1 **GenVectors: an integrative analytical tool for spatial genetics**

2 *Leandro Duarte^{1*}, Jacqueline Souza Lima¹, Renan Maestri¹, Vanderlei Debastiani¹ &*
 3 *Rosane Garcia Collevatti²*

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

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

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

9 **Running head:** Genetic eigenvectors for molecular ecologists

10 **Abstract**

11 Metapopulations are sets of local populations connected by dispersal. While genetic
 12 turnover informs about the number of alleles shared by (meta)populations, a set of
 13 populations that do not share alleles with a second set may still show low genetic
 14 divergence to it. Recent secondary contact driven by anthropogenic habitat
 15 fragmentation and/or current climate change, for instance, may erase the historical track
 16 of genetic turnover. On the other hand, genetic turnover among sets of populations is
 17 expected to be related to the degree of genetic divergence among them if
 18 metapopulations become isolated from others due to vicariance or ancient dispersal.
 19 Yet, current analytical tools do not permit direct inference about alternative processes
 20 underlying spatial, environmental and/or biogeographic correlates of genetic turnover
 21 among populations. We introduce *GenVectors*, a new R package that offers flexible
 22 analytical tools that allow evaluating biogeographic or environmental correlates of
 23 genetic turnover among sets of local populations based on fuzzy set theory. Analyses
 24 implemented in *GenVectors* allow exploring the distribution of haplotypes or SNPs

25 across sets of local populations. Moreover, *GenVectors* provides tools to analyze
 26 environmental or biogeographic correlates of haplotype or SNP turnover among sets of
 27 local populations by applying appropriate null models, which enable to discriminate
 28 history-driven genetic turnover (vicariance, ancient dispersal) from non-historical ones
 29 (recent secondary contact). Finally, we demonstrate the application of *GenVectors* in
 30 two empirical datasets, one based on single-locus marker (haplotypes) and other based
 31 on multi-loci marker (SNPs).

32 **Keywords:** Fuzzy set theory, genetic eigenvectors, haplotypes, single nucleotide
 33 polymorphism, SNP

34

35 **Introduction**

36 Metapopulations are sets of local populations connected by dispersal (Hanski & Gilpin
 37 1991). Its very nature is dynamic, which implies that a local population may belong to a
 38 given metapopulation for a time period, and later becomes isolated from any other
 39 previously connected population due to the creation of a natural barrier preventing
 40 dispersal, such as river beds (Simões *et al.* 2014), or anthropogenic barriers (Hale *et al.*
 41 2013). The degree of isolation among sets of local populations can be properly
 42 estimated by evaluating the amount of gene flow among them using a proper measure
 43 such as F_{ST} index (Wright 1951), which allows ultimately mapping metapopulations
 44 across space (García-Verdugo *et al.* 2010; Murphy *et al.* 2010). Nonetheless,
 45 identifying sets of interconnected populations does not indicate *per se* the processes
 46 underlying the degree of gene flow among them. While genetic turnover informs about
 47 the number of alleles shared by (meta)populations, it may or may not be related to the
 48 degree of genetic divergence among those populations. In the latter case, a set of
 49 populations that do not share alleles with a second set (high genetic turnover) may still
 50 show low genetic divergence to it whenever alleles present in the first set, while being
 51 not the same present in the second one, show a lower number of mutational steps to the
 52 second set than expected by the overall genetic variability present in both sets. On one
 53 hand, recent anthropogenic habitat fragmentation and/or current climate change, for
 54 instance, may favor dispersal events among sets of populations formerly isolated,
 55 erasing to some extent the historical track of genetic turnover among populations. Under
 56 such scenario, recent admixture from secondary contact between previously allopatric
 57 populations may completely erase genetic turnover (Fig. 1b), or promote genetic
 58 turnover among sets of populations that is unrelated to the degree of genetic divergence
 59 among populations (Fig 1c). On the other hand, metapopulations may become isolated

60 from others due to interruption of gene flow caused by vicariance driven by physical
61 barriers, past environmental changes or by ancient colonization of new sites via
62 dispersal from an ancestral area (Hewitt 2000; Queiroz 2005). In this scenario, genetic
63 turnover among sets of populations is expected to be related to the degree of genetic
64 divergence among them (Fig. 1d).

65 A major goal of spatial genetics consists in identifying major drivers of genetic
66 turnover across space, which can be expressed as the spatial distribution of alleles,
67 genotypes, haplotypes, or single nucleotide polymorphisms (SNPs) among sets of local
68 populations (Avice 2009; Balkenhol *et al.* 2017; Manel *et al.* 2003; Storfer *et al.* 2006).
69 Whenever these drivers are environmental or biogeographic factors underlying the
70 genealogical history of single-species populations, proper hypothesis testing may shed
71 light on metapopulation connections from landscape to regional scales (Hale *et al.* 2013;
72 Hand *et al.* 2016), or even biogeographic connections across entire regions, biomes or
73 continents (Carnaval *et al.* 2014; Diniz-Filho *et al.* 2008). In this sense, statistical
74 phylogeographical approaches (Papadopoulou & Knowles 2016; Templeton 2004)
75 helped moving that field beyond the essential descriptive nature present in its infancy.
76 Analysis of molecular variance (AMOVA, Excoffier *et al.* 1992) has been widely used
77 for analyzing genetic dissimilarities (turnover) among local populations based on
78 haplotypes (Fitzpatrick 2009; Maestri *et al.* 2016; Raffini *et al.* 2018). AMOVA, which
79 is a permutation procedure akin to approaches often used in ecological studies, such as
80 PERMANOVA (Anderson 2001; Pillar & Orlóci 1996) allows analyzing whether
81 pairwise genetic dissimilarities between individuals of different populations or sets of
82 populations are higher than expected by chance given within (meta)population
83 dissimilarities. It usually complements exploratory methods to visualize haplotype links
84 and distribution among populations such as haplotype networks (Mardulyn 2012) (Goal

1 in Fig. 2). Since AMOVA is based on overall genetic dissimilarities, it does not permit direct inference about alternative processes underlying spatial, environmental and/or biogeographic correlates of genetic turnover among populations (Goal 2 in Fig. 2).

Nowadays the scenario is still more puzzling given the ever-increasing number of molecular loci discovered and the numerous tools focused on spatial genetics (Epperson 2005; Miller 2005; Templeton 2004). Indeed, 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 *et al.* 2013). The number of molecular markers used increased from single locus to multiple genome regions (Blom *et al.* 2016). Non *a priori* individual-based clustering approaches, such as the Bayesian method implemented in the software STRUCTURE (Pritchard *et al.* 2000), and Discriminant Analysis of Principal Components (DAPC, Jombart *et al.* 2010), are widely used to define genetic groups based on SNPs (Van Cise *et al.* 2019; Viana *et al.* 2018; see Goal 3 in Fig. 2). While these methods allow evaluating genetic turnover among sets of populations and may provide indirect inference on potential majors of spatial genetic patterns, they do not allow directly inferring alternative processes underlying genetic turnover among populations (Goal 4 in Fig. 2). In the era of next-generation sequencing (NGS) technology in landscape genetics/genomics and phylogeographic studies, new analytical tools are imperative to better understand the increasingly complex phylogeographical patterns, to compare results from different loci, and to uncover environmental and spatial processes correlates of genetic turnover in the 'twilight zone' (Balkenhol *et al.* 2017; Diniz & Filho *et al.* 2008; Hand *et al.* 2016).

In classical set theory, the membership function assumes binary values, which means that an element either belongs or does not belong to a given set. According to

110 fuzzy set theory (Zadeh 1965), any potential element of a set may be described by a
 111 membership function defining its degree of belonging to that set ranging in the interval
 112 between zero and one. Given the intrinsically complex nature of biological phenomena,
 113 fuzzy sets might provide a more realistic categorization of biological entities, such as
 114 populations. Fuzzy set theory has been employed in community ecology and
 115 biogeography in order to define species clusters or composition gradients (Dale 1977;
 116 Feoli & Zuccarello 1986; Olivero *et al.* 2013; Pillar & Orlóci 1991; Roberts 1986).
 117 Fuzzy set theory has been demonstrated to be a useful framework for the analysis of
 118 metacommunity assembly patterns considering both functional (Pillar *et al.* 2009) and
 119 phylogenetic (Duarte *et al.* 2016; Pillar & Duarte 2010) dimensions. To our knowledge,
 120 no attempt to apply fuzzy set theory to explore genetic turnover among populations and
 121 to analyze underlying environmental or biogeographic factors driving genetic turnover
 122 of single-species populations has been made to the moment. Our main goal in this study
 123 is to fulfill this knowledge gap by offering a flexible and robust tool for spatial genetics.

124 We introduce *GenVectors*, a new R package freely available at
 125 <https://github.com/vanderleidebastiani/GenVectors>, that offers flexible analytical tools
 126 for spatial geneticists and molecular ecologists to disentangle biogeographic or
 127 environmental factors driving genetic turnover among sets of local populations.
 128 *GenVectors* applies Zadeh's fuzzy set theory in order to describe the distribution of
 129 haplotypes or SNPs across sets of local populations, allowing flexible exploratory
 130 analysis of genetic turnover among sets of populations. Moreover, by applying
 131 appropriate null models, *GenVectors* provides tools to analyze environmental or
 132 biogeographic correlates of haplotype or SNP turnover among sets of local populations
 133 (Goals 2 and 4 in Fig. 2). In this sense, the analysis is suitable for detecting which
 134 populations are connected to each other given an environmental or biogeographic

factor, and therefore to infer metapopulation structure. Furthermore, the null model approach also enables to discriminate history-driven genetic turnover among metapopulations, determined by vicariance or ancient dispersal, from non-historical processes, such as recent secondary contact. We demonstrate the application of *GenVectors* in two empirical datasets, one based on single-loci marker (haplotypes) and other based on multi-loci marker (SNPs).

Materials and Methods

Genetic fuzzy weighting

A fuzzy set denotes the degree of belonging (or membership), varying between zero and one, of a given element (say, an individual) to a set (Zadeh 1965). If that set is another individual in the population, the application of fuzzy set theory might provide a pairwise fuzzy set matrix describing the degree of belonging of each individual to every other in a sample defined by a number of individuals distributed across different locations (populations), and for which we have gathered genetic information (haplotypes, SNPs). Combining the information on the distribution of individuals among populations, and the degree of genetic belonging of each individual to every other provides the numerical basis of the method we describe over the next paragraphs.

Consider a sample of individuals for which we gathered information on the expression of 1) a single-locus marker, allowing us to classify each individual based on a set of haplotypes, or 2) genome-wide markers, allowing to estimate the pairwise amount of genetic similarity between individuals based on SNPs. For the sake of simplicity, our explanation below is based only on haplotypes, but the logic is also valid for SNPs. Each haplotype shows a variable number of mutational steps from all other haplotypes. Translating this information into fuzzy set terms, a given haplotype showing the highest number of mutations from any other haplotype in the sample will

160 have the lowest degree of belonging to between any pair of haplotypes in the sample.
 161 Note, however, that fuzzy belonging between two individuals will be not necessarily
 162 symmetric (Duarte *et al.* 2016). The belonging of a haplotype A to another haplotype B
 163 will depend not only on the number of mutational steps between them, but also on the
 164 number of mutations between each of them and all other haplotypes in the sample.
 165 Depending on the configurations of the haplotype network connecting the haplotypes,
 166 the belonging of haplotype A to haplotype B may be lower than the belonging of B to
 167 A. The reasons of such asymmetry in the degrees of belonging between two objects is
 168 discussed in more detail in Pillar and Duarte (2010) and Duarte *et al.* (2016), where
 169 analysis of phylogenetic composition across metacommunities based on fuzzy set
 170 theory is introduced. Although in that case the application of fuzzy sets focuses on
 171 phylogenetic relationships among species, which has somewhat different properties of
 172 genetic structure of single-species populations, the mathematical basis of the method
 173 remains valid and might help readers interested in a deeper understanding of the
 174 application of fuzzy logic to characterize biological systems.

175 In the package *GenVectors* we implement functions to perform genetic fuzzy
 176 weighting either for datasets containing information of distribution of haplotypes
 177 (function ‘HaploVectors’) or SNPs (function ‘SNPVectors’) among sets of populations.
 178 The first analytical step implemented in the ‘HaploVectors’ function consists of
 179 defining haplotypes for a set of samples and computing the frequency (number of
 180 individuals) of each haplotype per population. For this, two input datasets are required:
 181 (1) a *.fas file containing aligned genetic sequences for each sampled individual, and
 182 (2) a matrix describing the occurrence of each individual (columns) in a given
 183 population (rows). Based on these two datasets, the function ‘HaploDist’ extracts the
 184 haplotypes for the *.fas file using the function ‘haplotype’ implemented in the R

185 package *pegas* (Paradis 2010), and computes matrix **D** describing all pairwise distances
 186 between haplotypes based on the number of mutational steps between any pair of
 187 haplotypes. ‘HaploVectors’ then computes the frequency of each haplotype per
 188 population (matrix **W**) and submits **W** to genetic fuzzy weighting based on **D**. For this,
 189 matrix **D** is further converted into a similarity matrix **S** describing genetic similarities
 190 between all pairs of haplotypes *i* and *j* (s_{ij} , ranging between 0 and 1), as follows:

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

191 d_{ij} is the number of mutations separating the haplotypes *i* and *j* in **D**, and max
 192 (d_{ij}) is the maximum number of mutations between any pair of haplotypes. Matrix **S** is
 193 then standardized by marginal total within columns, generating a matrix **Q** that
 194 describes fuzzy belongings between haplotypes (Duarte *et al.* 2016; Pillar & Duarte
 195 2010). Matrix **Q**, containing haplotypic fuzzy sets are then employed to weight the
 196 frequencies of haplotypes per population (matrix **W**) by their genetic relatedness,
 197 generating matrix **P**, which describes haplotype turnover among populations. The
 198 output of ‘HaploVectors’ function provides matrices **W**, **D**, **Q** and **P**. Performing
 199 principal coordinates analysis on **P** generates haplotype eigenvectors, which decompose
 200 the total variation in haplotype turnover among the set of populations into independent
 201 fractions proportional to its respective eigenvalue λ (Pielou 1984). Those eigenvectors
 202 representing the higher amount of variation in **P** can be used to explore major trends in
 203 haplotype turnover across the populations. Those populations sharing most haplotypes
 204 will show similar scores, and therefore will group to each other in the scatter plot. Thus,
 205 this scatter plot allows simultaneously exploring genetic links among haplotypes and
 206 populations.

207 The function ‘SNPVectors’ is quite similar to ‘HaploVectors’, but it
 208 accommodates the peculiarities of SNP data. In this case, the two input datasets are

209 required: (1) a *.gen file containing individual SNPs genotypes, and (2) a matrix **W**,
 210 which in this case describes the occurrence of individuals (columns) in each population
 211 (rows). In ‘SNPVectors’, matrix **D** is defined by computing the proportion of shared
 212 alleles in a set of genotypes (individuals) using the function ‘propShared’ of package
 213 *adeigenet* (Jombart 2008). After computing matrix **S**, we obtain matrix **Q** in a similar
 214 way as described for haplotypes. Matrix **Q** containing SNP-based fuzzy sets are then
 215 employed to weight the occurrences of individuals per population (matrix **W**) by their
 216 genetic relatedness, generating matrix **P**, which describes SNP turnover among sets of
 217 populations. Performing principal coordinates on **P** generates SNP eigenvectors, which
 218 decompose the total variation in the SNP turnover across the set of populations into
 219 independent fractions proportional to its respective eigenvalue λ (Pielou 1984). Those
 220 eigenvectors representing the higher amount of variation in **P** can be used to explore
 221 major trends in SNP turnover across the populations. All further steps involving null
 222 model tests and outputs are similar to those previously described for ‘HaploVectors’.

223 Both functions ‘HaploVectors’ and ‘SNPVectors’ allow analyzing multiple
 224 environmental, biogeographic or spatial correlates of haplotype or SNP turnover among
 225 sets of populations, and therefore are useful tools for robust hypothesis test in spatial
 226 genetics. Both functions implement two null model tests, adapted from Duarte *et al.*
 227 (2016) and designed to test the following hypotheses:

228 Hypothesis 1: *A given environmental or biogeographic factor **E** explains genetic*
 229 *turnover among sets of populations.* This hypothesis is tested by means of a null model
 230 called *turnover test*, which is a classical permutation-based procedure assuming
 231 independence between haplotypes/SNPs and populations. The test can be performed
 232 based on either haplotype/SNP-based pairwise dissimilarities between populations
 233 computed from **P** (hereafter **D_P**) using an appropriate resemblance measure, such as

234 Euclidean distances or Bray-Curtis dissimilarities (Legendre & Legendre 2012), or
 235 directly on single haplotype/SNP eigenvectors. The test consists of 1) computing a F_{Obs} .
 236 statistic. If D_P is modeled on E , the test is based on a dissimilarity regression on
 237 distance matrices (hereafter called ADONIS; see (McArdle & Anderson 2001). For
 238 single haplotypic/SNP eigenvectors modeled on E , the test is an ordinary linear squares
 239 (OLS) model; 2) freely permuting the populations a number of times (say 999); 3) at
 240 each permutation step, computing F_{null} ; and 4) defining the probability of obtaining the
 241 observed statistic by chance ($H_0 = F_{Obs.} \leq F_{null}$), as the proportion of permutations in
 242 which F_{null} exceeded $F_{Obs.}$. Using this procedure, the test simultaneously addresses the
 243 influence of E on genetic turnover among sets of populations (for either haplotypes or
 244 SNPs) via both the frequency of haplotypes/SNPs shared between pairs of populations,
 245 and the genetic relatedness between the haplotypes/SNPs (the number of mutational
 246 steps between the pairs of haplotypes or the proportion of alleles shared between SNPs).
 247 Therefore, whenever this first hypothesis is corroborated (H_0 is rejected), the test allows
 248 to conclude that populations occurring under similar levels of E are more genetically
 249 connected to each other than to populations occurring under different levels of E .
 250 Nonetheless, genetic connectedness or isolation among populations might be
 251 determined either by 1) non-historical processes, such as recent secondary contact, or 2)
 252 history-related processes related to vicariance or an ancient dispersal event, which make
 253 genetic turnover among populations to be related to the amount of genetic divergence
 254 among them. To corroborate (or not) this second scenario, it is necessary to test a
 255 second hypothesis, which is conditioned on the validity of the first one:

256 Hypothesis 2: *The influence of E on genetic turnover among sets of populations*
 257 *is related to the genetic divergence among them.* To test this hypothesis, a second round
 258 of permutations (*divergence test*) is needed in order to compute F_{null} . Accordingly, the

frequency of haplotypes/SNPs in **W** is kept constant across the populations while genetic divergence between them is permuted by shuffling haplotype/SNP labels (Kembel *et al.* 2010). After computing $F_{Obs.}$ (step 1, as described for turnover test), the procedure consists of 2) freely permuting haplotype/SNP labels to generate random genetic divergences among haplotypes/SNPs and computing null matrices **D**, **Q** and **P**. The procedure is repeated a number of times (say 999); 3) at each permutation procedure, computing null D_P and, if necessary, null haplotype/SNP eigenvectors. In this latter case, null haplotype/SNP eigenvectors are submitted to procrustean adjustment (Jackson 1995) and fitted values between observed and null eigenvectors are obtained; 4) take null D_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). For the turnover test, whenever the null hypothesis is rejected, we conclude that **E** explains genetic turnover among sets of populations (hypothesis 1). Then we proceed to test for the hypothesis 2 (via divergence test). If the null hypothesis is rejected, we conclude that the influence of **E** on genetic turnover among sets of populations is proportional to the genetic divergence among them. Possible biological processes underlying alternative test outcomes are explained using a hypothetical example based on haplotypes (Fig. 1).

Application using empirical datasets

We demonstrate the application of ‘HaploVectors’ and ‘SNPVectors’ in spatial genetics analyses using two empirical datasets for the tree species *Handroanthus impetiginosus*

(Bignoniaceae): a set of cpDNA sequences from 258 adult individuals (Collevatti *et al.* 2012), and a set of putatively neutral SNPs loci from 128 adult individuals (Collevatti *et al.* 2019). In order to compare results obtained from different datasets, in this study only populations common to both datasets (12 populations) were kept in further analyses (Fig. 3).

Dataset 1: cpDNA sequences

Three chloroplast regions were sequenced from 258 individuals sampled from 17 populations across the entire distribution range of the species (for more details about sampling and genetic data see Collevatti *et al.* (2012). We used the ‘HaploVectors’ function to evaluate if the variation in the distribution of haplotypes might be related to either two present-day bioclimatic variables: temperature seasonality (hereafter TS) or precipitation of the driest quarter (hereafter PDQ), obtained from ecoClimate database (<http://www.ecoclimate.org>; Lima-Ribeiro *et al.* 2015). We hypothesized that (1) climatic variables explained the genetic turnover of haplotypes among sets of populations (metapopulations), and that (2) climate-haplotype associations were mediated by the amount of genetic divergence among haplotypes. We tested these hypotheses for each bioclimatic variable separately using ADONIS (for haplotypic dissimilarities between 12 localities computed from matrix $\mathbf{P} - \mathbf{D_P}$), and OLS (for haplotypic eigenvectors). For ADONIS, we computed matrix $\mathbf{P}$ using frequencies of haplotypes per site (matrix $\mathbf{W}$) and square-rooted Hamming distances between haplotypes (matrix $\mathbf{D}$). Haplotypic dissimilarities between sites were computed using Bray-Curtis dissimilarities. Either bioclimatic variables were taken as continuous predictor in the analysis (matrix $\mathbf{E}$). For OLS, we first computed haplotypic eigenvectors (haplo vectors) based on $\mathbf{D_P}$. Haplo vectors 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: neutral SNPs

Collevatti *et al.* (2019) genotyped 75,838 SNPs in 128 individuals distributed across 13 populations of *H. impetiginosus*. We used the ‘SNPVectors’ function to evaluate whether SNP turnover among sets of populations was related to either bioclimatic variables (TS or PDQ). Steps involving null model tests and outputs are similar to those previously described for ‘HaploVectors’.

Results

For the cpDNA dataset, the four first eigenvectors contained each at least five percent of total information in matrix **P**. ADONIS indicated that PDQ explained genetic turnover among populations (Table 1); in other words, sets of populations located at lower PDQ levels share more haplotypes with each other than with sets of populations located at higher PDQ levels ($P_{\text{turnover}} = 0.030$). Moreover, such association between haplotype turnover and PDQ was mediated by the amount of genetic divergence among haplotypes ($P_{\text{divergence}} = 0.022$). That is, haplotypes of populations at lower PDQ levels showed lower genetic divergence from each other than from populations at higher PDQ levels, and vice-versa. On the other hand, only OLS performed for first haplovector (30% of total variation in haplotype turnover) showed association with PDQ ($P_{\text{turnover}} = 0.012$; see Table 1). No other haplovector showed association with either environmental variables (Table 1; see Supporting Information). A clear separation among SUM and all other populations can be observed in the haplovector scatter plot (Fig. 4a). Nonetheless, this analysis did not reveal any evidence of haplotype-environment relationship being

332 mediated by the amount of genetic divergence among haplotypes ($P_{\text{divergence}} = 0.128$;
333 Table 1).

334 For the SNP dataset, the three first eigenvectors contained each at least five
335 percent of total information in matrix **P**. The null hypothesis of no association among
336 genetic turnover and climatic variation among populations tested using ADONIS was
337 rejected only for TS ($P_{\text{turnover}} = 0.012$; Table 1), and such association was mediated by
338 the amount of genetic divergence among populations ($P_{\text{divergence}} = 0.001$). Accordingly,
339 the scatter plot of SNPvectors split populations showing higher TS values (ALT, MOC,
340 SUM) from all others, which tended to show lower TS values (Fig. 4b). Although OLS
341 performed using the first SNPvector (59% of total genetic turnover) confirmed the
342 association between genetic turnover and TS ($P_{\text{turnover}} = 0.014$; Table 1), the influence of
343 the amount of genetic divergence among SNPs on this association was not so clear
344 ($P_{\text{divergence}} = 0.076$; Table 1). Neither second nor third SNPvector showed association
345 with either environmental variables (Table 1; see Supporting Information).

346 Discussion

347 Analyses implemented in *GenVectors* allow to explore genetic turnover across
348 (meta)populations and analytically infer the role of recent secondary contact versus
349 long-term split (driven by vicariance or ancient dispersal) on spatial genetic patterns by
350 means of appropriate null model tests. Moreover, *GenVectors* was demonstrated to be a
351 flexible tool in the sense it accommodates different genetic data (haplotypes, SNPs).
352 These features make the tool useful for several different purposes in spatial genetics,
353 and complementary to classic analytical tools widely used by molecular ecologists, such
354 as AMOVA (Excoffier *et al.* 1992), individual-based Bayesian clustering (Pritchard *et*
355 *al.* 2000), or DAPC (Jombart *et al.* 2010). Fundamental questions in ecology and
356 evolution, such as the role of ancient versus recent biogeographic events, the

relationship between morphological divergence and genetic turnover and divergence, and the effects of abiotic variables on genetic variation can easily be investigated using analyses implemented in *GenVectors*.

Exploring patterns of genetic turnover among sets of populations based on genetic eigenvectors revealed clusters of populations, indicating metapopulation structure, similarly to the patterns observed elsewhere (Hand *et al.* 2016). Nonetheless, population clustering differed depending on the genetic marker used to estimate genetic turnover, which is not surprising (Larsson *et al.* 2009; Ryman *et al.* 2006). In Collevatti *et al.* (2012), AMOVA pointed out genetic differentiation among populations, while evolutionary relationships among haplotypes of *H. impetiginosus* inferred using median-joining network analysis visually suggested that the phylogenetic relationships among haplotypes did not match the geographical distribution of the lineages. Nonetheless, those analyses did not enable evaluating possible environmental correlates of haplotype turnover across populations. Analyzing the same dataset, ADONIS-based null model tests demonstrated that haplotype turnover among populations of *H. impetiginosus* was associated with the amount of precipitation of the driest quarter (PDQ), and that such association was mediated by genetic divergence among haplotypes. This finding permits a new inference: variation in PDQ may prescribe long-term turnover that resulted in genetic divergence between populations along the PDQ gradient.

Noteworthy, results for OLS-based null model tests performed using the first haplovector only partly corroborated those for the entire matrix of genetic turnover. While a consistent association between haplotype turnover and PDQ was again observed, according to null model tests such relationship was not mediated by genetic divergence. Such discrepancy is probably related to the relatively small amount of

information contained in the first haplovector ($\cong 30\%$ of total information in matrix **P**). While haploectors are useful to visualize general similarity patterns among populations and their association with different haplotypes, ADONIS-based hypothesis testing has been previously found to be more effective for that purpose in terms of type I error and statistical power (Duarte *et al.* 2016).

Interestingly, SNP information for the same species and populations told us a different story. Collevatti *et al.* (2019) found similar spatial patterns of SNP-based genetic differentiation among populations when compared to the previous haplotype-based study (Collevatti *et al.* 2012), using Bayesian genetic clustering method implemented in STRUCTURE (Pritchard *et al.* 2000). Despite such spatial congruence in genetic turnover, our analyses showed that the environmental condition mostly associated with genetic turnover differed between haplotype and SNP-based analyses. While genetic turnover based on SNPs was associated with temperature seasonality (TS), haplotype-based genetic turnover was associated with PDQ. Moreover, in SNP-based analysis results obtained using either the entire matrix **P** (ADONIS) or OLS performed for the first SNPvector were much more congruent to each other than for haplotype analyses. Such congruence may be related to the first SNPvector have concentrated more information (60% of total information in matrix **P**) than the first haplovector ($\cong 30\%$ of total information in matrix **P**). A general recommendation is to preferably perform ADONIS using overall dissimilarities computed from matrix **P** whenever the main goal is hypothesis testing, while genetic eigenvectors should be used mostly to explore population clustering patterns across the environmental/biogeographic gradient.

Although we have demonstrated the usefulness of analyses implemented in *GenVectors* to investigate genetic turnover based on different types of genetic datasets,

the analytical framework was developed focusing on neutral loci. Yet, the analytical procedures described here may be easily applied to investigate non-neutral loci as well, which may open totally new avenues for future research (Kirk *et al.* 2011; Schoville *et al.* 2012). For instance, null model tests implemented in *GenVectors* might be used to prospect for associations between non-neutral genetic data and environmental variables in order to investigate the influence of ecological factors on the spatial distribution of adaptive genetic variability across sets of populations (Schoville *et al.* 2012; Sork *et al.* 2013), thereby allowing new insights on the influence of multiple environmental or biogeographic factors drive selection-based genetic turnover across (meta)populations.

Acknowledgements

The authors thank Jorge Sebastião Bernardo-Silva for valuable discussion on the topic addressed in this paper. LD research activities have been supported by CNPq Productivity Fellowship (grant 307527/2018-2). RGC research activities have been supported by CNPq Productivity Fellowship. JSL received a CNPq fellowship (grant 439051/2016-9). RM thanks to CAPES and CNPq (grant 406497/2018-4) for funding. The authors have no conflict of interests to declare. LD, 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 MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 32-46.
- Avice JC (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography* **36**, 3-15.
- Balkenhol N, Dudaniec RY, Krutovsky KV, *et al.* (2017) Landscape genomics: Understanding relationships between environmental heterogeneity and genomic characteristics of populations. In: *Population Genomics* (ed Rajora O), pp. 261-322. Springer, Cham.

- 435 Beheregaray LB (2008) Twenty years of phylogeography: the state of the field and the
436 challenges for the Southern Hemisphere. *Molecular Ecology* **17**, 3754-3774.
- 437 Blom MP, Horner P, Moritz C (2016) Convergence across a continent: adaptive
438 diversification in a recent radiation of Australian lizards. *Proceedings of the*
439 *Royal Society B: Biological Sciences* **283**, 20160181.
- 440 Carnaval AC, Waltari E, Rodrigues MT, *et al.* (2014) Prediction of phylogeographic
441 endemism in an environmentally complex biome. *Proceedings of the Royal*
442 *Society B: Biological Sciences* **281**, 20141461.
- 443 Collevatti RG, Novaes E, Silva-Junior OB, *et al.* (2019) A genome-wide scan shows
444 evidence for local adaptation in a widespread keystone Neotropical forest tree.
445 *Heredity* (DOI: 10.1038/s41437-019-0188-0).
- 446 Collevatti RG, Terribile LC, Lima Ribeiro MS, *et al.* (2012) A coupled
447 phylogeographical and species distribution modelling approach recovers the
448 demographical history of a Neotropical seasonally dry forest tree species.
449 *Molecular Ecology* **21**, 5845-5863.
- 450 Dale MB (1977) Planning an adaptive numerical classification. *Vegetatio* **35**, 131-136.
- 451 Diniz-Filho JAF, de Campos Telles MP, Bonatto SL, *et al.* (2008) Mapping the
452 evolutionary twilight zone: molecular markers, populations and geography.
453 *Journal of Biogeography* **35**, 753-763.
- 454 Duarte LDS, Debastiani VJ, Freitas AVL, Pillar VD (2016) Dissecting phylogenetic
455 fuzzy weighting: theory and application in metacommunity phylogenetics.
456 *Methods in Ecology and Evolution* **7**, 937-946.
- 457 Epperson B (2005) Estimating dispersal from short distance spatial autocorrelation.
458 *Heredity* **95**, 7-15.
- 459 Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred
460 from metric distances among DNA haplotypes: application to human
461 mitochondrial DNA restriction data. *Genetics* **131**, 479-491.
- 462 Feoli E, Zuccarello V (1986) Ordination based on classification: yet another solution?!
463 *Abstracta Botanica* **10**, 203-219.
- 464 Fitzpatrick BM (2009) Power and sample size for nested analysis of molecular variance.
465 *Molecular Ecology* **18**, 3961-3966.
- 466 García-Verdugo C, Forrest AD, Fay MF, Vargas P (2010) The relevance of gene flow
467 in metapopulation dynamics of an oceanic island endemic, *Olea europaea* subsp.
468 *guanchica*. *Evolution: International Journal of Organic Evolution* **64**, 3525-
469 3536.
- 470 Hale JM, Heard GW, Smith KL, *et al.* (2013) Structure and fragmentation of growling
471 grass frog metapopulations. *Conservation Genetics* **14**, 313-322.
- 472 Hand BK, Muhlfeld CC, Wade AA, *et al.* (2016) Climate variables explain neutral and
473 adaptive variation within salmonid metapopulations: the importance of
474 replication in landscape genetics. *Molecular Ecology* **25**, 689-705.
- 475 Hanski I, Gilpin M (1991) Metapopulation dynamics: brief history and conceptual
476 domain. *Biological Journal of the Linnean Society* **42**, 3-16.
- 477 Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907-913.
- 478 Jackson DA (1995) Protest: A Procrustean Randomization test of community
479 environment concordance. *Ecoscience* **2**, 297-303.
- 480 Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic
481 markers. *Bioinformatics* **24**, 1403-1405.
- 482 Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal
483 components: a new method for the analysis of genetically structured
484 populations. *BMC Genetics* **11**, 94.

485 Kembel SW, Cowan PD, Helmus MR, *et al.* (2010) Picante: R tools for integrating
486 phylogenies and ecology. *Bioinformatics* **26**, 1463-1464.

487 Kirk, H, Freeland JR (2011) Applications and implications of neutral versus non-neutral
488 markers in ecology. *International Journal of Molecular Sciences* **12**, 3966-3988.

489 Larsson LC, Charlier J, Laikre L, *et al.* (2009) Statistical power for detecting divergence
490 – organelle versus nuclear markers. *Conservation Genetics* **10**, 1255–1264.

491 Legendre P, Legendre L (2012) *Numerical Ecology*, 3 edn. Elsevier, Amsterdam.

492 Lima-Ribeiro MS, Varela S, González-Hernández J, *et al.* (2015) EcoClimate: a
493 database of climate data from multiple models for past, present, and future for
494 macroecologists and biogeographers. *Biodiversity Informatics* **10**.

495 Maestri R, Fornel R, Gonçalves GL, *et al.* (2016) Predictors of intraspecific
496 morphological variability in a tropical hotspot: comparing the influence of
497 random and non-random factors. *Journal of Biogeography* **43**, 2160-2172.

498 Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining
499 landscape ecology and population genetics. *Trends in Ecology & Evolution* **18**,
500 189-197.

501 Mardulyn P (2012) Trees and/or networks to display intraspecific DNA sequence
502 variation? *Molecular Ecology* **21**, 3385-3390.

503 McArdle BH, Anderson MJ (2001) Fitting multivariate models to community data: a
504 comment on distance-based redundancy analysis. *Ecology* **82**, 290-297.

505 Miller M (2005) Alleles In Space (AIS): computer software for the joint analysis of
506 interindividual spatial and genetic information. *Journal of Heredity* **96**, 722-724.

507 Murphy MA, Dezzani R, Pilliod DS, Storfer A (2010) Landscape genetics of high
508 mountain frog metapopulations. *Molecular Ecology* **19**, 3634-3649.

509 Olivero J, Márquez AL, Real R (2013) Integrating fuzzy logic and statistics to improve
510 the reliable delimitation of biogeographic regions and transition zones.
511 *Systematic Biology* **62**, 1-21.

512 Papadopoulou A, Knowles LL (2016) Toward a paradigm shift in comparative
513 phylogeography driven by trait-based hypotheses. *Proceedings of the National
514 Academy of Sciences* **113**, 8018-8024.

515 Paradis E (2010) pegas: an R package for population genetics with an integrated–
516 modular approach. *Bioinformatics* **26**, 419-420.

517 Pielou EC (1984) *The interpretation of ecological data: a primer on classification and
518 ordination* John Wiley & Sons.

519 Pillar VD, Duarte LDS (2010) A framework for metacommunity analysis of
520 phylogenetic structure. *Ecology Letters* **13**, 587-596.

521 Pillar VD, Duarte LDS, Sosinski EE, Joner F (2009) Discriminating trait-convergence
522 and trait-divergence assembly patterns in ecological community gradients.
523 *Journal of Vegetation Science* **20**, 334-348.

524 Pillar VD, Orlóci L (1991) Fuzzy components in community level comparisons. In:
525 *Computer Assisted Vegetation Analysis* (eds. Feoli E, Orlóci L), pp. 87–93.
526 Kluwer, Dordrecht.

527 Pillar VD, Orlóci L (1996) On randomization testing in vegetation science: multifactor
528 comparisons of relevé groups. *Journal of Vegetation Science* **7**, 585-592.

529 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using
530 multilocus genotype data. *Genetics* **155**, 945-959.

531 Queiroz AD (2005) The resurrection of oceanic dispersal in historical biogeography.
532 *Trends in Ecology & Evolution* **20**, 68-73.

533 Raffini F, Fruciano C, Meyer A (2018) Morphological and genetic correlates in the left–
534 right asymmetric scale-eating cichlid fish of Lake Tanganyika. *Biological*
535 *Journal of the Linnean Society* **124**, 67-84.

536 Roberts DW (1986) Ordination on the basis of fuzzy set theory. *Vegetatio* **66**, 123-131.

537 Ryman N, Palm S, André C, *et al.* (2006) Power for detecting genetic divergence:
538 differences between statistical methods and marker loci. *Molecular Ecology* **15**,
539 2031-2045.

540 Schoville SD, Bonin A, François O, *et al.* (2012) Adaptive genetic variation on the
541 landscape: methods and cases. *Annual Review of Ecology, Evolution and*
542 *Systematics* **43**, 23–43.

543 Simões PI, Stow A, Hödl W, *et al.* (2014) The value of including intraspecific measures
544 of biodiversity in environmental impact surveys is highlighted by the
545 Amazonian brilliant-thighed frog (*Allobates femoralis*). *Tropical Conservation*
546 *Science* **7**, 811-828.

547 Sork VL, Aitken SN, Dyer RJ, *et al.* (2013). Putting the landscape into the genomics of
548 trees: approaches for understanding local adaptation and population responses to
549 changing climate. *Tree Genetics & Genomes* **9**, 901-911.

550 Storfer A, Murphy MA, Evans JS, *et al.* (2006) Putting the ‘landscape’ in landscape
551 genetics. *Heredity* **98**, 128-142.

552 Templeton AR (2004) Statistical phylogeography: methods of evaluating and
553 minimizing inference errors. *Molecular Ecology* **13**, 789-809.

554 Turchetto-Zolet A, Pinheiro F, Salgueiro F, Palma-Silva C (2013) Phylogeographical
555 patterns shed light on evolutionary process in South America. *Molecular*
556 *Ecology* **22**, 1193-1213.

557 Van Cise AM, Baird RW, Baker CS, *et al.* (2019) Oceanographic barriers, divergence,
558 and admixture: Phylogeography and taxonomy of two putative subspecies of
559 short-finned pilot whale. *Molecular Ecology* (DOI: 10.1111/mec.15107).

560 Viana JPG, Siqueira MVBM, Araujo FL, *et al.* (2018) Genomic diversity is similar
561 between Atlantic Forest restorations and natural remnants for the native tree
562 *Casearia sylvestris* Sw. *PLoS ONE* **13**, e0192165.

563 Wright S (1951) The genetical structure of populations. *Annals of eugenics* **15**, 323-354.

564 Zadeh LA (1965) Fuzzy sets. *Information and Control* **8**, 338-353.

565 **Data Availability Statement**

566 The data that support the findings of this study are openly available in Genbank

567 (cpDNA dataset - <http://doi.org/10.1111/mec.12071>; SNP dataset -

568 <http://doi.org/10.1038/s41437-019-0188-0>).

569 **Author contributions**

570 LD designed research; All authors performed research; LD and JL analyzed data; all

571 authors wrote the paper.

572

Table 1. Results obtained by null model tests implemented in *GenVectors* for haplotype and SNP datasets and based on either ADONIS or OLS models. The tests were designed to analyze the influence of present-day bioclimatic variables (temperature seasonality – TS, or precipitation of the driest quarter – PDQ) on genetic turnover among sets of populations of *Handroanthus impetigenosus*. D_P – haplotype/SNP dissimilarities between sets of populations computed based on matrix P .

Source	Data type	Method	Results
Collevatti et al. (2012)	Haplotype	ADONIS: D_P on TS	$F_{Obs.} = 1.56; P_{turnover} = 0.137, P_{divergence} = 0.073$
		ADONIS: D_P on PDQ	$F_{Obs.} = 2.30; P_{turnover} = 0.030, P_{divergence} = 0.022$
			Haplovector 1 (31%):
		OLS: Haplovector on TS	$F_{Obs.} = 2.03; P_{turnover} = 0.186, P_{divergence} = 0.295$
			Haplovector 2 (27%):
			$F_{Obs.} = 3.11; P_{turnover} = 0.111, P_{divergence} = 0.224$
			Haplovector 1 (31%):
		OLS: Haplovector on PDQ	$F_{Obs.} = 8.60; P_{turnover} = 0.012, P_{divergence} = 0.128$
			Haplovector 2 (27%):
			$F_{Obs.} = 1.36; P_{turnover} = 0.274, P_{divergence} = 0.404$
Collevatti et al. (2019)	SNP	ADONIS: D_P on TS	$F_{Obs.} = 4.89; P_{turnover} = 0.012, P_{divergence} = 0.001$
		ADONIS: D_P on PDQ	$F_{Obs.} = 1.58; P_{turnover} = 0.175, P_{divergence} = 0.165$
			Haplovector 1 (60%):
		OLS: SNPvectors on TS	$F_{Obs.} = 10.63; P_{turnover} = 0.014, P_{divergence} = 0.076$
			Haplovector 2 (19%):
			$F_{Obs.} = 0.15; P_{turnover} = 0.698, P_{divergence} = 0.497$
			Haplovector 1 (60%):
		OLS: SNPvectors on PDQ	$F_{Obs.} = 1.78; P_{turnover} = 0.166, P_{divergence} = 0.294$
			Haplovector 2 (19%):
			$F_{Obs.} = 0.49; P_{turnover} = 0.481, P_{divergence} = 0.459$

578

579

Fig. 1. Hypothetical scenarios driving haplotype-based genetic turnover among two sets of local populations (identified as roman numerals I-IV) located at opposite margins of a river, which may represent a potential barrier to dispersal of individuals among populations. The same rationale may be applied for SNPs. A) Haplotype network showing the number of mutational steps connecting six haplotypes identified by numerals 1 to 6. B) Gene flow occurs throughout all populations, which means that the river does not represent a barrier to dispersal among them. In this scenario, both null model tests implemented in *GenVectors* package should return P -values > 0.05 . C) Lack of gene flow among sets of populations located at opposite margins of the river is demonstrated since local populations located at the same margin share haplotypes while populations located at opposite margins do not. Since genetic divergence between the two metapopulations is not correlated to the dispersal barrier between them, we may conclude that gene flow has been interrupted recently, and genetic turnover between metapopulations was most likely caused by secondary contact. While ‘turnover’ null model test should return P -value ≤ 0.05 , ‘divergence’ null model test is expected to show P -value > 0.05 . D) As in C), we also observe lack of gene flow among sets of populations located at opposite margins of the river. However, in this scenario genetic turnover between the two metapopulations is correlated to the dispersal barrier between them. Thus, we may conclude that interruption of gene flow has historical origin, and genetic divergence was most likely caused by vicariance. Under such scenario, both null model tests should return P -values ≤ 0.05 .

598

599 **Fig. 2.** Schematic representation of classical analytical approaches used in spatial genetics and molecular
600 ecology to explore and analyze genetic turnover among populations (Goals 1 and 3), and the new fuzzy
601 set-based approach introduced in this study (Goals 2 and 4). *Analyses based on haplotypes:* GenVectors
602 (Goal 2) allows to explore the distribution of haplotypes across populations by means of scatterplots
603 where the most informative haplotype eigenvectors (haplo vectors) show which populations are more
604 similar from each other, and which haplotypes are more important to determine similarities among
605 populations. Furthermore, haplo vectors can be taken as response variables in hypothesis tests designed to
606 analyze the influence of environmental or biogeographic factors on genetic turnover among sets of
607 populations, revealing metapopulation structure. Although AMOVA is also appropriate for this purpose
608 (Goal 1), fuzzy set-based analyses have the advantage of evaluating different processes driving genetic
609 turnover via null model tests. *Analyses based on SNPs:* Commonly used methods such as DAPC and
610 Bayesian clustering (STRUCTURE) allow to explore genetic similarities among population based on
611 clustering methods (Goal 3), but appropriate hypothesis testing aiming to reveal drivers of genetic
612 turnover using these approaches is not straightforward. In Goal 4 we see that a similar approach
613 developed to analyze haplotype turnover among sets of populations and reveal processes underlying
614 genetic turnover can be used to analyze SNP turnover and infer processes underlying metapopulation
615 structure.

616

617 **Fig. 3.** Map showing the geographical distribution of 12 sampled populations of *H. impetiginosus*
618 analyzed with both datasets used in this study (cpDNA and SNP). The letter codes denote sampling
619 localities: ALT, Altamiro Pacheco Forest Park, GO; BAG, Barra do Garça, MT; CAC, Cáceres, MT;
620 MIM, Mimoso, GO; MOC, Montes Claros, MG; NIQ, Niquelândia, GO; PAN, Januária, MG; POF, Porto
621 Ferro, MA; POT, Portelândia, GO; POS, Posse, GO; SEC, Campo Maior, PI; SUM, Lagoa Santa, MG.
622 Grey area represents the distribution of Cerrado biome.
623

Fig. 4. Scatter plots for the two first haplovecors and SNPvecors computed for different datasets. a) cpDNA dataset. Grey circles indicate the populations, while three-letter codes indicate population names (see Fig. 3). Open circles indicate the haplotypes. PDQ – Precipitation of the driest quarter. b) neutral SNP dataset. Grey circles indicate the populations, while three-letter codes indicate population names (see Fig. 3). Asterisks indicate each individual genotype. TS – Temperature seasonality.