HaploVectors: an integrative analytical tool for phylogeography

Leandro Duarte[1], Jacqueline de Souza Lima[1], Renan Maestri[1], Vanderlei Debastiani[1] & Rosane Garcia Collevatti[2]*

[1] Departamento de Ecologia, Universidade Federal do Rio Grande do Sul
Av. Bento Gonçalves 9500 CP 15007, Porto Alegre, RS, 91501-970, Brazil

[2] Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal de Goiás. 74690-900, Goiânia, GO, Brazil

* E-mail corresponding author: *duarte.ldas@gmail.com*

Running head: Haplotypic eigenvectors for phylogeography

Abstract

Phylogeographic approaches are commonly used to understand historical-biogeographic patterns in the distribution of haplotypes. However, the emphasis of most tools lies on describing spatial patterns of genetic variation and assess how large are haplotypic differences among populations. An evaluation of the relative influence of environmental factors compared to pure neutral process of haplotypic distribution - a question of great interest for molecular ecologists - is less investigated, in part because appropriate tools are lacking. Here, we introduce HaploVectors, a flexible tool that allows exploring phylogeographical patterns and discriminating biogeographic, neutral and environmental factors acting to shape genetic distribution across space. Haplovectors are variables that summarize the major gradients of haplotypic distribution across a set of localities and allow weighting haplotypic frequencies by the number of mutational steps using a fuzzy weighting approach. HaploVectors is presented as an R package for computing haplotypic eigenvectors and performing null model-based tests. Investigation of HaploVectors using empirical datasets showed that the method is useful to uncover hidden patterns of haplotypic distribution, not easily detected using traditional methods. Using a plant species as study case, we demonstrate by means of HaploVectors that, even though the distribution of plant haplotypes was associated with different biogeographic regions of the Brazilian Cerrado biome, such association was not mediated by evolutionary relationships among haplotypes. The applicability of HaploVectors is broad, ranging from the pure pattern exploration and discrimination of genetic populations, to a hypothesis-testing framework that uses null-models to understand the influence of environmental factors on haplotypic distribution.

Keywords: Fuzzy set theory, haplotypic eigenvectors, haplotype network, principal coordinates of genetic structure

Introduction

A complete understanding of the historical and biogeographic patterns of species distribution benefits from the connection between micro and macroevolution, a major goal of the field of phylogeography (Avise, 1987; Avise, 2009). Since the middle 1990's, the number of studies using molecular markers to understand phylogeographical patterns is increasing at astonishing rates (Beheregaray, 2008; Turchetto-Zolet, Pinheiro, Salgueiro & Palma-Silva, 2013). Accordingly, the number of molecular markers used increased from single locus to multiple genome regions (Freeland, 2014; Blom, Horner & Moritz, 2016). Testing hypotheses about ecological influences on the genealogical history of populations from a single species, and the comparison of such patterns across multiple species (comparative phylogeography), has the potential to shed light on process of species diversification and on the geological and biogeographic connections of entire regions (Diniz-Filho et al., 2008; Carnaval et al., 2014). However, hypothesis testing in a rigorous statistical framework were only latter incorporated into phylogeography, with the advent of statistical phylogeographical approaches (Templeton et al., 1998; Papadopoulou & Knowles, 2016) that helped moving the field beyond the essential descriptive nature present in its infancy. Still, the development of analytical tools to integrate the ecological thinking into phylogeography has lagged behind the ever-increasing number of molecular loci discovered and the numerous tools focused on spatial genetics (e.g., Templeton, 2004; Miller, 2005; Epperson, 2005). In the era of multiple molecular markers and genomics in phylogeography, new analytical tools are imperative to better understand the increasingly complex phylogeographical patterns, to compare results from different loci, and to uncover environmental correlates of genetic distribution in the 'twilight zone' (Diniz-Filho et al., 2008).

A principal goal in studies of molecular ecology and phylogeography is to understand to what extent genetic variation of a species, expressed as the spatial distribution of alleles, genotypes and haplotypes, might be affected by different factors, such as dispersal mechanism, history of populations, climate changes or just by isolation by distance (Manel, Schwartz, Luikart & Tarbelet, 2003, Storfer et al., 2007, Avise, 2009). To explore processes underlying genetic diversity distribution, a representation of evolutionary relationships among organisms or populations by means of phylogenetic trees or networks is often performed. In this sense, networks are recognized to be more appropriate than trees, since the former allow visualization of reticulation events, such as hybridization and recombination (Kong, Sánchez-Pacheco & Murphy, 2015). Nonetheless, a network representation per se does not allow robust hypothesis tests to evaluate the interplay between genetic variation across space and alternative explanatory factors. Analysis of molecular variance (AMOVA, Excoffier, Smouse & Quattro, 1992) has been widely used for this purpose (Fitzpatrick, 2009; Maestri et al., 2016; Raffini et al., 2018). AMOVA, which is a permutation procedure akin to approaches often used in ecological studies, such as PERMANOVA (Pillar & Orlóci, 1996; Anderson, 2001) allows analyzing whether pairwise genetic dissimilarities between individuals distributed across different sites or regions defined by groups of sites is higher than expected by chance given within site (or region) dissimilarities. Note that AMOVA is based on overall genetic dissimilarities, and therefore does not permit direct inference about effects of spatial, environmental and/or biogeographic factors on the distribution of haplotypes across space.

Despite debates about different methods to construct haplotypes networks (see Mardulyn et al., 2012), phylogenetic median-joining network (MJN, Bandelt, Forster & Röhl, 1999) is the method normally used in phylogeographical studies. The MJN method is based in the minimum spanning trees and shows the relationships between haplotypes obtained by the distance measured among them (e.g. number of character differences - Hamming distance). The use of MJN has been grown exponentially since its development (Kong et al., 2015). Moreover, the network representation used to explore evolutionary relatedness among haplotypes do not allow neither a clear visualization of haplotype co-occurrence within sites nor general trends in haplotype distribution across space.

Phylogenetic eigenvectors have been used to express the variation of phylogenetic beta diversity (or simply phylobetadiversity) among an array of localities (Duarte, 2011; Duarte, Debastiani, Freitas & Pillar, 2016) based on phylogenetic fuzzy weighting (Pillar & Duarte, 2010). Phylogenetic fuzzy weighting allows describing sites by their phylogenetic composition based on species incidences/abundances and the phylogenetic relatedness among species (see also Duarte et al., 2016). The phylogenetic composition of a set of sites can be thereby decomposed into independent phylogenetic eigenvectors using Principal Coordinates of Phylogenetic Structure, or simply PCPS (Duarte, 2011; Duarte et al., 2016). Each eigenvector describes the sites by scores that position them along a phylogenetic gradient expressing a fraction of the total phylogenetic relatedness among the species distributed across the sites. By doing so, PCPS analysis renders single variables that synthesize phylogenetic gradients, and thereby can be used to analyses addressing environmental, spatial or biogeographic determinants of phylogeny-mediated species distribution (Duarte, Bergamin, Marcilio-Silva, Seger & Marques, 2014; Carlucci et al., 2016).

In this study we introduce HaploVectors, a new flexible analytical tool for molecular ecologists to disentangle biogeographic or environmental factors driving haplotypic distribution across space. For this, HaploVectors extends the application of phylogenetic fuzzy weighting in order to describe the distribution of haplotypes across sets of localities, allowing flexible exploratory analysis. Moreover, by applying appropriate null models, HaploVectors allows to analyze multiple determinants of the frequencies of haplotypes, as well as the number of mutations separating different haplotypes, across sets of localities, providing robust hypotheses tests for phylogeographical studies. We demonstrate the application of HaploVectors in two empirical datasets.

Materials and Methods

Haplotypic eigenvectors

Haplotypic eigenvector analysis and hypotheses tests based on null models were implemented in the R package *HaploVectors* (freely available at https://github.com/vanderleidebastiani/HaploVectors). The function 'HaploVectors' allows defining the frequency of each haplotype across a set of localities from where the individuals were sampled and to weight those frequencies across localities according to the number of mutational steps between all pairs of haplotypes arranged into a network based on fuzzy set theory (Zadeh, 1965). Further, the function implements null model tests to analyze the influence of environmental, biogeographic or spatial variables on the distribution of haplotypes across sets of localities, as well as to estimate to what extent such influence is mediated by the evolutionary distance between haplotypes. The rationale underlying the HaploVectors approach, including the null model tests, was adapted from phylogenetic fuzzy weighting (Pillar & Duarte, 2010), a method originally developed to analyze multiple determinants of phylogenetic composition across metacommunities based on fuzzy set theory (see details on the method in Duarte et al., 2016).

The first analytical step implemented in HaploVectors consists of defining haplotypes for a set of samples and computing the frequency (number of individuals) of each haplotype per locality. For this, two input datasets are required: (1) a *.fas file containing aligned genetic sequences for each sampled individual, and (2) a matrix describing the incidence of each individual (rows) in a given locality (column). Based on these two datasets, the function 'HaploNetDist' extracts the haplotypes for the *.fas file using the function 'haplotype', and computes an haplotypic network using the function 'haploNet', functions originally implemented in the R package *pegas* (Paradis, 2010).

Further, 'HaploNetDist' computes the frequency of each haplotype per locality (matrix **W**), and submit **W** to fuzzy weighting as follows: From the matrix **D** describing all pairwise distances between haplotypes based on the number of mutational steps between any pair of haplotypes, remove all distances between haplotypes *not* connected in the network, replacing them by the mutations separating two haplotypes in the network plus one. This procedure generates matrix $\mathbf{D}_{\mathrm{N}}$, which reproduces haplotype connections described in the network. Matrix $\mathbf{D}_{\mathrm{N}}$ is further converted into a similarity matrix **S** describing the similarities between all pairs of haplotypes *i* and *j* (s_{ij}, ranging between 0 and 1), as follows:

$$s_{ij} = 1 - \left(\frac{d_{ij}}{\max\left(d_{ij}\right)} \right)$$

d_{ij} is the number of mutations separating the haplotypes *i* and *j* in $\mathbf{D}_{\mathrm{N}}$, and max (d_{ij}) is the maximum number of mutations between any pair of haplotypes in the network. Matrix **S** is then standardized by marginal total within columns, generating a matrix **Q** that describes fuzzy belongings between haplotypes (Pillar & Duarte, 2010; Duarte et al., 2016). Matrix **Q**, containing haplotypic fuzzy sets are then employed to weight the frequencies of haplotypes per locality described in matrix **W** by their evolutionary relatedness, generating matrix **P**, which describes localities by their haplotypic composition, that is, haplotype frequencies per locality weighted by their evolutionary relatedness. The output of HaploVectors function provides matrices **W**, **D**, $\mathbf{D}_{\mathrm{N}}$, **Q** and **P**.

Performing principal coordinates on **P** generates haplotypic eigenvectors, which decompose the total variation in the haplotypic composition across the set of localities into independent fractions proportional to its respective eigenvalue λ. Those eigenvectors representing the higher amount of variation in **P** can be used to explore major trends in haplotype distribution across the localities. Those localities sharing most haplotypes will show similar scores, and therefore will group to each other in the scatter plot. Thus, this scatter plot allows simultaneously explore evolutionary links among haplotypes and localities.

The function 'HaploVectors' also allows analyzing multiple environmental, biogeographic or spatial determinants of haplotype distribution across a set of localities, and therefore is a useful tool for robust hypothesis test in phylogeography. For this, the function implements two null model tests, adapted from Duarte et al. (2016) and designed to test the following hypotheses:

Hypothesis 1: *A given environmental, biogeographic or spatial factor* ***E*** *affects the distribution of haplotypes across a set of localities.* This hypothesis is tested by means of a null model called *site shuffle*, which is a classical permutation-based procedure assuming independence between haplotypes and localities. The test can be performed based on either haplotypic dissimilarities between localities computed based on **P** (hereafter D_{P}) using an appropriate resemblance measure, such as Euclidean distances or Bray-Curtis dissimilarities (Legendre & Legendre, 2012), or directly on single haplotypic eigenvectors. The test consists of 1) computing a F_{Obs} statistic. If D_{P} is modeled on **E**, the test is based on a dissimilarity regression on distance matrices (hereafter called ADONIS; see McArdle & Anderson, 2001). For single haplotypic eigenvectors modeled on **E**, the test is an ordinary linear squares (OLS) model; 2) freely permuting the localities a number of times (say 999); 3) at each permutation step, computing F_{null}; and 4) defining the probability of obtaining the observed statistic by chance ($H_0 = F_{Obs} \leq F_{null}$), as the proportion of permutations in which F_{null} exceeded F_{Obs}. Using this procedure, the test simultaneously addresses the influence of **E** on the distribution of haplotypes across the localities (the number of haplotypes shared between pairs of localities) and to what extent such influence is mediated by the evolutionary relatedness between the haplotypes (the number of mutational steps between the pairs of haplotypes). Therefore, even if this first hypothesis is corroborated (H_0 is rejected), such test does not allow us to conclude that the influence of **E** on haplotype distribution is or is not dependent on the evolutionary relatedness among haplotypes. To accomplish that, it is necessary to test a second hypothesis, which is conditioned on the validity of the first one:

Hypothesis 2: *The influence of **E** on the distribution of haplotypes across a set of localities depends on the evolutionary relatedness among them.* To test this hypothesis, a second round of permutations (*network shuffle*) is needed in order to compute F_{null}. Accordingly, the frequency of haplotypes in **W** is kept constant across the localities while evolutionary relatedness between them is permuted by haplotype label shuffling (Kembel et al., 2010). After computing F_{Obs} (step 1, as described for site shuffle), the procedure consists of 2) freely permuting haplotype labels in the network to generate random evolutionary relatedness among haplotypes and computing null matrices **D**, $\mathbf{D_N}$, **Q** and **P**. The procedure is repeated 999 times; 3) at each permutation procedure, computing null $D_\mathbf{P}$ and, if necessary, null haplotypic eigenvectors. In this later case, null haplotypic eigenvectors are submitted to procrustean adjustment (Jackson, 1995) and fitted values between observed and null eigenvectors are obtained; 4) take null $D_\mathbf{P}$ or selected adjusted null eigenvectors as response variable in ADONIS or OLS, respectively, using **E** as predictor, and compute F_{null} values; 5) generating a set of F_{null} to get the probability under the null hypothesis ($H_0 = F_{Obs} \leq F_{null}$); 6) defining a probability under the null hypothesis.

By performing both null model tests, two probability values are generated. Previous analyses using simulated data demonstrated that both null models show appropriate type I error and statistical power (Duarte et al., 2016). Using site shuffle, whenever the null hypothesis is rejected, we conclude that **E** affects the distribution of haplotypes across a set of localities (hypothesis 1). Then we proceed to test for the hypothesis 2 (via network shuffle). If the null hypothesis is rejected, we conclude that the influence of **E** on the distribution of haplotypes across the localities depends on the evolutionary relatedness among them. In Fig. 1 we illustrate the expected distribution of haplotypes and the respective probabilities generated under site and network shuffle models.

Application using empirical datasets

We demonstrate the application of HaploVectors in phylogeographical analyses through two empirical data sets: a set of cpDNA sequences from 333 adult individuals of *Eugenia dysenterica*, a tree species from Myrtaceae family, sampled from 23 localities (Lima, Telles, Chaves, Lima-Ribeiro & Collevatti, 2017), and a set of cpDNA sequences from 257 adult individuals of *Mauritia flexuosa*, a palm tree sampled from 26 localities (Lima, Lima-Ribeiro, Tinoco, Terribile & Collevatti, 2014). Both data sets are available at GenBank (accession numbers: MF752706- MF753038 and KC527837-KC528609, respectively).

Dataset 1: *Eugenia dysenterica*

Using a phylogeographical approach, Lima et al. (2017) investigated the spatial pattern of genetic diversity on *E. dysenterica*, a widely distributed tree species in the Brazilian savanna. Four regions of the chloroplast were sequenced from individuals sampled at 23 populations throughout its distribution (for details about sampling and genetic data see Lima et al*.*, 2017). The authors inferred the phylogenetic relationships among haplotypes using median-joining network analysis. Furthermore, Analysis of Molecular Variance (AMOVA) was used to analyze spatial patterns of genetic variation among biogeographic regions of the Cerrado biome (Table 1). Although AMOVA pointed out genetic differentiation among sites located at three different Cerrado regions (Central, Northeast and Southeastern; $P < 0.001$), the results of the network analysis visually suggested that the phylogenetic relationships among haplotypes did not match the geographical distribution of the lineages (Fig. 2a).

We analyzed the variation in the distribution of haplotypes across the biogeographic regions of the Cerrado using HaploVectors approach. Our hypotheses propose that (1) the spatial distribution of haplotypes varies among the different biogeographic regions of the Cerrado biome, and that (2) the biogeographic distribution of haplotypes across the Cerrado regions depends on the evolutionary relationships among them. We tested these hypotheses using ADONIS, based on haplotypic dissimilarities between localities (D_{P}), and OLS using haplotypic eigenvectors. For ADONIS, we computed matrix **P** using log-transformed frequencies of haplotypes per site (matrix **W**) and square-rooted Hamming distances between haplotypes (matrix $\mathbf{D}_{\mathrm{N}}$). Haplotypic dissimilarities between sites were computed using Euclidean distances. The same three biogeographic regions of Cerrado were taken as a categorical predictor in the analysis (matrix **E**). For OLS, we first computed haplotypic eigenvectors (haplovectors) based on D_{P}. Haplovectors containing more than 5% of total information in **P** were taken as response variables in linear models, while **E** was used as predictor.

Dataset 2: *Mauritia flexuosa*

Mauritia flexuosa is a dioecious palm species distributed widely across South America, occurring in Brazilian savannas and Amazonia (Lima et al., 2014). Because its occurrence is associated with wetlands, Lima et al. (2014) investigated spatial patterns in chloroplast genome regions among populations of *M. flexuosa* occurring across four different river basins (Amazon, Araguaia/Tocantins, Paraná and São Francisco). For this purpose, the authors estimated phylogenetic relationships among haplotypes using the median-joining network analysis and performed a hierarchical analysis of molecular variance (AMOVA) to analyze the genetic differentiation among populations from different river basins. The haplotypic network (Fig. 2c) did not allow a clear congruence between the geographic distribution of phylogenetic lineages and river basins, although AMOVA found significant genetic differentiation among river basins (Table 1). For this dataset we performed similar analyses as described for *E. dysenterica*.

Results

For analyses performed using *E. dysenterica* dataset, ADONIS indicated that sites occurring at the same biogeographic region of the Cerrado biome share more haplotypes with each other than with sites located at different regions ($P_{\text{site shuffle}} = 0.001$, Table 1); nonetheless, such difference in haplotype composition is not related with evolutionary relatedness among haplotypes ($P_{\text{network shuffle}} = 0.973$). The two first haplovectors (Fig. 2b), containing 32% and 17% of total variation in haplotype composition of sites, respectively, indicated association between haplotype distribution across sites and biogeographic regions ($P_{\text{site shuffle}} < 0.02$). The first haplovector shows a separation of populations from Southeast Brazil (green circles) from other regions (Fig. 2b), which relies only on haplotype composition, but not evolutionary relatedness among haplotypes ($P_{\text{network shuffle}} = 0.936$). On the other hand, the second haplovector discriminated some populations from Northeast Brazil from other regions (Fig. 2b). In this case, an evolutionary signal in the association between haplotype composition of sites and biogeographic regions was detected ($P_{\text{network shuffle}} = 0.024$), based mostly on the evolutionary path from haplotype one to 19 to 11 (Figs. 2a, 2b).

For the *Mauritia flexuosa* dataset, ADONIS rejected null hypothesis for both site shuffle and network shuffle null models (Table 1), indicating that haplotype composition differed between river basins ($P_{\text{site shuffle}} = 0.01$), and that such difference was mediated by evolutionary relatedness between haplotypes ($P_{\text{network shuffle}} = 0.005$). For this dataset, only the first haplovector (Fig. 2d), containing 42% of total variation in haplotype composition of sites, indicated association between haplotype distribution across sites and river basins ($P_{\text{site shuffle}} = 0.021$), which was mediated by evolutionary relatedness between haplotypes ($P_{\text{network shuffle}} = 0.007$). The first haplovector showed a clear separation between populations located in Amazon (left side of the plot in Fig. 2d), which were related to haplotypes seven and eight (Fig. 2c), and Araguaia/Tocantins basins (right side of the plot in Fig. 2d), mostly associated with haplotypes one to five.

Discussion

Current implemented methods for phylogeographical analyses treat haplotypic frequency across localities over a given environmental gradient/factor and phylogenetic relationships among those haplotypes in disconnected manners, lacking a clear conceptual framework to integrate both. Haplovectors provide such integration, allowing disentangling the environmental or biogeographic influence on haplotypic distribution and assessing whether that distribution is resulting from the evolutionary relationship among haplotypes. In cases where AMOVA (Excoffier et al., 1992) and analyses using haplotype networks (e.g., Templeton, 1998) reveal contradicting results of haplotypic distributions, we propose that HaploVectors can elucidate the conundrum.

In the first example (*E. dysenterica* dataset) we found that the influence of biogeographic regions at structuring haplotypes is independent from the evolutionary relatedness among them (i.e. the number of mutational steps separating haplotypes). This shows that biogeographic regions are indeed structuring haplotypes (i.e. different haplotypes can be found in different regions, implying few haplotypes occurring in more than one region), however, inside any given biogeographic region, haplotypes are not the closely related to each other based on the phylogenetic relationships among haplotypes. This means that phylogenetic closely related haplotypes occur in distinct biogeographic regions, and each biogeographic region comprises exclusive haplotypes from multiple evolutionary origins.

Different from that observed in *E. dysenterica* dataset, for the *M. flexuosa* dataset we found that haplotype composition differed between river basins and this difference was associated with the evolutionary relatedness among haplotypes. This reveals that haplotypes that occur in the same river basins are more phylogenetically related than those that occur in different river basins. These interpretations provided by HaploVectors solve the apparent paradox found in the results of previous analyses in both cases: the AMOVA found haplotypic differences among biogeographic regions and river basins; however, the haplotypic network failed to revel a clear structured haplotypic distribution over the same regions. The null model tests implemented in HaploVectors permit treating the haplotype frequency across localities on a given environmental factor, independently from the phylogenetic similarities among haplotypes. The combination of both tests in a joint approach allows for tracing a complete picture of the evolutionary history of populations.

We hope that this approach will be useful in all cases where the distribution of haplotypes is hypothesized to be under the influence of an environmental or biogeographic factor. These questions are likely to be encountered with increasingly frequency by molecular ecologists and phylogeographers.

Acknowledgements

The authors thank Jorge Sebastião Bernardo-Silva for valuable discussion on the topic addressed in this paper. LDSD research activities have been supported by CNPq Productivity Fellowship (grant 307886/2015-8). RGC research activities have been supported by CNPq Productivity Fellowship. JSL receives a fellowship from CNPq (project 439051/2016-9). The authors have no conflict of interests to declare. LDSD, JSL and RGC are members of the National Institutes for Science and Technology (INCT) in Ecology, Evolution and Biodiversity Conservation, supported by MCTIC/CNPq (proc. 465610/2014-5) and FAPEG.

References

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*, 32-46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Avise, J. C., Arnold, J. & Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., … Saunders, N.C. (1987). Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, *18*, 489-522. doi: 10.1146/annurev.es.18.110187.002421

Avise, J. C. (2009). Phylogeography: retrospect and prospect. *Journal of Biogeography*, *36*, 3-15. doi: 10.1111/j.1365-2699.2008.02032.x

Bandelt, H. J., Forster, P. & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, *16*, 37-48. doi: 10.1093/oxfordjournals.molbev.a026036

Beheregaray, L. B. (2008). Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere. *Molecular Ecology*, *17*, 3754-3774. doi: 10.1111/j.1365-294X.2008.03857.x

Blom, M. P. K., Horner, P. & Moritz, C. (2016). Convergence across a continent: adaptive diversification in a recent radiation of Australian lizards. *Proceedings of the Royal Society B: Biological Sciences*, *283*, 20160181. doi: 10.1098/rspb.2016.0181

Carlucci, M. B., Seger, G. D. S., Sheil, D., Amaral, I. L., Chuyong, G. B., Ferreira, L. V., … Duarte, L. S. (2016). Phylogenetic composition and structure of tree communities shed light on historical processes influencing tropical rainforest diversity. Ecography, *40*, 521-530. doi: 10.1111/ecog.02104

Carnaval, A. C., Waltari, E., Rodrigues, M. T., Rosauer, D., VanDerWal, J., Damasceno, R., ... Moritz, C. (2014) Prediction of phylogeographic endemism in an environmentally complex biome. *Proceedings of the Royal Society B: Biological Sciences*, *281*, 20141461. doi: 10.1098/rspb.2014.1461

Diniz-Filho, J. A. F., Telles, M. P. C., Bonatto, S. L., Eizirik, E., Freitas, T. R. O., De Marco Jr, P., … Soares, T. N. (2008). Mapping the evolutionary twilight zone: molecular markers, populations and geography. *Journal of Biogeography*, *35*, 753-763. doi: 10.1111/j.1365-2699.2008.01912.x

Duarte, L. D. S. (2011). Phylogenetic habitat filtering influences forest nucleation in grasslands. *Oikos*, 120, 208-215. doi: 10.1111/j.1600-0706.2010.18898.x

Duarte, L. D. S., Bergamin, R. S., Marcilio-Silva, V., Seger, G. D. S. & Marques, M. C. M. (2014). Phylobetadiversity among forest types in the Brazilian Atlantic Forest complex. PLoS ONE 9:e105043. doi: 10.1371/journal.pone.0105043

Duarte, L. D. S., Debastiani, V. J., Freitas, A. V. L. & Pillar, V. D. (2016). Dissecting phylogenetic fuzzy weighting: theory and application in metacommunity phylogenetics. *Methods in Ecology and Evolution*, 7, 937-946. doi: 10.1111/2041-210X.12547

Epperson, B. K. (2005). Estimating dispersal from short distance spatial autocorrelation. *Heredity*, *95*, 7-15. doi: 10.1038/sj.hdy.6800680

Excoffier, L., Smouse, P. E. & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, *131*, 479-491.

Fitzpatrick, B. M. (2009). Power and sample size for nested analysis of molecular variance. *Molecular Ecology*, *18*, 3961-6. doi: 10.1111/j.1365-294X.2009.04314.x

Freeland, J. R. (2014). *Molecular Ecology*. In: eLS. John Wiley & Sons, Ltd: Chichester. doi: 10.1002/9780470015902.a0003268.pub2

Jackson, D. A. (1995). Protest: A Procrustean Randomization test of community environment concordance. *Ecoscience.*, *2*, 297-303. doi: 10.1080/11956860.1995.11682297

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., … Webb, C. O. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, *26*, 1463-4. doi: 10.1093/bioinformatics/btq166

Kong, S., Sánchez-Pacheco, S. J. & Murphy, R. W. (2015). On the use of median-joining networks in evolutionary biology. *Cladistics*, *32*, 691-699. doi: 10.1111/cla.12147

Legendre, P. & Legendre, L. (2012) *Numerical Ecology* (3rd ed.). Elsevier, Amsterdam.

Lima, N. E., Lima-Ribeiro, M. S., Tinoco, C. F., Terribile, L. C. & Collevatti, R. G. (2014). Phylogeography and ecological niche modelling, coupled with the fossil pollen record, unravel the demographic history of a Neotropical swamp palm through the Quaternary. *Journal of Biogeography*. doi: 10.1111/jbi.12269

Lima, J. S., Telles, M. P. C., Chaves, L. C., Lima-Ribeiro, M. S. & Collevatti, R. G. (2017). Demographic stability and high historical connectivity explain the diversity of a savanna tree species in the Quaternary. *Annals of Botany*, *119*, 645-657. doi: 10.1093/aob/mcw257

Maestri, R., Fornel, R., Gonçalves, G. L., Geise, L., Freitas, T. R. O. & Carnaval, A. C. (2016). Predictors of intraspecific morphological variability in a tropical hotspot: comparing the influence of random and non-random factors. *Journal of Biogeography*, *43*, 2160-2172. doi: 10.1111/jbi.12815

Mardulyn, P. (2012). Trees and⁄or networks to display intraspecific DNA sequence variation? *Molecular Ecology*, 21, 3385-3390. doi: 10.1111/j.1365-294X.2012.05622.x

Manel, M., Scwartz, M. K., Luikart, G. & Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, *18*, 189-197. doi: 10.1016/S0169-5347(03)00008-9

McArdle, B. H. & Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, *82*, 290-297. doi: 10.1890/0012-9658(2001)082[0290:FMMTCD]2.0.CO;2

Miller, M. P. (2005). Alleles in Space: computer software for the joint analysis of interindividual spatial and genetic information. *Journal of Heredity*, *96*, 722-724. doi: 10.1093/jhered/esi119

Papadopoulou, A. & Knowles, L. L. (2016). Toward a paradigm shift in comparative phylogeography driven by trait-based hypotheses. *PNAS*, *113*, 8018-8024. doi: 10.1073/pnas.1601069113

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

Pillar, V. & Orlóci, L. (1996). On randomization testing in vegetation science: multifactor comparisons of relevé groups. *Journal of Vegetation Science*, *7*, 585-592. doi: 10.2307/3236308

Pillar, V. & Duarte, L. S. (2010). A framework for metacommunity analysis of phylogenetic structure. *Ecology Letters*, 13, 587–596. doi: 10.1111/j.1461-0248.2010.01456.x

Raffini, F., Fruciano, C. & Meyer, A. (2018). Morphological and genetic correlates in the left–right asymmetric scale-eating cichlid fish of Lake Tanganyika. *Biological Journal of the Linnean Society*, *124*, 67-84. doi: 10.1093/biolinnean/bly024

Storfer, A., Murphy, M. A., Evans, J. S., Goldberg, C. S., Robinson, S., Spear, S. F., …Waits, L. P. (2007). Putting the 'landscape' in landscape genetics. *Heredity*, *98*, 128-142. doi: 10.1038/sj.hdy.6800917

Templeton, A. R. (1998). Nested clade analysis of phylogeographical data: testing hypotheses about gene flow and population history. *Molecular Ecology*, *7*, 381-397. doi: 10.1046/j.1365-294x.1998.00308.x

Templeton, A. R. (2004). Statistical phylogeography: methods of evaluating and minimizing inference errors. *Molecular Ecology*, *13*, 789-809. doi: 10.1046/j.1365-294X.2003.02041.x

Turchetto-Zolet, A. C., Pinheiro, F., Salgueiro, F. & Palma-Silva, C. (2013). Phylogeographical patterns shed light on evolutionary process in South America. *Molecular Ecology*, *22*, 1193-1213. doi: 10.1111/mec.12164

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338-353. doi: 10.1016/S0019-9958(65)90241-X

Data accessibility

The data are archived in GenBank (accession numbers: MF752706-MF753038 and KC527837-KC528609).

Author contributions

LD designed research; All authors performed research; LD and JL analyzed data; all authors wrote the paper.

Table 1. Comparison between results obtained by Analysis of Molecular Variance (AMOVA) and null model tests implemented in HaploVectors. D_P - haplotypic dissimilarities between localities computed based on matrix **P**.

Source	Case species	Question	Method	Results
Lima et al. (2017)	*Eugenia dysenterica*	Differences among Cerrado regions?	AMOVA	F_{CT}= 0.164; $P < 0.001$
			ADONIS on D_P	$R^2 = 0.20$; $F_{Obs} = 5.38$; $P_{\text{site shuffle}} = 0.001$, $P_{\text{network shuffle}} = 0.973$
			OLS on haplovectors	*Haplovector 1 (32%)*: $R^2 = 0.58$; $F_{Obs} = 11.44$; $P_{\text{site shuffle}} = 0.001$, $P_{\text{network shuffle}} = 0.936$ *Haplovector 2 (17%)*: $R^2 = 0.15$; $F_{Obs} = 2.36$; $P_{\text{site shuffle}} = 0.016$, $P_{\text{network shuffle}} = 0.024$ *Haplovector 3 (11%)*: $R^2 < 0.01$; $F_{Obs} = 0.34$; $P_{\text{site shuffle}} = 0.608$, $P_{\text{network shuffle}} = 0.854$ *Haplovector 4 (9%)*: $R^2 < 0.01$; $F_{Obs} = 0.67$; $P_{\text{site shuffle}} = 0.432$, $P_{\text{network shuffle}} = 0.111$
Lima et al. (2014)	*Mauritia flexuosa*	Differences among basins?	AMOVA	$F_{CT} = 0.387$; $P < 0.050$
			ADONIS on D_P	$R^2 = 0.23$; $F_{Obs} = 2.24$; $P_{\text{site shuffle}} = 0.010$, $P_{\text{network shuffle}} = 0.005$
			OLS on haplovectors	*Haplovector 1 (42%)*: $R^2 = 0.24$; $F_{Obs} = 2.78$; $P_{\text{site shuffle}} =0.021$, $P_{\text{network shuffle}} = 0.007$ *Haplovector 2 (19%)*: $R^2 = 0.07$; $F_{Obs} = 1.23$; $P_{\text{site shuffle}} = 0.211$, $P_{\text{network shuffle}} = 0.460$ *Haplovector 3 (13%)*: $R^2 = 0.03$; $F_{Obs} = 0.93$; $P_{\text{site shuffle}} = 0.304$, $P_{\text{network shuffle}} = 0.695$ *Haplovector 4 (6%)*: $R^2 = 0.03$; $F_{Obs} = 1.28$; $P_{\text{site shuffle}} = 0.204$, $P_{\text{network shuffle}} = 0.680$

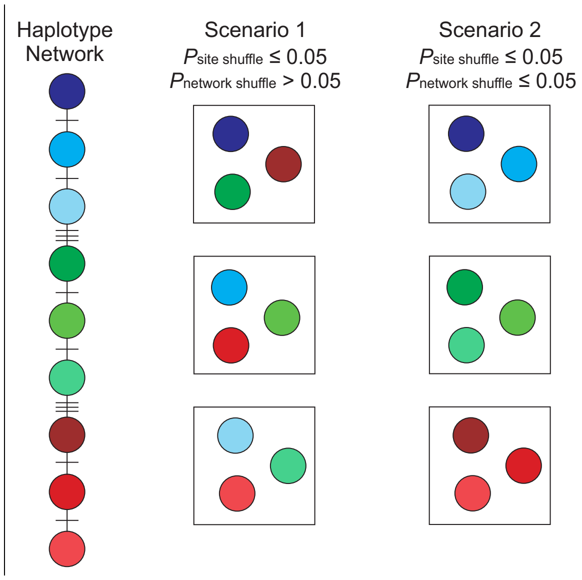
Haplotype Network
Scenario 1
$P_{\text{site shuffle}} \leq 0.05$
$P_{\text{network shuffle}} > 0.05$
Scenario 2
$P_{\text{site shuffle}} \leq 0.05$
$P_{\text{network shuffle}} \leq 0.05$

Fig. 1. Expected distribution of haplotypes (colors denote different haplotypes) across localities (squares) and the respective probabilities of homogeneity of haplotype composition among localities, under site and network shuffle null models. Scenario 1): Localities contain different haplotypes ($P_{\text{site shuffle}} \leq 0.05$), but variation of haplotype composition among sites is not associated with evolutionary structure depicted by the haplotype network ($P_{\text{network shuffle}} > 0.05$). Scenario 2): Localities contain different haplotypes ($P_{\text{site shuffle}} \leq 0.05$), and haplotype distribution across sites is mediated by evolutionary relatedness among them ($P_{\text{network shuffle}} \leq 0.05$).

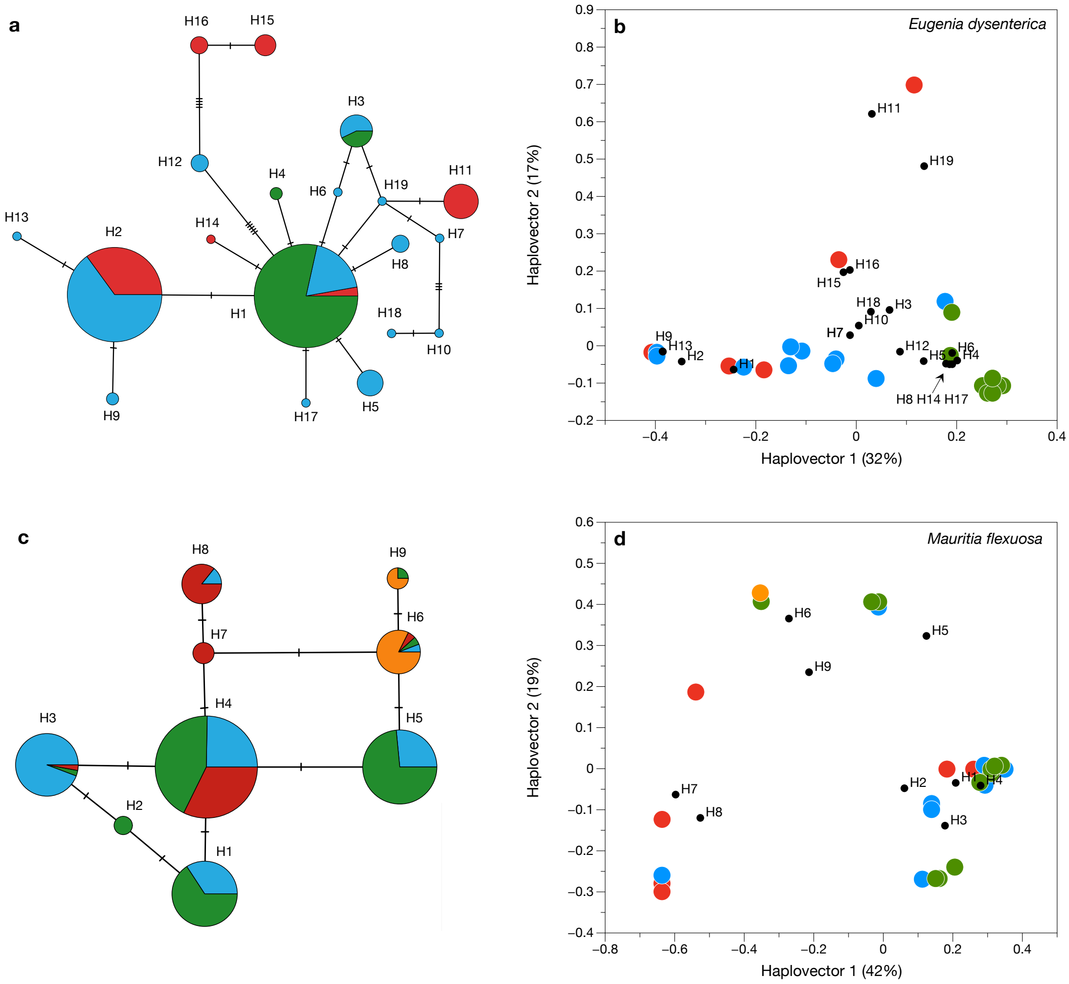

Fig. 2. Haplotype networks and their respective scatter plots of haplovectors computed for two datasets. 1a-b) *Eugenia dysenterica* (Lima et al. 2017). Red, blue and green circles indicate Northeast, Central and Southeastern biogeographic regions of the Brazilian Cerrado biome, respectively. a) MJN haplotypic network; b) Scatter plot for the two first haplovectors. Black circles indicate haplotypes (H1-H19). 1c-d) *Mauritia flexuosa* (Lima et al. 2014). Red, blue, orange and green circles indicate Amazon, Paraná, São Francisco and Araguaia/Tocantins river basins in South America, respectively. c) MJN haplotypic network; d) Scatter plot for the two first haplovectors. Black circles indicate haplotypes (H1-H9).