Investigating the relative role of dispersal and demographic traits in predictive phylogeography

Rilquer Mascarenhas[1] and Ana Carnaval[1]

[1]The City College of New York

October 10, 2023

Abstract

Many studies suggest that aside from environmental variables, such as topography and climate, species-specific ecological traits are relevant to explain the geographic distribution of intraspecific genetic lineages. Here, we investigated whether and to what extent incorporating such traits systematically improves the accuracy of random forest models in predicting genetic differentiation among pairs of localities. We leveraged available ecological datasets for birds and tested the inclusion of two categories of ecological traits: dispersal-related traits (i.e., morphology and foraging ecology) and demographic traits (such as species survival rate and generation length). We estimated genetic differentiation from published mitochondrial DNA sequences for 31 species of birds (1,801 total genetic samples, 526 localities) in the Atlantic Forest of South America. Aside from the aforementioned ecological traits, we included geographic, topographic and climatic distances between localities as environmental predictors. We then created models using all available data to evaluate model uncertainty both across space and across the different categories of predictors. Finally, we investigated model uncertainty in predicting genetic differentiation individually for each species (a common challenge in conservation biology). Our results show that while environmental conditions are the most important predictors of genetic differentiation, model accuracy largely increases with the addition of ecological traits. Additionally, the inclusion of dispersal traits improves model accuracy to a larger extent than the inclusion of demographic traits. Similar results are observed in models for individual species, although model accuracy is highly variable. We conclude that ecological traits improve predictive models of genetic differentiation, refining our ability to predict phylogeographic patterns from existing data. Additionally, demographic traits may not be as informative as previously hypothesized. Finally, prediction of genetic differentiation for species with conservation concerns may require further careful assessment of the environmental and ecological variations within the species range.

1. Introduction

Documenting and predicting spatial patterns of biodiversity at different scales remains the major goal of biogeography and a crucial step in proposing appropriate conservation policies (Cadotte and Tucker 2018, Burbano-Girón et al. 2022). Efforts to map and conserve intraspecific genetic diversity, in particular, are relevant whenever one's goal is to preserve evolutionary history (Tucker et al. 2019), maintain population connectivity (Schoville et al. 2018, Bracco et al. 2019) and ensure adaptation potential in the face of future environmental changes (Hoelzel et al. 2019). Identifying regions of lineage turnover within the spatial range of species is therefore important when delimiting conservation areas that account for cryptic genetic diversity (Crandall et al. 2000, D'Amen et al. 2013). However, performing such task for multiple species at a time requires extensive field work, time and resources - especially in megadiverse communities such as those in highly threatened tropical systems. Approaches that predict the distribution of genetic diversity, without the need of additional intensive fieldwork, are therefore desirable and may contribute novel insights to inform conservation (Manel and Holderegger 2013, Pollock et al. 2020, Green et al. 2022). For instance, predictive models of intraspecific genetic differentiation can be useful in conservation biology by summarizing genetic patterns within multiple co-distributed species in a region and by providing a map of genetic barriers within

a community. Moreover, if these models are both reliable and transferable across a group of species, one may be able to use them to predict the location of genetic breaks in species for which environmental and ecological information are available, but genetic data are scarce or inexistent. This would be the case of models built from community-level environmental, ecological, and genetic data, yet devoted to species-specific predictions for target, endangered, or data-deficient taxa for which molecular data are not widely available. With phylogeographic surveys increasing in numbers across the world (Hickerson et al. 2010), and community-level datasets becoming more common, these exercises are now possible.

Contributing toward this goal, evolutionary biologists have been generating models that predict levels of genetic diversity and connectivity between populations from environmental information, especially topographic and climatic gradients across space and time (van Strien et al. 2014, Brown et al. 2016, Espíndola et al. 2016). This is based on widespread observations that both climate and geography are highly correlated with the spatial distribution of genetic diversity (Carstens and Richards 2007, Carnaval et al. 2014, Cabanne et al. 2016). Not surprisingly, it has been shown that levels of genetic differentiation within a species are also impacted by ecological traits, especially those characteristics thought to correlate with dispersal capacity, such as morphological attributes (e.g., body size; Pabijan et al. 2012), reproductive strategies (Paz et al. 2015), habitat occupancy (Burney and Brumfield 2009) and foraging ecology (Miller et al. 2021). These elements of the ecology of species may be especially important for conservation genetics by helping us understand whether, how and why the patterns of genetic differentiation within one specific taxon may differ from the more common (or general) pattern detected in the overall community (Fortuna et al. 2009, Porto et al. 2013).

Less explored, however, are the roles of demographic traits such as fecundity, mortality and generation length, especially at large spatial scales. Such traits have been hypothesized to indirectly affect dispersal capacity through their effect on the number of individuals in a population, the number of offspring per generation, the local extinction rate and the frequency of reproductive cycles, all of which can influence the probability of dispersal events (Perry et al. 2005, Stevens et al. 2013, Castorani et al. 2017, Bonte and Dahirel 2017, Weil et al. 2022). For instance, dispersal distances were shown to be correlated with fast life-history strategies (i.e., high fecundity and low survival rates) in plants (Beckman et al. 2018). Additionally, it has been hypothesized that species with shorter generation length have more dispersal opportunities per time unit, which has been supported in butterflies (Stevens et al. 2012). However, tests of these relationships using genetic data are still sparse, mostly because demographic traits are costly to estimate for many taxa. Additionally, the question remains regarding to what extent these demographic traits are correlated with morphological and foraging ecology traits (e.g., body size), and therefore whether they would be informative in predictive models of genetic differentiation.

Here, we create a machine learning model that draws from landscape genetics and predictive phylogeography (Espíndola et al. 2016, Pelletier and Carstens 2018, Sullivan et al. 2019) to predict the magnitude of genetic differentiation among populations, using environmental descriptors and both dispersal-related and demographic traits within multiple co-distributed bird taxa in the Atlantic forest of Brazil. Machine learning techniques have been showing great promise in population genetics as a tool to leverage available data to understand and predict geographic patterns (Schrider and Kern 2018). We used such approach to combine published mtDNA data with available ecological datasets and evaluate model accuracy in predicting regions that concentrate high levels of genetic differentiation, representing phylogeographic breaks in this ecosystem. We specifically ask: 1) can this machine learning approach be used to accurately predict both global (assembly-wide) and species-specific genetic differentiation? 2) does the inclusion of species-specific ecological traits improve model accuracy? 3) what is the relative importance of dispersal traits and demographic traits in aiding model prediction? The first question aims to evaluate whether a machine learning approach can summarize existing information and learn enough about the spatial correlates of genetic breaks to predict which areas may function as barriers to gene flow in a focal region or for a focal species. The second question addresses whether features of the abiotic environment alone are enough to explain these correlations and to what extent data on ecological traits can help predictive approaches. Finally, the third question helps address the aforementioned knowledge gap about the correlation of demographic traits and dispersal.

2. Methods

2.1 *Study system and available genetic data*

We created models using available genetic, environmental and trait data for 31 species of birds from the Atlantic Forest of South America, a biogeographic region with an accumulating number of phylogeographic studies over the past decades (Peres et al. 2020). We summarized and retrieved all intra-specific mitochondrial DNA (mtDNA) data made available across 20 published studies to date (Table S1 in Supporting File 1). While not representing the entire genome, we decided to use mtDNA loci for the purpose of this study since they 1) have been largely used in phylogeographic studies over the past two decades (Hickerson et al. 2010), especially in the Atlantic Forest (Peres et al. 2020); 2) represent a good approximation of intraspecific genetic differentiation due to higher mutation rates and smaller effective population size (Avise et al. 2016). From each available study, we listed all mtDNA loci sequenced across all individuals, as well as their associated georeferenced locality and GenBank accession number. When the accession number was missing from the original manuscript, we performed a search of the GenBank database using the species name, locus, museum voucher and specimen ID as keywords, through the NCBI API function *entrez_search* implemented in the R package *rentrez*(Winter 2017). For sequences not found through this automated search, we performed a manual search on the NCBI website. Sequences were downloaded using the*entrez_fetch* function of the *rentrez* R package. Mitochondrial DNA loci included in this analyses are Cytochrome b (CytB), NADH dehydrogenase 2 (ND2) and the D-loop (or control) region. All sequence data were aligned per species and locus, using the Muscle algorithm (Edgar 2004) in the *muscle* R package.

Sequences with missing coordinates were georeferenced using Google Maps API within the R package *tidygeocoder*(Kahle and Wickham 2013). We performed each search using all provided information about the locality in each manuscript, which in most cases included a locality name, state and country. We then plotted all localities per species in geographic space to visualize possible errors, and manually corrected erroneous localities. Although this approach may not retrieve the exact location where specimens were collected, the accuracy of the geocoding tool is sufficient to guide our measurements of genetic differences given the spatial scale of the data (Singh 2017).

2.2 *Response and predictor variables*

To summarize intraspecific genetic differentiation, we calculated the average number of nucleotide differences (D_{XY}; Nei 1987) between pairs of localities for each locus alignment, in each species. D_{XY} was calculated with the R package *PopGenome*(Pfeifer et al. 2014). Pairwise values of D_{XY} were then used as the response variable in our models. Since different loci will yield different ranges of D_{XY} values due to different mutation rates, the information about the locus used to calculate this metric was included in our models as a control variable.

Environmental predictors were employed as distances between every pair of localities for which sequences were available. Specifically, we measured distance as the landscape resistance between the two localities in each pair: this estimates the effective resistance of the landscape to organism movement by calculating the relative probability of individuals to move from any one point in space to adjacent points, based on characteristics of the environment (McRae 2006). For that, all movement probabilities across a specific region were summarized in resistance matrices, and a resistance distance between two localities was then calculated as the average probability of individuals to move from one locality to another. We created resistance matrices for the Atlantic Forest based on topographic variation and 19 bioclimatic variables retrieved from the WorldClim database (Fick and Hijmans 2017) at a 2.5 arc-min resolution. Additionally, we created a resistance matrix to be a proxy of plain geographic distance by assuming equal probability of individuals moving in all directions in space (i.e., no effect of topography or environment variation). Resistance matrices were created using the function *transition* and resistance distances were calculated using the function *costDistance* , both from the R package *gdistance* (van Etten 2017).

To test the importance of ecological traits in the predictive models, we summarized phenotypic characteristics shown to be related to dispersal capacity (hereinafter referred to as dispersal traits) and reproductive rates

known to influence population dynamic (hereinafter referred to as demographic traits). Dispersal traits included three morphological measurements (body size, wing length and tarsus length), two descriptors of foraging ecology (foraging stratum and diet) and a measure of propensity to disperse through open areas (forest sensitivity). Foraging stratum and diet were transformed into ordinal variables to incorporate a gradient of ecological variation: values ranged from low to high to indicate a gradient from understorey towards exclusively canopy birds, and from diets consisting of one item (arthropods) to more generalized diets (i.e., arthropods along with other items). Diet based on nectar (observed only for *Thalurania glaucopis*) was quantified as the highest value to indicate an entirely different diet resource. Demographic traits included annual adult survival, age at first reproductive event, maximum longevity and generation length. Body size, foraging stratum and diet were obtained from the Handbook of Birds of the World (Billerman et al. 2022), whereas wing length and tarsus length were obtained from the AVONET database (Tobias et al. 2022). Forest sensitivity was obtained from (Stotz et al. 1996). Finally, demographic traits were obtained from (Bird et al. 2020). The values of all ecological traits of each species included in this study are summarized in Table 1.

2.3 *Model fitting and evaluation*

We created models using the Random Forest technique (Breiman 2001), a machine learning approach particularly efficient in investigating relationships when many predictor variables are present (Schrider and Kern 2018). This approach is well suited to predict values of genetic differentiation since 1) it has been shown to perform well in regression analyses to predict continuous data (which is the case of our response variable; Boehmke and Greenwell 2019, Barrow et al. 2021), and 2) it has been successfully used in previous studies predicting genetic breaks (Sullivan et al. 2019). We implemented the random forest algorithm in the R package *ranger*(Wright and Ziegler 2015). The set of parameters best suited for our dataset was estimated using the *tune* package (Kuhn 2023), which implements cross-validations to explore different combinations of parameters. We performed such exploration by combining different values for the following parameters: number of trees (varying from 1 to 2000), number of variables per tree (from 1 to 22) and minimum node size (from 2 to 40). To choose the best parameter values for our models, we calculated the root mean squared error for each parameter combination.

To test the efficiency of a random forest model in predicting pairwise D_{XY} values, we first created a global model by pooling together all pairwise locality data from all species. The goal of this model was to verify if one is able to predict levels of D_{XY} across the entire region and highlight regions of possible lineage turnover in space. To evaluate model uncertainty, we created 100 replicates of this global model. In each replicate, we retained 70% of the data points as a training dataset, and evaluated the model by performing predictions of the 30% data points left out as testing dataset. To avoid over-fitting, we performed what we called a "species cross-validation", i.e. we forced each species to be either entirely included or entirely excluded from the training dataset, forcing the model to always be evaluated on a set of species that was not present in the training dataset. We estimated variable importance using the impurity metric, which measures the change in prediction accuracy when a predictor variable is removed from a decision tree (Wright and Ziegler 2015). Since predictor variables may be correlated to different extents, we also calculated the Spearman correlation coefficient among all predictor variables and used that information to discuss their relative importance.

To test the transferability of the model and evaluate whether we can extrapolate correlations detected in a group of well-studied species to taxa with little or no available genetic data, we additionally created species-specific models to predict values of D_{XY} in each species. These models were created by leaving out the genetic data of one target species at a time and training our model on the remaining data. This approach differs from the one described above in that species-specific models directly evaluate how a model trained on all available data performs when predicting intraspecific genetic breaks within a single species instead of globally throughout the Atlantic Forest. We created one species-specific model for each combination of locus and species in our dataset (n = 37; Table 2).

Both approaches described above (global and species-specific models) were implemented with four different sets of predictor variables: 1) environmental data only ; 2) environmental data along with dispersal trait data; 3) environmental data along with demographic trait data; and 4) all available predictor variables (i.e.,

environmental data along with both types of ecological trait data). We compared the predictive power of these different sets of predictors by correlating observed and predicted values of D_{XY} and reporting the R^2 square of such correlations. For both global and species-specific models, we tested if the distributions of R^2 values across replicates (in global models) or species (in species-specific models) differed among models using different sets of predictors, by using a Kruskal-Wallis test. Additionally, we calculated the difference in R^2between models that included versus models that did not include ecological traits, and used a Wilcoxon Test to evaluate which set of traits (i.e., dispersal or demographic traits) led to a larger increase in R^2.

2.4 *Visualizing the location of modeled genetic barriers in space*

To visualize observed and predicted values of genetic differentiation in space, we mapped D_{XY} values from each pair of locality to the geographic point located at the center of the shortest path connecting those two sites (hereafter referred to as midpoint). Because we expect that high values of D_{XY} will occur whenever two localities are on opposite sides of a barrier to gene flow, the use of a color legend facilitates the identification of genetic breaks in space. For global models, we calculated and mapped for each midpoint the mean and standard deviation of the predicted values of D_{XY}, as well as the average difference between predicted and observed, across the total number of replicates (n of replicates = 100). For species-specific models, we plotted the observed and predicted value of D_{XY} for each species. To visualize values continuously over the landscape, we performed an inverse distance weighted (IDW) interpolation of D_{XY} values using the function *idw* from the package *gstat*(Gräler et al. 2016). Interpolated maps were created for observed values as well as values predicted by models using each different set of predictors. To focus on the relative levels of genetic differentiation across space, rather than the absolute values, we re-scaled the predicted values of D_{XY} in our interpolation based on the maximum and minimum values of observed D_{XY}. Additionally, in order to visualize genetic differentiation from different loci in the same map, we re-scaled values across loci to the same range. These re-scaling procedures allow us to remove the effect of different locus when visualizing genetic differentiation in geographic space, as well as emphasize the relative differences in model prediction across the area.

3. Results

The final dataset consisted of 1,801 individual sequences across 526 localities (Figure 1A). A total of 6,350 pairwise values of D_{XY} were derived from this dataset. D_{XY}values ranged from 0 to 56.33 (mean = 9.685; median = 4.5, see Figure S1 in Supporting File 1 for the range of D_{XY} values per locus for each species). When the observed D_{XY} values are plotted in space (Figure 1B), it becomes evident that genetic breaks (represented by midpoints between localities with relatively high values of D_{XY}) accumulate around three regions of the Atlantic Forest: 1) lowland valleys within the Serra do Mar mountain range and Paraíba do Sul river, in the southern range of the forest; 2) the Doce river and nearby regions; 3) northern regions near the São Francisco river.

Global models including only environmental predictors performed worse on average than models that included both environmental predictors and ecological traits (Figure 2A). Models based solely on environmental predictors had mean $R^2 = 0.14$ (ranging from 0.0007 to 0.45), whereas those included environmental and dispersal data had mean $R^2 = 0.53$ (ranging from 0.04 to 0.81) and those that included environment and demographic data had mean $R^2= 0.43$ (ranging from 0.003 to 0.77). Finally, models including environmental data and both types of ecological traits (i.e., dispersal and demographic traits) had mean $R^2 = 0.54$ (ranging from 0.06 to 0.81). A Kruskal-Wallis test suggests that the distribution of R^2 differs among all four sets of predictors ($X^2 = 170.81$, p -value < 0.01) and Wilcoxon tests suggest that all models that including traits have consistently higher predictive accuracy than models based solely on environmental data (p -value $<$ 0.001 for each set of predictors including ecological traits). In addition, the inclusion of dispersal traits led to a higher increase in R^2 values (when compared to models based solely on environmental data) than the inclusion of demographic traits (Figure 2B).

Correlation indexes across predictor variables revealed that geographic, topographic and bioclimatic resistance distances were highly correlated (Table S2). Additionally, body size was highly correlated with wing

length ($\rho = 0.873$) and adult survival ($\rho = 0.872$). Environmental distances, represented mainly by temperature seasonality and precipitation of coldest quarter, consistently had the highest impact in model accuracy (Figure 3). Morphological traits, represented mainly by wing length, were equally important whenever they were included. Adult survival and longevity were important ecological traits in models based solely on environmental data and demographic traits, but were surpassed by environmental data and dispersal traits whenever those were also present. Finally, the mtDNA locus used to calculate D_{XY}values was always present among the five most important variables across all models.

Species-specific predictions show a larger variation in R^2 within each set of predictors (values ranging from 0.0001 to 0.9; Figure 4). However, models including ecological traits tend to have higher mean R^2 (Table 2; Figure 5A). A Kruskal-Wallis test moderately supports that the distribution of R^2 differs among all four sets of predictors ($X^2 = 9.53$, p -value = 0.02). Similar to global models, Wilcoxon tests of R^2 values for species-specific models suggest that all models including traits have consistently higher predictive accuracy than models based solely on environmental data (p -value < 0.001 for each set of predictors including ecological traits). When considering only the model with highest predictive power for each combination of species and locus, it becomes clear that models including only environmental data tend to have low predictive power ($R^2 < 0.17$) even when they are the best model across the four sets of predictors (Figure 5B). An exception to this pattern is the Cytb dataset for species*Sclerurus scansor* , where the model based solely on environmental data simultaneously was the best model and showed high accuracy ($R^2 = 0.71$; Figure 4). Finally, similar to global models, the inclusion of dispersal traits led to a higher increase in R^2 values (when compared to models based solely on environmental data) than the inclusion of demographic traits (Figure S2).

Maps of the interpolated values of predicted D_{XY} reveal that, although models generally agree with maps of observed values (Figure 6A), model uncertainty is higher in the northern Atlantic Forest (hereinafter, northern AF), especially in models based solely on environmental data (Figure 6B). Additionally, models tend to overpredict genetic differentiation in northern AF (i.e., above the Doce River) and underpredict differentiation in the southern Atlantic Forest (hereinafter, southern AF; Figure 6C). Both over and underprediciton decreases when ecological traits are added.

4. Discussion

By incorporating data from multiple species in a supervised machine learning framework, we were able to explore different correlates of the spatial distribution of genetic differentiation as well as our ability to predict patterns across space. Model predictive accuracy showed large variation but was highly dependent on the set of predictors utilized: models including species-specific ecological traits led to consistently higher accuracy (Fig. 2). This result is in line with previous studies suggesting that species ecological characteristics interact with the abiotic environment in driving observed patterns (Burney and Brumfield 2009, Pabijan et al. 2012, Paz et al. 2015, Sullivan et al. 2019, Miller et al. 2021). Importantly, even though model accuracy increases when ecological traits are included, environmental predictors are still the most important variables in our models, particularly differences in temperature seasonality (bio4) and precipitation during cold periods (bio19; Fig. 3), which is highly correlated with additional variables describing temperature ranges and extremes of lower precipitation (Table S2). This suggests that genetic differentiation increases when the geographic cells along the path connecting two localities are similar in values of temperature range and precipitation extremes. This important environmental effect is not surprising since the biogeographic literature largely supports the abiotic environment as a major driver of spatial patterns on several scales (Davies et al. 2007, Stein et al. 2014, Voskamp et al. 2017, French et al. 2023). However, the increase in prediction accuracy (Fig. 2) clearly shows that the abiotic environment alone cannot account for all the observed intraspecific genetic variation. This is expected to be especially true on spatial scales where environmental variation is large, as in the present study (Peres et al. 2020). In these cases, the combination of abiotic predictors with species-specific traits helps decrease the amount of unexplained observed genetic variation.

We additionally found that different categories of ecological traits have varying predictive power. Dispersal traits, especially morphological measurements, were more informative than demographic traits in our pre-

dictive models (Fig. 2 and 3D). This makes sense considering the large support in the bird literature to the close relationship between body size and wing length with dispersal ability (e.g., Dawideit et al. 2009, Claramunt et al. 2012). Demographic traits also improve model accuracy, but to a lesser extent (Fig. 2), and we believe three aspects may explain their weaker effect on genetic differentiation: the spatial scale of our study, possible correlations with other predictors and their effect on changes in standing genetic variation rather than spatial differentiation. First, at the regional scale encompassed by our dataset, the effect of environmental differences (Manel and Holderegger 2013) and long range dispersal dictated mainly by morphology (Claramunt et al. 2012, Sheard et al. 2020, Claramunt 2021) may prevail over the effect of demographic traits, which can be more pronounced in local scales (Castorani et al. 2017, Drake et al. 2022), especially in regions with high environmental heterogeneity. Second, we find that survival was highly correlated to body size in our dataset (Table S2), and was also among the best predictors when morphological traits were not included (Fig. 3). This suggests that some of the biological importance of survival as a correlate of genetic differentiation may be accounted by body size, especially in models where both traits are included (Fig. 3D). We highlight that (Sullivan et al. 2019) found clutch size, a demographic trait, to be an important predictor of intraspecific divergence, but this trait is also thought to be correlated with body size models (Tuomi 1980, Ford and Seigel 1989, McGinley 1989, Sibly and Brown 2007, Werner and Griebeler 2011) and models where that trait was shown to be important did not include body size (Sullivan et al. 2019). In such cases, where dispersal and demographic traits are correlated, the relative contribution of the two categories is hard to disentangle. Finally, demographic traits may be more important to explain changes in effective population size, through their effect on demographic rates such as growth and recruitment rates (Saether et al. 2013, Waples 2016). Therefore, they may contribute mainly to the relative amount of genetic variation present in different populations (i.e, standing genetic variation) and contribute only indirectly to the relative differences in landscape connectivity across species.

Mapping model uncertainty (i.e., variance and error in predicted values) further allows us to discuss the relative importance of different drivers of genetic differentiation. Aside from the predictive variance inherent to the modeling procedure (Boehmke and Greenwell 2019) and to the stochasticity of the evolutionary process (Lenormand et al. 2009), we assume additional variance stems from the proportion of genetic variation that is not explained by predictors in our model. First, we observe that variance and error is higher when predictor traits are absent (Fig. 6), further emphasizing their relevance to predict genetic differentiation. Additionally, even in maps incorporating all of our predictors, variance and error is higher in the northern Atlantic Forest. We suggest two possible reasons for this result, the first being the absence of predictors that reflect past environmental conditions. Mitochondrial DNA genetic variation is expected to reflect relatively recent spatial and temporal changes in populations (Avise 2009), and it has been suggested that the distribution of genetic diversity in the northern region of the Atlantic Forest is better explained by past climate dynamics (Carnaval et al. 2014). In the framework we follow here, where genetic differentiation is calculated across pairs of localities, we believe incorporating past climatic conditions to explain current environmental differences is problematic because of the uncertainty in the past distribution of the species, which is expected to have suffered significant changes in the last 100 thousand years (Hofreiter and Stewart 2009, Baker et al. 2020). The past environmental distance between two present localities is not a good proxy of the effect of historical climate because individuals in each locality do not necessarily represent the genetic diversity observed in that locality in the past.

A second and equally plausible reason for higher uncertainty in northern AF is the fact that most sampled localities are distributed in the southern AF (Fig. 1A). In fact, southern AF (south of latitude 19 ^{o}S) encompasses 83% of the data points in our dataset. This means that even though we observe both low and high values of genetic differentiation across the entire region, most of the variation in our response variable is concentrated in the southern AF. This raises the question of whether the relationship between environmental and genetic data in southern AF (which dominate the training of our data) can be extrapolated to the northern AF. If that is a safe extrapolation, we could conclude that variance and predictive error in northern AF stems from undocumented phylogeographic structure. However, we believe that case is unlikely since spatial autocorrelation suggests these two regions will tend to have different environmental characteristics

(Keitt et al. 2002, Carnaval et al. 2014). We therefore believe that uncertainty in northern AF would mostly stem from lack of representation of that environment in our models. The same rationale can be applied to the variation in ecological traits across species: most species in our dataset are distributed entirely in the southern AF (Table 1). This means ecological differences across species might not be high in northern AF data points and therefore do not contribute to increasing model accuracy. Overall, these results point to the need for uniform geographic representation of genetic variation when implementing predictive models.

The high variation observed in predictive accuracy of species-specific models further emphasizes the relevance of evaluating the ability of the model to extrapolate learned relationships. In species-specific models, accuracy is dependent on how much of the environmental variation in the species range is present in the set of species in which the model was trained. We observed low accuracy prediction when there is little overlap of the species range with the ranges from species in the training dataset. That is the case, for instance, for *Cacicus chrysopterus* and *Synallaxis cinerea* , which occur in the southern extreme of the Atlantic Forest and in the Diamantina mountains, respectively, locations where few other species are sampled. Finally, we also observe low predictive accuracy in species where range outside of training combines with sparse geographic sampling (e.g.,*Phylloscartes ventralis* and *Poecilotriccus fumifrons*) or in small ranged species that have fewer points to be predicted (such as*Synallaxis cinerea*). Models based solely on environment still perform worse than those that include traits (Fig. 5B). Combined, these results suggest that the use of predictive models to infer distribution of genetic diversity in unsampled species require careful evaluation of how represented the species is within the training variation, and that information on morphological traits might still be relevant to increase prediction accuracy.

Our results highlight the relevance of balancing the goals of explanation and prediction in predictive biogeography: by exploring models with different sets of predictors, we show that environmental variation best explains genetic differentiation but is not enough to perform accurate predictions. Additionally, we show how mapping predictions and the related uncertainty allows for further investigation of model accuracy over space and gives directions to improve prediction. Finally, the goal of predicting is readily applicable to conservation biology. Policies aiming to create a network of preserved areas can use machine learning algorithms to predict areas of turnover and feed this information into approaches like systematic conservation planning (Margules and Pressey 2000, Nielsen et al. 2023). We suggest that, at least for birds, morphological traits should be included given their relevance for model accuracy. When the aim is to make predictions on a focal species based on all available data for a community, it is necessary to: 1) make sure the available data has good genetic sampling covering the area our focal species exist in; 2) include dispersal traits whenever possible to give more realistic predictions. Even though demographic traits did not lead to the highest observed increase in accuracy, they may also be included especially in species for which population connectivity is thought to be more correlated to life history strategies such as strong philopatry or unique social structures (Drake et al. 2022). As our results show, the use of machine learning approaches in predictive biogeography gains from incorporating extra predictor information but careful evaluation is needed to assess what type of information leads to the highest increase in prediction accuracy.

5. References

Avise, J. C. 2009. Phylogeography: retrospect and prospect. - J. Biogeogr. 36: 3–15.

Avise, J. C. et al. 2016. In the light of evolution X: Comparative phylogeography. - Proc. Natl. Acad. Sci. U. S. A. 113: 7957–7961.

Baker, P. A. et al. 2020. Beyond Refugia: New Insights on Quaternary Climate Variation and the Evolution of Biotic Diversity in Tropical South America. - In: Rull, V. and Carnaval, A. C. (eds), Neotropical Diversification: Patterns and Processes. Springer International Publishing, pp. 51–70.

Barrow, L. N. et al. 2021. Predicting amphibian intraspecific diversity with machine learning: Challenges and prospects for integrating traits, geography, and genetic data. - Mol. Ecol. Resour. 21: 2818–2831.

Batalha-Filho, H. and Miyaki, C. Y. 2016. Late Pleistocene divergence and postglacial expansion in the

Brazilian Atlantic Forest: multilocus phylogeography of Rhopias gularis (Aves: Passeriformes). - J. Zoolog. Syst. Evol. Res. 54: 137–147.

Batalha-Filho, H. et al. 2012. Phylogeography of an Atlantic forest passerine reveals demographic stability through the last glacial maximum. - Mol. Phylogenet. Evol. 65: 892–902.

Batalha-Filho, H. et al. 2019. Historical climate changes and hybridization shaped the evolution of Atlantic Forest spinetails (Aves: Furnariidae). - Heredity 123: 675–693.

Beckman, N. G. et al. 2018. High dispersal ability is related to fast life-history strategies. - J. Ecol. 106: 1349–1362.

Billerman, S. M. et al. 2022. Birds of the World - Cornell Laboratory of Ornithology.

Bird, J. P. et al. 2020. Generation lengths of the world's birds and their implications for extinction risk. - Conserv. Biol. 34: 1252–1261.

Bocalini, F. et al. 2021. Comparative phylogeographic and demographic analyses reveal a congruent pattern of sister relationships between bird populations of the northern and south-central Atlantic Forest. - Mol. Phylogenet. Evol. 154: 106973.

Boehmke, B. and Greenwell, B. M. 2019. Hands-On Machine Learning with R. - CRC Press.

Bolivar-Leguizamon, S. D. et al. 2020. Phylogeographic and demographic history of the Variable Anthsrike (Thamnophilidae: Thamnophilus caerulescens), a widespread South American passerine distributed along multiple environmental gradients. - Mol. Phylogenet. Evol.: 106810.

Bonte, D. and Dahirel, M. 2017. Dispersal: a central and independent trait in life history. - Oikos 126: 472–479.

Bracco, A. et al. 2019. Integrating physical circulation models and genetic approaches to investigate population connectivity in deep-sea corals. - J. Mar. Syst. 198: 103189.

Breiman, L. 2001. Random Forests. - Mach. Learn. 45: 5–32.

Brown, J. L. et al. 2016. Predicting the genetic consequences of future climate change: The power of coupling spatial demography, the coalescent, and historical landscape changes. - Am. J. Bot. 103: 153–163.

Burbano-Giron, J. et al. 2022. An assessment of spatial conservation priorities for biodiversity attributes: Composition, structure, and function of Neotropical biodiversity. - Biol. Conserv. 265: 109421.

Burney, C. W. and Brumfield, R. T. 2009. Ecology predicts levels of genetic differentiation in neotropical birds. - Am. Nat. 174: 358–368.

Cabanne, G. S. et al. 2008. Nuclear and mitochondrial phylogeography of the Atlantic forest endemic Xiphorhynchus fuscus (Aves: Dendrocolaptidae): biogeography and systematics implications. - Mol. Phylogenet. Evol. 49: 760–773.

Cabanne, G. S. et al. 2011. Evolution of Dendrocolaptes platyrostris (Aves: Furnariidae) between the South American open vegetation corridor and the Atlantic forest. - Biol. J. Linn. Soc. Lond. 103: 801–820.

Cabanne, G. S. et al. 2013. Matrilineal evidence for demographic expansion, low diversity and lack of phylogeographic structure in the Atlantic forest endemic Greenish Schiffornis Schiffornis virescens (Aves: Tityridae). - J. Ornithol. 154: 371–384.

Cabanne, G. S. et al. 2016. Effects of Pleistocene climate changes on species ranges and evolutionary processes in the Neotropical Atlantic Forest. - Biol. J. Linn. Soc. Lond. 119: 856–872.

Cabanne, G. S. et al. 2019. Phylogeographic variation within the Buff-browed Foliage-gleaner (Aves: Furnariidae: Syndactyla rufosuperciliata) supports an Andean-Atlantic forests connection via the Cerrado. - Mol. Phylogenet. Evol. 133: 198–213.

Cadotte, M. W. and Tucker, C. M. 2018. Difficult decisions: Strategies for conservation prioritization when taxonomic, phylogenetic and functional diversity are not spatially congruent. - Biol. Conserv. 225: 128–133.

Carnaval, A. C. et al. 2014. Prediction of phylogeographic endemism in an environmentally complex biome. - Proc. Biol. Sci.: 20141461

Carstens, B. C. and Richards, C. L. 2007. Integrating coalescent and ecological niche modeling in comparative phylogeography. - Evolution 61: 1439–1454.

Castorani, M. C. N. et al. 2017. Fluctuations in population fecundity drive variation in demographic connectivity and metapopulation dynamics. - Proc. Biol. Sci. 284: 10.1098/rspb.2016.2086.

Claramunt, S. 2021. Flight efficiency explains differences in natal dispersal distances in birds. - Ecology 102: e03442.

Claramunt, S. et al. 2012. High dispersal ability inhibits speciation in a continental radiation of passerine birds. - Proc. Biol. Sci. 279: 1567–1574.

Crandall, K. A. et al. 2000. Considering evolutionary processes in conservation biology. - Trends Ecol. Evol. 15: 290–295.

D'Amen, M. et al. 2013. Conservation of phylogeographic lineages under climate change. - Glob. Ecol. Biogeogr. 22: 93–104.

Dantas, G. P. M. et al. 2015. Population genetic structure of the Atlantic Forest endemic Conopophaga lineata (Passeriformes: Conopophagidae) reveals a contact zone in the Atlantic Forest. - J. Ornithol. 156: 85–99.

Davies, R. G. et al. 2007. Topography, energy and the global distribution of bird species richness. - Proc. Biol. Sci. 274: 1189–1197.

Dawideit, B. A. et al. 2009. Ecomorphological predictors of natal dispersal distances in birds. - J. Anim. Ecol. 78: 388–395.

d'Horta, F. M. et al. 2011. The genetic effects of Late Quaternary climatic changes over a tropical latitudinal gradient: diversification of an Atlantic Forest passerine. - Mol. Ecol. 20: 1923–1935.

Drake, J. et al. 2022. The value of considering demographic contributions to connectivity: a review. - Ecography: 10.1111/ecog.05552.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. - Nucleic Acids Res. 32: 1792–1797.

Espindola, A. et al. 2016. Identifying cryptic diversity with predictive phylogeography. - Proc. Biol. Sci. 283: 10.1098/rspb.2016.1529.

Fick, S. E. and Hijmans, R. J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas: new climate surfaces for global land areas. - Int. J. Climatol. 37: 4302–4315.

Ford, N. B. and Seigel, R. A. 1989. Relationships among Body Size, Clutch Size, and Egg Size in Three Species of Oviparous Snakes. - Herpetologica 45: 75–83.

Fortuna, M. A. et al. 2009. Networks of spatial genetic variation across species. - Proc. Natl. Acad. Sci. U. S. A. 106: 19044–19049.

French, C. M. et al. 2023. Global determinants of insect mitochondrial genetic diversity. - Nat. Commun. 14: 5276.

Graler, B. et al. 2016. Spatio-Temporal Interpolation using gstat. - The R Journal 8: 204–218.

Green, S. J. et al. 2022. Trait-based approaches to global change ecology: moving from description to prediction. - Proceedings of the Royal Society B: Biological Sciences 289: 20220071.

Hickerson, M. J. et al. 2010. Phylogeography's past, present, and future: 10 years after. - Mol. Phylogenet. Evol. 54: 291–301.

Hoelzel, A. R. et al. 2019. Conservation of adaptive potential and functional diversity. - Conserv. Genet. 20: 1–5.

Hofreiter, M. and Stewart, J. 2009. Ecological change, range fluctuations and population dynamics during the Pleistocene. - Curr. Biol. 19: R584–94.

Kahle, D. and Wickham, H. 2013. Ggmap: Spatial visualization with ggplot2. - R J. 5: 144.

Keitt, T. H. et al. 2002. Accounting for spatial pattern when modeling organism-environment interactions. - Ecography 25: 616–625.

Kuhn, M. 2023. Tune: Tidy tuning tools.: https://tune.tidymodels.org/.

Lenormand, T. et al. 2009. Stochasticity in evolution. - Trends Ecol. Evol. 24: 157–165.

Manel, S. and Holderegger, R. 2013. Ten years of landscape genetics. - Trends Ecol. Evol. 28: 614–621.

Margules, C. R. and Pressey, R. L. 2000. Systematic conservation planning. - Nature 405: 243–253.

Mascarenhas, R. et al. 2019. Late Pleistocene climate change shapes population divergence of an Atlantic Forest passerine: a model-based phylogeographic hypothesis test. - J. Ornithol. 160: 733–748.

Mata, H. et al. 2009. Molecular phylogeny and biogeography of the eastern Tapaculos (Aves: Rhinocryptidae: Scytalopus, Eleoscytalopus): cryptic diversification in Brazilian Atlantic Forest. - Mol. Phylogenet. Evol. 53: 450–462.

McGinley, M. A. 1989. The influence of a positive correlation between clutch size and offspring fitness on the optimal offspring size. - Evol. Ecol. 3: 150–156.

McRae, B. H. 2006. Isolation by resistance. - Evolution 60: 1551–1561.

Miller, M. J. et al. 2021. Demographic consequences of foraging ecology explain genetic diversification in Neotropical bird species. - Ecol. Lett.: 10.1111/ele.13674.

Nei, M. 1987. Molecular Evolutionary Genetics. - Columbia University Press.

Nielsen, E. S. et al. 2023. Molecular ecology meets systematic conservation planning. - Trends Ecol. Evol. 38: 143–155.

Pabijan, M. et al. 2012. Small body size increases the regional differentiation of populations of tropical mantellid frogs (Anura: Mantellidae). - J. Evol. Biol. 25: 2310–2324.

Paz, A. et al. 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.

Pelletier, T. A. and Carstens, B. C. 2018. Geographical range size and latitude predict population genetic structure in a global survey. - Biol. Lett. 14: 10.1098/rsbl.2017.0566.

Peres, E. A. et al. 2020. Patterns of Species and Lineage Diversity in the Atlantic Rainforest of Brazil. - In: Rull, V. and Carnaval, A. C. (eds), Neotropical Diversification: Patterns and Processes. Springer International Publishing, pp. 415–447.

Perry, A. L. et al. 2005. Climate change and distribution shifts in marine fishes. - Science 308: 1912–1915.

Pfeifer, B. et al. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. - Mol. Biol. Evol. 31: 1929–1936.

Pollock, L. J. et al. 2020. Protecting Biodiversity (in All Its Complexity): New Models and Methods. - Trends Ecol. Evol. 35: 1119–1128.

Porto, T. J. et al. 2013. Evaluating forest refugial models using species distribution models, model filling and inclusion: a case study with 14 Brazilian species. - Divers. Distrib. 19: 330–340.

Pulido-Santacruz, P. et al. 2016. Multiple evolutionary units and demographic stability during the last glacial maximum in the Scytalopus speluncae complex (Aves: Rhinocryptidae). - Mol. Phylogenet. Evol. 102: 86–96.

Raposo do Amaral, F. et al. 2013. Multilocus tests of Pleistocene refugia and ancient divergence in a pair of Atlantic Forest antbirds (Myrmeciza). - Mol. Ecol. 22: 3996–4013.

Raposo do Amaral, F. et al. 2021. Rugged relief and climate promote isolation and divergence between two neotropical cold-associated birds. - Evolution 75: 2371–2387.

Ribeiro, T. da S. et al. 2020. Life history and ecology might explain incongruent population structure in two co-distributed montane bird species of the Atlantic Forest. - Mol. Phylogenet. Evol.: 106925.

Saether, B.-E. et al. 2013. How life history influences population dynamics in fluctuating environments. - Am. Nat. 182: 743–759.

Schoville, S. D. et al. 2018. Preserving genetic connectivity in the European Alps protected area network. - Biol. Conserv. 218: 99–109.

Schrider, D. R. and Kern, A. D. 2018. Supervised Machine Learning for Population Genetics: A New Paradigm. - Trends Genet. 34: 301–312.

Sheard, C. et al. 2020. Ecological drivers of global gradients in avian dispersal inferred from wing morphology. - Nat. Commun. 11: 2463.

Sibly, R. M. and Brown, J. H. 2007. Effects of body size and lifestyle on evolution of mammal life histories. - Proc. Natl. Acad. Sci. U. S. A. 104: 17707–17712.

Singh, S. K. 2017. Evaluating two freely available geocoding tools for geographical inconsistencies and geocoding errors. - Open Geospatial Data, Software and Standards 2: 1–8.

Stein, A. et al. 2014. Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. - Ecol. Lett. 17: 866–880.

Stevens, V. M. et al. 2012. How is dispersal integrated in life histories: a quantitative analysis using butterflies. - Ecol. Lett. 15: 74–86.

Stevens, V. M. et al. 2013. Dispersal syndromes and the use of life-histories to predict dispersal. - Evol. Appl. 6: 630–642.

Stotz, D. F. et al. 1996. Neotropical Birds: Ecology and Conservation. - University of Chicago Press.

Sullivan, J. et al. 2019. Integrating life history traits into predictive phylogeography. - Mol. Ecol. 28: 2062–2073.

Tobias, J. A. et al. 2022. AVONET: morphological, ecological and geographical data for all birds. - Ecol. Lett. 25: 581–597.

Trujillo-Arias, N. et al. 2018. Forest corridors between the central Andes and the southern Atlantic Forest enabled dispersal and peripatric diversification without niche divergence in a passerine. - Mol. Phylogenet. Evol. 128: 221–232.

Trujillo-Arias, N. et al. 2020. Evolution between forest macrorefugia is linked to discordance between genetic and morphological variation in Neotropical passerines. - Mol. Phylogenet. Evol.: 106849.

Tucker, C. M. et al. 2019. Assessing the utility of conserving evolutionary history. - Biol. Rev. Camb. Philos. Soc. 94: 1740–1760.

Tuomi, J. 1980. Mammalian reproductive strategies: A generalized relation of litter size to body size. - Oecologia 45: 39–44.

van Etten, J. 2017. R Package gdistance: Distances and Routes on Geographical Grids. - Journal of Statistical Software 76: 1–21.

van Strien, M. J. et al. 2014. Landscape genetics as a tool for conservation planning: predicting the effects of landscape change on gene flow. - Ecol. Appl. 24: 327–339.

Voskamp, A. et al. 2017. Global patterns in the divergence between phylogenetic diversity and species richness in terrestrial birds. - J. Biogeogr. 44: 709–721.

Waples, R. S. 2016. Life-history traits and effective population size in species with overlapping generations revisited: the importance of adult mortality. - Heredity 117: 241–250.

Weil, S.-S. et al. 2022. Chameleon biogeographic dispersal is associated with extreme life history strategies. - Ecography 2022: 10.1111/ecog.06323.

Werner, J. and Griebeler, E. M. 2011. Reproductive biology and its impact on body size: comparative analysis of mammalian, avian and dinosaurian reproduction. - PLoS One 6: e28442.

Winter, D. J. 2017. rentrez: An R package for the NCBI eUtils API.

Wright, M. N. and Ziegler, A. 2015. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. - arXiv [stat.ML]

6. Figure captions

Figure 1. A) Map of all localities utilized in this study (black dots) along with the distribution of the Atlantic Forest (gray polygon). Blue lines represent three major rivers in the Atlantic Forest, from north to south: the Sao Francisco River, the Doce River and the Paraiba do Sul river. B) Interpolated observed (empirical) D_{XY} values across all midpoints used for the global model. The black arrow indicates the region of the Serra do Mar valley.

Figure 2. A) Values of R^2 for global models across different sets of predictors. B) Difference in values of R^2 between global models for each set of predictors that include ecological traits in comparison to models using only environmental predictors.

Figure 3. Distribution of importance values for each predictor variable across 100 replicates of global models. A) Models including environmental predictors only; B) models including environmental and dispersal predictors; C) models including environmental and demographic predictors. D) models including all predictors.

Figure 4. Values of R^2 for species-specific models, for each combination of species and locus utilized and each set of predictors. Dashed lines indicate the location in the graph where $R^2 = 0$ (red line) and $R^2 = 0.5$ (orange line). For each species and locus combination, four points are plotted across the x-axis, representing the four different sets of predictors.

Figure 5. A) Values of R^2 for species-specific models across all four sets of predictors and all combinations of species and locus. B) Values of R^2 for models with highest R^2 across the four sets of predictors within each combination of species and locus.

Figure 6. Interpolated predicted values of D_{XY} in the Atlantic Forest across 100 replicates of the global model. A) Mean of predicted D_{XY}; B) standard deviation of predicted D_{XY}; C) Difference between the predicted and observed values of D_{XY}.

7. Table captions

Table 1. Ecological traits information utilized for each species in this study.

Table 2. Values of R2 for species-specific models across all sets of predictors. Sample size and the best set of predictors for each model is included.

Hosted file

`Table_1.xlsx` available at https://authorea.com/users/673058/articles/671897-investigating-the-relative-role-of-dispersal-and-demographic-traits-in-predictive-phylogeography

Hosted file

`Table_2.xlsx` available at https://authorea.com/users/673058/articles/671897-investigating-the-relative-role-of-dispersal-and-demographic-traits-in-predictive-phylogeography