GenVectors: an integrative analytical tool for spatial genetics

Leandro Duarte[1*], *Jacqueline 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: Genetic eigenvectors for molecular ecologists

Abstract

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

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

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

Introduction

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

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

A major goal of spatial genetics consists in identifying major drivers of genetic turnover across space, which can be expressed as the spatial distribution of alleles, genotypes, haplotypes, or single nucleotide polymorphisms (SNPs) among sets of local populations (Avise 2009; Balkenhol *et al.* 2017; Manel *et al.* 2003; Storfer *et al.* 2006). Whenever these drivers are environmental or biogeographic factors underlying the genealogical history of single-species populations, proper hypothesis testing may shed light on metapopulation connections from landscape to regional scales (Hale *et al.* 2013; Hand *et al.* 2016), or even biogeographic connections across entire regions, biomes or continents (Carnaval *et al.* 2014; Diniz□Filho *et al.* 2008). In this sense, statistical phylogeographical approaches (Papadopoulou & Knowles 2016; Templeton 2004) helped moving that field beyond the essential descriptive nature present in its infancy. Analysis of molecular variance (AMOVA, Excoffier *et al.* 1992) has been widely used for analyzing genetic dissimilarities (turnover) among local populations based on haplotypes (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 (Anderson 2001; Pillar & Orlóci 1996) allows analyzing whether pairwise genetic dissimilarities between individuals of different populations or sets of populations are higher than expected by chance given within (meta)population dissimilarities. It usually complements exploratory methods to visualize haplotype links 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

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

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

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

In the package *GenVectors* we implement functions to perform genetic fuzzy weighting either for datasets containing information of distribution of haplotypes (function 'HaploVectors') or SNPs (function 'SNPVectors') among sets of populations. The first analytical step implemented in the 'HaploVectors' function consists of defining haplotypes for a set of samples and computing the frequency (number of individuals) of each haplotype per population. 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 occurrence of each individual (columns) in a given population (rows). Based on these two datasets, the function 'HaploDist' extracts the haplotypes for the *.fas file using the function 'haplotype' implemented in the R

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

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

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

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

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

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

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

Euclidean distances or Bray-Curtis dissimilarities (Legendre & Legendre 2012), or directly on single haplotype/SNP eigenvectors. The test consists of 1) computing a $F_{Obs.}$ statistic. If $D_{\mathbf{P}}$ 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/SNP eigenvectors modeled on **E**, the test is an ordinary linear squares (OLS) model; 2) freely permuting the populations 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 genetic turnover among sets of populations (for either haplotypes or SNPs) via both the frequency of haplotypes/SNPs shared between pairs of populations, and the genetic relatedness between the haplotypes/SNPs (the number of mutational steps between the pairs of haplotypes or the proportion of alleles shared between SNPs). Therefore, whenever this first hypothesis is corroborated (H_0 is rejected), the test allows to conclude that populations occurring under similar levels of **E** are more genetically connected to each other than to populations occurring under different levels of **E**. Nonetheless, genetic connectedness or isolation among populations might be determined either by 1) non-historical processes, such as recent secondary contact, or 2) history-related processes related to vicariance or an ancient dispersal event, which make genetic turnover among populations to be related to the amount of genetic divergence among them. To corroborate (or not) this second scenario, 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 genetic turnover among sets of populations is related to the genetic divergence among them.* To test this hypothesis, a second round 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 $D_{\mathbf{P}}$ 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 $D_{\mathbf{P}}$ 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). 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 **P** - $D_{\mathbf{P}}$), and OLS (for haplotypic eigenvectors). For ADONIS, we computed matrix **P** using frequencies of haplotypes per site (matrix **W**) and square-rooted Hamming distances between haplotypes (matrix **D**). Haplotypic dissimilarities between sites were computed using Bray-Curtis dissimilarities. Either bioclimatic variables were taken as continuous predictor in the analysis (matrix **E**). For OLS, we first computed haplotypic eigenvectors (haplovectors) based on $D_{\mathbf{P}}$. 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: 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

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

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

Discussion

Analyses implemented in *GenVectors* allow to explore genetic turnover across (meta)populations and analytically infer the role of recent secondary contact versus long-term split (driven by vicariance or ancient dispersal) on spatial genetic patterns by means of appropriate null model tests. Moreover, *GenVectors* was demonstrated to be a flexible tool in the sense it accommodates different genetic data (haplotypes, SNPs). These features make the tool useful for several different purposes in spatial genetics, and complementary to classic analytical tools widely used by molecular ecologists, such as AMOVA (Excoffier *et al.* 1992), individual-based Bayesian clustering (Pritchard *et al.* 2000), or DAPC (Jombart *et al.* 2010). Fundamental questions in ecology and 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 (≅ 30% of total information in matrix **P**). While haplovectors 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 (≅ 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.

Avise 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.

Beheregaray LB (2008) Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere. *Molecular Ecology* **17**, 3754-3774.

Blom MP, 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.

Carnaval AC, Waltari E, Rodrigues MT, *et al.* (2014) Prediction of phylogeographic endemism in an environmentally complex biome. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20141461.

Collevatti RG, Novaes E, Silva-Junior OB, *et al.* (2019) A genome-wide scan shows evidence for local adaptation in a widespread keystone Neotropical forest tree. *Heredity* (DOI: 10.1038/s41437-019-0188-0).

Collevatti RG, Terribile LC, Lima□Ribeiro MS, *et al.* (2012) A coupled phylogeographical and species distribution modelling approach recovers the demographical history of a N eotropical seasonally dry forest tree species. *Molecular Ecology* **21**, 5845-5863.

Dale MB (1977) Planning an adaptive numerical classification. *Vegetatio* **35**, 131-136.

Diniz□Filho JAF, de Campos Telles MP, Bonatto SL, *et al.* (2008) Mapping the evolutionary twilight zone: molecular markers, populations and geography. *Journal of Biogeography* **35**, 753-763.

Duarte LDS, Debastiani VJ, Freitas AVL, Pillar VD (2016) Dissecting phylogenetic fuzzy weighting: theory and application in metacommunity phylogenetics. *Methods in Ecology and Evolution* **7**, 937-946.

Epperson B (2005) Estimating dispersal from short distance spatial autocorrelation. *Heredity* **95**, 7-15.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479-491.

Feoli E, Zuccarello V (1986) Ordination based on classification: yet another solution?! *Abstracta Botanica* **10**, 203-219.

Fitzpatrick BM (2009) Power and sample size for nested analysis of molecular variance. *Molecular Ecology* **18**, 3961-3966.

García□Verdugo C, Forrest AD, Fay MF, Vargas P (2010) The relevance of gene flow in metapopulation dynamics of an oceanic island endemic, Olea europaea subsp. guanchica. *Evolution: International Journal of Organic Evolution* **64**, 3525-3536.

Hale JM, Heard GW, Smith KL, *et al.* (2013) Structure and fragmentation of growling grass frog metapopulations. *Conservation Genetics* **14**, 313-322.

Hand BK, Muhlfeld CC, Wade AA, *et al.* (2016) Climate variables explain neutral and adaptive variation within salmonid metapopulations: the importance of replication in landscape genetics. *Molecular Ecology* **25**, 689-705.

Hanski I, Gilpin M (1991) Metapopulation dynamics: brief history and conceptual domain. *Biological Journal of the Linnean Society* **42**, 3-16.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907-913.

Jackson DA (1995) Protest: A Procrustean Randomization test of community environment concordance. *Ecoscience* **2**, 297-303.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94.

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

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

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

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

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

Maestri R, Fornel R, Gonçalves GL, *et al.* (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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

Turchetto□Zolet A, Pinheiro F, Salgueiro F, Palma□Silva C (2013) Phylogeographical patterns shed light on evolutionary process in S outh A merica. *Molecular Ecology* **22**, 1193-1213.

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

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

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

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

Data Availability Statement

The data that support the findings of this study are openly available in Genbank (cpDNA dataset - http://doi.org/10.1111/mec.12071; SNP dataset - http://doi.org/10.1038/s41437-019-0188-0).

Author contributions

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

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*. $D_{\mathbf{P}}$ – haplotype/SNP dissimilarities between sets of populations computed based on matrix **P**.

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

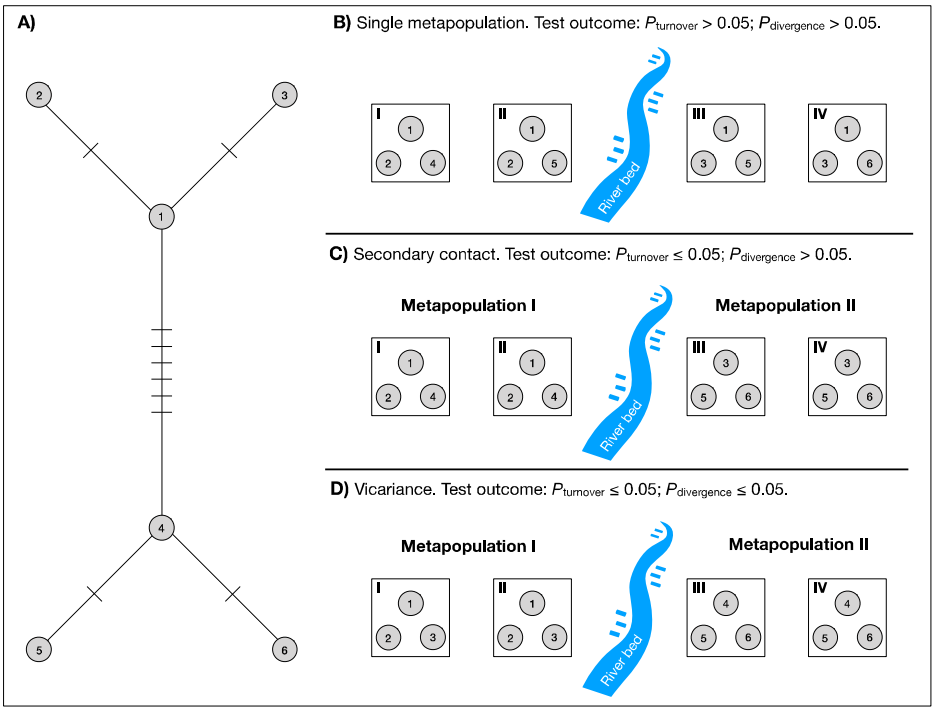

Fig. 1. Hypothetical scenarios driving haplotype-based genetic turnover among two sets of local populations (identified as roman numerals I-IV) located at opposite margins of a river, which may represent a potential barrier to dispersal of individuals among populations. The same rationale may be applied 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.

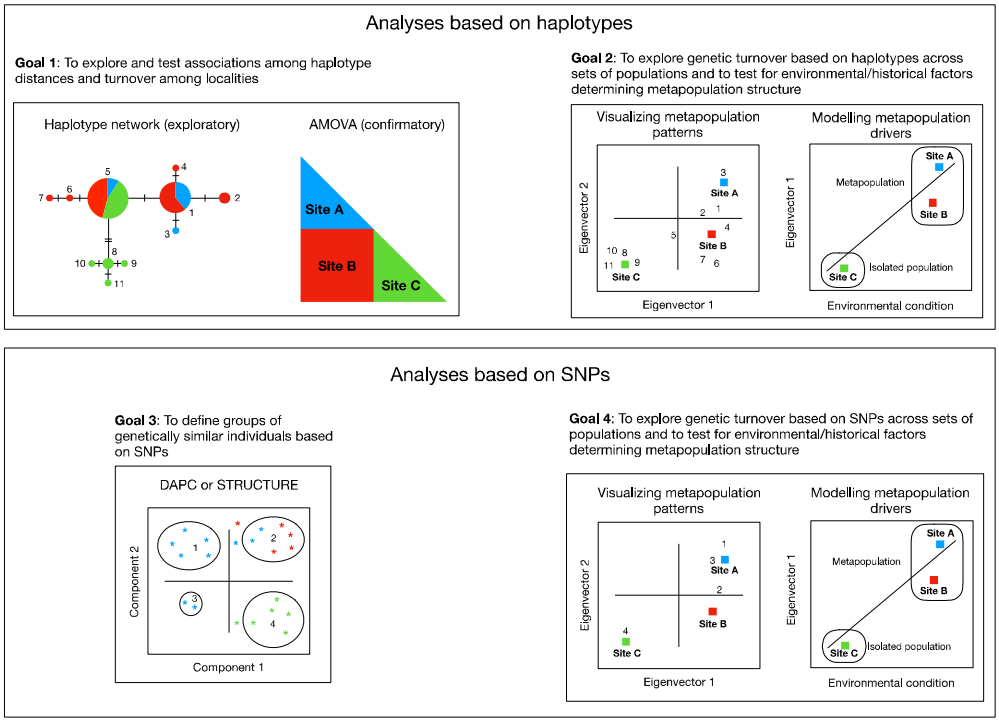

Fig. 2. Schematic representation of classical analytical approaches used in spatial genetics and molecular ecology to explore and analyze genetic turnover among populations (Goals 1 and 3), and the new fuzzy set-based approach introduced in this study (Goals 2 and 4). *Analyses based on haplotypes*: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

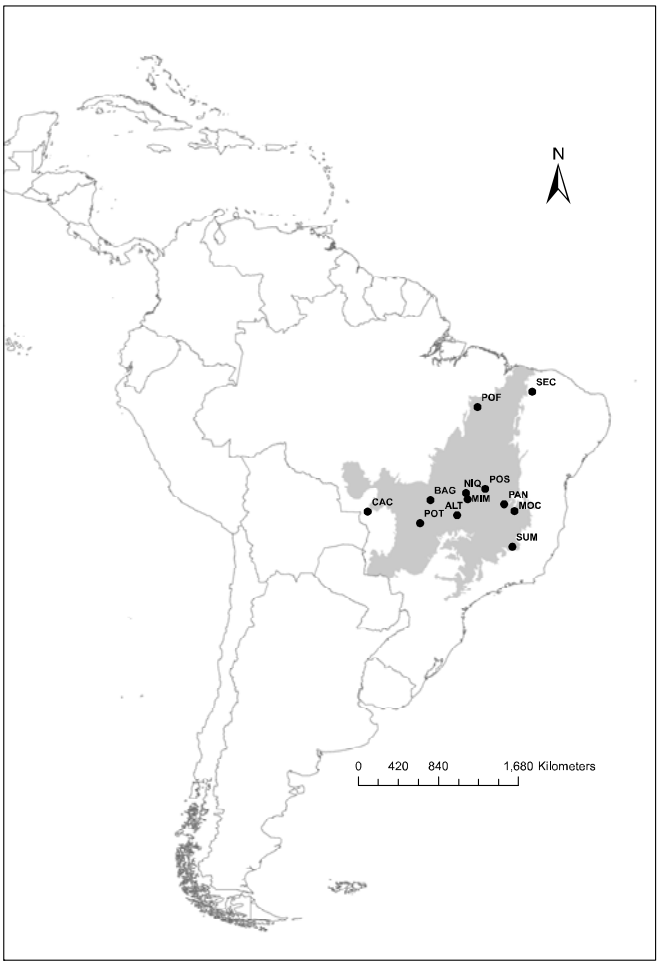

Fig. 3. Map showing the geographical distribution of 12 sampled populations of *H. impetiginosus* analyzed with both datasets used in this study (cpDNA and SNP). The letter codes denote sampling localities: ALT, Altamiro Pacheco Forest Park, GO; BAG, Barra do Garça, MT; CAC, Cáceres, MT; MIM, Mimoso,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

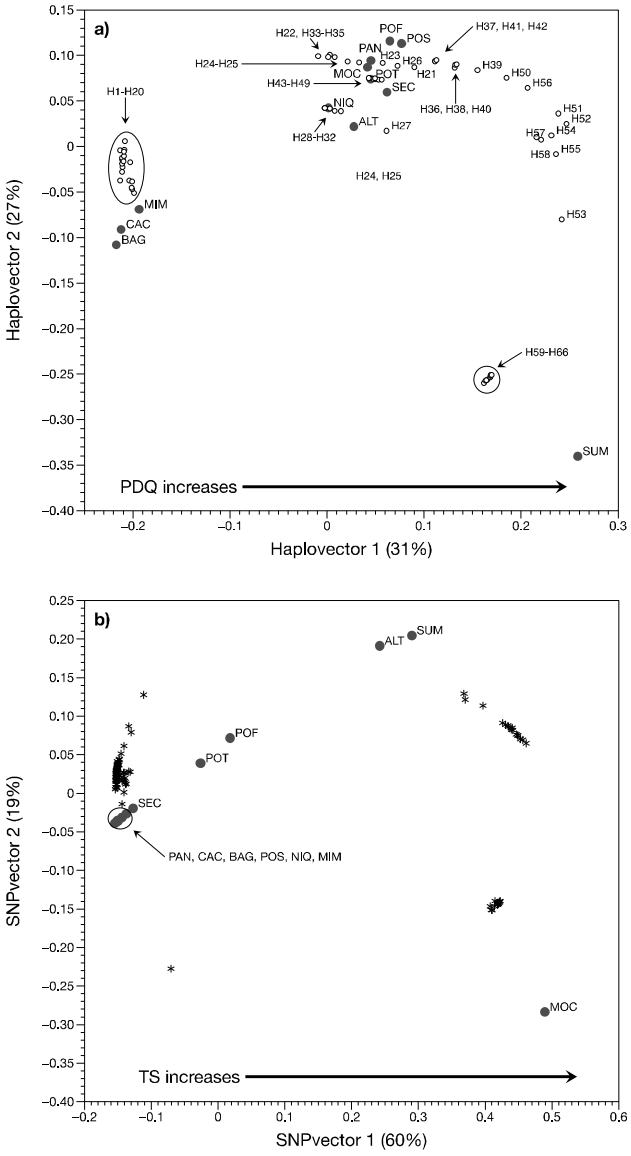

Fig. 4. Scatter plots for the two first haplovectors and SNPvectors computed for different datasets. a) cpDNA dataset. Grey circles indicate the populations, while three-letter codes indicate population names (see Fig. 3). Open circles indicate the haplotypes. PDQ – Precipitation of the driest quarter. 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.