https://doi.org/10.3389/fgene.2018.00068.
Sullivan, J., Smith, M.L., Espíndola, A., Ruffley, M., Rankin, A., Tank, D., Carstens, B., 2019. Integrating life history traits into predictive phylogeography. Mol. Ecol. 28, 2062–2073. https://doi.org/10.1111/mec.15029.
Thom, G., Gehara, M., Smith, B.T., Miyaki, C.Y., do Amaral, F.R., 2021. Microevolutionary dynamics show tropical valleys are deeper for montane birds of the Atlantic Forest. Nat. Comm. 12, 6269. https://doi.org/10.1038/s41467-021-26537-9.
Thomé, M.T.C., Carstens, B.C., Rodrigues, M.T., Alexandrino, J., Haddad, C.F.B., 2021a. Genomic data from the Brazilian sibilator frog reveal contrasting pleistocene dynamics and regionalism in two South American dry biomes. J. Biogeog. 48, 1112–1123. https://doi.org/10.1111/jbi.14064.
Thomé, M.T.C., Carstens, B.C., Rodrigues, M.T., Galetti, P.M., Alexandrino, J., Haddad, C.F.B., 2021b. A role of asynchrony of seasons in explaining genetic differentiation in a Neotropical toad. Heredity 127, 363–372. https://doi.org/10.1038/s41437-021-00460-7.
Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S., Fumagalli, M., 2019. ImaGene: A convolutional neural network to quantify natural selection from genomic data. BMC Bioinf. 20, 1–12. https://doi.org/10.1186/s12859-019-2927-x.
van Etten, J., 2017. R package gdistance: Distances and routes on geographical grids. J. Stat. Soft. 76, 1–21. https://doi.org/10.18637/jss.v076.i13.
Vasconcellos, M.M., Colli, G.R., Weber, J.N., Ortiz, E.M., Rodrigues, M.T., Cannatella, D. C., 2019. Isolation by instability: historical climate change shapes population structure and genomic divergence of treefrogs in the Neotropical Cerrado savanna. Mol. Ecol. 28, 1748–1764. https://doi.org/10.1111/mec.15045.
Waits, L.P., Storfer, A., 2015. Basics of Population Genetics: Quantifying Neutral and Adaptive Genetic Variation for Landscape Genetic Studies. In: Landscape Genetics. John Wiley & Sons, Ltd, pp. 35–57. https://doi.org/10.1002/9781118525258.ch03.
Wang, I.J., 2010. Recognizing the temporal distinctions between landscape genetics and phylogeography. Cons. Genet. 19, 2605–2608. https://doi.org/10.1111/j.1365-294X.2010.04715.x.
Wells, K.D., 2007. The ecology and behavior of amphibians. University of Chicago Press, Chicago.
Wieringa, J.G., Boot, M.R., Dantas-Queiroz, M.V., Duckett, D., Fonseca, E.M., Glon, H., Hamilton, N., Kong, S., Lanna, F.M., Mattingly, K.Z., Parsons, D.J., Smith, M.L., Stone, B.W., Thompson, C., Zuo, L., Carstens, B.C., 2020. Does habitat stability structure intraspecific genetic diversity? It's complicated. Front. Biogeog. 12 e45377. 10.21425/F5FBG45377.
Wright, S., 1943. Isolation by distance. Genetics 28, 114.
Zamudio, K.R., Bell, R.C., Mason, N.A., 2016. Phenotypes in phylogeography: Species' traits, environmental variation, and vertebrate diversification. Proc. Nat. Acad. Sci. USA 113, 8041–8048. https://doi.org/10.1073/pnas.1602237113.