Caudata macrogenetics: Geographic attributes predict mitochondrial genetic variation in salamanders on a global scale

Short running title: Macrogenetics in salamanders worldwide

Abstract

Aim: Genetic diversity contains valuable information about ecological and evolutionary aspects of species, such as historical demography. Intraspecific genetic variation is shaped by species' natural history traits and by characteristics of geography and climate within their ranges. Salamanders (Order Caudata) are amphibians of ecological and conservation interest because of their global distribution, trait diversity, and roles within ecological communities. Here, we studied genetic variation within salamanders to investigate predictors of nucleotide diversity and spatial patterns of genetic differentiation.

Location: Global.

Time Period: Present.

Major Taxa Studied: Salamanders.

Methods: We repurposed mitochondrial DNA sequences, natural history traits, and geographic and environmental data from open-access databases for 214 salamander species. We calculated nucleotide diversity (π) and tested for isolation by distance (IBD) and isolation by environment (IBE) within each species. We analyzed these three response variables with random forest analyses and phylogenetic comparative methods using 27 predictors expected to be associated with genetic variation (e.g., number of offspring, total range size).

Results: We recovered 8,175 Cytb sequences with their associated geographic coordinates, of which 7,007 sequences were manually curated by us. Species range size and sample size (the number of sequences) were important predictors of genetic variation. We found higher diversity in regions including the Neotropics and central-eastern Europe. The absence of strong phylogenetic signal in π, IBD, and IBE suggests that genetic variation is shaped by local ecological and geographical factors rather than by shared ancestry.

Main Conclusions: Our finding that range size was an important predictor of genetic variation aligns with theoretical expectations that species with larger ranges tend to harbor more genetic diversity. Our results underscore the importance of integrating spatial data (i.e., DNA sequences

For Peer Review

and their associated coordinates) into macrogenetic studies, providing valuable information for future studies and conservation strategies targeting regions with high or low genetic diversity.

Keywords

Amphibians, isolation-by-distance, isolation-by-environment, nucleotide diversity, species range size.

For Peer Review

Introduction

Genetic variation plays a foundational role in maintaining biological diversity (Kardos et al., 2021). One significant evolutionary factor that predicts genetic variation within populations is the number of individuals that contribute DNA to the next generation (the effective population size; Wright, 1931). Genetic variation is proportional to population size, as explained by the neutral theory of molecular evolution (Kimura, 1968). Gene flow between populations is another key factor influencing variation because it can add and maintain novel alleles, counteract the effects of genetic drift and selection, and enable population and species persistence (Park et al., 2024). Natural selection also has an important role in affecting neutral diversity of populations through background or linked selection (Corbett-Detig et al., 2015). Understanding this variation and which factors influence it across different geographic and environmental contexts is essential for elucidating population dynamics and will help us identify populations with conservation needs (Pauls et al., 2013). For example, declining levels of genetic variation may be further exacerbated under global change scenarios due to the reduction of suitable habitat for species (Schierenbeck, 2017).

Climatic and geographic factors shape patterns of dispersal and population size through mechanisms such as habitat connectivity, environmental stability, and barriers to gene flow. These processes, in turn, determine the extent and distribution of genetic variation within species (e.g., Hanson et al., 2017). For example, geographic distance can limit gene flow among populations, resulting in patterns of isolation-by-distance (IBD), where genetic dissimilarity increases as populations become more spatially separated (Wright, 1943). Similarly, environmental differences can restrict gene flow by selecting against migrants non-adapted to the new environmental conditions, leading to isolation-by-environment (IBE), which reflects the influence of habitat heterogeneity and ecological gradients on genetic variation (Wang and Bradburd, 2014). Historical climatic fluctuations and complex topography often act together to promote cycles of population contraction and expansion, which generate and maintain genetic variation. In the alpine toad *Scutiger ningshanensis*, for instance, repeated glacial cycles during the Pleistocene, combined with uplifts of mountain ranges, likely fragmented populations and created refugia, increasing opportunities for genetic differentiation (Meng et al., 2014). Life

For Peer Review

history traits, particularly those related to reproduction, also influence genetic diversity by affecting effective population size and gene flow (Ellegren and Galtier, 2016). For example, species with high fecundity, widespread dispersal, or external fertilization can maintain large and connected populations, which reduces the effects of genetic drift and promotes diversity. Broad-scale analyses across hundreds of plant and animal species further support the idea that reproductive traits and climatic variables jointly shape genetic diversity by influencing demographic stability and dispersal potential (De Kort et al., 2021), suggesting that these variables act as consistent predictors of genetic variation across taxa.

Studies on focal taxa within a limited geographic range provide the resolution needed to understand local genetic patterns and idiosyncrasies (e.g., Fonseca et al., 2021). It is equally important to investigate genetic variation on broader geographic and taxonomic scales, to have a global perspective of the role of ecological and geographic factors on evolutionary patterns. Macrogenetics is an approach to analyze open-access DNA sequences across large scales and investigate patterns of intraspecific genetic diversity across higher taxonomic groups (Blanchet et al. 2017). Macrogenetic research has gained popularity over the past decade, driven by the increasing availability of molecular resources (e.g., GenBank), computational tools (e.g., machine learning), and trait databases (Leigh et al., 2021). Despite limitations, particularly the genetic markers represented by available data (see Millette et al., 2021; Paz-Vinas et al., 2021), efforts to understand large-scale patterns and predictors of intraspecific genetic diversity have been undertaken across various taxonomic groups. A study of global genetic diversity in insects attributed high levels of mitochondrial DNA (mtDNA) genetic diversity in the subtropics to temperature and climatic stability (French et al., 2023). Another study of nucleotide diversity using mtDNA for animals and chloroplast DNA for plants in over 38,000 species found that latitude significantly correlates with genetic diversity, with tropical species exhibiting higher intraspecific diversity (Fonseca et al., 2023). These studies highlight the potential of macrogenetics to uncover important predictors of intraspecific genetic diversity across broad sets of taxa globally.

Amphibians contain remarkable diversity in life history strategies and the geographic and climatic variability of their habitats. Recent studies of amphibian genetic diversity indicate the

For Peer Review

biogeographic region can influence which factors predict intraspecific genetic variation. For example, in Nearctic amphibians, taxonomic family, the number of sequences, and latitude were key predictors of intraspecific variation (Barrow et al., 2021), whereas range size, elevation, latitude, and precipitation predicted genetic variation in Neotropical amphibians (Amador et al., 2024). In the Americas, Lawrence et al. (2023) found a strong relationship between environmental variables and amphibian expected heterozygosity (H_E), and that amphibian H_E decreased with sample size (number of individuals). Within amphibians, salamanders are particularly valuable for macroevolutionary studies because of their broad distribution (excluding Oceania, the Afrotropics, and Antarctica), manageable number of species, and unique traits. They display remarkable ecological and life history diversity that can directly shape population dynamics and, consequently, genetic structure. For instance, body size variation, from the tiny *Thorius arboreus* (~20 mm) to the 1.8-meter-long *Andrias japonicus,* can influence dispersal ability and home range size, thereby affecting gene flow and population structure. Likewise, salamanders include generalist species like the tiger salamander, *Ambystoma tigrinum*, versus habitat specialist species adapted to aquatic, terrestrial, or arboreal environments, which can promote or constrain connectivity across landscapes.

One trait that may influence patterns of genetic diversity in salamanders is developmental mode. Species with direct development (i.e., without a larval stage) often have limited dispersal, which could lead to smaller, isolated populations favoring population differentiation (Paz et al. 2015; Liedtke et al. 2022). In contrast, species with a larval stage and dispersing terrestrial adults may maintain more interconnected populations, reducing genetic differentiation (e.g., Zamudio and Wieczorek 2007). Some salamanders exhibit paedomorphosis, a reproductive strategy in which adults retain larval characteristics, which reduces gene flow by allowing persistent aquatic life cycles in species that could otherwise be terrestrial (Denoël et al., 2005). Recent studies have begun to untangle the relationships between species-level traits and genetic variation. For example, Segovia-Ramírez et al. (2023) used genomic data from 62 Neotropical salamander species (one individual per species) to demonstrate that precipitation variability and snout-vent length (body size) were significant predictors of genomic diversity, perhaps related to their roles in shaping population persistence and abundance. Parsons et al. (2024) combined genetic, geographic, climatic, and life history data from salamanders and used a machine learning model

For Peer Review

to identify predictors of unrecognized genetic lineages. They found that Caudata hidden diversity is the result of variation in climatic variables. A more comprehensive, georeferenced view of intraspecific genetic variation in salamanders, encompassing more species and potential predictors (e.g., climatic and natural history traits), would provide useful insights into the evolutionary and demographic processes driving variation in this group.

Here, we combine molecular, phenotypic, and environmental open-access data to investigate the determinants of intraspecific genetic variation in salamanders. We used machine learning techniques and phylogenetic comparative methods to determine whether natural history traits or climatic or environmental variables can predict genetic variation within species on a global scale. We hypothesized that reproductive traits, such as the total number of eggs (a proxy for fitness) and reproductive mode (e.g., larval vs. direct development), would predict intraspecific genetic variation in salamanders. Species that produce more offspring may maintain larger, more stable populations, which can preserve higher levels of genetic diversity over time (Lacy, 1987; Reed and Frankham, 2003). Additionally, reproductive mode could influence dispersal, leading to lower expected connectivity and more genetic differentiation in species with direct development compared to those with aquatic larvae. We also predicted that traits associated with total species population size (e.g., range size) would be positively correlated with genetic variation, as larger ranges may support more numerous and genetically diverse populations (Frankham, 1996). Finally, we expected high spatial autocorrelation in nucleotide diversity of salamanders overall based on their limited dispersal capabilities, which can lead to genetic clustering at small scales (Fusco et al., 2021).

For Peer Review

Methods

Data collection

We repurposed Caudata DNA sequences, natural history traits, and geographic and environmental data from open-access databases. Following the approach of recent research on the determinants of genetic diversity of multiple amphibian species (e.g., Barrow et al., 2021; Amador et al., 2024), we assembled sequences of the cytochrome-b (Cytb) mitochondrial gene. This gene is suitable for genetic diversity studies because it contains both conserved and variable regions, is commonly used in amphibian single-locus phylogenetic and population genetic studies, and is therefore the most abundant gene in open-access databases for amphibians. Sequences were obtained from GenBank (National Center for Biotechnology Information, NCBI), phylogatR (Pelletier et al. 2022), the *phruta* R package (Román-Palacios, 2023), and Amador et al. (2024). We included species with at least 10 sequences, seeking a better representation of intraspecific genetic variation. Our dataset contained species with high variation in sampling effort, such as those with more than 200 sequences. To account for this sampling bias, we also conducted a set of analyses on randomly subsampled datasets that included 10–20 sequences per species. Alignments were generated, visualized, and edited with the MUSCLE aligner v.3.8.31 (Edgar, 2004) in AliView v.1.28 (Larsson, 2014) and saved in FASTA format.

The geographic coordinates (latitude and longitude in decimal degrees format) previously associated with each Cytb sequence were recovered from GenBank, phylogatR, and Amador et al., (2024); corresponding to 31 species (14.5% of the total species) and 1,169 sequences (14.3% of the total sequences). We then manually retrieved the remaining geographic coordinates associated with the sequences (corresponding to 183 species and 7,007 sequences) following a tutorial prepared for this purpose (see Supplementary Material 1). We georeferenced localities using either the function geocode() of the *tidygeocoder* R package (Cambon et al., 2021) or the GEOLocate web application (Rios and Bart, 2010). We only retained sequences that were georeferenced with an error of less than 20 km, resulting in 12 species with fewer than 10 georeferenced sequences.

For Peer Review

We calculated both the total range size and the sampling range size of DNA sequences for each species as follows. For total range size, we obtained the geographic range map in shapefile format for each species from the International Union for Conservation of Nature (IUCN) portal using the getIUCN() function of the *rasterSp* R package (RS-eco, 2023). Species ranges not available in the IUCN portal (24 species) were generated with minimum convex polygons using the sequence coordinates and the function mcp() in the R package *adehabitatHR* (Calenge, 2024; see Supplementary Material). This same approach was also used to obtain the geographic range of the sampled sequences for all species (i.e., using only the sequence occurrences). Range size was calculated from shapefiles using the areaPolygon() function of the *geosphere* R package (Hijmans, 2022). Elevation and latitude (min, max, and mean) were obtained from the retrieved geographic coordinates using the function get_elev_point() of the *elevatr* R package (Hollister, 2023) for elevation, and custom R scripts for latitude (see Supplementary Material 2). We also obtained precipitation and temperature data (mean and standard deviation) from WorldClim v2.1 (Fick and Hijmans, 2017) using the function worldclim_global() of the *geodata* R package v0.6-2 (Hijmans et al., 2024).

We used the same natural history and geographic traits as in Amador et al., (2024) with slight additions and omissions, choosing predictors we expected to be related to genetic variation (e.g., body size, reproductive mode, elevation, latitude, species range size, precipitation and temperature; Table S1). We omitted the variable activity because nearly all salamanders (except three of the 214 species) included in our dataset were nocturnal. Seven new predictors were added to the matrix (27 predictors in total), which included sampling effort/coordinates (i.e., number of sequences with associated coordinates), sequence length/coordinates (i.e., number of base pairs of sequences with associated coordinates), snout-vent length in millimeters (SVL), paedomorphism (whether or not juvenile features are retained as an adult), litter size min (minimum number of offspring or eggs per clutch), litter size max (maximum number of offspring or eggs per clutch), and latitudinal midpoint (mean of the minimum and maximum latitudinal values) (Table S1). We obtained trait information from the open-access database AmphiBIO (Oliveira et al. 2017), species accounts in AmphibiaWeb (AmphibiaWeb, 2024), and relevant literature.

For Peer Review

Measures of genetic variation

We evaluated overall genetic diversity (range-wide nucleotide diversity) and spatial patterns of genetic variation (IBD and IBE) within salamander species. We calculated nucleotide diversity (π) with the function nuc.div() of the *pegas* R package (Paradis, 2010) for all sequences (π_A) and sequences with associated coordinates only (π_S). Genetic distances (*gendist*) for each species were calculated based on the raw difference between sequences, using the dist.dna() function of the *ape* R package (Paradis and Schliep, 2019). We tested IBD and IBE for each salamander species as follows. Geographic distance (*geodist*) was calculated between each pair of sequence coordinates using Euclidean distance with the dist() function of the *stats* package (R Core Team, 2023) in R. Before calculating environmental distances, we used the function scale() in R to standardize variables measured in different units (e.g., temperature - °C or precipitation - mm). Environmental distance (*envdist*) was calculated between each pair of sequence coordinates based on the 19 variables of the WorldClim v2.1 database using the vegdist() function of the *vegan* R package (Oksanen et al., 2024). This function calculates a distance matrix, selecting all environmental variables using Euclidean distances. We performed Multiple Matrix Regression with Randomization (MMRR) analyses for each species (Wang, 2013). MMRR quantifies the relative effects of IBD and IBE, represented as distance matrices (*geodist* and *envdist*, respectively), in explaining genetic variation, represented by genetic distances. We compared the observed correlations between *geodist* vs. *gendist* (IBD) and *envdist* vs. *gendist* (IBE) to the correlations from 1,000 permutations. Species were coded as 'Yes' if they did present IBD and IBE or 'No' if they did not based on an assessment of the associated p-values from the permutation tests. To account for multiple comparisons, we applied a strict Bonferroni correction (Bonferroni, 1936) to reduce the p-value for assessing significance. We used the output (yes or no IBD; yes or no IBE) as binary response variables for further analyses.

We mapped π_S to visualize the distribution of genetic diversity across the globe and assess differences among biogeographic realms (Nearctic, Neotropic, Oriental, and Palearctic). Briefly, following Amador et al. (2024), we combined sequences within a species based on a maximum distance of 100 km between coordinates, and calculated π_S for each locality (see Supplementary

For Peer Review

Material 3). We then mapped π_S for four different grid sizes (1–4), which correspond with one to four decimal degree grids, or approximately 110 km^2 to 440 km^2. We also visualized the number of sequences per grid cell for the four resolutions. For these visualizations, we used the R packages *ggplot2* (Wickham, 2016), *ggspatial* (Dunnington, 2023), and *sf* (Pebesma, 2018; Pebesma and Bivand, 2023). Finally, we tested for spatial autocorrelation for π and the number of sequences separately with the Moran's I test using the moran.test() function in the *spdep* R package (Bivand et al., 2013).

To further understand the potential link between historical demography and the patterns of genetic variation we found, we used a common statistic for evaluating population size change, Tajima's D (Tajima, 1989), using the tajima.test() function of the *pegas* package. This test is compatible with different interpretations related to natural selection versus demographic change, making it difficult to distinguish among hypotheses (Yang, 2014). Therefore, we evaluate these results for comparison rather than include the statistic as a variable in our primary analyses. We followed Fonseca et al. (2023) in assuming that the demographic change interpretation is likely more relevant for global comparisons. Thus, in our study, we interpret Tajima's D values close to zero as representative of neutrality (no population size change over time), positive values as evidence of recent population contraction, and negative values as recent population expansion.

Predicting genetic variation using random forests (RF)

We used random forest (RF) regression to predict π and RF classification to predict IBD and IBE based on a set of geographic range characteristics and ecological variables. We built RF models with the function randomForest() in the *randomForest* R package (Liaw & Wiener, 2002), with 10,000 trees and 100 permutations per model. We used the tuneRF() function to obtain the optimal mtry parameter (number of variables that are randomly sampled as candidates at each split) and then ran a new RF regression or classification analysis using the best mtry value. The data was split into a training set (70% of the data) for building models and a testing set (30%) for making predictions. For RF regression, we assessed models using the R-squared (R^2) statistic, which represents the proportion of the variance explained by the model, with a higher R^2 value

For Peer Review

indicating a better fit and more predictive power. Relative importance of the predictor variables for π_A and π_S was evaluated based on the percentage increase in mean squared error (%IncMSE).

In RF classification, in addition to the original models, we also tried upsampling and downsampling to address the large class imbalance within our datasets (yes IBD = 55 species, no IBD = 156 species; yes IBE = 39 species, no IBE = 172 species; based on MMRR results with Bonferroni correction). For downsampling, the majority cases (no IBD/IBE) were randomly subsampled to equal the number of minority cases (yes IBD/IBE). For upsampling, minority cases were randomly duplicated to equal the number of majority cases. We chose the best model based on the lowest out-of-bag (OOB) error rate. To assess overfitting and improve cross-validation in the model, we split our data into a training set and a test set (70% train, 30% test). To evaluate predictions of the RF classification models, we created a confusion matrix with the confusionMatrix() function of the *caret* R package (Kuhn, 2008) and tested whether model accuracy was significantly better than the no information rate (NIR). Finally, we assessed the relative importance of the variables based on the Mean Decrease Accuracy (MDA) for the RF classification models of IBD and IBE.

Evaluating predictors of genetic variation with phylogenetic comparative methods

For phylogenetic analyses, we downloaded a phylogeny subset from VertLife.org for the species in our dataset, which represents phylogenetic relationships inferred by Jetz and Pyron (2018). We used a subset of 180 species for these analyses because 34 species from our dataset are not included in the available phylogeny. To assess the potential effect of phylogenetic history in our models, we tested for phylogenetic signal in π using Pagel's λ (lambda; Pagel, 1999) and Blomberg's K (Blomberg et al., 2003) implemented in the *phytools* R package (Revell, 2012) with the function phylosig(). For IBD and IBE, we used the *D* statistic implemented in the package *caper* with the function phylo.d() to assess phylogenetic signal.

We used phylogenetic independent contrasts (PIC; Felsenstein, 1985) to investigate the evolutionary correlation between π and the most important predictors identified in the RF analyses. We used the pic() function of the *ape* R package to compute PIC for both response and

For Peer Review

explanatory variables. We then fit linear models in R for both the uncorrected and phylogenetically corrected (PIC) response and predictor variables. We also used phylogenetic generalized linear mixed models (PGLMMs) to analyze binary traits as described in Ives and Helmus (2011). PGLMMs allow us to account for phylogenetic covariance while reducing Type I error rates. This method was used to analyze IBD and IBE as binary dependent variables (0 for no IBD/IBE, 1 for yes IBD/IBE) versus the most important predictors based on RF results. Because of the nature of our data (binary outcomes), we choose binomial as the error family with a logit link function. We compared IBD vs. sampling range size, number of sequences, maximum latitude, and mean precipitation; and IBE vs. sampling range size, number of sequences, mean elevation, and mean latitude. We implemented this analysis using the function pglmm() in the *phyr* R package (Li et al., 2020), including a phylogenetic random effect to account for the non-independence among species due to shared evolutionary history.

For Peer Review

Results

Data summary

We obtained 13,265 Cytb sequences from 214 species distributed globally, of which 8,963 sequences were retrieved directly from GenBank, 2,349 sequences from phylogatR, 1,657 from the *phruta* package, and 296 from Amador et al. (2024). Of these, we obtained geographic coordinates for 8,175 sequences from 213 species. A single species, *Plethodon sherando*, did not have associated coordinates, therefore was not included in IBD and IBE analyses; and 11 species had between four and nine sequences with coordinates. The final dataset included 214 species in nine families (representing >25% of species and 90% of families globally), three response variables (π, IBD, and IBE), and 26 predictors (Table S2). All sequences associated with geographic coordinates and R scripts used to retrieve sequences and coordinates are available as Supplementary Material 3.

Patterns of genetic variation in global salamanders

Nucleotide diversity within species ranged from $\pi_A = 0$ to $\pi_A = 0.087$ (*Pseudoeurycea lineola*) when we analyzed all sequences, and from $\pi_S = 0$ to $\pi_S = 0.089$ (*Desmognathus amphileucus*) using only the sequences with associated coordinates. We found that nucleotide diversity for all sequences (π_A) and for the sequences with associated coordinates (π_S) were highly correlated (Figure S1; Pearson's correlation $r = 0.94$, $t(212) = 40.12$, p-value < 0.001). Based on this validation, we conducted further analyses using sequences with associated coordinates (π_S). *Bolitoglossa yariguiensis* was the only species with a value of $\pi = 0$ (Table S2; Figure S2). We found that π and the number of sequences were not randomly distributed across the world map (Figure 1, Figures S3 – S5). For example, regions with high genetic diversity included the Ecuadorian Amazon and central-southern Mexico in the Neotropics; and central-eastern Europe in the Palearctic. Areas with low genetic diversity were identified in northern North America (Nearctic), western Europe, and central-western China in the Oriental realm (Figure 1). Furthermore, we found statistically significant, but weak, positive spatial autocorrelation in both

For Peer Review

π_S (Moran's I = 0.1173, p-value = 0.0003) and the number of sequences (Moran's I = 0.2053, p-value < 0.0001).

Most species did not present IBD and IBE patterns, with 156 species (73.9%) showing no IBD and 172 species (81.5%) showing no IBE (Table S2; Figure S6). We found that only 18 species had significant patterns for both IBD and IBE (Table S2). The Neotropical and Oriental realms had the lowest percentages of species that showed IBD and IBE (IBD: 13.3% and 24.3%, IBE: 13.3% and 9.8% respectively; Figure S7). The Palearctic region had the highest percentage of species that showed IBD (31.4%) and IBE (25.7%) patterns (Figure S7).

Genetic variation was randomly distributed across the phylogeny (Figure 2). For example, we found species in the same genus with dissimilar π values, such as *Plethodon* (range = 0.0008 – 0.0698) and *Eurycea* (range = 0.0015 – 0.0768) in Plethodontidae, *Paramesotriton* (range = 0.0041 – 0.0767) in Salamandridae, and *Hynobius* (range = 0.0025 – 0.0853) in Hynobiidae. We found no or low phylogenetic signal in π based on the method used, for example, with Pagel's λ = 0.029, p-value (based on LR test) = 0.425; and Blomberg's K = 0.106, p-value (based on 1000 randomizations) = 0.046 (Figure S8). We found no significant phylogenetic signal in IBD (Estimated D_{IBD} = 0.908, p-value = 0.146), and significant but moderate phylogenetic signal in IBE (Estimated D_{IBE} = 0.741, p-value = 0.005) (Figure S9).

Our results from assessing Tajima's D indicate that many species in our dataset (n = 107) have undergone recent population expansion (i.e., negative Tajima's D, more recent mutations than expected under neutrality; Figure S10). Positive Tajima's D values indicate more alleles at intermediate frequencies than expected under neutrality, possibly because of a recent population contraction or bottleneck, as we observed in 49 salamander species (Figure S10). Fifty-seven species do not show strong evidence for population expansion or contraction (Tajima's D ≈ 0) suggesting no deviation from expectations under neutrality (Table S2; Figure S10).

Predictors of genetic variation in global salamanders

For Peer Review

Our random forest analyses showed an influence of sampling effort on genetic variation estimates, with the number of sequences (with associated geographic coordinates) as a top predictor of π, IBD, and IBE (Figure 3). Additionally, we compared both species range size variables in our dataset, total range size versus sampling range size, and we found a strong, positive linear relationship (Figure S11; Pearson's correlation $r = 0.78$, $t(212) = 18.41$, p-value < 0.001). These results suggested the sampled range for most species adequately represented the total range size. Nevertheless, we ran RF analyses using either total range size or sampling range size and found similar results between the two variables (Figures 3, S12). Because the range size variables were highly correlated with one another and captured similar aspects of the species' biology, we subsequently describe the results using the sampling range size.

We found that sampling range size was the best predictor of π, and the second-best predictor of IBD and IBE in random forest analyses (Figure 3). Random forest regressions showed that taxonomic family was the third-best predictor of π (Figure 3a, S13). Validation of the RF regression results for π based on the R^2 score indicated good model performance ($R^2 = 0.94$). Mean precipitation, elevation, and latitude were other important predictors of IBD (Figure 3b, S14). For IBE, RF showed again that mean precipitation was a top predictor, and body size, and latitude were also important predictors (Figure 3c, S15). In RF classification analyses, the best IBD and IBE models were obtained when we downsampled our dataset. Our IBD RF model had high accuracy (93.65%; 95% Confidence Interval CI: 84.53% – 98.24%), sensitivity (80.00%), and specificity (97.92%); our model was statistically better than predicting only the majority class, p-value [Acc > NIR] = 0.0001. The IBE RF model had high accuracy (94.64%; CI: 85.13% – 98.88%), sensitivity (71.43%), and high specificity (97.96%); however, the model's accuracy was not statistically better than the NIR, p-value [Acc > NIR] = 0.0686.

Sampling range size had a significant, positive relationship with π, showing that genetic diversity within species increases with the geographic area sampled (Figure 4A; $R^2 = 0.09$; p-value < 0.001). We found a similar pattern using phylogenetic independent contrasts (PIC), where sampling range size and π have a strong, positive relationship (Figure 4B; $R^2 = 0.25$; p-value < 0.001). We found a significant, positive relationship between mean temperature and π, suggesting that species living in areas with higher temperatures have more genetic diversity

For Peer Review

(Figure S16A). However, PIC did not show a significant correlation between temperature (SD) or other predictors (body size and elevation) and π after correcting for phylogenetic relatedness (Figures S16 – S19).

The number of sequences associated with geographic coordinates and sampling range size were top predictors for IBD and IBE models. There were clear differences between the number of sequences sampled and IBD/IBE patterns (p-value < 0.01), in which species with more sequences tended to show significant IBD/IBE patterns (Figure 5B, D). Species with significant IBD/IBE had slightly larger sampling range sizes on average (Figure 5A, C). When temperate and tropical species were considered separately, we found a similar trend in which species with larger ranges were more likely to exhibit IBD and IBE, but the only significant difference was for temperate species with IBE patterns (Figure S20). Phylogenetic generalized linear mixed models (PGLMM) found that the number of sequences had a statistically significant, positive relationship with IBD ($\beta = 0.017161$, SE = 4.16×10^{-3}, $p < 0.001$), and IBE ($\beta = 0.016539$, SE = 4.16×10^{-3}, $p < 0.001$). These results suggest that as sampling effort increases, the likelihood of detecting an IBD or IBE pattern increases (Table S3). Coefficients, standard errors, Z-scores, and p-values of the other predictors (i.e., sampling range size for IBD, and mean Latitude for IBE) are shown in Table S3.

For Peer Review

Discussion

In this study, we analyzed 13,265 Cytb sequences from 214 global salamander species and identified key predictors of intraspecific diversity (π) and patterns of spatial genetic variation (IBD and IBE). Species range size was an important predictor of all three aspects of genetic variation. Specifically, species occupying larger geographic areas tended to have greater overall nucleotide diversity and to be more geographically and environmentally structured. The results for π are consistent with population genetics theory (Wright, 1931), in that large geographic ranges can support larger populations that harbor more genetic diversity because of reduced genetic drift (Lacy, 1987). The significant relationship found between range size and spatial genetic differentiation (i.e., IBD and IBE) could be explained by the greater potential for geographic and environmental isolation between populations at distant locations in species with large ranges (Eckert et al., 2008). However, species range size alone does not explain much of the variation in IBD or IBE presence (small effect sizes recovered in logistic regression models), and the strong influence of the number of sequences sampled highlights one of the major challenges for macrogenetic studies—limited data availability for many species worldwide.

Global patterns of nucleotide diversity (π)

Our results provide evidence of the uneven distribution of nucleotide diversity (π_S) across the globe. We observed that certain hotspots of genetic diversity, like the Ecuadorian Amazon and central Europe, southwestern and central North America, and southeast Asia, align with regions of high biodiversity (e.g., Mittermeier et al., 2011). Contrastingly, low genetic diversity in regions such as northern North America or western Europe could reflect recent recolonizations or historical bottlenecks (e.g., Riberon et al., 2002; Bonato et al., 2018; Auteri et al., 2022). Our results contrasted with Miraldo et al. (2016), who found that amphibian genetic diversity decreased at higher latitudes. In our study, latitude variables (mean, maximum, or minimum) were not important predictors of genetic diversity in salamanders. The high diversity we found in the Neotropics (16 species) could be explained by a combination of factors including tropical climate stability, high topographic heterogeneity, or long evolutionary history in the Neotropics (Antonelli, 2022).

For Peer Review

Another plausible explanation for the observed high levels of π in some regions is that they coincide with glacial refugia, areas where species persisted and maintained high levels of genetic diversity through time (e.g., Ramírez-Barahona and Eguiarte, 2013). The Andean foothills were a major glacial refugium, particularly the central Andes in the Neotropical region (e.g., Escobar et al., 2021). In the Nearctic, the Southern Appalachians and coastal areas of the Pacific Northwest are considered potential glacial refugia (e.g., Shafer et al., 2010). In the Palearctic, the Caucasus Mountain range has been identified as a potential refugium during glacial periods (e.g., *Triturus* newts; Wielstra et al., 2013). The Korean Peninsula and the mountainous regions of China are two examples of glacial refugia in the Oriental region (Fu and Wen, 2023). Low levels of π in northern North America are consistent with the impact of historical and demographic processes in the region, such as Pleistocene glaciations and associated range contractions in salamanders (Rovito and Schoville, 2017; López-Delgado and Meirmans, 2022). We showed that species experiencing expansion, based on Tajima's D, presented low levels of nucleotide diversity (Figure S21). This result is expected when recent expansion from a small population occurred after a bottleneck, leading to a limited gene pool carried by a few founding individuals, such as range expansions after the last glacial maximum (Mayr, 2001).

Global patterns of isolation-by-distance (IBD) and environment (IBE)

The limited evidence for both IBD and IBE across species (only 21 species, or ~10%, showed both patterns) suggests that gene flow in many salamander species may not be constrained by geography or environment. Instead, the genetic variation in these species is shaped more by demographic processes such as bottlenecks or population expansions than current spatial or ecological gradients (Sexton et al., 2024). For instance, some salamander species may maintain high gene flow that can homogenize genetic variation, reducing patterns of IBD or IBE. In contrast, others with strong philopatry may exhibit spatial genetic structure influenced more by historical processes that disrupt patterns of genetic isolation associated with IBD and IBE (Wake, 2009; Sexton et al., 2014). For example, montane populations isolated in separate refugia in response to climatic changes might differentiate and exhibit highly restricted gene flow even at

For Peer Review

short distances or among similar environments, as reported in plethodontid salamanders (Kozak and Wiens, 2006; Rovito, 2017).

These results underscore the importance of considering biogeographic and ecological context when evaluating the drivers of intraspecific genetic variation on a global scale. The proportions of species showing IBD and IBE differed slightly across biogeographic realms, with the Palearctic including the most species with these patterns. One example is the fire salamander (*Salamandra salamandra*), one of the most common and widespread salamander species in Europe, which was found to be strongly structured in a landscape genetics analysis (Bani et al., 2015). The Neotropical and Oriental regions had lower proportions of species with IBD or IBE patterns compared to the other regions, but these differences were relatively minor. This contrast between realms suggests that historical biogeography and environmental complexity are key to understanding spatial genetic variation. For example, the incidence of IBD/IBE in the Nearctic and Palearctic could reflect the consequence of glacial cycles (Schmitt, 2007), while the low incidence in tropical regions may reflect more recent colonization dynamics that occurred in the late Pleistocene (Carnaval et al., 2009).

Predictors of genetic variation in global salamanders

Two of the best predictors of nucleotide diversity in this study, the number of sequences and range size, could both be considered proxies of demographic processes. Species with more available sequences could truly be more abundant in nature, reflecting both a true demographic property as well as sampling bias towards species that are more frequently encountered. Either way, sampling more sequences could provide a greater likelihood of detecting genetic variation because more samples analyzed will capture the full range of genetic diversity present within a species. When we analyzed π, IBD, and IBE using a standardized sampling effort per species (10 to 20 sequences), we note that the number of sequences was not among the top predictors in the random forest models (it ranked 5th in π, and 4th in IBE) and was not an important predictor of IBD. Similarly, species sampling range size was not a key predictor for IBD and IBE but was the most important predictor for π in this subsampled dataset (Figure S22). Litter size, referred as the minimum number of offspring in a single birth event, was the best predictor for IBD, and

For Peer Review

elevation was for IBE (Figure S22); however, the error rate estimated in these RF classification models was too high compared to the original models (Figure 3). These results suggest that differences in sampling effort across species can provide challenges for macrogenetic studies and should be considered carefully in future studies.

The positive correlations found between number of sequences, range size, and π in our dataset are consistent with the prediction that total population size (with number of sequences and range size as potential proxies) corresponds with genetic diversity (Figure S23). Species with larger geographic ranges can have larger and more balanced populations, reflecting the relationship between effective population size and genetic variation (Waples, 2022). Geographic range size was also one of the best predictors of IBD and IBE, like previous findings in a wide range of taxonomic groups (Pelletier and Carstens, 2018). A large range can harbor more differentiation, where population subdivision could be influenced by different ecological and evolutionary pressures and undergo contrasting demographic histories (Lowe et al., 2017). The relationships between range size and patterns of genetic variation we identified in global salamanders is partially consistent with previous regional studies, such as Amador et al. (2024), which found a similar pattern in Neotropical amphibians for π, but not for IBD or IBE. Similar to previous meta-analyses of IBD and IBE (Crispo and Hendry 2005; Jenkins et al. 2010), we found that identifying global predictors of spatial genetic variation is challenging because of the many potential factors that influence the presence and detection of these patterns.

Climatic variables, including different aspects of temperature and precipitation, were among the top ten predictors of genetic variation in RF analyses, suggesting an important role of climate in diversity and population differentiation within salamanders. Mean temperature was the sixth best predictor for IBE and π. The association between higher genetic variation and warmer temperatures is consistent with the evolutionary speed hypothesis (Wright et al., 2010; Gillman and Wright, 2014), where genetic diversity is expected to increase with temperature. Variation in precipitation was also an important predictor for IBD and IBE. A possible explanation for this association is that current stable climates, such as those in tropical environments, promote spatial genetic variation, thus supporting the climate stability hypothesis (Janzen, 1967; Stevens, 1989). Although the relationships between these climatic variables and the response variables we

For Peer Review

examined were relatively weak, previous work has also highlighted the role of climate in explaining variation within salamanders. Specifically, phylogeographic structure measured using species delimitation methods was best explained by variation in climate (Parsons et al., 2024). Like other studies on birds (Smith et al., 2017) and amphibians (Barrow et al., 2021), our analyses suggest that life history and ecological traits are less important than geographic or climatic variation for predicting intraspecific diversity in salamanders.

Phylogenetic, taxonomic, and conservation implications

We did not find significant phylogenetic signal in π, although taxonomic family was the second-best predictor of π in the final RF model, suggesting that some variation can be explained by evolutionary history, but not all. Sample sizes were limited within some families, but diversity estimates in those families were consistently high (e.g., Sirenidae) or low (e.g., Cryptobranchidae), indicating family can be a useful predictor of intraspecific variation. On the other hand, discrepancies in π within genera of family Plethodontidae, such as observed in *Plethodon* or *Eurycea*, emphasize the influence of local ecological and recent demographic processes over shared ancestry for influencing variation within species. Processes leading to homoplasy (e.g., convergence, parallel evolution) dilute phylogenetic signal (Klingenberg and Gidaszewski, 2010). Homoplasy is a common phenomenon in salamander evolution; for example, paedomorphosis has evolved independently several times (Wake, 2009). Likewise, levels of genetic variation may be similar in salamander species that do not share a recent common ancestor (Kozak and Wiens, 2006). Phylogenetic signal in IBD was not statistically significant either, suggesting other factors independent of phylogenetic history influence these patterns. This result was consistent with a previous study on Neotropical amphibians (Amador et al. 2024), which found no phylogenetic signal in IBD or IBE. Significant phylogenetic signal in nucleotide diversity, however, was previously found in amphibian datasets including frogs (Barrow et al. 2021; Amador et al. 2024).

Species with higher π values could contain cryptic diversity and would be useful target species for exploring species boundaries in future studies (e.g., Parsons et al. 2022). Using a smaller dataset than the present study (83 salamander species total, most from the COI mitochondrial

For Peer Review

gene), Parsons et al. (2024) found that ~2/3 of species in the dataset showed strong phylogeographic structure, determined by single locus molecular species delimitation methods. Several species with high values of π in our dataset match species from their study, indicating they have hidden genetic lineages that should be further explored (*Mertensiella caucasica, Bolitoglossa rufescens, Plethodon cinereus, Eurycea cirrigera, Batrachoseps attenuatus, Ommatotriton vittatus, Aneides flavipunctatus*), ideally incorporating nuclear genomic data and thorough geographic sampling.

Salamanders are within one of the most threatened vertebrate groups (Luedtke et al., 2023) and require innovative syntheses of available information to highlight priorities for conservation strategies. Nucleotide diversity could be a useful proxy for conservation status (Petit-Marty, 2021). The relationship we identified between species range size and genetic variation highlights the vulnerability of species with small ranges. Conservation status was an important predictor for π, but sample sizes were limited within several categories. Seventy-five (35%) of the 214 species in our dataset are threatened based on the IUCN Red List; 11 as Critical Endangered (CR), 29 Endangered (EN), and 35 Vulnerable (VU). These salamanders should receive special attention, especially if they occupy small ranges and have low genetic diversity.

Limitations and challenges

One major limitation for macrogenetic studies is the uneven spatial sampling of available sequences, which can lead to an incomplete picture of the distribution of genetic diversity. There are also sampling biases among taxa, such that some species are well characterized while others are omitted from global comparisons. Another limitation is the use of a single mitochondrial gene to calculate genetic diversity and genetic distances. Although mtDNA is widely used in phylogenetic and population genetic studies because it evolves quickly (Galtier et al., 2009), it provides an incomplete picture of intraspecific diversity because it is a single locus, haploid, and solely maternally inherited (Moore, 1995). However, single locus macrogenetic studies are still valuable because the amount and availability of data elucidate patterns associated with historical processes and variability within species. Furthermore, the large genome size of salamanders hampers the use of nuclear genomic data in studies on this scale. These limitations reinforce the

For Peer Review

need for additional geographic sampling of high-quality tissues that include broader taxonomic coverage and facilitate new genomic sequencing efforts.

The most important challenge for our study was the time required to retrieve the geographic coordinates associated with each Cytb sequence. Most sequences in open access databases such as GenBank do not have their latitude and longitude associated with them, which limits studies of spatial genetic variation. Efforts such as phylogatR (Pelletier et al. 2022) aggregate genetic sequences associated with georeferenced occurrences from global databases using automated pipelines, but researchers providing the sequences must include the proper metadata when uploading their data for this approach to work. We encourage scientists to share spatial data and museum catalogue numbers when applicable along with sequences when they are published, ensuring data reproducibility and extendibility.

Conclusions

Our study underscores the value of integrating spatial information into macrogenetic studies. Our results support the hypotheses that variables associated with population size are positively correlated with genetic variation and that salamanders tend to exhibit spatial patterns of genetic variation that also differ across biogeographic realms. Based on our results, we reject the hypothesis that traits associated with reproductive strategies predict mitochondrial genetic variation in salamanders. Our results suggest the importance of range-wide climatic variation, local environmental heterogeneity, and historical demographic events for influencing genetic variation. Future amphibian macrogenetic and genomic studies can incorporate more explicit tests of range expansion and contraction to better understand these processes. Finally, our findings have broad implications for establishing conservation priorities and highlighting regions where sampling has been historically scarce, particularly in biodiversity hotspots that warrant further attention.

For Peer Review

References

Amador, L., Arroyo-Torres, I., & Barrow, L. N. (2024). Machine learning and phylogenetic models identify predictors of genetic variation in Neotropical amphibians. *Journal of Biogeography*, 51(5), 909–923. https://doi.org/10.1111/jbi.14795

AmphibiaWeb. (2024). <https://amphibiaweb.org> University of California, Berkeley, CA, USA. Accessed 3 October 2024.

Antonelli, A. (2022). The rise and fall of Neotropical biodiversity. *Botanical Journal of the Linnean Society*, 199(1), 8–24. https://doi.org/10.1093/botlinnean/boab061

Auteri, G. G., Marchán-Rivadeneira, M. R., Olson, D. H., & Knowles, L. L. (2022). Landscape connectivity among coastal giant salamander (*Dicamptodon tenebrosus*) populations shows no association with land use, fire frequency, or river drainage but exhibits genetic signatures of potential conservation concern. *PLOS ONE*, *17*(6), e0268882. https://doi.org/10.1371/journal.pone.0268882

Bani, L., Pisa, G., Luppi, M., Spilotros, G., Fabbri, E., Randi, E., & Orioli, V. (2015). Ecological connectivity assessment in a strongly structured fire salamander (*Salamandra salamandra*) population. *Ecology and Evolution*, *5*(16), 3472–3485. https://doi.org/10.1002/ece3.1617

Barrow, L. N., Masiero Da Fonseca, E., Thompson, C. E. P., & Carstens, B. C. (2021). Predicting amphibian intraspecific diversity with machine learning: Challenges and prospects for integrating traits, geography, and genetic data. *Molecular Ecology Resources*, 21(8), 2818–2831. https://doi.org/10.1111/1755-0998.13303

Bivand R, Pebesma E, Gómez-Rubio V. (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY. https://asdar-book.org/.

Blanchet, S., Prunier, J. G., & De Kort, H. (2017). Time to Go Bigger: Emerging Patterns in Macrogenetics. *Trends in Genetics*, *33*(9), 579–580. https://doi.org/10.1016/j.tig.2017.06.007

Bonferroni, C. E. (1936) Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8, 3-62, 1936.

Calenge C (2024). adehabitatHR: Home Range Estimation. R package version 0.4.22, <https://CRAN.R-project.org/package=adehabitatHR>.

Blomberg, S. P., Garland, T., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57(4), 717–745.

For Peer Review

Bonato, L., Corbetta, A., Giovine, G., Romanazzi, E., Šunje, E., Vernesi, C., & Crestanello, B. (2018). Diversity among peripheral populations: Genetic and evolutionary differentiation of *Salamandra atra* at the southern edge of the Alps. *Journal of Zoological Systematics and Evolutionary Research*, *56*(4), 533–548. https://doi.org/10.1111/jzs.12224

Cambon, J., Hernangómez, D., Belanger, C., & Possenriede, D. (2021). tidygeocoder: An R package for geocoding. *Journal of Open Source Software*, *6*(65), 3544. https://doi.org/10.21105/joss.03544Corbett-Detig, R. B., Hartl, D. L., & Sackton, T. B. (2015). Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLOS Biology*, *13*(4), e1002112. https://doi.org/10.1371/journal.pbio.1002112

Crispo, E., & Hendry, A. P. (2005). Does time since colonization influence isolation by distance? A meta-analysis. *Conservation Genetics*, *6*(5), 665–682. https://doi.org/10.1007/s10592-005-9026-4

De Kort, H., Prunier, J. G., Ducatez, S., Honnay, O., Baguette, M., Stevens, V. M., & Blanchet, S. (2021). Life history, climate and biogeography interactively affect worldwide genetic diversity of plant and animal populations. *Nature Communications*, *12*(1), 516. https://doi.org/10.1038/s41467-021-20958-2

Denoël, M., Joly, P., & Whiteman, H. H. (2005). Evolutionary ecology of facultative paedomorphosis in newts and salamanders. *Biological Reviews*, *80*(4), 663–671. https://doi.org/10.1017/S1464793105006858

Dunnington, D. (2023). ggspatial: Spatial Data Framework for ggplot2. R package version 1.1.9, <https://CRAN.R-project.org/package=ggspatial>.

Eckert, C. G., Samis, K. E., & Lougheed, S. C. (2008). Genetic variation across species' geographical ranges: The central–marginal hypothesis and beyond. *Molecular Ecology*, *17*(5), 1170–1188. https://doi.org/10.1111/j.1365-294X.2007.03659.x

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113. https://doi.org/10.1186/1471-2105-5-113

Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, *17*(7), 422–433. https://doi.org/10.1038/nrg.2016.58

Escobar, S., Helmstetter, A. J., Jarvie, S., Montúfar, R., Balslev, H., & Couvreur, T. L. P. (2021). Pleistocene climatic fluctuations promoted alternative evolutionary histories in *Phytelephas*

For Peer Review

aequatorialis , an endemic palm from western Ecuador. *Journal of Biogeography*, *48*(5), 1023–1037. https://doi.org/10.1111/jbi.14055

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, *125*(1,), 1–15.

Fick, S.E. and R.J. Hijmans, (2017). WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37 (12): 4302-4315.

Fonseca, E. M., Garda, A. A., Oliveira, E. F., Camurugi, F., Magalhães, F. D. M., Lanna, F. M., Zurano, J. P., Marques, R., Vences, M., & Gehara, M. (2021). The riverine thruway hypothesis: Rivers as a key mediator of gene flow for the aquatic paradoxical frog *Pseudis tocantins* (Anura, Hylidae). *Landscape Ecology*, *36*(10), 3049–3060. https://doi.org/10.1007/s10980-021-01257-z

Fonseca, E. M., Pelletier, T. A., Decker, S. K., Parsons, D. J., & Carstens, B. C. (2023). Pleistocene glaciations caused the latitudinal gradient of within-species genetic diversity. *Evolution Letters*, *7*(5), 331–338. https://doi.org/10.1093/evlett/qrad030

Frankham, R. (1996). Relationship of Genetic Variation to Population Size in Wildlife. *Conservation Biology*, *10*(6), 1500–1508. https://doi.org/10.1046/j.1523-1739.1996.10061500.x

French, C. M., Bertola, L. D., Carnaval, A. C., Economo, E. P., Kass, J. M., Lohman, D. J., Marske, K. A., Meier, R., Overcast, I., Rominger, A. J., Staniczenko, P. P. A., & Hickerson, M. J. (2023). Global determinants of insect mitochondrial genetic diversity. *Nature Communications*, *14*(1), 5276. https://doi.org/10.1038/s41467-023-40936-0

Fu, J., & Wen, L. (2023). Impacts of Quaternary glaciation, geological history and geography on animal species history in continental East Asia: A phylogeographic review. *Molecular Ecology*, *32*(16), 4497–4514. https://doi.org/10.1111/mec.17053

Fusco, N. A., Pehek, E., & Munshi-South, J. (2021). Urbanization reduces gene flow but not genetic diversity of stream salamander populations in the New York City metropolitan area. *Evolutionary Applications*, *14*(1), 99–116. https://doi.org/10.1111/eva.13025

Galtier, N., Nabholz, B., Glémin, S., & Hurst, G. D. D. (2009). Mitochondrial DNA as a marker of molecular diversity: A reappraisal. *Molecular Ecology*, 18(22), 4541–4550. https://doi.org/10.1111/j.1365-294X.2009.04380.x

Gillman, L. N., & Wright, S. D. (2014). Species richness and evolutionary speed: The influence of temperature, water and area. *Journal of Biogeography*, *41*(1), 39–51. https://doi.org/10.1111/jbi.12173

For Peer Review

Hanson, J. O., Rhodes, J. R., Riginos, C., & Fuller, R. A. (2017). Environmental and geographic variables are effective surrogates for genetic variation in conservation planning. *Proceedings of the National Academy of Sciences*, *114*(48), 12755–12760. https://doi.org/10.1073/pnas.1711009114

Hijmans R (2022). geosphere: Spherical Trigonometry. R package version 1.5-18, <https://CRAN.R-project.org/package=geosphere>.

Hijmans RJ, Barbosa M, Ghosh A, Mandel A (2024). geodata: Download Geographic Data. R package version 0.6-2, <https://CRAN.R-project.org/package=geodata>.

Ives, A. R., & Helmus, M. R. (2011). Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, *81*(3), 511–525. https://doi.org/10.1890/10-1264.1

Janzen, D. H. (1967). Why Mountain Passes are Higher in the Tropics. *The American Naturalist*, *101*(919), 233–249. https://doi.org/10.1086/282487

Jenkins, D. G., Carey, M., Czerniewska, J., Fletcher, J., Hether, T., Jones, A., Knight, S., Knox, J., Long, T., Mannino, M., McGuire, M., Riffle, A., Segelsky, S., Shappell, L., Sterner, A., Strickler, T., & Tursi, R. (2010). A meta-analysis of isolation by distance: Relic or reference standard for landscape genetics? *Ecography*, *33*(2), 315–320. https://doi.org/10.1111/j.1600-0587.2010.06285.x

Jetz, W., & Pyron, R. A. (2018). The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. *Nature Ecology & Evolution*, *2*(5), 850–858. https://doi.org/10.1038/s41559-018-0515-5

Kardos, M., Armstrong, E. E., Fitzpatrick, S. W., Hauser, S., Hedrick, P. W., Miller, J. M., Tallmon, D. A., & Funk, W. C. (2021). The crucial role of genome-wide genetic variation in conservation. *Proceedings of the National Academy of Sciences*, *118*(48), e2104642118. https://doi.org/10.1073/pnas.2104642118

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129), 624-626. https://doi.org/10.1038/217624a0

Klingenberg, C. P., & Gidaszewski, N. A. (2010). Testing and Quantifying Phylogenetic Signals and Homoplasy in Morphometric Data. *Systematic Biology*, 59(3), 245–261. https://doi.org/10.1093/sysbio/syp106

For Peer Review

Kozak, K. H., & Wiens, J. J. (2006). Does niche conservatism promote speciation? A case study in North American salamanders. *Evolution*, *60*(12), 2604–2621. https://doi.org/10.1111/j.0014-3820.2006.tb01893.x

Lacy, R. C. (1987). Loss of Genetic Diversity from Managed Populations: Interacting Effects of Drift, Mutation, Immigration, Selection, and Population Subdivision. *Conservation Biology*, *1*(2), 143–158.

Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Lawrence, E. R., Pedersen, E. J., & Fraser, D. J. (2023). Macrogenetics reveals multifaceted influences of environmental variation on vertebrate population genetic diversity across the Americas. *Molecular Ecology*, *32*(16), 4557–4569. https://doi.org/10.1111/mec.17059

Lee-Yaw, J. A., & Irwin, D. E. (2012). Large geographic range size reflects a patchwork of divergent lineages in the long-toed salamander (*Ambystoma macrodactylum*). *Journal of Evolutionary Biology*, *25*(11), 2276–2287. https://doi.org/10.1111/j.1420-9101.2012.02604.x

Leigh, D. M., Van Rees, C. B., Millette, K. L., Breed, M. F., Schmidt, C., Bertola, L. D., Hand, B. K., Hunter, M. E., Jensen, E. L., Kershaw, F., Liggins, L., Luikart, G., Manel, S., Mergeay, J., Miller, J. M., Segelbacher, G., Hoban, S., & Paz-Vinas, I. (2021). Opportunities and challenges of macrogenetic studies. *Nature Reviews Genetics*, *22*(12), 791–807. https://doi.org/10.1038/s41576-021-00394-0

Li, D., Dinnage, R., Nell, L. A., Helmus, M. R., & Ives, A. R. (2020). phyr: An r package for phylogenetic species-distribution modelling in ecological communities. Methods in Ecology and Evolution, 11(11), 1455–1463. https://doi.org/10.1111/2041-210X.13471

Liaw, A., Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3), 18-22. <https://CRAN.R-project.org/doc/Rnews/>.

López-Delgado, J., & Meirmans, P. G. (2022). History or demography? Determining the drivers of genetic variation in North American plants. *Molecular Ecology*, 31(7), 1951–1962. https://doi.org/10.1111/mec.16230

Lowe, W. H., Kovach, R. P., & Allendorf, F. W. (2017). Population Genetics and Demography Unite Ecology and Evolution. *Trends in Ecology & Evolution*, *32*(2), 141–152. https://doi.org/10.1016/j.tree.2016.12.002

For Peer Review

Luedtke, J. A., Chanson, J., Neam, K., Hobin, L., Maciel, A. O., Catenazzi, A., Borzée, A., Hamidy, A., Aowphol, A., Jean, A., Sosa-Bartuano, Á., Fong G., A., De Silva, A., Fouquet, A., Angulo, A., Kidov, A. A., Muñoz Saravia, A., Diesmos, A. C., Tominaga, A., … Stuart, S. N. (2023). Ongoing declines for the world's amphibians in the face of emerging threats. *Nature*, *622*(7982), 308–314. https://doi.org/10.1038/s41586-023-06578-4

Malcher, G., Amorim, A.L., Ferreira, P. *et al.* (2023). First evaluation of the population genetics and aspects of the evolutionary history of the Amazonian snook, *Centropomus irae*, and its association with the Amazon plume. *Hydrobiologia* 850, 2115–2125. https://doi.org/10.1007/s10750-023-05223-5

Mayr, E. (2001). *What Evolution Is*. New York, NY: Basic Books.

Meng, H., Li, X., & Qiao, P. (2014). Population Structure, Historical Biogeography and Demographic History of the Alpine Toad *Scutiger ningshanensis* in the Tsinling Mountains of Central China. *PLoS ONE*, *9*(6), e100729. https://doi.org/10.1371/journal.pone.0100729

Millette, K. L., Fugère, V., Debyser, C., Greiner, A., Chain, F. J. J., & Gonzalez, A. (2021). Refining analyses of existing data sets is valuable for macrogenetics: A response to Paz-Vinas, Jensen et al., (2021). *Ecology Letters*, *24*(6), 1285–1286. https://doi.org/10.1111/ele.13733

Liedtke, H.C., Wiens, J.J. & Gomez-Mestre, I. The evolution of reproductive modes and life cycles in amphibians. Nat Commun 13, 7039 (2022). https://doi.org/10.1038/s41467-022-34474-4

Miraldo, A., Li, S., Borregaard, M. K., Flórez-Rodríguez, A., Gopalakrishnan, S., Rizvanovic, M., Wang, Z., Rahbek, C., Marske, K. A., & Nogués-Bravo, D. (2016). An Anthropocene map of genetic diversity. *Science*, 353(6307), 1532–1535. https://doi.org/10.1126/science.aaf4381

Mittermeier, R. A., Turner, W. R., Larsen, F. W., Brooks, T. M., & Gascon, C. (2011). Global Biodiversity Conservation: The Critical Role of Hotspots. In F. E. Zachos & J. C. Habel (Eds.), *Biodiversity Hotspots* (pp. 3–22). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20992-5_1

Moore, W. S. (1995). Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. Evolution, 49(4), 718-726.

Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P, Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux B,

For Peer Review

Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette M, Ribeiro Cunha E, Smith T, Stier A, Ter Braak C, Weedon J (2024). vegan: Community Ecology Package. R package version 2.6-8, <https://CRAN.R-project.org/package=vegan>

Oliveira, B. F., São-Pedro, V. A., Santos-Barrera, G., Penone, C., & Costa, G. C. (2017). AmphiBIO, a global database for amphibian ecological traits. *Scientific Data*, 4(1), 170123. https://doi.org/10.1038/sdata.2017.123

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*(6756), 877–884. https://doi.org/10.1038/44766

Paradis, E. (2010). pegas: An R package for population genetics with an integrated–modular approach. *Bioinformatics*, *26*(3), 419–420. https://doi.org/10.1093/bioinformatics/btp696

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Parsons, D. J., Pelletier, T. A., Wieringa, J. G., Duckett, D. J., & Carstens, B. C. (2022). Analysis of biodiversity data suggests that mammal species are hidden in predictable places. *Proceedings of the National Academy of Sciences*, *119*(14), e2103400119. https://doi.org/10.1073/pnas.2103400119

Parsons, D. J., Green, A. E., Carstens, B. C., & Pelletier, T. A. (2024). Predicting genetic biodiversity in salamanders using geographic, climatic, and life history traits. *PLOS ONE*, *19*(10), e0310932. https://doi.org/10.1371/journal.pone.0310932

Park, K. Y., Lucas, M., Chaulk, A., Matter, S. F., Roland, J., & Keyghobadi, N. (2024). Immigration allows population persistence and maintains genetic diversity despite an attempted experimental extinction. *Royal Society Open Science*, *11*(7), 240557. https://doi.org/10.1098/rsos.240557

Pauls, S. U., Nowak, C., Bálint, M., & Pfenninger, M. (2013). The impact of global climate change on genetic diversity within populations and species. *Molecular Ecology*, *22*(4), 925–946. https://doi.org/10.1111/mec.12152

Paz, A., Ibáñez, R., Lips, K.R. and Crawford, A.J. (2015), Testing the role of ecology and life history in structuring genetic variation across a landscape: a trait-based phylogeographic approach. Mol Ecol, 24: 3723-3737. https://doi-org.libproxy.unm.edu/10.1111/mec.13275

For Peer Review

Paz-Vinas, I., Jensen, E. L., Bertola, L. D., Breed, M. F., Hand, B. K., Hunter, M. E., Kershaw, F., Leigh, D. M., Luikart, G., Mergeay, J., Miller, J. M., Van Rees, C. B., Segelbacher, G., & Hoban, S. (2021). Macrogenetic studies must not ignore limitations of genetic markers and scale. *Ecology Letters*, *24*(6), 1282–1284. https://doi.org/10.1111/ele.13732

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

Pebesma, E., & Bivand, R. (2023). Spatial Data Science: With Applications in R. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

Pelletier, T. A., & Carstens, B. C. (2018). Geographical range size and latitude predict population genetic structure in a global survey. *Biology Letters*, 14(1), 20170566. https://doi.org/10.1098/rsbl.2017.0566

Pelletier, T. A., Parsons, D. J., Decker, S. K., Crouch, S., Franz, E., Ohrstrom, J., & Carstens, B. C. (2022). phylogatR: Phylogeographic data aggregation and repurposing. *Molecular Ecology Resources*, 22(8), 2830–2842. https://doi.org/10.1111/1755-0998.13673

Petit-Marty, N., Vázquez-Luis, M., & Hendriks, I. E. (2021). Use of the nucleotide diversity in COI mitochondrial gene as an early diagnostic of conservation status of animal species. *Conservation Letters*, *14*(1), e12756. https://doi.org/10.1111/conl.12756

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Ramírez-Barahona, S., & Eguiarte, L. E. (2013). The role of glacial cycles in promoting genetic diversity in the Neotropics: The case of cloud forests during the Last Glacial Maximum. *Ecology and Evolution*, *3*(3), 725–738. https://doi.org/10.1002/ece3.483

Reed, D. H., & Frankham, R. (2003). Correlation between Fitness and Genetic Diversity. *Conservation Biology*, *17*(1), 230–237. https://doi.org/10.1046/j.1523-1739.2003.01236.x

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things): *phytools: R package*. *Methods in Ecology and Evolution*, *3*(2), 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Riberon, A., Sotiriou, E., Miaud, C., Andreone, F., & Taberlet, P. (2002). Lack of genetic diversity in *Salamandra lanzai* revealed by Cytochrome b gene sequences. *Copeia*, *2002*(1), 229–232.

For Peer Review

Rios, N. E. & Bart, H. L. (2010). GEOLocate (Version 3.22). [Computer software]. Belle Chasse, LA: Tulane University Museum of Natural History.

Román-Palacios, C. (2023). The phruta r package: Increasing access, reproducibility and transparency in phylogenetic analyses. *Methods in Ecology and Evolution*, 14(9), 2284–2299. https://doi.org/10.1111/2041-210X.14147

Rovito, S. M. (2017). The Geography of Speciation in Neotropical Salamanders. *Herpetologica*, *73*(3), 229–241. https://doi.org/10.1655/HERPETOLOGICA-D-16-00077.1

Rovito, S. M., & Schoville, S. D. (2017). Testing models of refugial isolation, colonization and population connectivity in two species of montane salamanders. *Heredity*, *119*(4), 265–274. https://doi.org/10.1038/hdy.2017.31

RS-eco (2023). rasterSp: R Package to rasterize and summarise IUCN range maps. R package version 0.0.1, <https://github.com/RS-eco/rasterSp>.

Schierenbeck, K. A. (2017). Population-level genetic variation and climate change in a biodiversity hotspot. *Annals of Botany*, *119*(2), 215–228. https://doi.org/10.1093/aob/mcw214

Segovia-Ramírez, M. G., Ramírez-Sánchez, O., Decena Segarra, L. P., Rios-Carlos, H., & Rovito, S. M. (2023). Determinants of genetic diversity in Neotropical salamanders (Plethodontidae: Bolitoglossini). *Ecology and Evolution*, *13*(11), e10707. https://doi.org/10.1002/ece3.10707

Sexton, J. P., Hangartner, S. B., & Hoffmann, A. A. (2014). Genetic isolation by environment or distance: which pattern of gene flow is most common? Special Section. *Evolution*, *68*(1), 1–15. https://doi.org/10.1111/evo.12258

Shafer, A. B. A., Cullingham, C. I., Côté, S. D., & Coltman, D. W. (2010). Of glaciers and refugia: A decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, *19*(21), 4589–4621. https://doi.org/10.1111/j.1365-294X.2010.04828.x

Smith, B. T., Seeholzer, G. F., Harvey, M. G., Cuervo, A. M., & Brumfield, R. T. (2017). A latitudinal phylogeographic diversity gradient in birds. *PLOS Biology*, *15*(4), e2001073. https://doi.org/10.1371/journal.pbio.2001073

Stevens, G. C. (1989). The Latitudinal Gradient in Geographical Range: How so Many Species Coexist in the Tropics. *The American Naturalist*, *133*(2), 240–256. https://doi.org/10.1086/284913

For Peer Review

Tajima, F. (1989). Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595.

Wake, D. B. (2009). What Salamanders Have Taught Us About Evolution. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 333–352. https://doi.org/10.1146/annurev.ecolsys.39.110707.173552Wang, I. J. (2013). Examining the full effects of landscape heterogeneity on spatial genetic variation: a multiple matrix regression approach for quantifying geographic and ecological isolation: special section. *Evolution*, *67*(12), 3403–3411. https://doi.org/10.1111/evo.12134

Wang, I. J., & Bradburd, G. S. (2014). Isolation by environment. *Molecular Ecology*, *23*(23), 5649–5662. https://doi.org/10.1111/mec.12938

Waples, R. S. (2022). What is N_e, anyway? *Journal of Heredity*, 113(4), 371–379. https://doi.org/10.1093/jhered/esac023

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wielstra, B., Crnobrnja-Isailović, J., Litvinchuk, S. N., Reijnen, B. T., Skidmore, A. K., Sotiropoulos, K., Toxopeus, A. G., Tzankov, N., Vukov, T., & Arntzen, J. W. (2013). Tracing glacial refugia of *Triturus* newts based on mitochondrial DNA phylogeography and species distribution modeling. *Frontiers in Zoology*, *10*(1), 13. https://doi.org/10.1186/1742-9994-10-13

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*(2), 97–159. https://doi.org/10.1093/genetics/16.2.97

Wright, S. (1943). Isolation by distance. *Genetics*, *28*(2), 114–138. https://doi.org/10.1093/genetics/28.2.114

Wright, S. D., Gillman, L. N., Ross, H. A., & Keeling, D. J. (2010). Energy and the tempo of evolution in amphibians. *Global Ecology and Biogeography*, *19*(5), 733–740. https://doi.org/10.1111/j.1466-8238.2010.00549.x

Yang, Z. (2014). Molecular evolution: a statistical approach. Oxford University Press, Oxford.

Zamudio, K. R.. & Wieczorek, A. M. (2007), Fine-scale spatial genetic structure and dispersal among spotted salamander (*Ambystoma maculatum*) breeding populations. Molecular Ecology, 16: 257-274. https://doi-org.libproxy.unm.edu/10.1111/j.1365-294X.2006.03139.x

For Peer Review

Figures

Figure 1. Maps of A) mean nucleotide diversity (π) and B) mean number of sequences with coordinates (log scale) of global salamanders. Biogeographic realms are indicated with shading: Nearctic (brown), Neotropic (green), Oriental (blue), Palearctic (purple). Boxplots of nucleotide diversity for each biogeographic realm are shown in map A. The number of species and mean number of sequences with coordinates for each biogeographic realm are shown in map B. For visualization purposes, we show grid cell = 4° (approximately 440 km^2) for mapping.

Figure 2. Phylogeny of Caudata species included in this study, showing the values of nucleotide diversity (π) as blue barplots to the right and the presence (gray) or absence (yellow) of isolation-by-distance (IBD) and isolation-by-environment (IBE) per species. Relationships are based on Jetz et al. (2018). *Plethodon cinereus* silhouette was taken from phylopic.org.

Figure 3. Random Forest (RF) results for A) nucleotide diversity (π, using RF regression), B) isolation-by-distance (IBD, using RF classification), and C) isolation-by-environment (IBE, using RF classification). All predictors except number of base pairs, and total range size were used in these RF models, but only the top 10 best predictors are shown here. Range size* = sampling range size based on localities with associated DNA sequences; # of sequences$^+$ = number of sequences associated with geographic coordinates; SVL = snout-vent length; mm = millimeters; BIO4 = Temperature Seasonality (standard deviation ×100); BIO12 = Annual Precipitation.

Figure 4. A) Comparison of sampling range size (in log scale) with nucleotide diversity (π) for each species. B) Comparison of phylogenetic independent contrast (PIC) results between the sampling range size (log scale) and nucleotide diversity (π).

Figure 5. Boxplots comparing the sampling range size (A, C) and the number of sequences (B, D) between species with (Yes, gray color) and without (No, yellow color) isolation-by-distance (IBD) and isolation by environment (IBE). Each panel includes the result of a chi-squared test. Black dots represent outliers.

For Peer Review

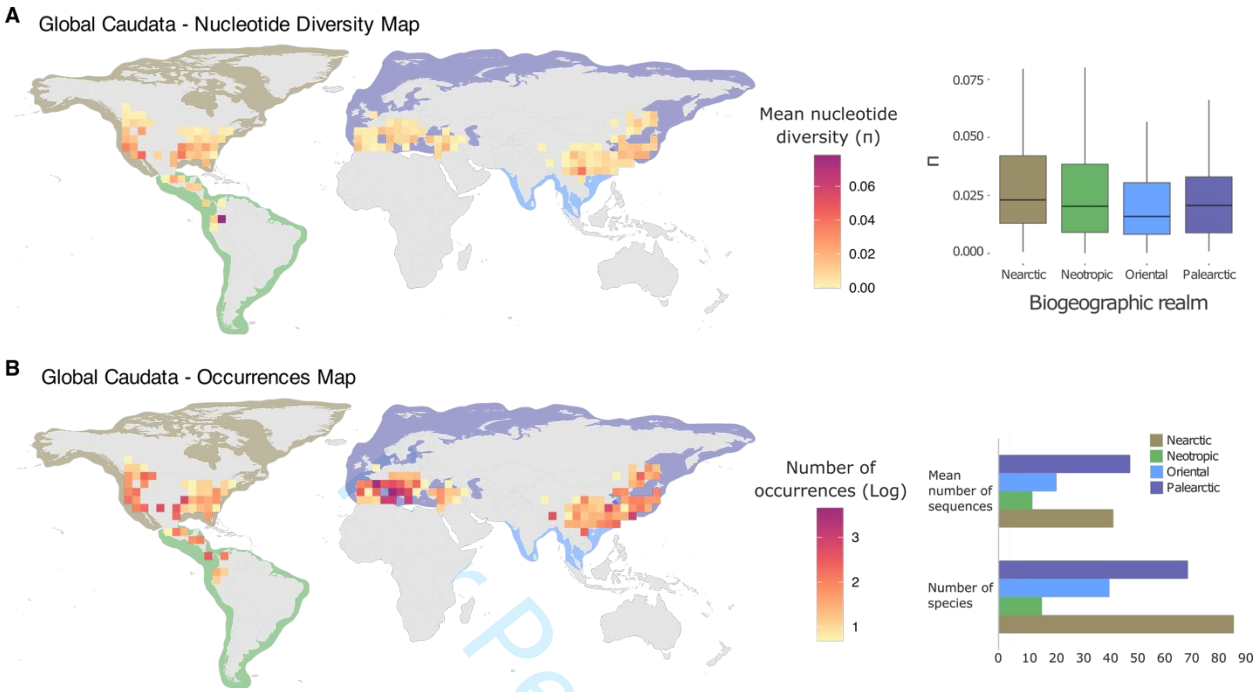


Figure 1. Maps of A) mean nucleotide diversity (π) and B) mean number of sequences with coordinates (log scale) of global salamanders. Biogeographic realms are indicated with shading: Nearctic (brown), Neotropic (green), Oriental (blue), and Palearctic (purple). Boxplots of nucleotide diversity for each biogeographic realm are shown in map A. The number of species and mean number of sequences with coordinates for each biogeographic realm are shown in map B. For visualization purposes, we show grid cell = 4° (approximately 440 km^2) for mapping.

For Peer Review

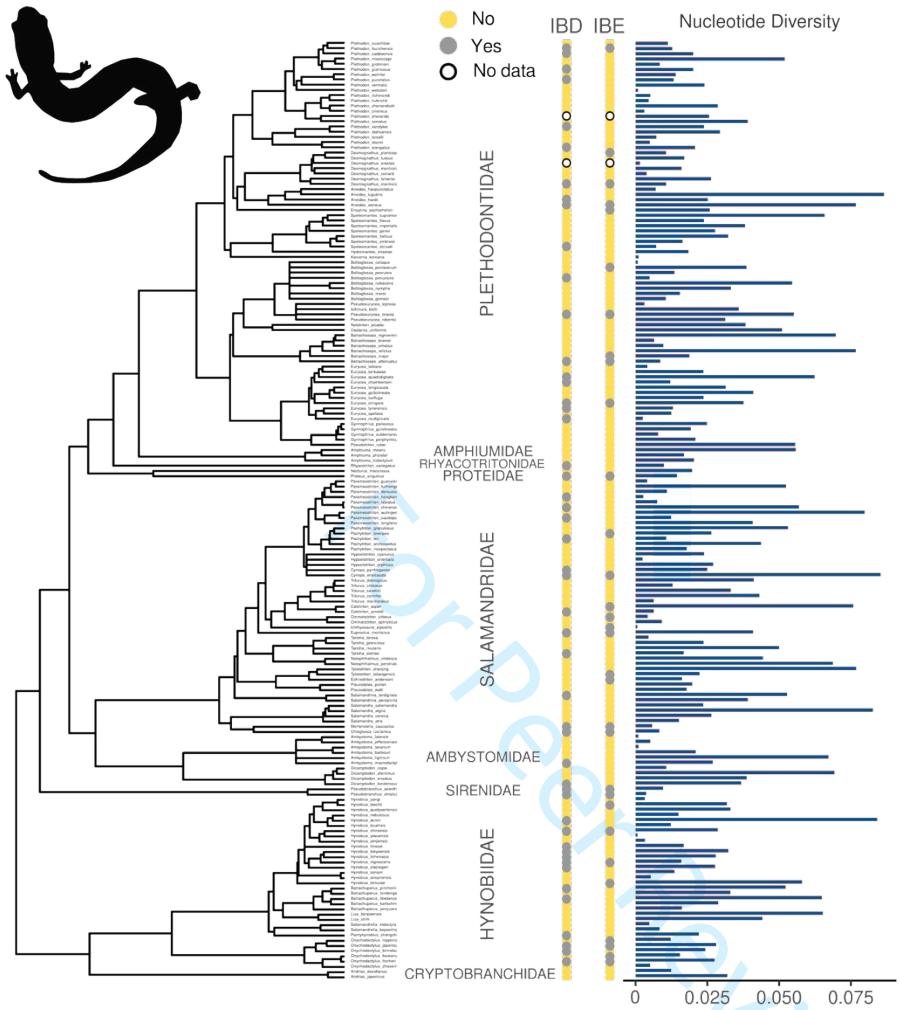


Figure 2. Phylogeny of Caudata species included in this study, showing the values of nucleotide diversity (π) as blue barplots to the right and the presence (gray) or absence (yellow) of isolation-by-distance (IBD) and isolation-by-environment (IBE) per species. Relationships are based on Jetz et al. (2018). *Plethodon cinereus* silhouette was taken from phylopic.org.

For Peer Review

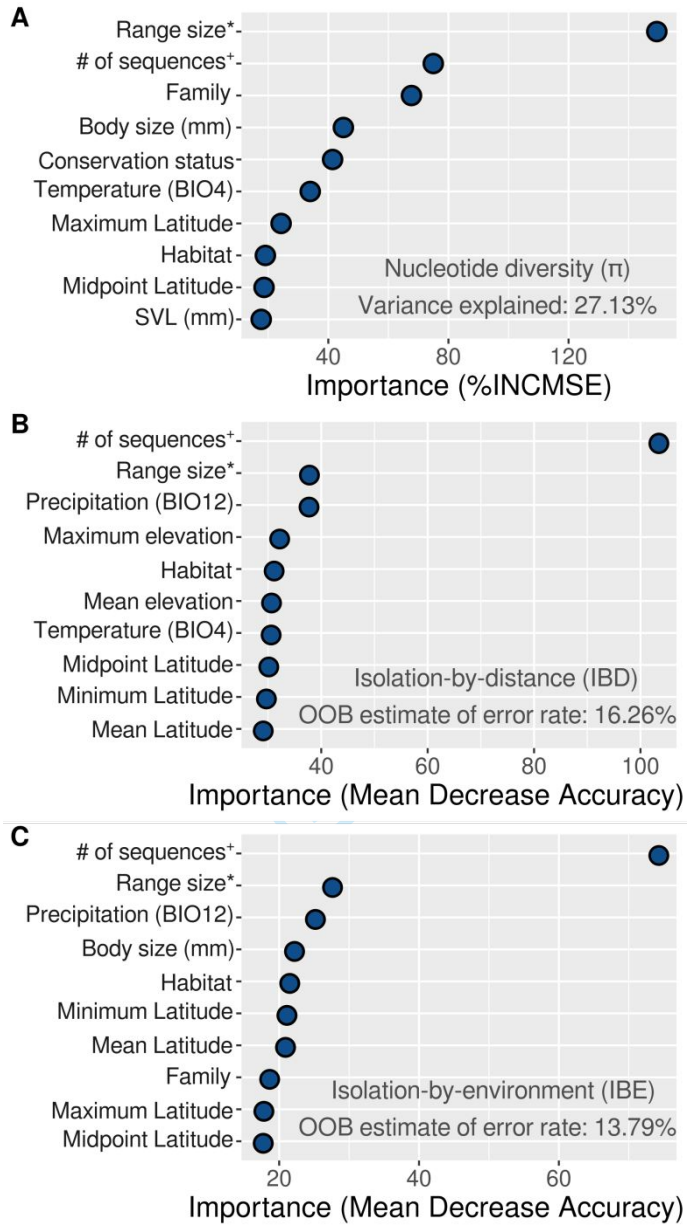


Figure 3. Random Forest (RF) results for A) nucleotide diversity (π, using RF regression), B) isolation-by-distance (IBD, using RF classification), and C) isolation-by-environment (IBE, using RF classification). All predictors except number of base pairs, and total range size were used in these RF models, but only the top 10 best predictors are shown here. Range size* = sampling range size based on localities with associated DNA sequences; # of sequences^{+} = number of sequences associated with geographic coordinates; SVL = snout-vent length; mm = millimeters; BIO4 = Temperature Seasonality (standard deviation ×100); BIO12 = Annual Precipitation.

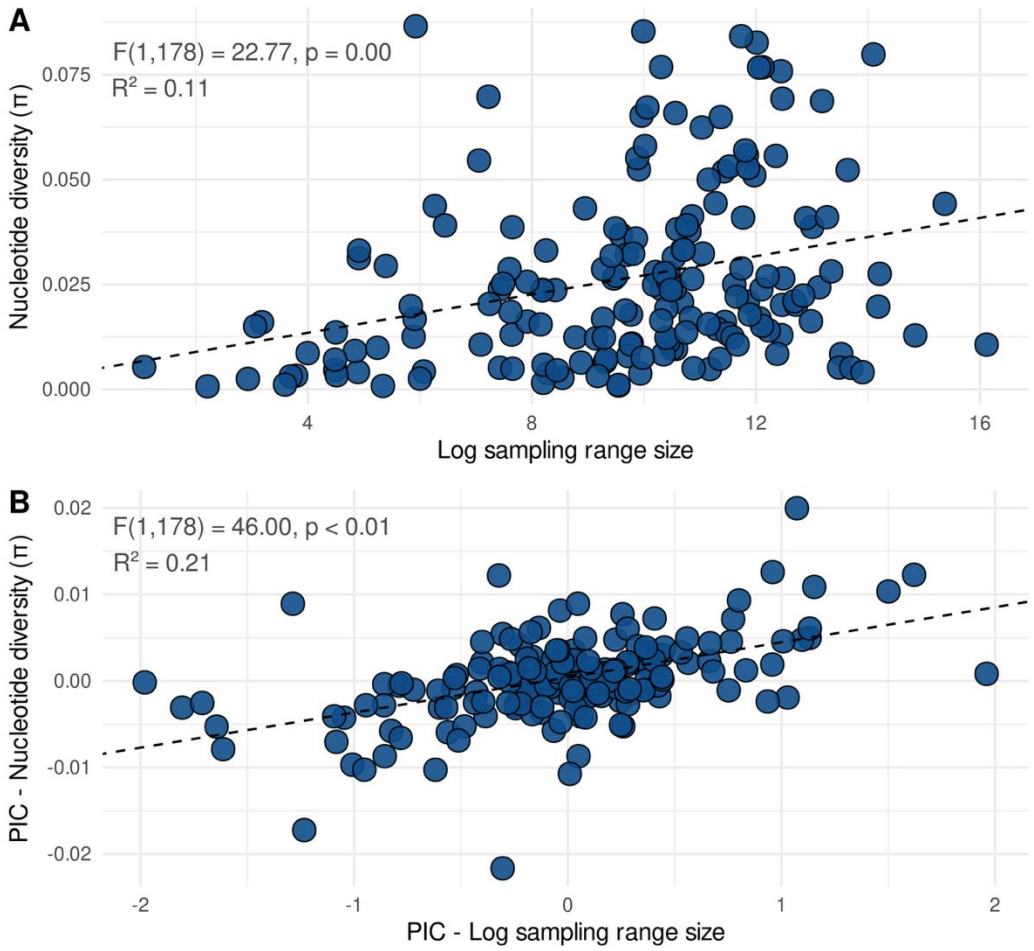


Figure 4. A) Comparison of sampling range size (in log scale) with nucleotide diversity (π) for each species. B) Comparison of phylogenetic independent contrast (PIC) results between the sampling range size (log scale) and nucleotide diversity (π).

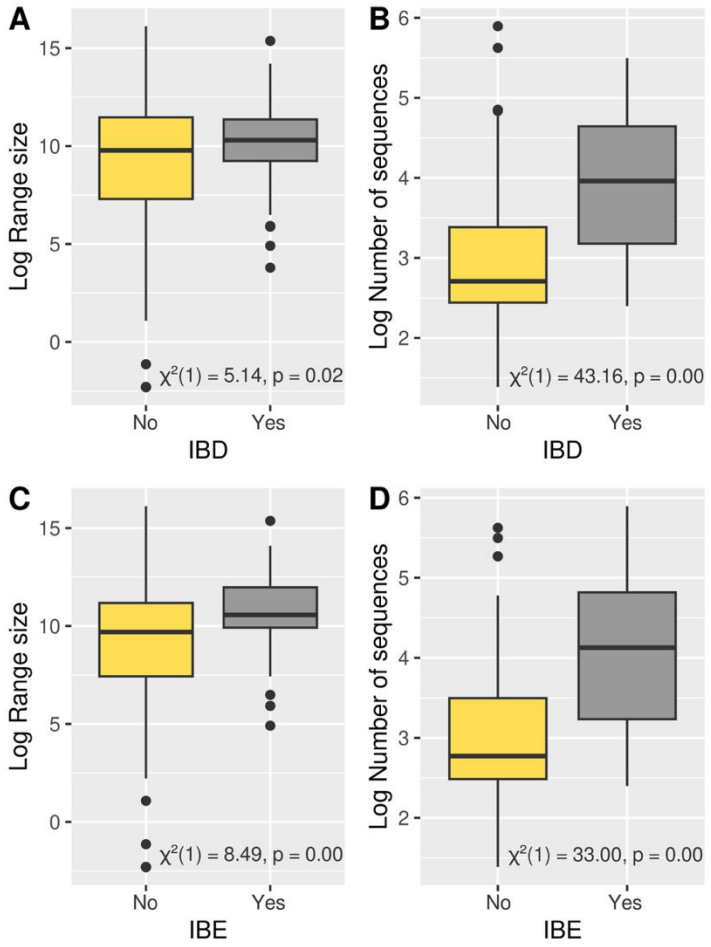


Figure 5. Boxplots comparing the sampling range size (A, C) and the number of sequences (B, D) between species with (Yes, gray color) and without (No, yellow color) isolation-by-distance (IBD) and isolation by environment (IBE). Each panel includes the result of a chi-squared test. Black dots represent outliers.

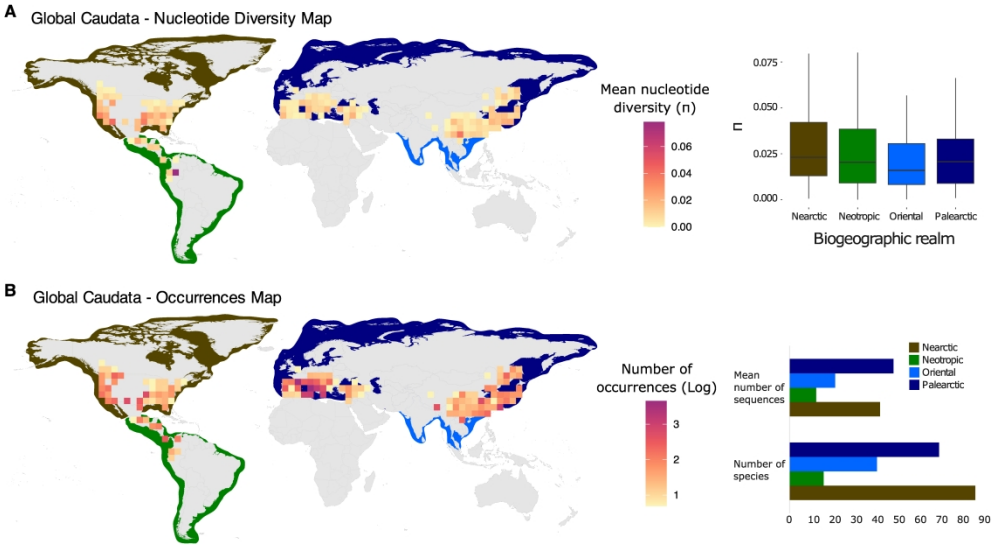
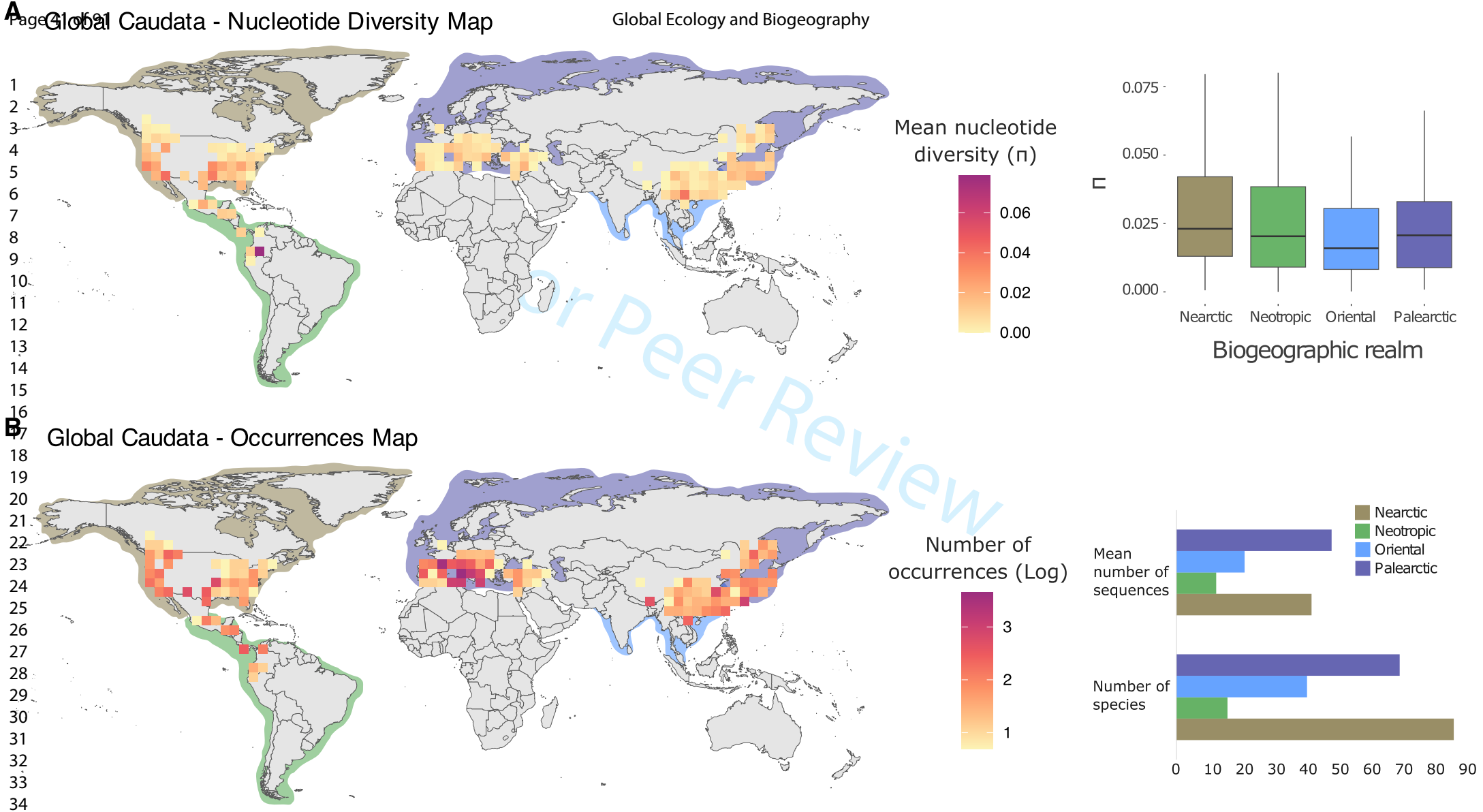


Figure 1. Maps of A) mean nucleotide diversity (π) and B) mean number of sequences with coordinates (log scale) of global salamanders. Biogeographic realms are indicated with shading: Nearctic (brown), Neotropic (green), Oriental (blue), and Palearctic (purple). Boxplots of nucleotide diversity for each biogeographic realm are shown in map A. The number of species and mean number of sequences with coordinates for each biogeographic realm are shown in map B. For visualization purposes, we show grid cell = 4° (approximately 440 km2) for mapping.

491x271mm (354 x 354 DPI)



For Peer Review

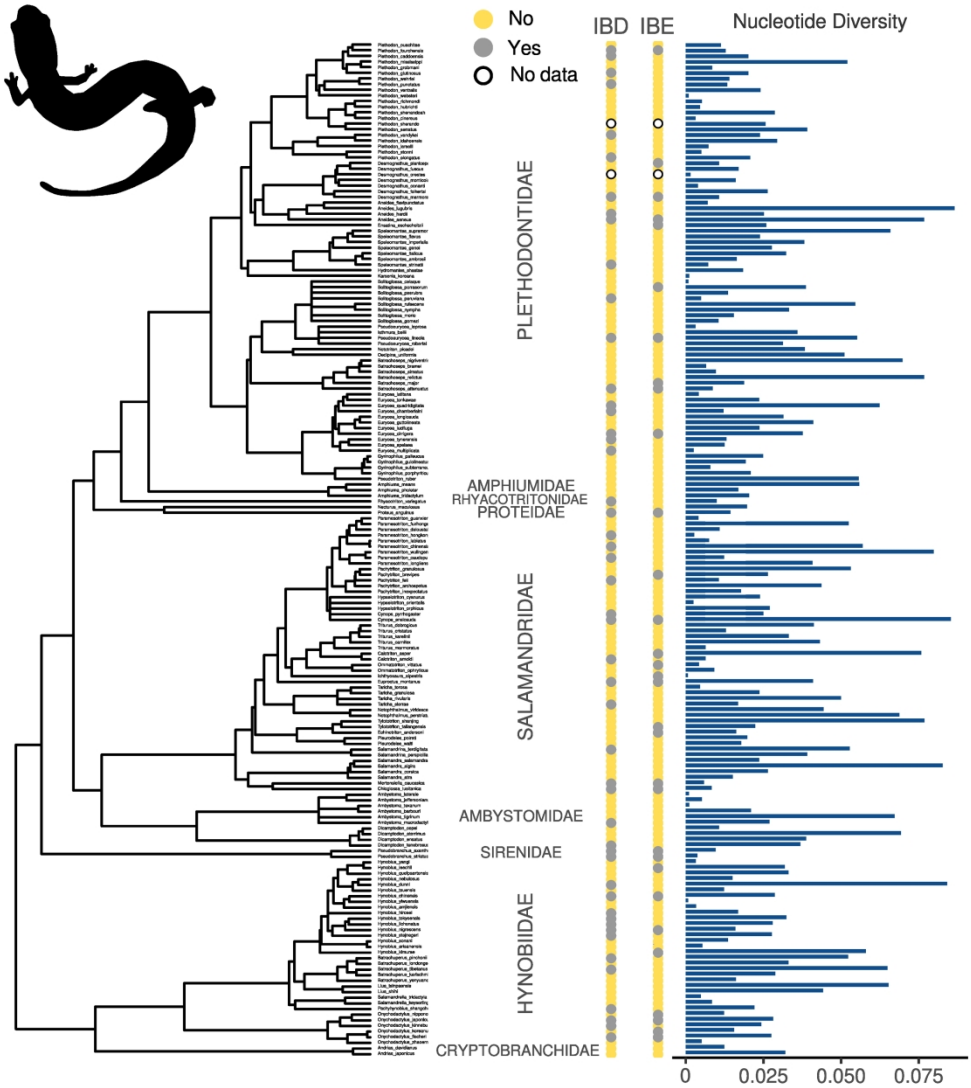
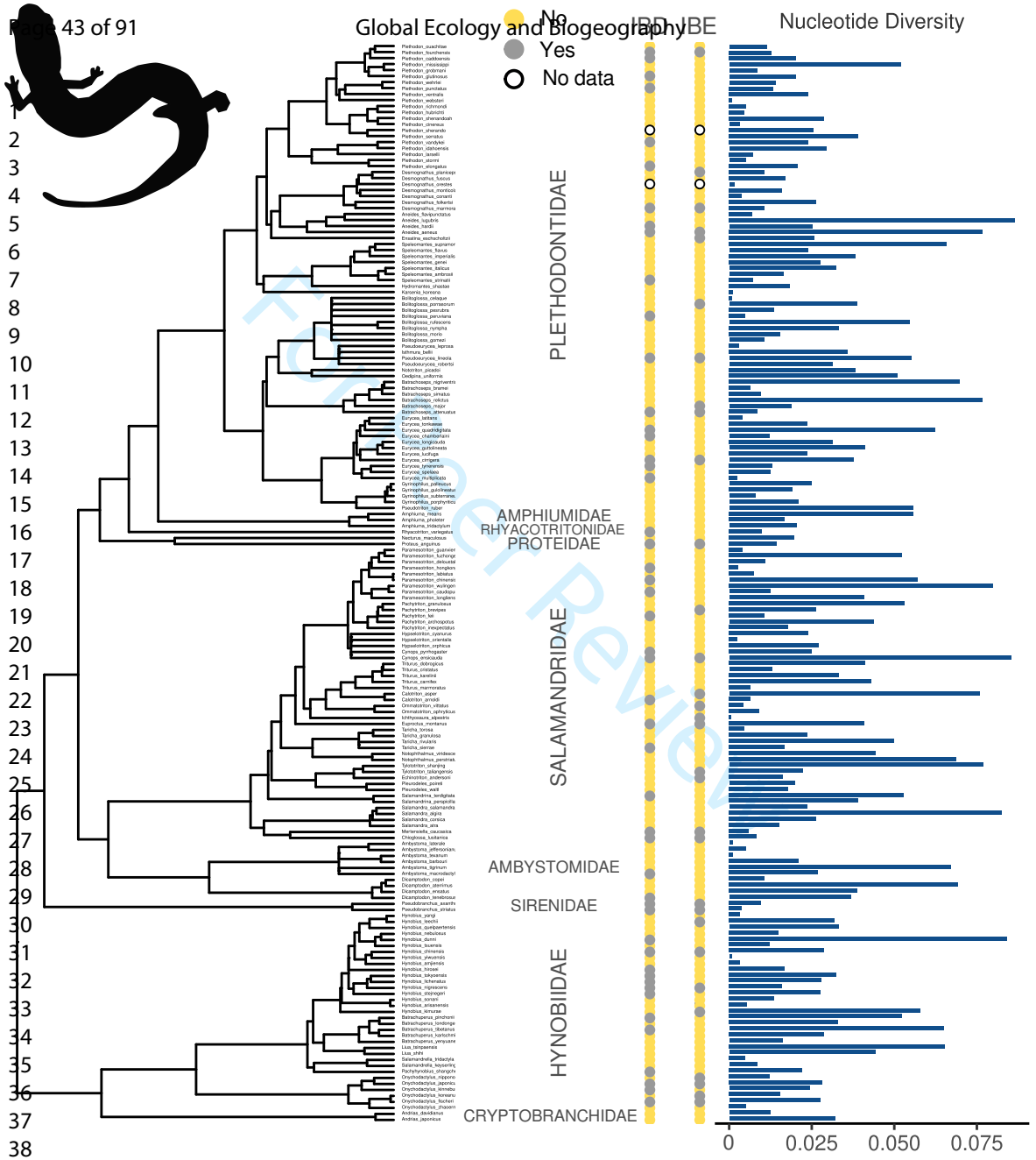


Figure 2. Phylogeny of Caudata species included in this study, showing the values of nucleotide diversity (π) as blue barplots to the right and the presence (gray) or absence (yellow) of isolation-by-distance (IBD) and isolation-by-environment (IBE) per species. Relationships are based on Jetz et al. (2018). Plethodon cinereus silhouette was taken from phylopic.org.

271x299mm (300 x 300 DPI)



For Peer Review

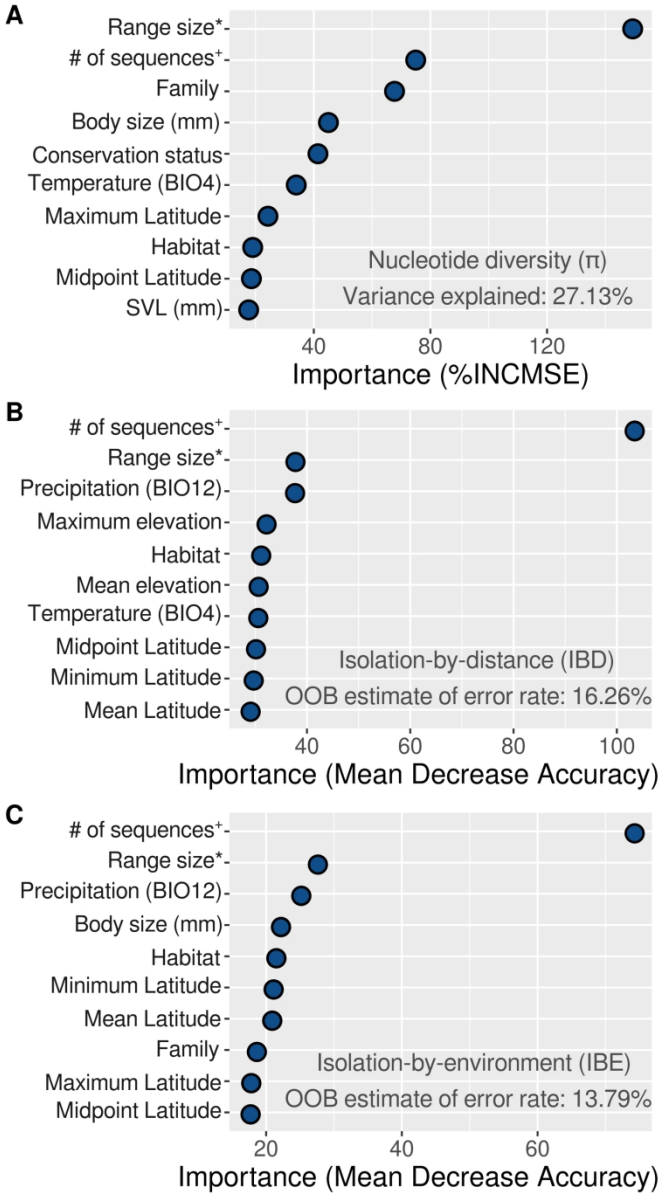
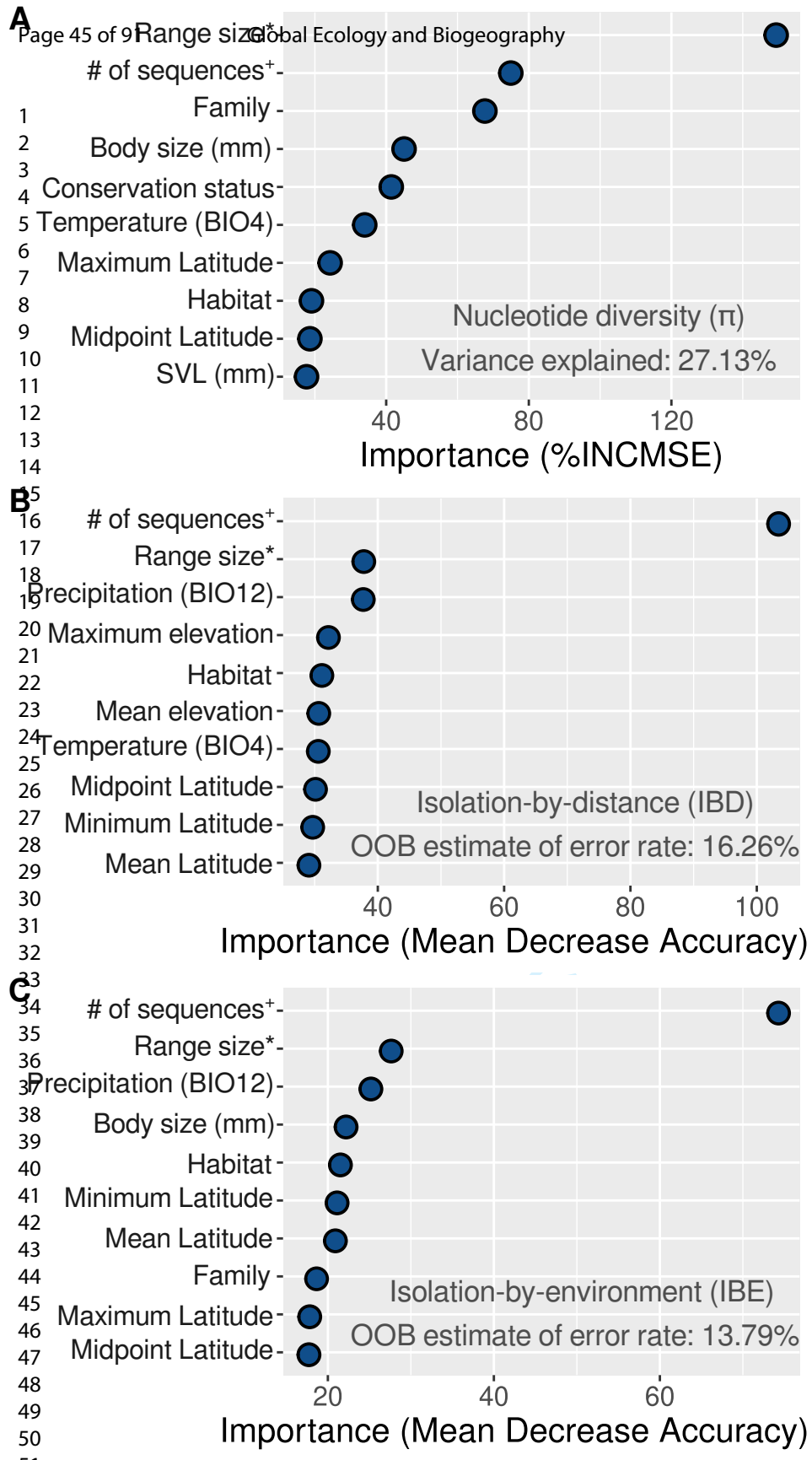


Figure 3. Random Forest (RF) results for A) nucleotide diversity (π, using RF regression), B) isolation-by-distance (IBD, using RF classification), and C) isolation-by-environment (IBE, using RF classification). All predictors except number of base pairs, and total range size were used in these RF models, but only the top 10 best predictors are shown here. Range size* = sampling range size based on localities with associated DNA sequences; # of sequences+ = number of sequences associated with geographic coordinates; SVL = snout-vent length; mm = millimeters; BIO4 = Temperature Seasonality (standard deviation ×100); BIO12 = Annual Precipitation.

143x258mm (300 x 300 DPI)


A
Range size*
# of sequences+
Family
Body size (mm)
Conservation status
Temperature (BIO4)
Maximum Latitude
Habitat
Midpoint Latitude
SVL (mm)
Nucleotide diversity (π)
Variance explained: 27.13%
40
80
120
Importance (%INCMSE)
B
# of sequences+
Range size*
Precipitation (BIO12)
Maximum elevation
Habitat
Mean elevation
Temperature (BIO4)
Midpoint Latitude
Minimum Latitude
Mean Latitude
Isolation-by-distance (IBD)
OOB estimate of error rate: 16.26%
40
60
80
100
Importance (Mean Decrease Accuracy)
C
# of sequences+
Range size*
Precipitation (BIO12)
Body size (mm)
Habitat
Minimum Latitude
Mean Latitude
Family
Maximum Latitude
Midpoint Latitude
Isolation-by-environment (IBE)
OOB estimate of error rate: 13.79%
20
40
60
Importance (Mean Decrease Accuracy)

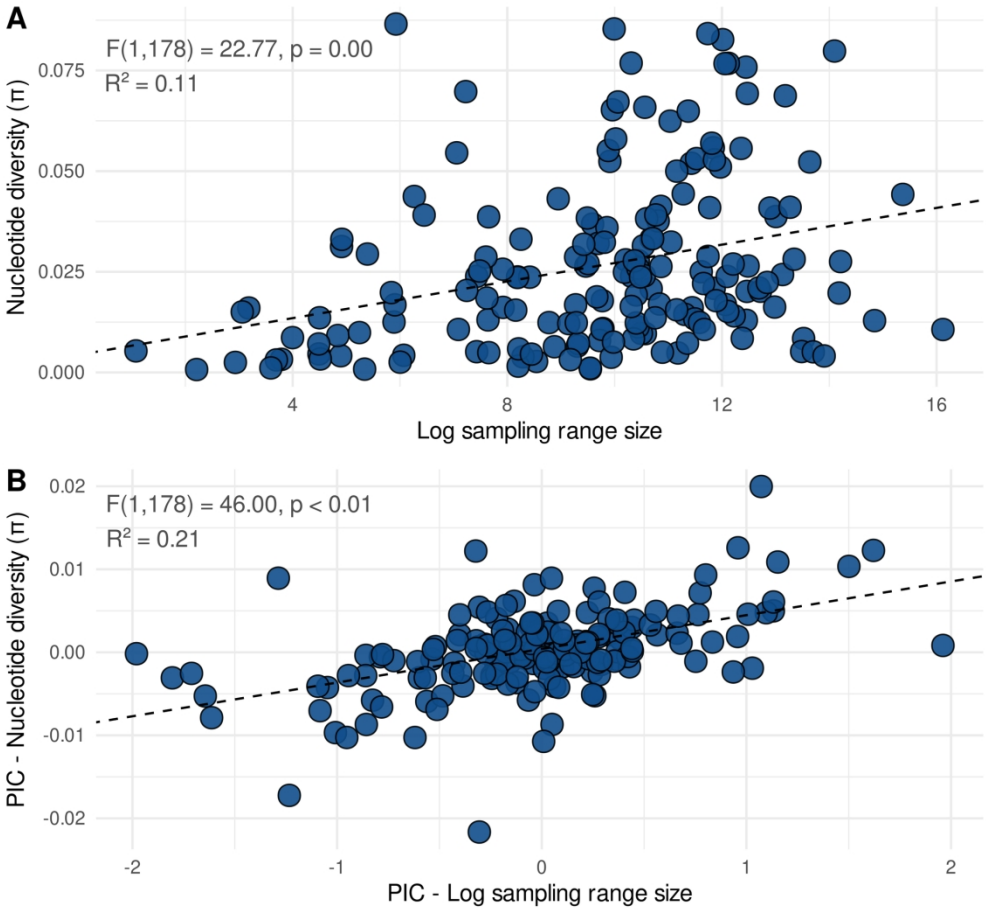
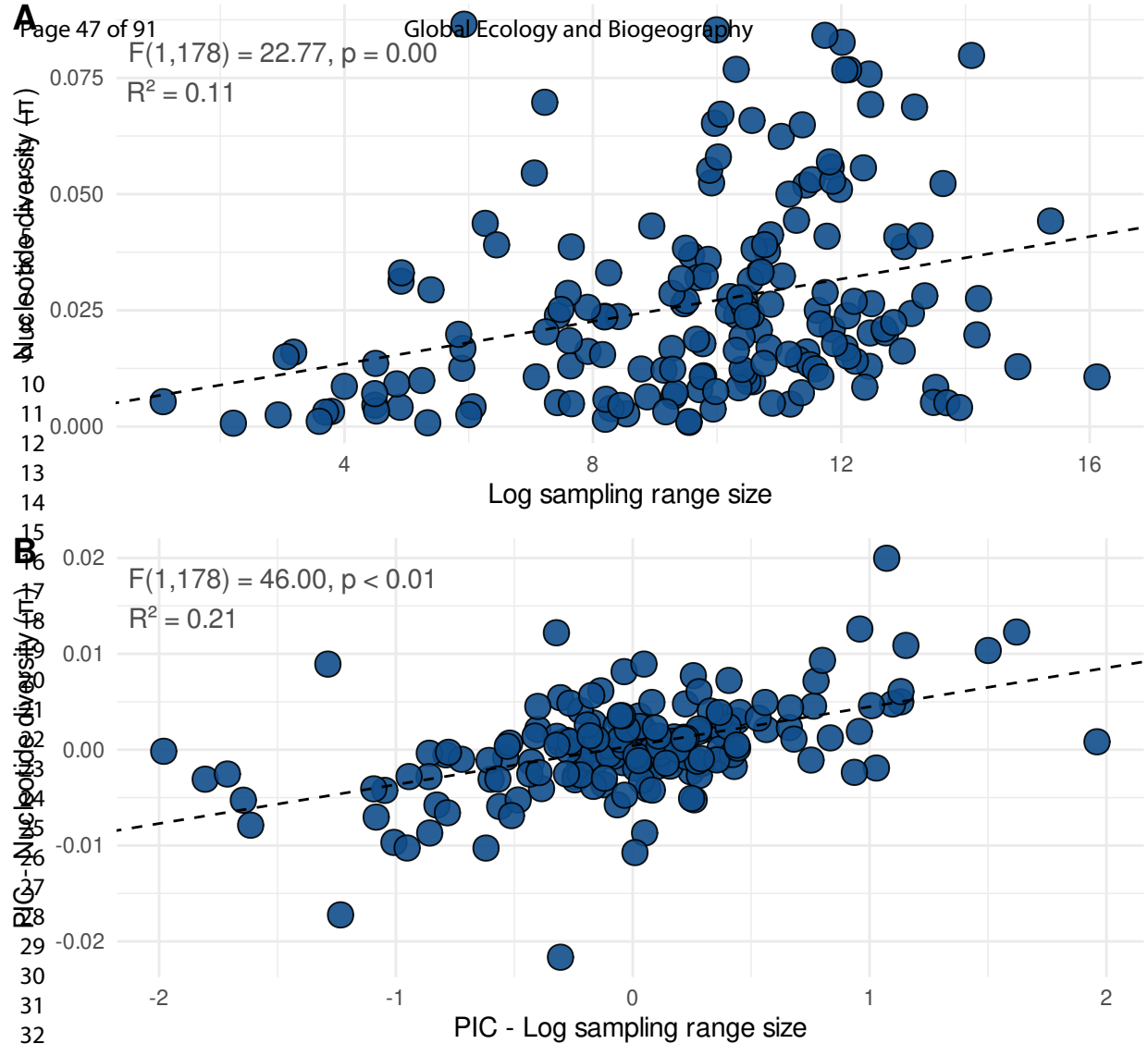


Figure 4. A) Comparison of sampling range size (in log scale) with nucleotide diversity (π) for each species. B) Comparison of phylogenetic independent contrast (PIC) results between the sampling range size (log scale) and nucleotide diversity (π).

186x172mm (300 x 300 DPI)


A
F(1,178) = 22.77, p = 0.00
R² = 0.11
Nucleotide diversity (π)
0.075
0.050
0.025
0.000
4
8
12
16
Log sampling range size
B
F(1,178) = 46.00, p < 0.01
R² = 0.21
PIC - Nucleotide diversity (π)
0.02
0.01
0.00
-0.01
-0.02
-2
-1
0
1
2
PIC - Log sampling range size

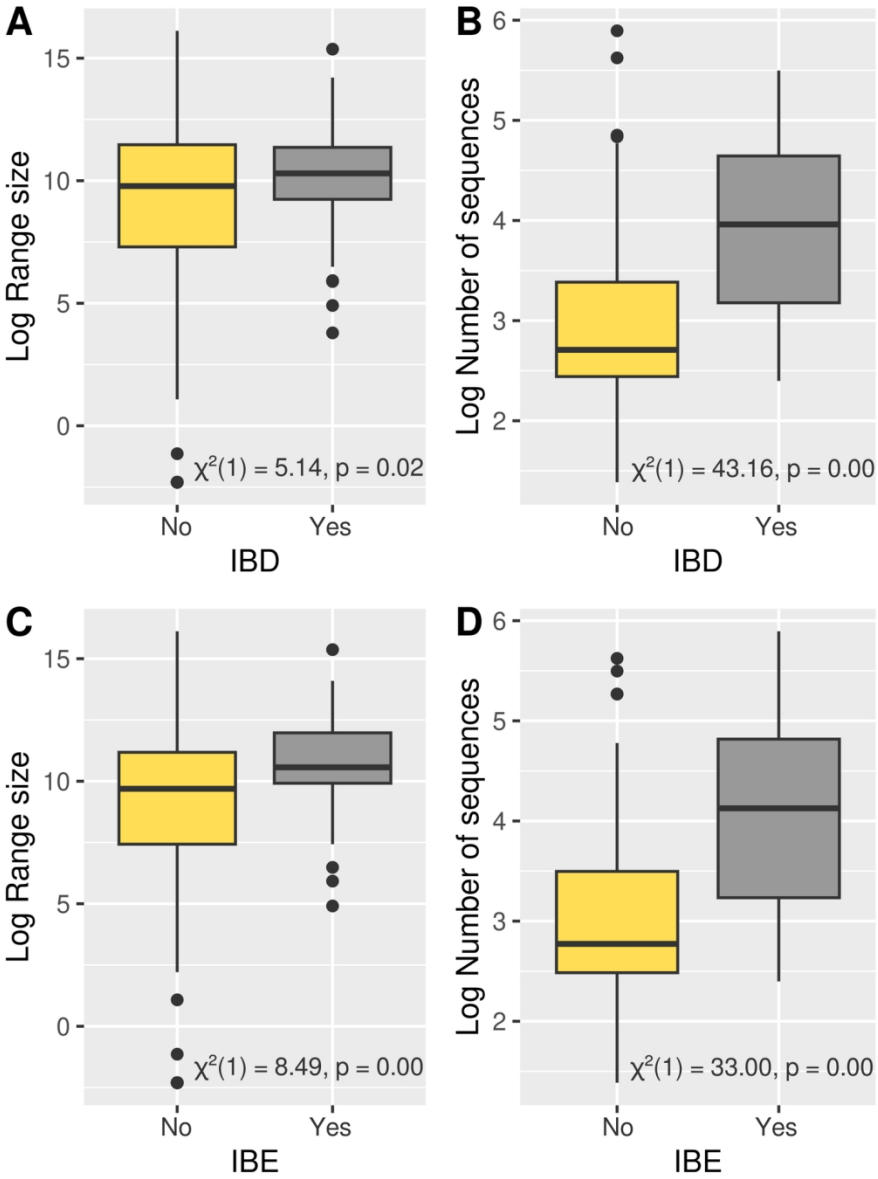
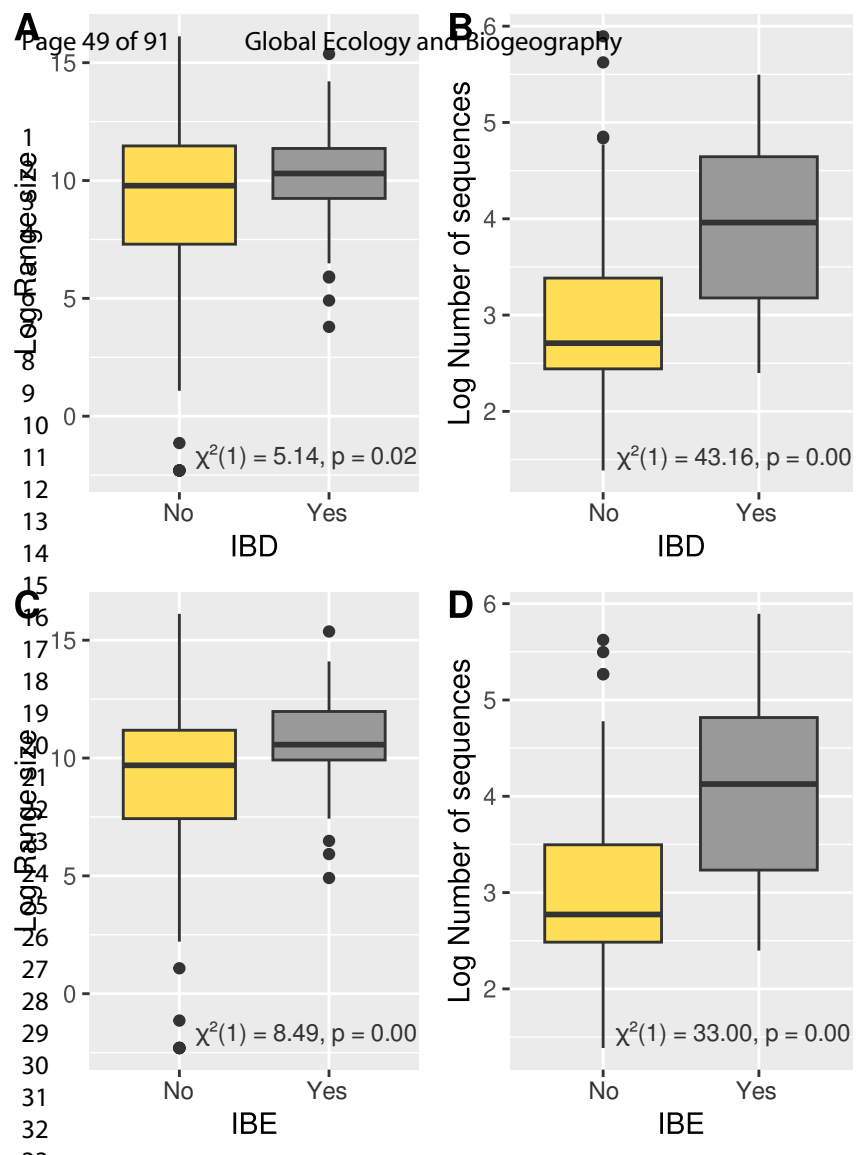


Figure 5. Boxplots comparing the sampling range size (A, C) and the number of sequences (B, D) between species with (Yes, gray color) and without (No, yellow color) isolation-by-distance (IBD) and isolation by environment (IBE). Each panel includes the result of a chi-squared test. Black dots represent outliers.

129x172mm (300 x 300 DPI)


A
Log Range size
15
10
5
0
χ²(1) = 5.14, p = 0.02
No
Yes
IBD
B
Log Number of sequences
6
5
4
3
2
χ²(1) = 43.16, p = 0.00
No
Yes
IBD
C
Log Range size
15
10
5
0
χ²(1) = 8.49, p = 0.00
No
Yes
IBE
D
Log Number of sequences
6
5
4
3
2
χ²(1) = 33.00, p = 0.00
No
Yes
IBE

Sheet1

Table S1. Variables used in th

Variable
Biogeographic Realm
Body size
bp CytB all
bp CytB coord
Family
Genus
Habitat
IBD
IBE
IUCN
Land cover
Litter_size_max_n
Litter_size_min_n
Max_Elevation
Max_Latitude
Mean precipitation
Mean temperature
Mean_Elevation
Mean_Latitude
Min_Elevation
Min_Latitude
Midpoint Latitude
Number of sequences
Paedomorphic
pi CytB all
pi CytB coord
Precipitation (SD)
Reproductive mode
Sampling_Range_size
Sequences w/ coordinates
Species
SVL
Temperature (SD)
Total_Range_size

For Peer Review

Sheet1

ıe study, variables in italics indicate predictor variables. Abbreviations: Aqu = Aquatic, Arb = Ar

Definition
Terrestrial realm/division where species occurs
Maximum adult body size. Body size is reported as total length (TL) in millimeters
Number of base pairs per species/alignment (all species – all sequences)
Number of base pairs per species/alignment (only species with coordinates)
Taxonomic Family (AmphibiaWeb, 2024)
Genus (AmphibiaWeb, 2024)
Aqu, Arb, Fos, Ter
Isolation by distance (YES, NO)
Isolation by environment (YES, NO)
Conservation status: CR, EN, VU, NT, LC, DD, NE
30-meter resolution that shows tree canopy coverage per pixel indicated by a single value between 0 and 1
Maximum number of offspring or eggs per clutch.
Minimum number of offspring or eggs per clutch.
Maximum elevation in meters
Maximum Latitude in decimal degrees
Average annual precipitation measured in millimeters (mm). BIO12 WorldClim
Average annual temperature in Celsius. BIO1 WorldClim
Mean elevation in meters
Mean Latitude in decimal degrees
Minimum elevation in meters
Minimum Latitude in decimal degrees
Mean of the minimum and maximum latitudinal values
Number of cytochrome B sequences in the original fasta file
Retaining juvenile features as an adult (YES, NO)
Nucleotide diversity (all species – all sequences)
Nucleotide diversity (only species with coordinates)
Precipitation Seasonality (Coefficient of Variation). BIO15 WorldClim
Dir, Lar, Viv
Species range size in square meters using just geographic coordinates
Number of cytochrome B sequences with geographic coordinates
Scientific name (AmphibiaWeb, 2024)
Maximum snout-vent length in millimeters
Temperature Seasonality (standard deviation ×100). BIO4 WorldClim
Species range size in square meters

For Peer Review

Sheet1

rboreal, Fos = Fossorial, Ter = Terrestrial. CR = Critically Endangered, EN = Endangered, VU

For Peer Review

Sheet1

U = Vulnerable, NT = Near Threatened, LC = Least Concern, DD = Data Deficient, NE = Not

For Peer Review

Sheet1

t Evaluated. Dir = Direct development, Lar = Larval stage, Viv = Viviparous

For Peer Review

Sheet1

For Peer Review

Family	Genus	Species	Author	Nativeness	Status	Habit	Height	Leaf area	Leaf shape	Leaf type	Leaf arrangement	Leaf margin	Leaf texture	Leaf color	Leaf phyllotaxy	Thorns	Spines	Prickles	Glands	Growth form	Growth rate	Nitrogen fixation	Root type	Root depth	Root density	Root mycorrhizal type	Root nodule type	Root herbivory	Root herbivore type	Root pathogen	Root pathogen type	Root pathogen effect	Root vector	Root antibiotic	Root biomass	Root C:N ratio	Root N fixation	Root C4
Cucurbitaceae	Ampelopsis	Ampelopsis glandulosa	Walter	Nativeno	Tentative	10	5	0.00000000	-1.00000000	0.00000000	negative	NO	1.00	0.00000000	0.00000000	0.00000000	Yes	Yes	Yes	Yes	Yes	Yes	Tree_Shrub_Vine	40	2.00	1.00	200	0	1.00	320.00000000	27.00000000	50.00000000	41.00000000	11.00000000	64.00000000	0.00000000	0.00000000	0

Sheet1

Table S3. Results of the phylogenetic generalized linear mixed models (

Response	Model	Random effects		
			Variance	Std.Dev
IBD	pglmm(IBD ~ Sampling range size + Number of sequences + (1 | Family), covariance matrix for random terms = phylogeny	Family	0.1526	0.3906
			Variance	Std.Dev
IBE	pglmm(IBE ~ Sampling range size + Number of sequences + (1 | Family) + (1 | Genus), covariance matrix for random terms = phylogeny	Family	2.79E-06	1.67E-03
		Genus	5.10E-01	7.14E-01

For Peer Review

Sheet1

(PGLMM) of isolation by distance (IBD) and isolation by environment (IBE) showi

Fixed effects				
	Value	Std.Error	Zscore	Pvalue
(Intercept)	-1.63E+00	3.22E-01	-5.078	**3.81E-07**
Sampling range size	-2.46E-07	2.34E-01	-1.049	0.2944
Number of sequences	1.72E-02	4.16E-03	4.1268	**3.68E-05**

	Value	Std.Error	Zscore	Pvalue
(Intercept)	-2.24E+00	3.32E-01	-6.743	**1.55E-11**
Sampling range size	-1.03E-07	2.24E-07	-0.46	0.6455
Number of sequences	1.84E-02	4.31E-03	4.2649	**2.00E-05**

For Peer Review

Sheet1

ing the model, random and fixed effects and the respective coefficients.

For Peer Review

Caudata macrogenetics: Geographic attributes predict mitochondrial genetic variation in salamanders on a global scale

Supplementary Material

SM1. Tutorial used to retrieve the geographic coordinates associated with DNA sequences. https://github.com/luchoamador/tutorials/blame/main/Coords_DNAseqs.html.
SM2. R scripts used to retrieve species ranges, elevation, latitude, precipitation, and temperature data associated with each species. https://github.com/luchoamador/Caudata
SM3. A) All CytB DNA sequences and geographic coordinates from salamander species analyzed in this study. B) R scripts used to obtain some of the sequences and coordinates used in analyses. https://github.com/luchoamador/Caudata

Supplementary Tables

Table S1. Predictors used in this study and their respective definitions.
Table S2. The complete matrix of values for all variables analyzed in this study.
Table S3. Results of the phylogenetic generalized linear mixed models (PGLMM) of isolation-by-distance (IBD) and isolation-by-environment (IBE) showing the model, random and fixed effects, and the respective coefficients.

Supplementary Figures

Figure S1. The relationship between nucleotide diversity calculated for all repurposed sequences (y-axis – all, π_A) and nucleotide diversity calculated for sequences with associated geographic coordinates (x-axis – coords, π_S). A linear model suggests a highly statistically significant and strong positive correlation; $F(1,212) = 1609.62$, $p < 0.001$, $R^2 = 0.88$.
Figure S2. Values of nucleotide diversity (π) for each salamander species (n = 214). Nucleotide diversity was calculated using the sequences with associated coordinates (π_S).
Figure S3. Maps of A) mean nucleotide diversity (π), and B) mean number of sequences with coordinates (log scale) of global salamanders using a grid cell = 1° (approximately 110 km^2) for mapping.
Figure S4. Maps of A) mean nucleotide diversity (π), and B) mean number of sequences with coordinates (log scale) of global salamanders using a grid cell = 2° (approximately 220 km^2) for mapping.
Figure S5. Maps of A) mean nucleotide diversity (π), and B) mean number of sequences with coordinates (log scale) of global salamanders using a grid cell = 3° (approximately 330 km^2) for mapping.

Figure S6. The number of salamander species presenting either IBD/IBE (gray) or no IBD/IBE (yellow). Three species for IBD and two species for IBE were excluded because they did not have sufficient localities with sequences.

Figure S7. Salamander species with and without IBD and IBE patterns in each biogeographic realm. The number of species per realm is indicated in the bars for each isolation pattern.

Figure S8. Results of phylogenetic signal for nucleotide diversity (π). A) Observed Pagel's Lambda and its associated p-value. B) The observed value of Blomberg's K statistic compared to a null distribution of K.

Figure S9. Results of phylogenetic signal in IBD and IBE using D statistic. Estimated D values for IBD and IBE are indicated, and test-associated p-values are also shown.

Figure S10. The number of salamander species exhibiting either recent population contraction (tan), recent population expansion (red), or no population change size (green), based on Tajima's D analyses.

Figure S11. The relationship between Log sampling range size in km^2 and Log total range size in km^2. A linear model suggests a highly statistically significant and positive correlation; $F(1,212) = 338.83$, $p < 0.001$, $R^2 = 0.62$.

Figure S12. Random forest (RF) results for A) nucleotide diversity (π), B) isolation-by-distance (IBD), and C) isolation-by-environment (IBE), considering the total range size. Only the top 10 best predictors are shown. For π, we used the total number of sequences; for IBD and IBE, we used the number of sequences associated with geographic coordinates. SVL = snout-vent length, mm = millimeters.

Figure S13. Nucleotide diversity (π) values of nine salamander families. Boxplots display the distribution of π and visualize the median (bold lines), lower and upper quartiles (boxes), and outlier points. The results of a linear model performed in R with the lm() function are shown.

Figure S14. Relationship between isolation-by-distance (IBD) with two important predictors based on random forest. A) The mean precipitation vs IBD for temperate salamander species, B) mean precipitation vs IBD for tropical species, C) maximum latitude vs IBD patterns in temperate species, and D) maximum latitude vs IBD for tropical species. The results of the generalized linear models performed in R with the glm() function are shown in each panel.

Figure S15. Relationship between isolation-by-environment (IBE) and two important predictors based on random forest. A) Mean latitude vs IBE for temperate salamander species, B) mean latitude vs IBE for tropical species, C) mean elevation vs IBE patterns in temperate species, and D) mean elevation vs IBE for tropical species. The results of the generalized linear models performed in R with the glm() function are shown in each panel.

Figure S16. A) Comparison of mean temperature with π. B) Phylogenetic independent contrast (PIC) results between mean temperature and π.

Figure S17. A) Comparison of the standard deviation (SD) temperature with π. B) Phylogenetic independent contrast (PIC) results between SD temperature and π.

Figure S18. A) Comparison of species body size (in log scale) with π. B) Phylogenetic independent contrast (PIC) results between the log body size and π.

For Peer Review

Figure S19. A) Comparison of the mean elevation (in log scale) with π. B) Phylogenetic independent contrast (PIC) results between the log mean elevation and π.

Figure S20. Sampling range size (log) for species with and without A) isolation-by-distance (IBD) for temperate species, B) Sampling range size (log) vs IBD for tropical species, C) Sampling range size (log) vs isolation-by-environment (IBE) for temperate species, D) Sampling range size (log) vs IBE for tropical species. Each boxplot indicates the median (bold lines) and the interquartile range (boxes). The results of the generalized linear models performed in R with the glm() function are shown in each panel.

Figure S21. Relationship between nucleotide diversity (π) and Tajima's D (demographic change interpretation). Each boxplot indicates the median (bold lines), the interquartile range (boxes), and the observations (dots). The results of a linear model performed in R with the lm() function are shown.

Figure S22. Random Forest (RF) results for A) nucleotide diversity (π, using RF regression), B) isolation-by-distance (IBD, using RF classification), and C) isolation-by-environment (IBE, using RF classification) with a subsampled dataset using between 10 and 20 sequences per species. All predictors except number of base pairs, and total range size were used in these RF models, but only the top 10 best predictors are shown here. Range size* = sampling range size based on localities with associated DNA sequences; # of sequences+ = number of sequences associated with geographic coordinates; mm = millimeters.

Figure S23. Relationship between total range size (log) and number of sequences; the colors of the dots indicate nucleotide diversity (π) in each species.

For Peer Review

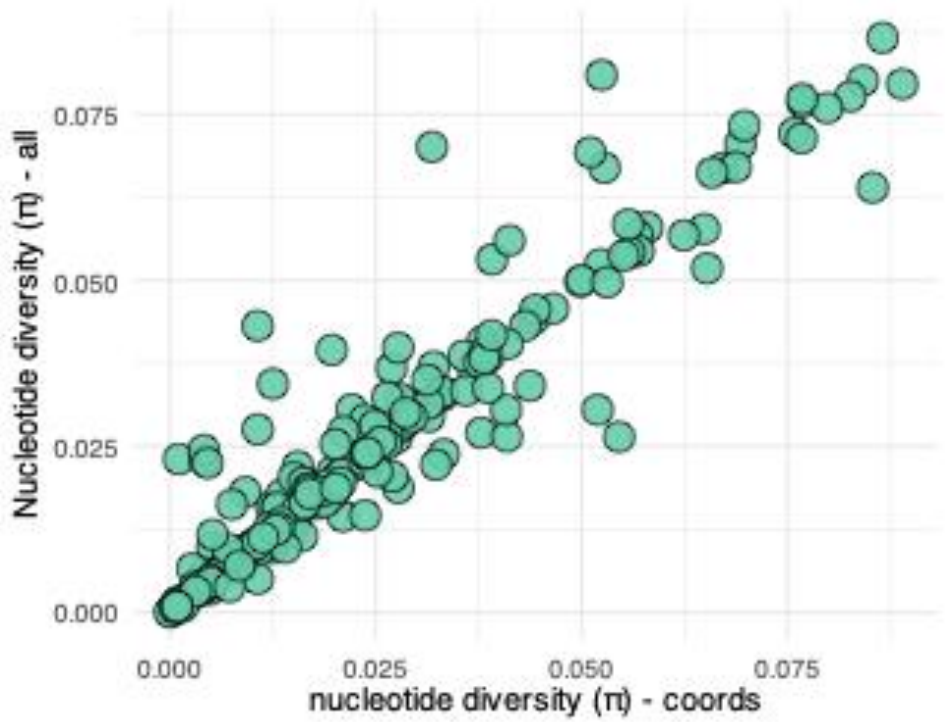


Figure S1. The relationship between nucleotide diversity calculated for all repurposed sequences (y-axis – all, π_A) and nucleotide diversity calculated for sequences with associated geographic coordinates (x-axis – coords, π_S). A linear model suggests a highly statistically significant and strong positive correlation; $F(1,212) = 1609.62$, $p < 0.001$, $R^2 = 0.88$.

For Review

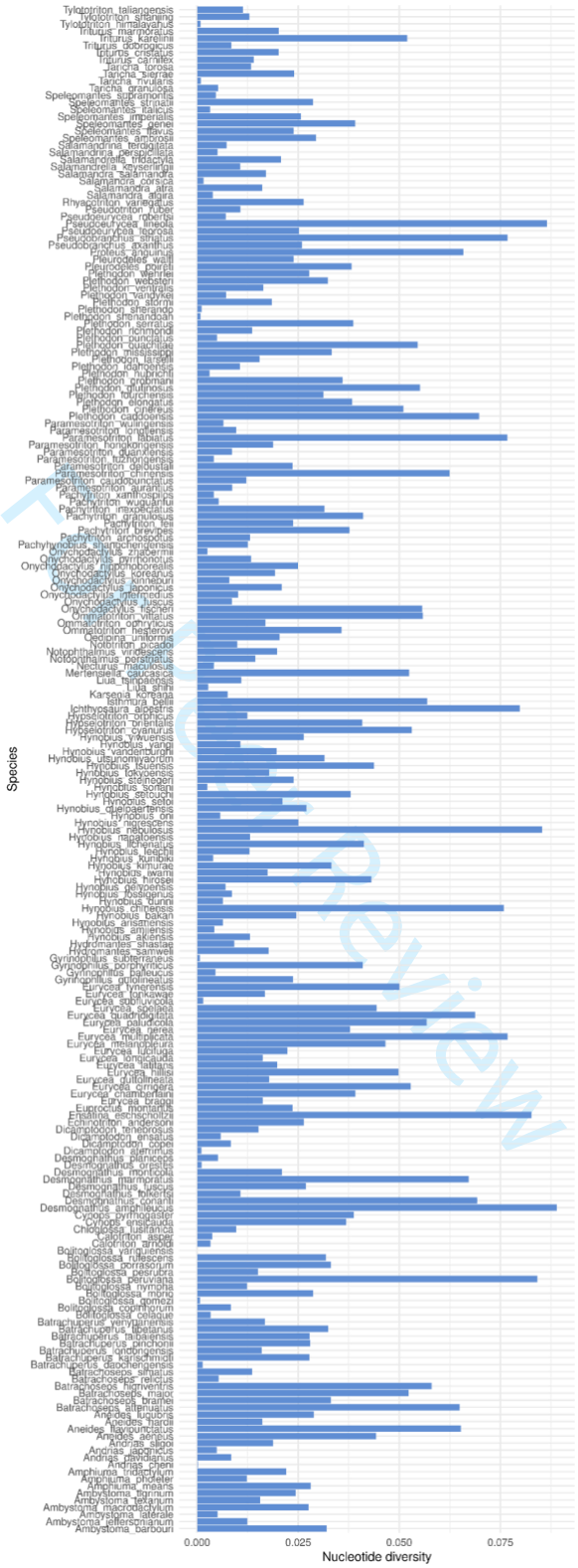


Figure S2. Values of nucleotide diversity (π) for each salamander species (n = 214). Nucleotide diversity was calculated using the sequences with associated coordinates (π_S).

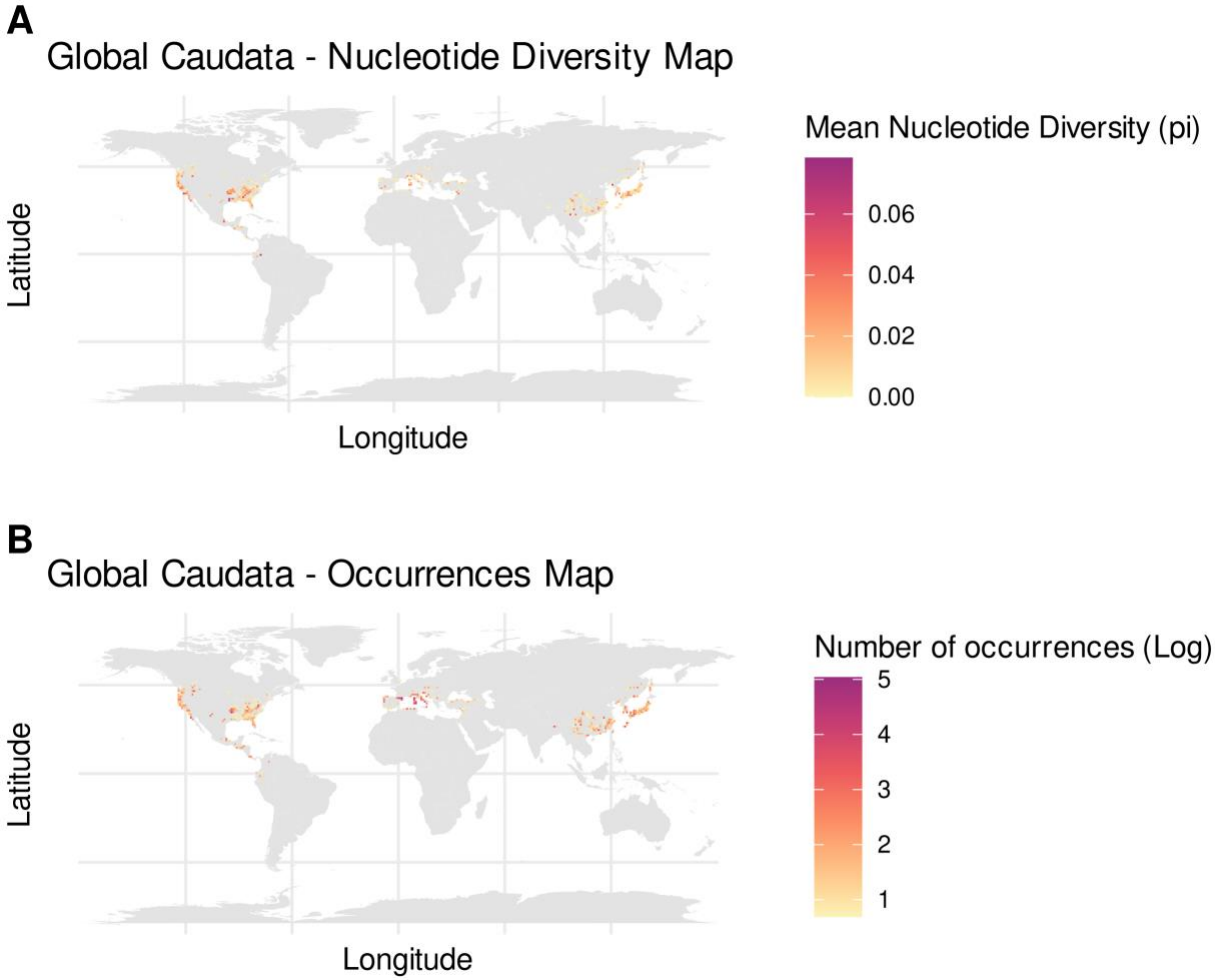


Figure S3. Maps of A) mean nucleotide diversity (π), and B) mean number of sequences with coordinates (log scale) of global salamanders using a grid cell = 1° (approximately 110 km^2) for mapping.

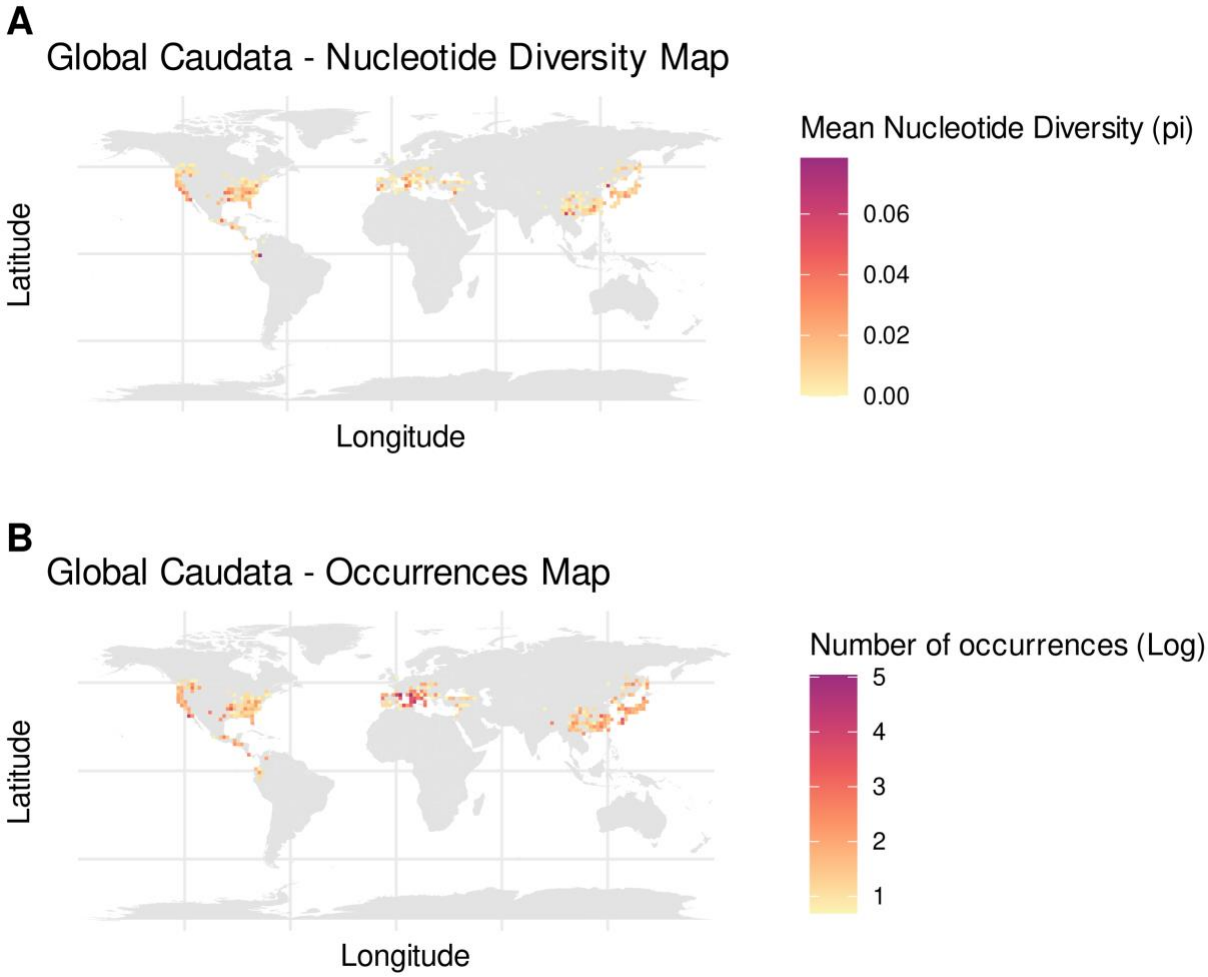


Figure S4. Maps of A) mean nucleotide diversity (π), and B) mean number of sequences with coordinates (log scale) of global salamanders using a grid cell = 2° (approximately 220 km^2) for mapping.

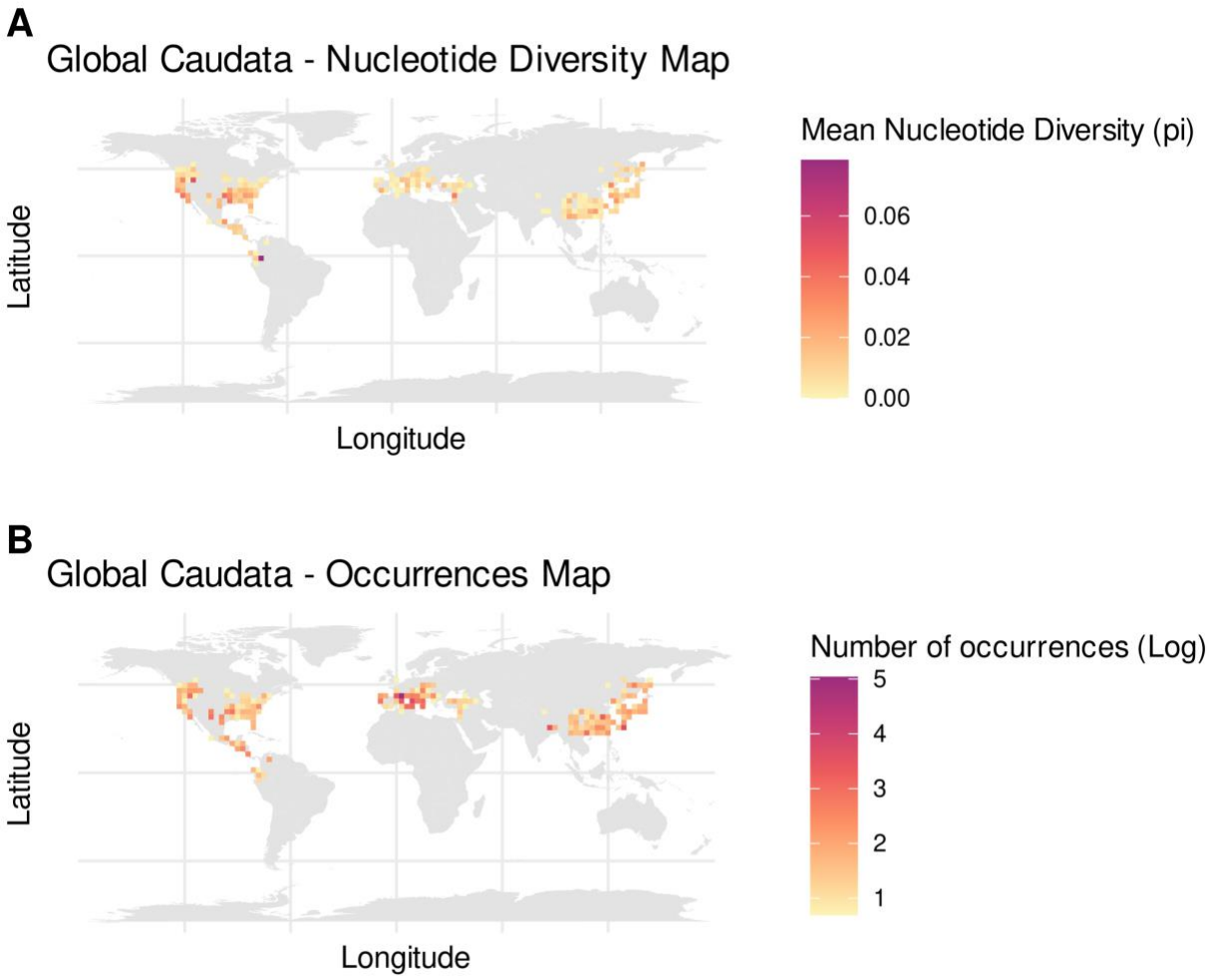


Figure S5. Maps of A) mean nucleotide diversity (π), and B) mean number of sequences with coordinates (log scale) of global salamanders using a grid cell = 3° (approximately 330 km^2) for mapping.

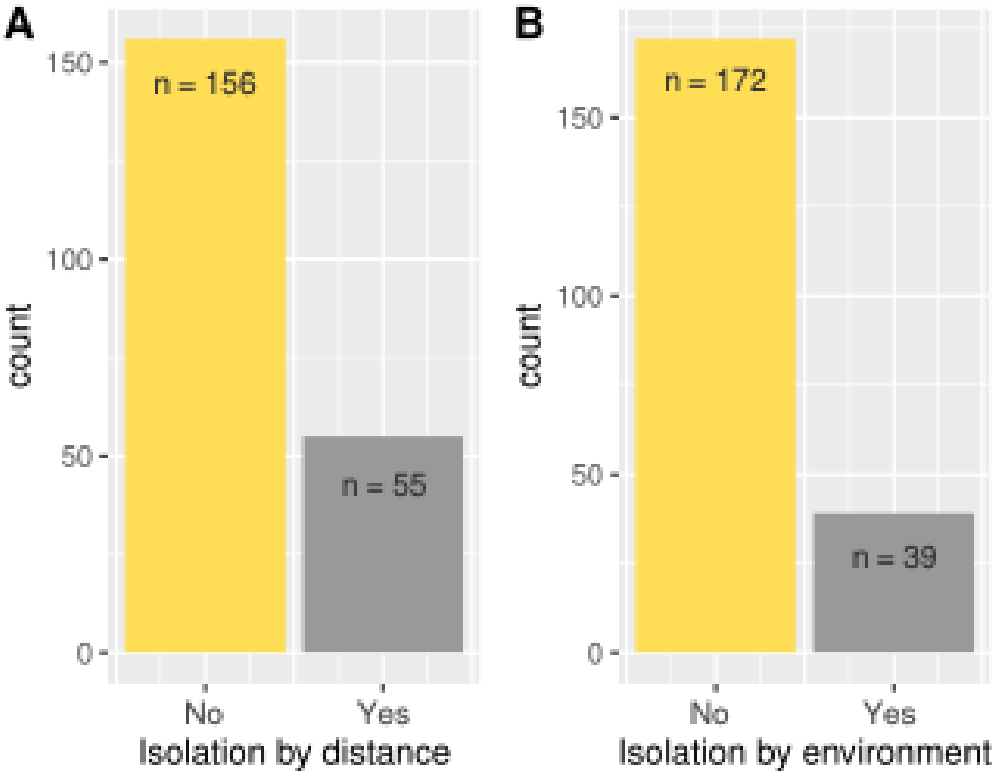


Figure S6. The number of salamander species presenting either IBD/IBE (gray) or no IBD/IBE (yellow). Three species for IBD and two species for IBE were excluded because they did not have sufficient localities with sequences.

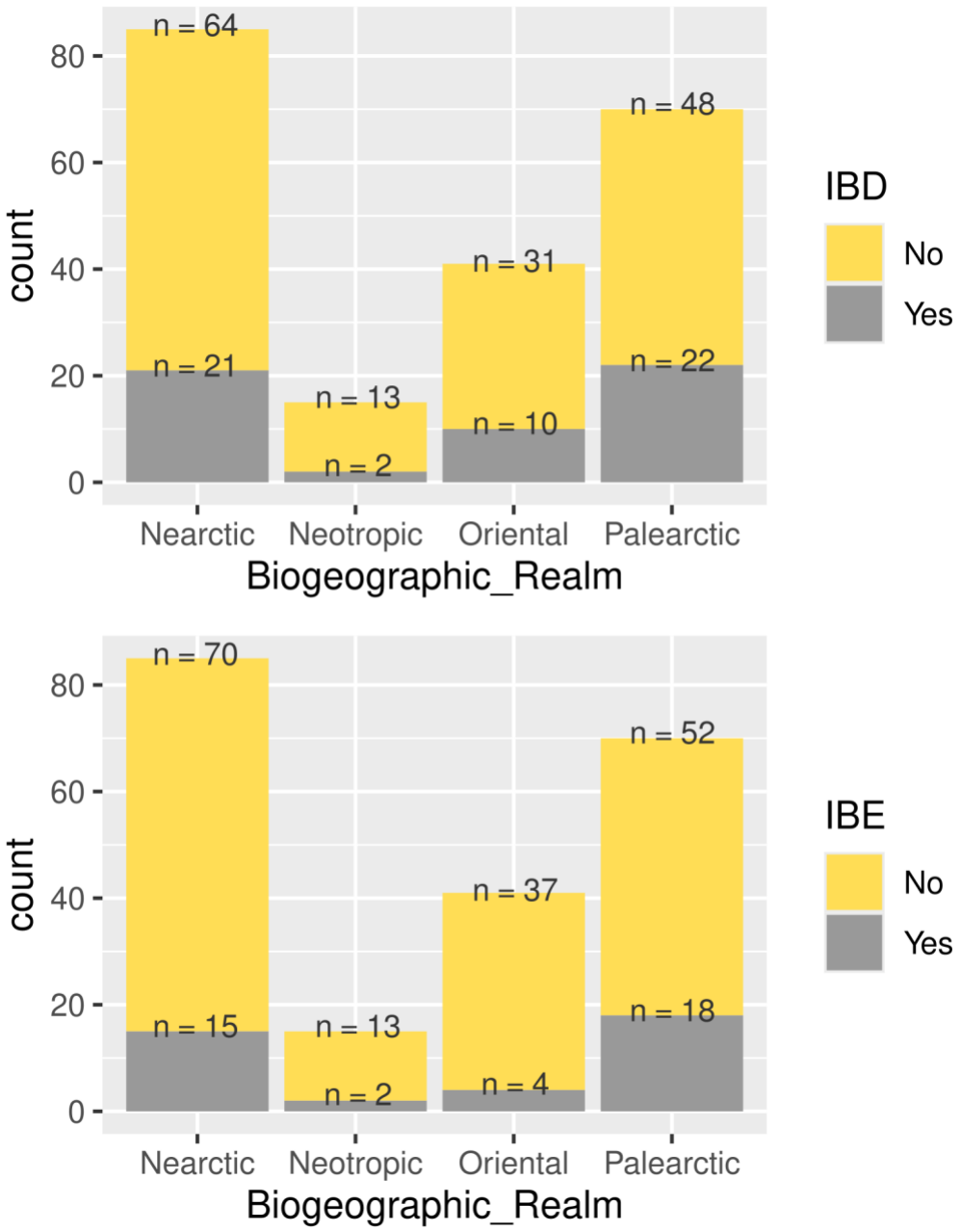


Figure S7. Salamander species with and without IBD and IBE patterns in each biogeographic realm. The number of species per realm is indicated in the bars for each isolation pattern.

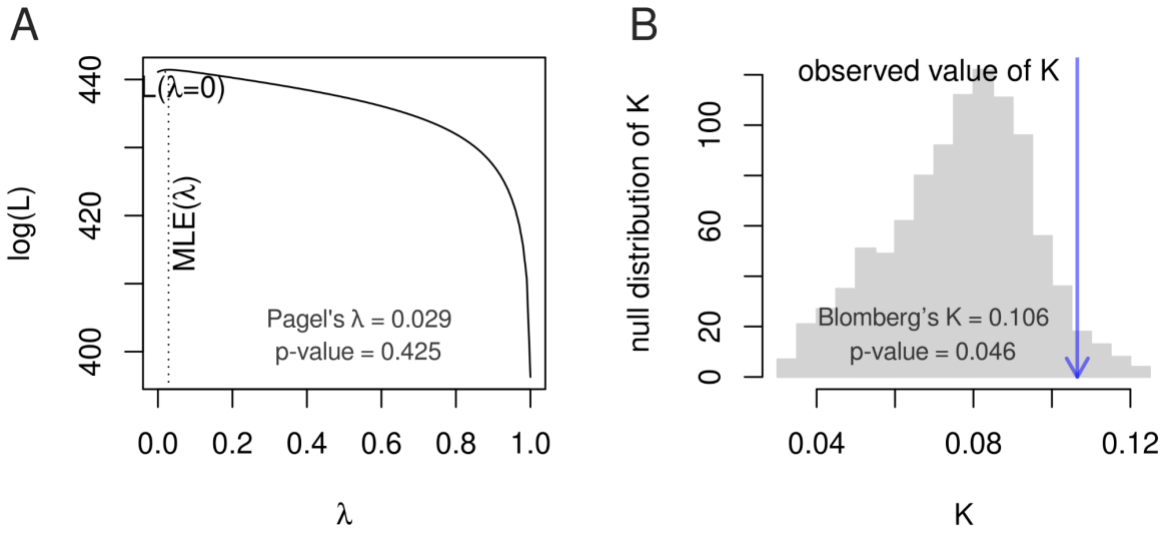


Figure S8. Results of phylogenetic signal for nucleotide diversity (π). A) Observed Pagel's Lambda and its associated p-value. B) The observed value of Blomberg's K statistic compared to a null distribution of K.

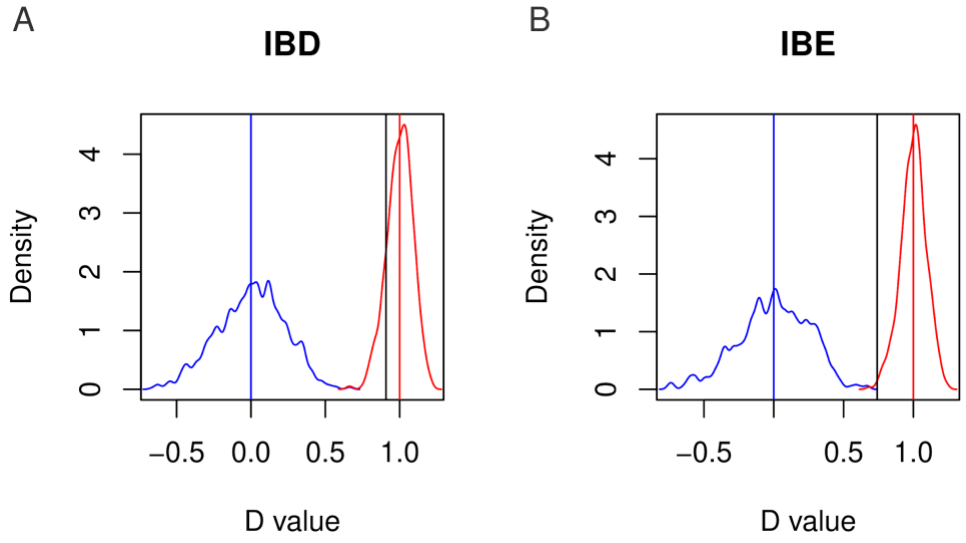


Figure S9. Results of phylogenetic signal in IBD and IBE using D statistic. Estimated D values for IBD and IBE are indicated, and test-associated p-values are also shown.

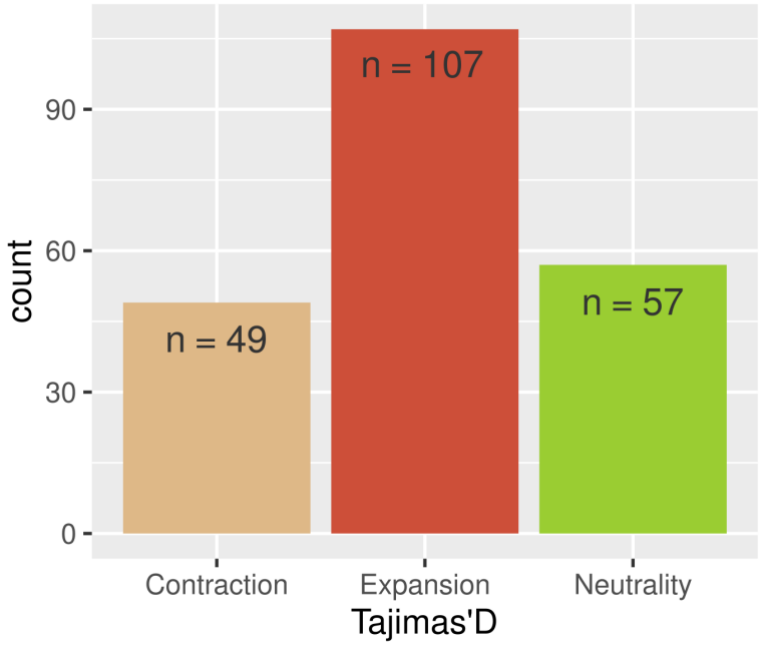


Figure S10. The number of salamander species exhibiting either recent population contraction (tan), recent population expansion (red), or no population change size (green), based on Tajima's D analyses.

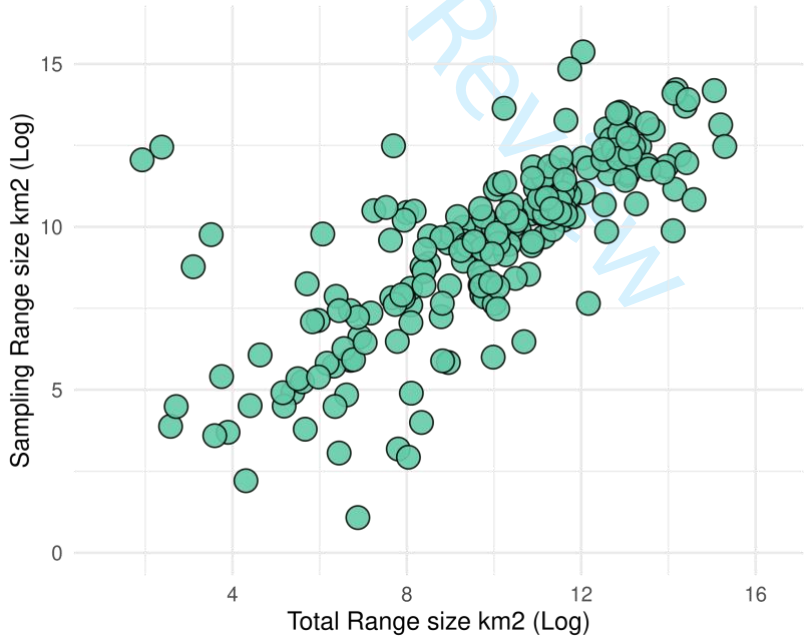


Figure S11. The relationship between Log sampling range size in km^2 and Log total range size km^2. A linear model suggests a highly statistically significant and positive correlation; $F(1,212) = 338.83$, $p < 0.001$, $R^2 = 0.62$.

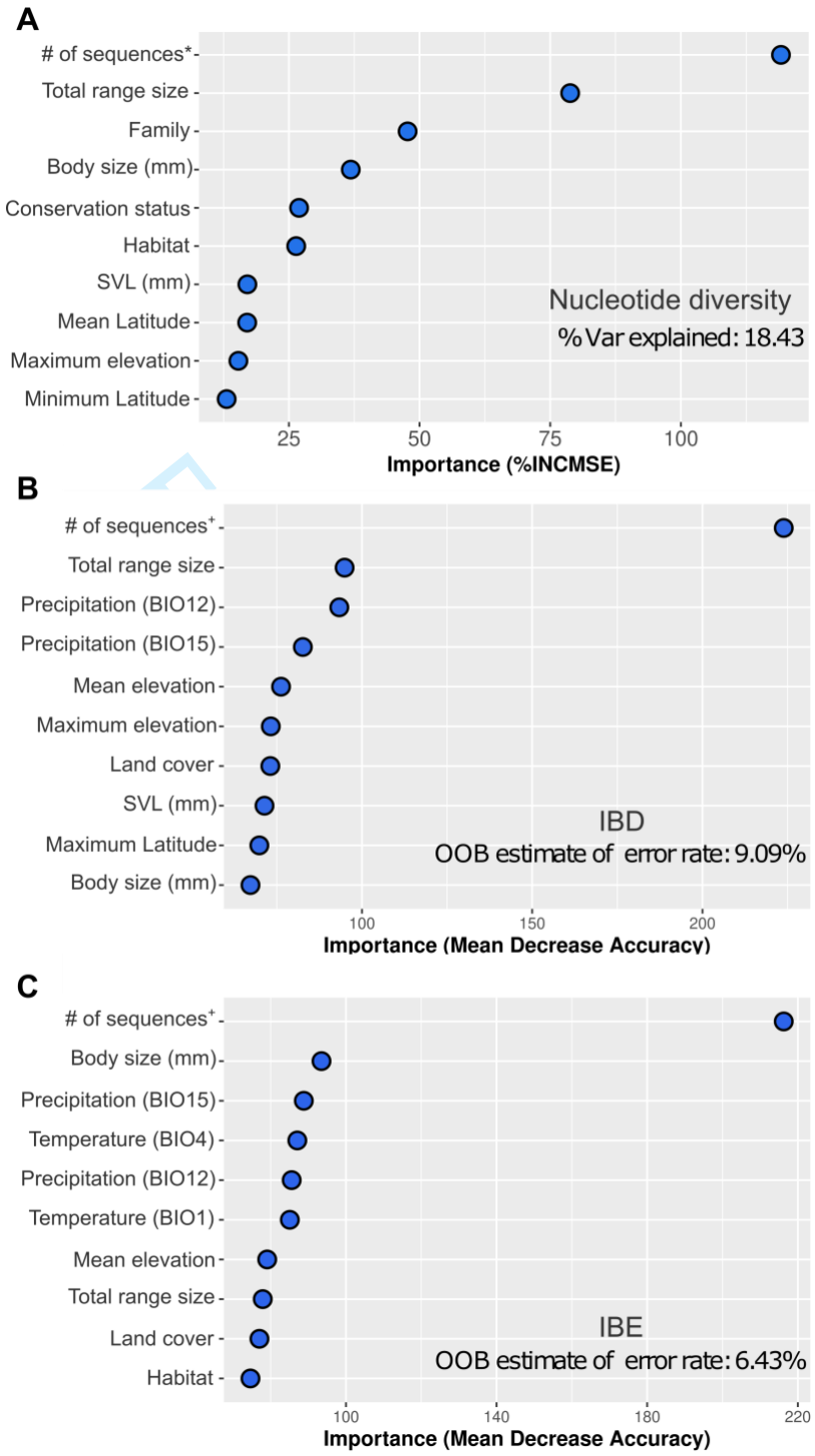


Figure S12. Random forest (RF) results for A) nucleotide diversity (π), B) isolation-by-distance (IBD), and C) isolation-by-environment (IBE), considering the total range size. Only the top 10 best predictors are shown. For π, we used the total number of sequences; for IBD and IBE, we used the number of sequences associated with geographic coordinates. SVL = snout-vent length, mm = millimeters.

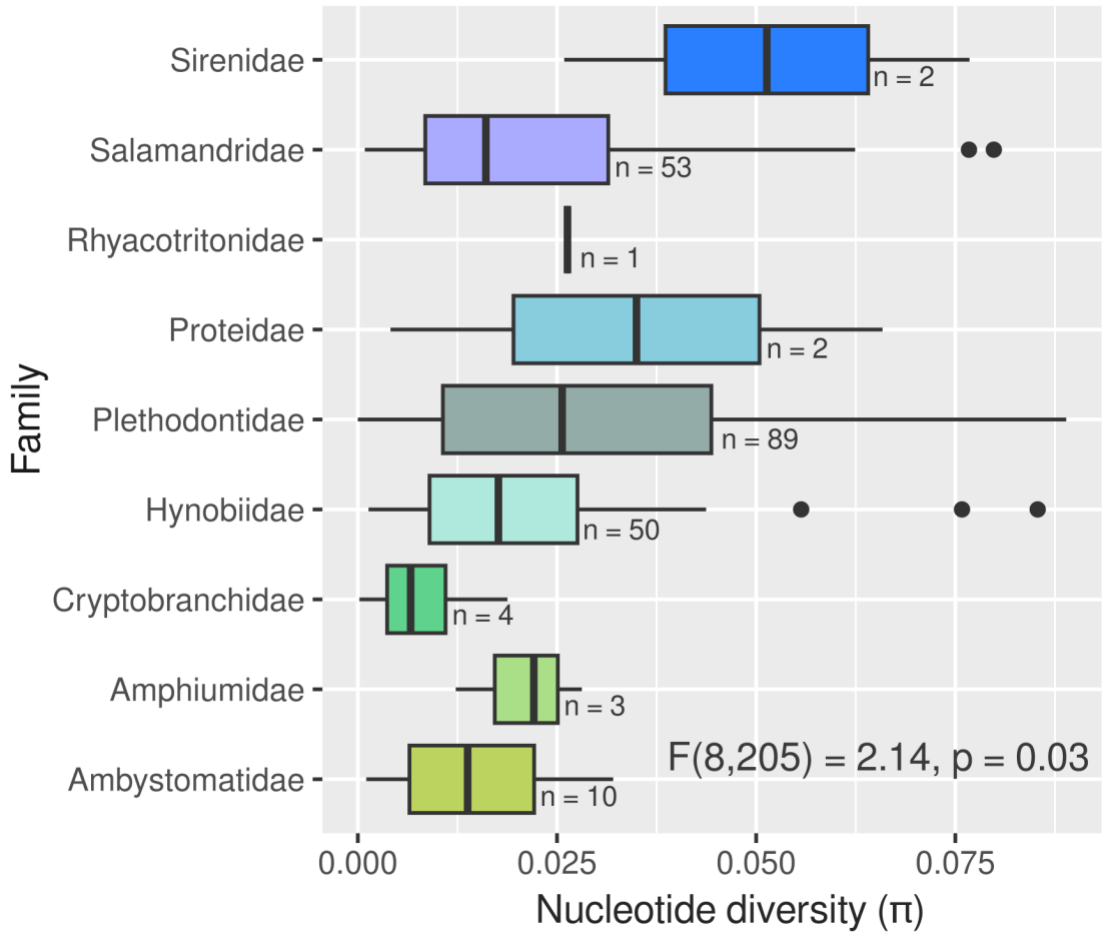


Figure S13. Nucleotide diversity (π) values of nine salamander families. Boxplots display the distribution of π and visualize the median (bold lines), lower and upper quartiles (boxes), and outlier points. The results of a linear model performed in R with the lm() function are shown.

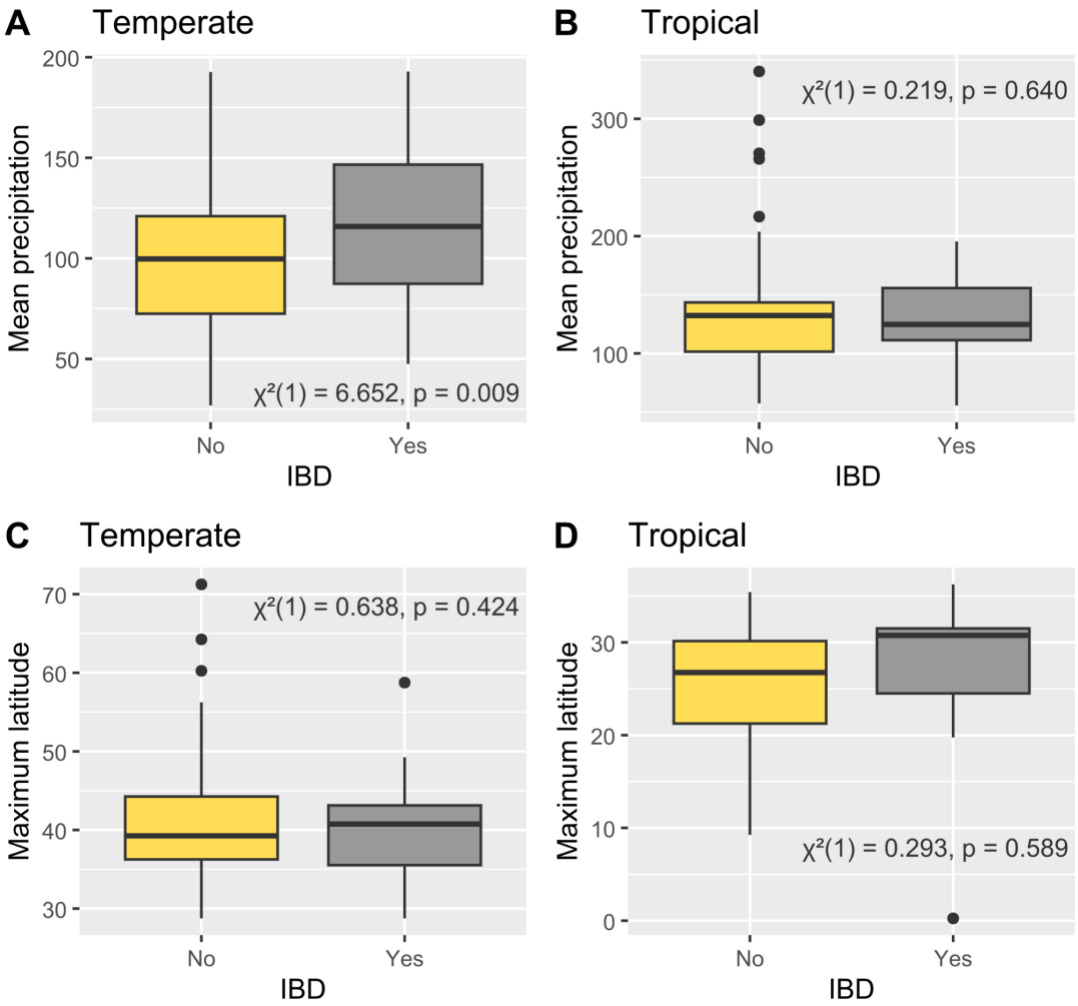


Figure S14. Relationship between isolation-by-distance (IBD) with two important predictors based on random forest. A) The mean precipitation vs IBD for temperate salamander species, B) mean precipitation vs IBD for tropical species, C) maximum latitude vs IBD patterns in temperate species, and D) maximum latitude vs IBD for tropical species. The results of the generalized linear models performed in R with the glm() function are shown in each panel.

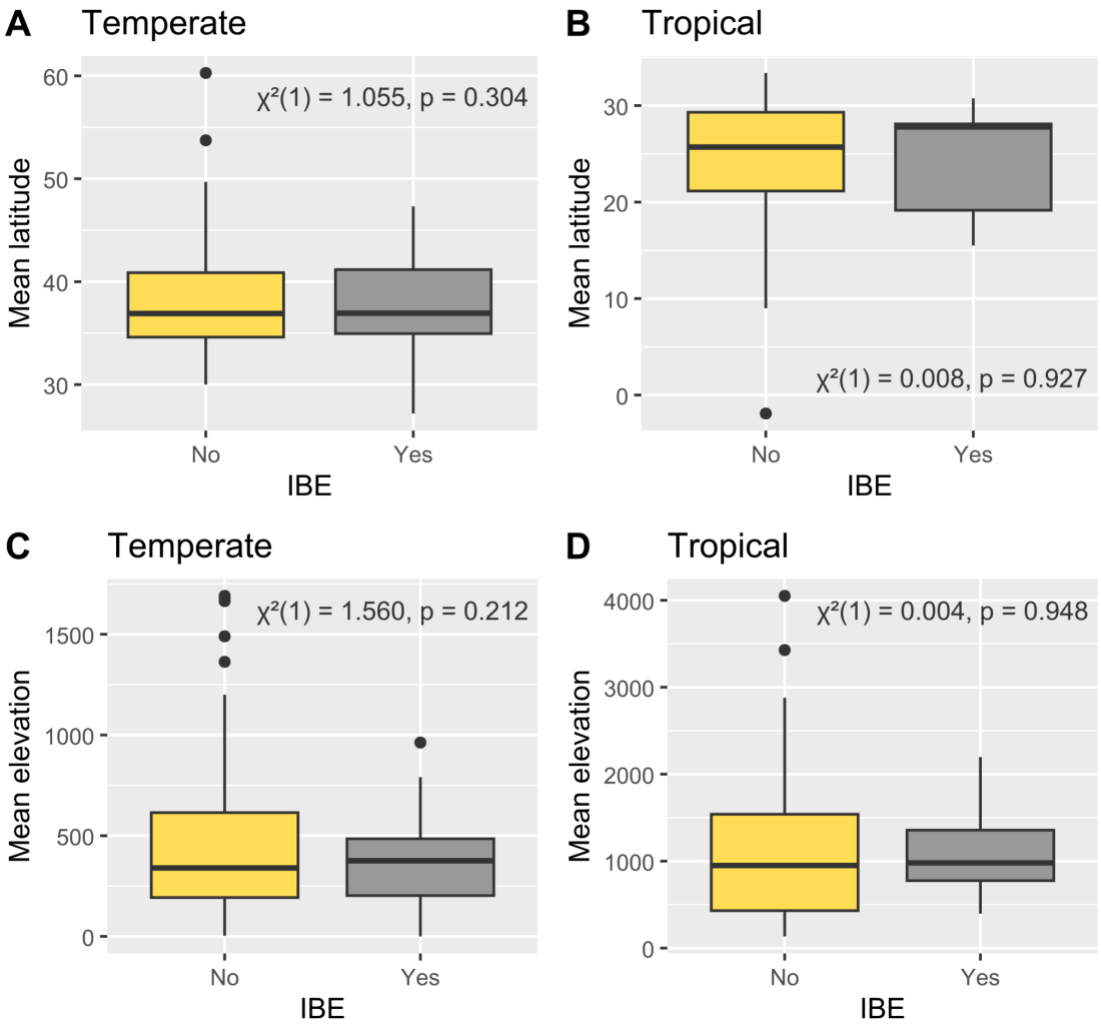


Figure S15. Relationship between isolation-by-environment (IBE) and two important predictors based on random forest. A) Mean latitude vs IBE for temperate salamander species, B) mean latitude vs IBE for tropical species, C) mean elevation vs IBE patterns in temperate species, and D) mean elevation vs IBE for tropical species. The results of the generalized linear models performed in R with the glm() function are shown in each panel.

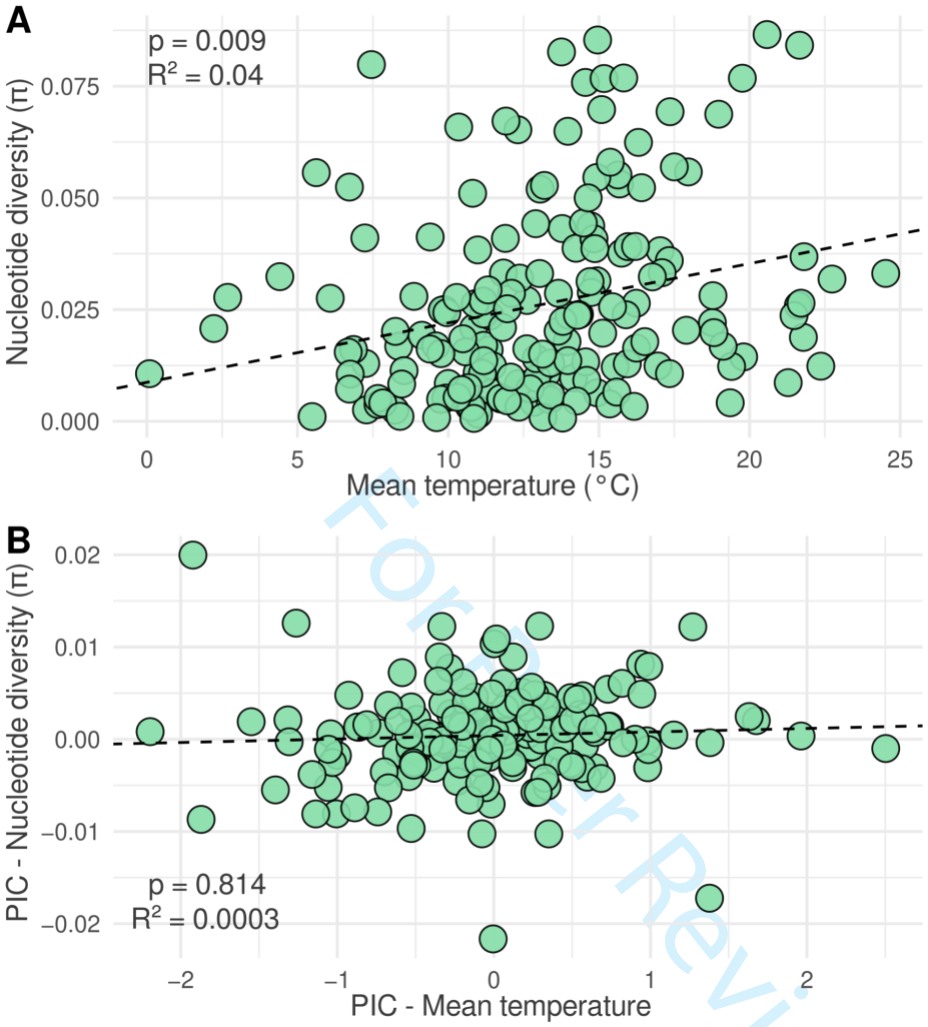


Figure S16. A) Comparison of mean temperature with π. B) Phylogenetic independent contrast (PIC) results between mean temperature and π.

For Peer Review

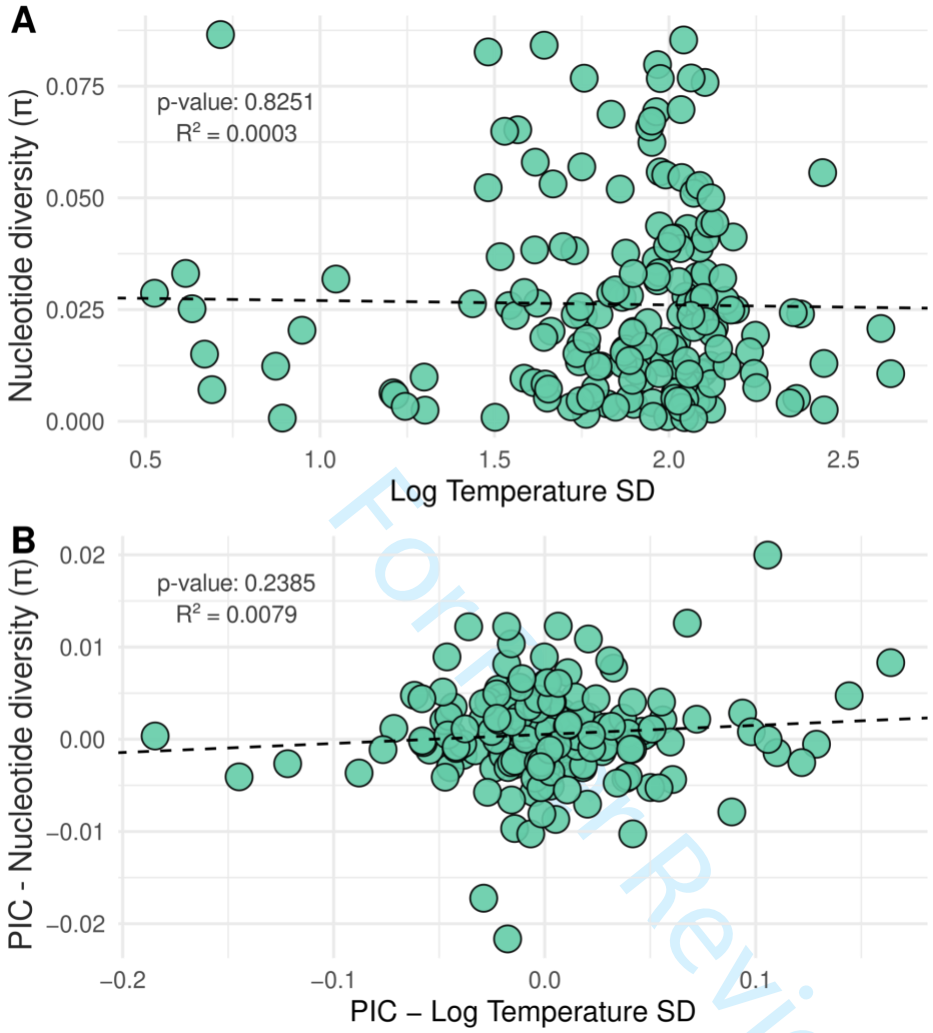


Figure S17. A) Comparison of the standard deviation (SD) temperature with π. B) Phylogenetic independent contrast (PIC) results between SD temperature and π.

For Peer Review

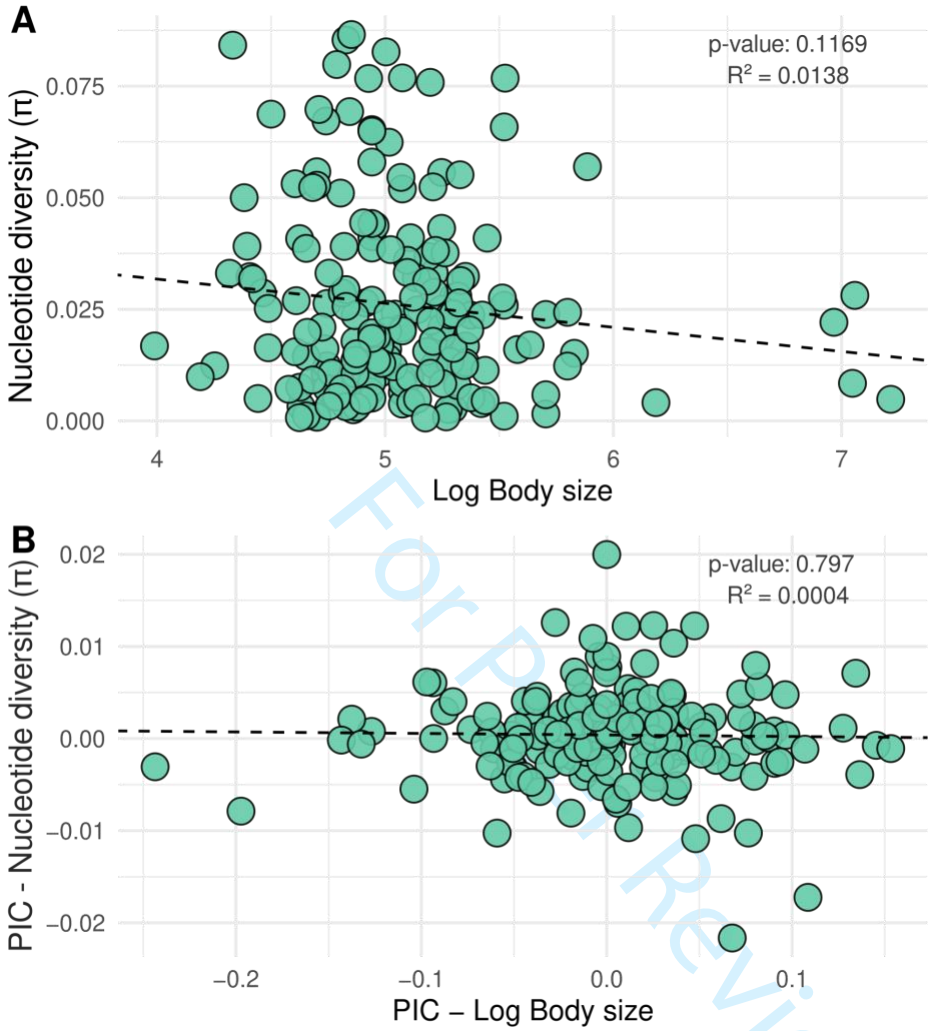


Figure S18. A) Comparison of species body size (in log scale) with π. B) Phylogenetic independent contrast (PIC) results between the log body size and π.

For Peer Review

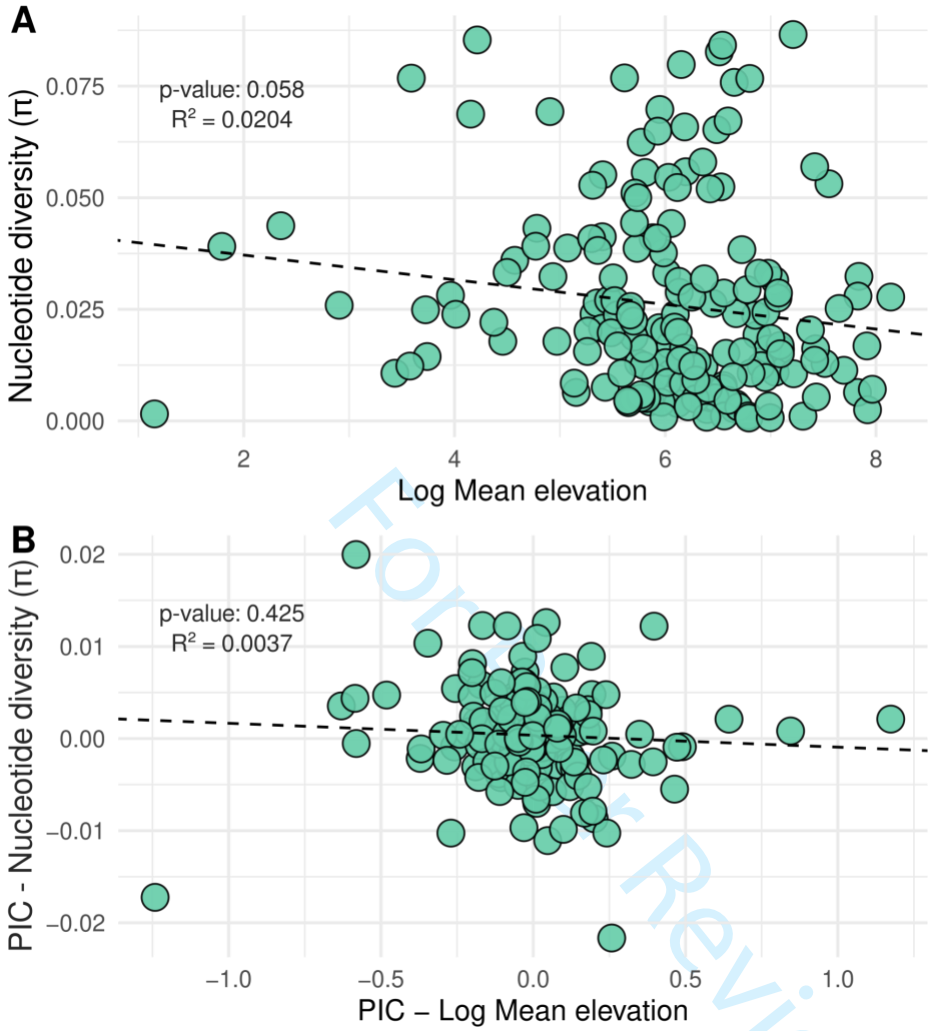


Figure S19. A) Comparison of the mean elevation (in log scale) with π. B) Phylogenetic independent contrast (PIC) results between the log mean elevation and π.

For Peer Review

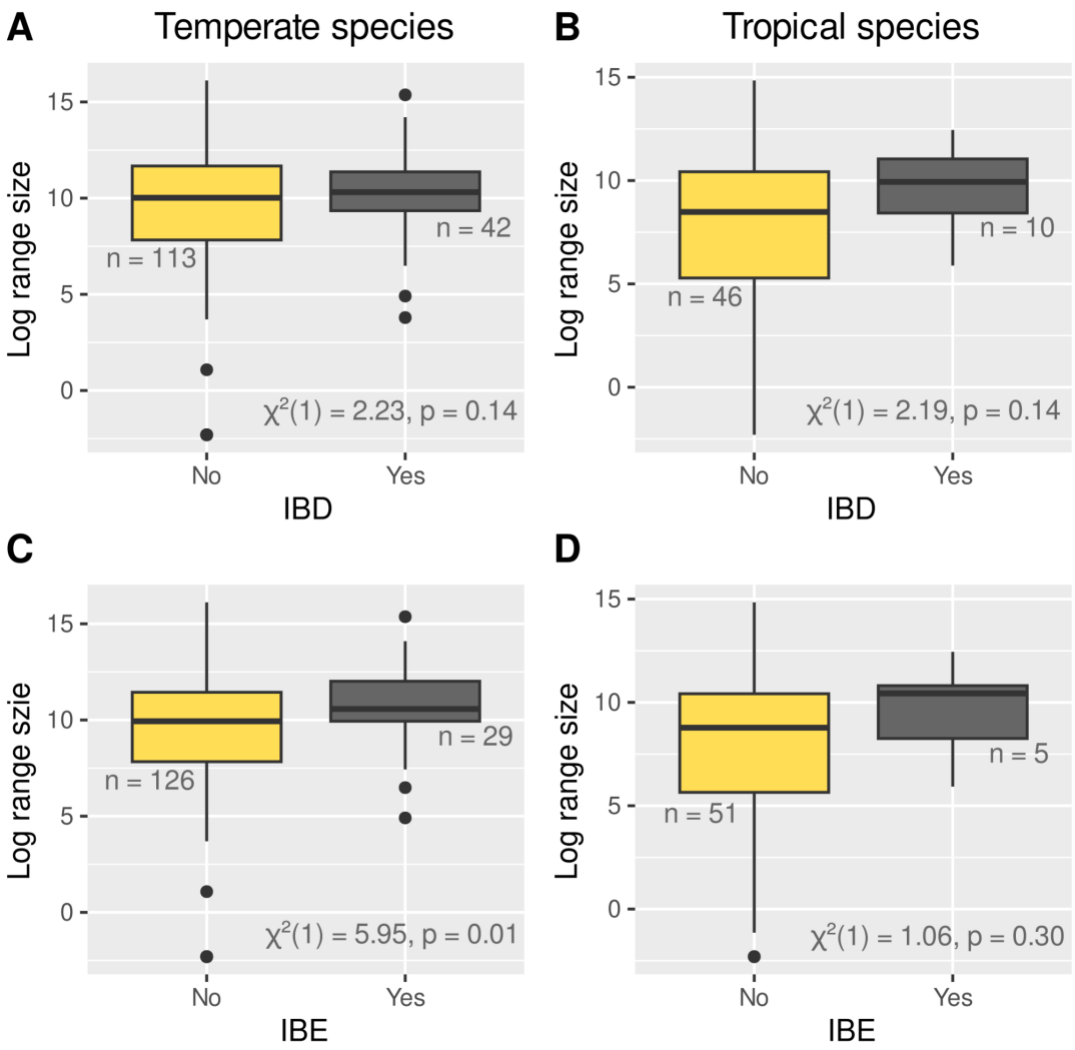


Figure S20. Sampling range size (log) for species with and without A) isolation-by-distance (IBD) for temperate species, B) Sampling range size (log) vs IBD for tropical species, C) Sampling range size (log) vs isolation-by-environment (IBE) for temperate species, D) Sampling range size (log) vs IBE for tropical species. Each boxplot indicates the median (bold lines) and the interquartile range (boxes). The results of the generalized linear models performed in R with the glm() function are shown in each panel.

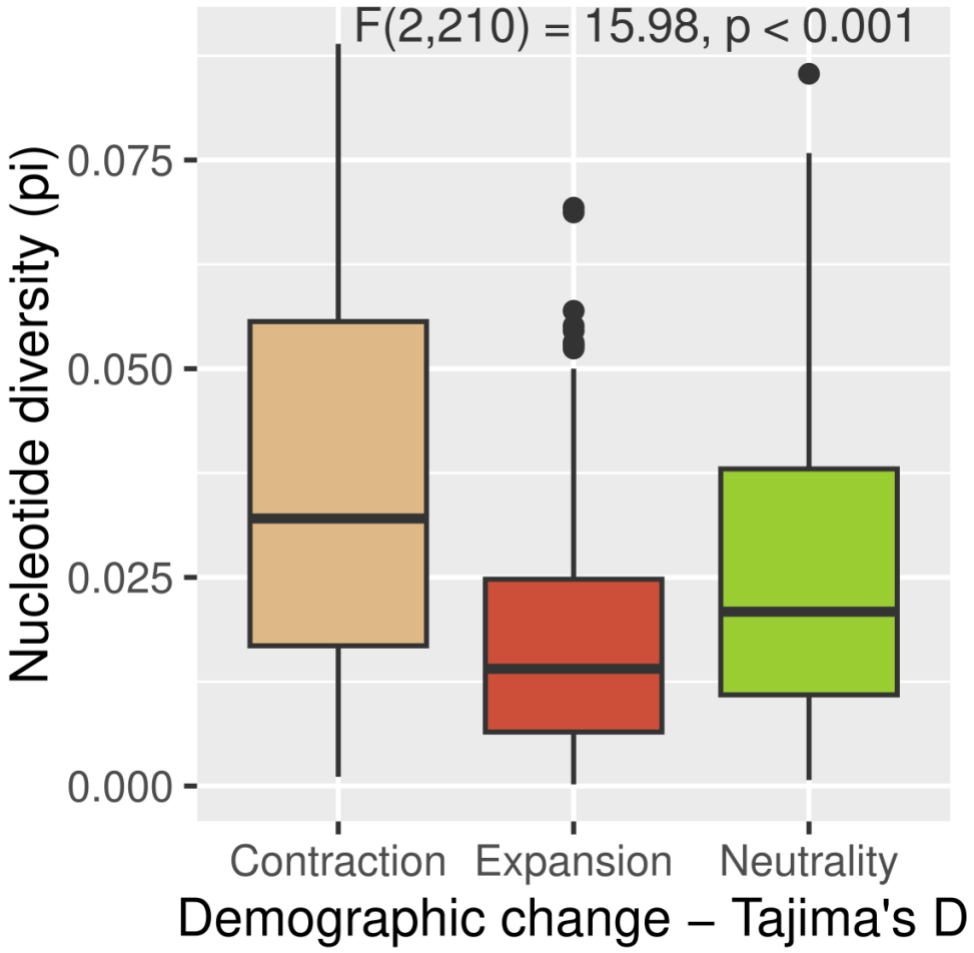


Figure S21. Relationship between nucleotide diversity (π) and Tajima's D (demographic change interpretation). Each boxplot indicates the median (bold lines), the interquartile range (boxes), and the observations (dots). The results of a linear model performed in R with the lm() function are shown.

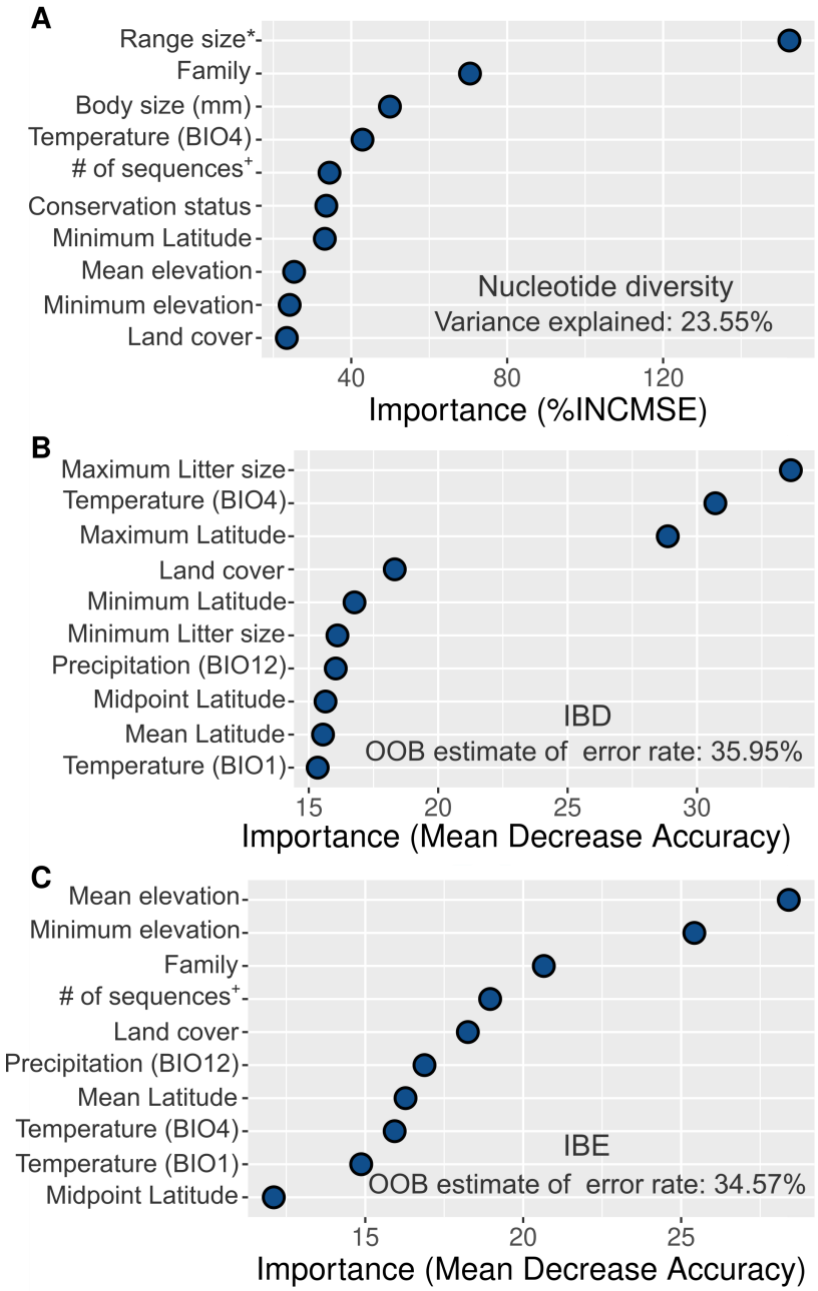


Figure S22. Random Forest (RF) results for A) nucleotide diversity (π, using RF regression), B) isolation-by-distance (IBD, using RF classification), and C) isolation-by-environment (IBE, using RF classification) with a subsampled dataset using between 10 and 20 sequences per species. All predictors except number of base pairs, and total range size were used in these RF models, but only the top 10 best predictors are shown here. Range size* = sampling range size based on localities with associated DNA sequences; # of sequences+ = number of sequences associated with geographic coordinates; mm = millimeters.

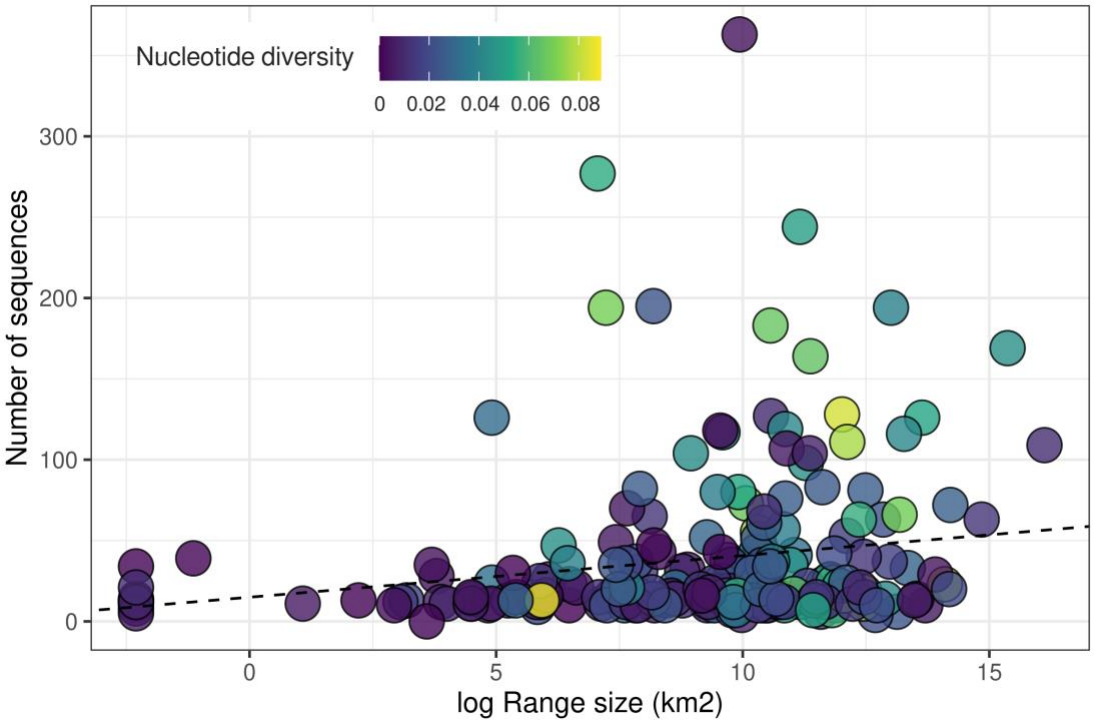


Figure S23. Relationship between total range size (log) and number of sequences; the colors of the dots indicate nucleotide diversity (π) in each species.

For Review