Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-026-00669-6

REGULAR ARTICLE

Approaches to biological species delimitation based on genetic and spatial dissimilarity

Gabriele d'Angella[1] · **Christian Hennig**[1]

Received: 13 March 2025 / Revised: 12 December 2025 / Accepted: 20 January 2026

Abstract
The delimitation of biological species, i.e., deciding which individuals belong to the same species and whether and how many different species are represented in a dataset, is key to the conservation of biodiversity. Much existing work uses only genetic data for species delimitation, often employing some kind of cluster analysis. This can be misleading, because geographically distant groups of individuals can be genetically quite different even if they belong to the same species. We investigate the problem of testing whether two potentially separated groups of individuals can belong to a single species or not, based on genetic and spatial data. Already existing methods such as the partial Mantel test and jackknife-based distance-distance regression are considered. New approaches, i.e., an adaptation of a mixed effects model, a bootstrap approach, and a jackknife version of partial Mantel, are proposed. All these methods address the issue that distance data violate the independence assumption for standard inference regarding correlation and regression. A standard linear regression is also considered. The approaches are compared on simulated meta-populations generated with the software packages SLiM and GSpace that can simulate spatially explicit genetic data at an individual level. Simulations show that the new jackknife version of the partial Mantel test provides a good compromise between power and respecting the nominal type I error rate. Mixed-effects models have larger power than jackknife-based methods, but in some situations they display type I error rates above the significance level. An application on brassy ringlets concludes the paper.

Keywords Distance-distance regression · Partial Mantel test · Mixed effects model · Jackknife · Bootstrap · Biodiversity

Mathematics Subject Classification 62P10 · 62F40 · 62J99

This work was supported by the Alma Mater Studiorum - University of Bologna.

Extended author information available on the last page of the article

Published online: 24 February 2026

1 Introduction

For the delimitation of biological species, empirical data is used to determine which groups of individual organisms constitute different populations of a single species and which constitute different species (Rannala and Yang 2020). Species delimitation is crucial for the preservation of biodiversity and has applications in several areas, such as ecology and medicine (Burbrink and Ruane 2021). The empirical data employed to delimit species can be molecular (see, e.g., Rannala and Yang (2020) for a review), morphological (Gratton et al. 2016), behavioural (Scapini et al. 2002), ecological (Raxworthy et al. 2007; Rissler and Apodaca 2007). There are also integrative approaches using different types of data (Edwards and Knowles 2014) and methods (Carstens et al. 2013).

Spatial information is key for this task, as witnessed by the increase in publications in the field of landscape genetics (Storfer et al. 2010), which combines population genetics and landscape ecology (Balkenhol et al. 2015). Neglecting geographical information when delimiting species can lead to misassessment of the genetic structure in the data (Frantz et al. 2009). This can particularly happen in the presence of spatial patterns of genetic differentiation, such as isolation by (geographical) distance (IBD; Ishida 2009): ignoring spatial information, individuals may be wrongly assigned to different species because their genetic dissimilarity tends to increase with geographical separation (Bradburd et al. 2018), violating the often involved assumption of random mating within the population.

A way to include spatial information in molecular species delimitation routines is to study the relationship between genetic dissimilarity and geographical distance. The investigation of this relationship has a long tradition in the population genetics literature, where it was pursued to study migration models (Kimura and Weiss 1964) or estimate demographic parameters (Slatkin 1993; Rousset 1997; Clarke et al. 2002). While isolation by distance assumes that the genetic dissimilarity between two individuals simply increases with a plain geographical distance (Euclidean distance, or the geodesic distance based on latitude and longitude), isolation by resistance (IBR; McRae 2006) takes into account landscape features such as rivers and heights: this translates to developments like the least-cost path approach (Adriaensen et al. 2003), the circuit-based framework (McRae et al. 2008) or the least-cost transect analysis (Van Strien et al. 2012), in which a more sophisticated "landscape" dissimilarity measure is used. We will work with the Euclidean distance as geographical distance here, but all the considered methods also work with other geographical distances and dissimilarities.

Consider a setup with two groups of individuals to be tested for conspecificity, i.e., being from the same species. Inference is based on checking whether the relationship between genetic and geographical dissimilarity differs between pairs of individuals in the same group and pairs in different groups (see Sects. 2 and 5 for details). Each of the three panels in Fig. 1 shows genetic and geographical dissimilarities of two groups for which a test for conspecificity is of interest. Distances within the two groups are black circles and red triangles, distances between the groups are green diamonds. The plots show some (albeit weak) tendency that larger genetic distance comes with larger geographical distance, also within groups. On the upper left side,

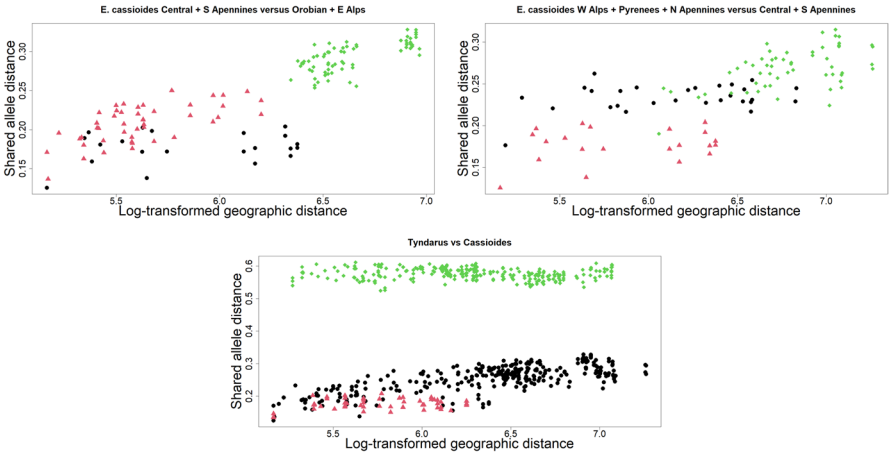

Fig. 1 Log-transformed geographical distances vs. Shared allele distances for three pairs of groups from the brassy ringlets data (upper left side *E. Cassioides central + S. Apennines* vs. *Orobian + E. Alps*; upper right side *E. Cassioides W. Alps, Pyrenees + N. Apennines* vs. *central + S. Apennines*; lower plot *E. Tyndarus* vs. *E. Cassioides*), see Sect. 5, for which conspecificity is of interest. The black circles (first group) and red triangles (second group) show distances between pairs of individuals belonging to the same group. The green diamonds show the distances between two individuals belonging to different groups

genetic ("shared allele") distances between groups seem slightly higher on average than genetic distances within groups, but also the geographical distances tend to be higher, and just from looking at the data it is not clear cut whether larger genetic distances between groups can be explained by the geographical distances only (in which case there is no reason to consider the two groups as different species), even less so on the upper right side. In the lower plot, it is clear that genetic distances between groups are much larger than they could be expected to be in case the two groups belonged to the same species.

The impact of grouping on the genetic dissimilarity can be quantified controlling for the effect of geographical distance. Medrano et al. (2014) used a permutation-based partial Mantel test (PMT; Smouse et al. 1986) to assess the significance of the partial correlation coefficient between genetic and grouping dissimilarities given the geographical distance, where the grouping dissimilarity is defined as 0 if a pair of individuals is in the same group and 1 otherwise. Hausdorf and Hennig (2020) suggested a jackknife test for whether a regression fitted on the within-group distances can also explain the between-groups distances. Clarke et al. (2002) employed individual random effects in order to model the dependence between dissimilarities of the same kind belonging to the same individual. This approach can be extended and adapted to the IBD problem by testing for an effect of the grouping dissimilarity as we do in the present paper. As further new approaches, we consider partial Mantel tests using jackknife or bootstrap instead of permutations. All these techniques have in common that they use statistics that were originally devised for situations with i.i.d. (independently identically distributed) data or residuals. Dissimilarities involving the same individual are certainly not independent, but the methods account for dependence by using permutations or jackknife at the individual rather than the dis-

similarity level, or by modelling dependence using a random effect. For exploring how much of a difference taking into account the dependence between dissimilarities actually makes, we also consider a multiple regression with genetic dissimilarities as response and geographical and grouping dissimilarities as explanatory variables, ignoring dependence between dissimilarities.

There exists only anecdotal evidence of the performance of the methods in Medrano et al. (2014); Spriggs et al. (2018) and Hausdorf and Hennig (2020); they have not been systematically assessed from a statistical perspective. The study presented here consists in a systematic comparison of the type I error rate and power of the aforementioned methods based on simulating spatially-explicit genetic datasets generated by the simulators GSpace (Virgoulay et al. 2021) and SLiM (Haller and Messer 2023).

A distinctive feature of the present study is that the genetic dissimilarities on which inference is based are much simpler than the data from which they are computed, see Sect. 2. To our knowledge, due to complexity, currently no tests for conspecificity exist that operate directly on the genetic information and take into account geographical distance. GSpace and SLiM provide sophisticated models for the genetic data, but the inference does not use such models. Instead it is based on much simpler models for the dissimilarities without taking into account how these were computed from the original genetic data, see Fig. 2. Here we confront such inference with the more complex genetic models for exploring its statistical characteristics. Methods and results may be relevant also for other problems where regression between dissimilarities is of interest.

The species concept in biology is somewhat controversial (Hausdorf 2011) and there is biological differentiation between populations at different levels: there are species that are more or less closely related, and there is differentiation also below the species level - see, e.g., De Queiroz (2007). Because of this, any result of the treated tests should not be taken as conclusive regarding conspecificity. The aim is just to formalise a key aspect of the information in the data. The tests developed here test for discontinuities in the gene flow. These may occur because different populations are actually different species, but could also be caused, for example, by geographical barriers.

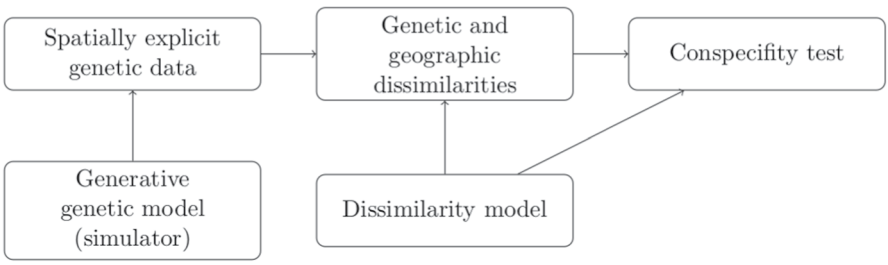

Fig. 2 The conspecificity tests treated here are computed on genetic and geographical dissimilarities assuming models for general dissimilarities (i.e., not taking into account how exactly the dissimilarities came about). The dissimilarities are computed from the originally observed spatially explicit genetic data, which in our simulation study are simulated from generative genetic models

This paper is organized as follows: the data and distances are introduced in Sect. 2, then all methods are discussed in Sect. 3. The two simulators used in this study and the results obtained with them are discussed in Sect. 4. Section 5 presents an application to the brassy ringlets data examined by Gratton et al. (2016). Section 6 concludes the paper.

2 Data and dissimilarities

Spatially explicit genetic data consists of individuals carrying information about their location and genetic make-up. Methods will be applied on individuals from two groups, with known membership. Hence, two columns will correspond to the unit's coordinates (northings and eastings, latitude and longitude, etc.), one column will report the group labels (either group 1 or 2) and the other P columns will be loci (locations on the DNA). Individual-level codominant data, such as SNPs or microsatellites, with diploid genotypes will be considered (Waits and Storfer 2015): this means that each locus will contain two alleles. Following Hausdorf and Hennig (2019) we represent alleles by single characters although elsewhere in the literature more elaborate coding is used (Rousset 2008). The resulting $n \times (P+3)$ data frame will be denoted by

$$\mathbf{Z} = \begin{pmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_n \end{pmatrix} = \begin{pmatrix} z_1^{(x)} & z_1^{(y)} & z_1^c & Z_1^1 & \cdots & Z_1^P \\ z_2^{(x)} & z_2^{(y)} & z_2^c & Z_2^1 & \cdots & Z_2^P \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ z_n^{(x)} & z_n^{(y)} & z_n^c & Z_n^1 & \cdots & Z_n^P \end{pmatrix},$$

where each observed locus Z_i^p, $p = 1, \ldots, P$, is a set of characters, like $\{A, B\}$ for heterozygous loci ("BA") or $\{B\}$ for homozygous ones ("BB"); also note that alleles are arranged in lexicographical order in the sets because a meaningful order is not normally observable. Each $\mathbf{z}_i$ is a $1 \times (P+3)$ vector representing the i^{th} individual. $z_i^{(x)}, z_i^{(y)}$ are geographical coordinates, and $z_i^c \in \{1, 2\}$ is a group indicator, where groups 1 and 2 represent the candidate species to be tested for conspecificity.

In this study, the Euclidean distance will be employed as geographical distance (subscript x):

$$d_x(\mathbf{z}_i, \mathbf{z}_j) = \sqrt{\left(z_i^{(x)} - z_j^{(x)}\right)^2 + \left(z_i^{(y)} - z_j^{(y)}\right)^2}.$$

As genetic dissimilarity (subscript y), the shared allele dissimilarity (Bowcock et al. 1994) will be used:

$$d_y(\mathbf{z}_i, \mathbf{z}_j) = 1 - \frac{1}{2P} \sum_{p=1}^{P} \left| Z_i^p \cap Z_j^p \right| \cdot \left[1 + \mathbb{1}\left(|Z_i^p| + |Z_j^p| = 2\right)\right], \tag{2.1}$$

where $\mathbb{1}(\text{condition}) = 1$ if the condition is true and zero otherwise. In real data occasionally there is missing data (missing loci). In this case d_y just averages over the loci that are non-missing in both $\mathbf{z}_i$ and $\mathbf{z}_j$. If there are no missing values, the shared allele dissimilarity is actually a distance (see proof in Supplementary Material), but missing values can cause a violation of the triangle inequality.

It is easy to see that, the larger P, the finer is the quantification of the genetic dissimilarity between two species, as more sites are available for the comparison of two individuals' genetic information.

Let $n_j = |\{i : z_i^c = j\}|$ be the number of individuals belonging to group $j = 1, 2$. In practice, this grouping information can be based on morphological, behavioural or even spatial grounds or can simply represent the researcher's hypothesis. The number of geographical distances in the dataset amounts to:

$$\frac{1}{2}(n_1+n_2)(n_1+n_2-1) = \underbrace{\frac{1}{2}n_1(n_1-1)}_{\text{within group 1}} + \underbrace{\frac{1}{2}n_2(n_2-1)}_{\text{within group 2}} + \underbrace{n_1 n_2}_{\text{between groups}},$$

to be stored in the following $n \times n$ block matrix, with $n = n_1 + n_2$,

$$\mathbf{D}_x = \left(\begin{array}{c|c} \mathbf{D}_x^{11} & \mathbf{D}_x^{12} \\ \hline \mathbf{D}_x^{21} & \mathbf{D}_x^{22} \end{array}\right) = \left(\begin{array}{cccc|ccc} 0 & d_x(\mathbf{z}_1,\mathbf{z}_2) & \cdots & d_x(\mathbf{z}_1,\mathbf{z}_{n_1}) & d_x(\mathbf{z}_1,\mathbf{z}_{n_1+1}) & \cdots & d_x(\mathbf{z}_1,\mathbf{z}_{n_1+n_2}) \\ d_x(\mathbf{z}_2,\mathbf{z}_1) & 0 & \cdots & d_x(\mathbf{z}_2,\mathbf{z}_{n_1}) & d_x(\mathbf{z}_2,\mathbf{z}_{n_1+1}) & \cdots & d_x(\mathbf{z}_2,\mathbf{z}_{n_1+n_2}) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ d_x(\mathbf{z}_{n_1},\mathbf{z}_1) & d_x(\mathbf{z}_{n_1},\mathbf{z}_2) & \cdots & 0 & d_x(\mathbf{z}_{n_1},\mathbf{z}_{n_1+1}) & \cdots & d_x(\mathbf{z}_{n_1},\mathbf{z}_{n_1+n_2}) \\ \hline d_x(\mathbf{z}_{n_1+1},\mathbf{z}_1) & d_x(\mathbf{z}_{n_1+1},\mathbf{z}_2) & \cdots & d_x(\mathbf{z}_{n_1+1},\mathbf{z}_{n_1}) & 0 & \cdots & d_x(\mathbf{z}_{n_1},\mathbf{z}_{n_1+n_2}) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ d_x(\mathbf{z}_{n_1+n_2},\mathbf{z}_1) & d_x(\mathbf{z}_{n_1+n_2},\mathbf{z}_2) & \cdots & d_x(\mathbf{z}_{n_1+n_2},\mathbf{z}_{n_1}) & d_x(\mathbf{z}_{n_1+n_2},\mathbf{z}_{n_1+1}) & \cdots & 0 \end{array}\right),$$

where matrix $\mathbf{D}_x^{11}$ stores the distances among the observations belonging to group 1, $\mathbf{D}_x^{22}$ those within group 2 and $\mathbf{D}_x^{12} = (\mathbf{D}_x^{21})^\intercal$ those among individuals of different groups. $\mathbf{D}_x$ carries redundant information: it is sufficient to work with the lower triangular matrix $\{d_x(\mathbf{z}_r, \mathbf{z}_c)\}_{r>c}$. Analogously, the $n \times n$ matrices $\mathbf{D}_y$ and $\mathbf{D}_g$ store the genetic and grouping dissimilarities, with $d_g(\mathbf{z}_i, \mathbf{z}_j) = \mathbb{1}(z_i^c \neq z_j^c)$.

We will later use the sets of all index pairs referring to within-group and between-group dissimilarities, respectively:

$$W = \Big\{(r,c)|\, c < r \leq n_1 \vee n_1 < c < r \leq n\Big\},\ B = \Big\{(r,c)|\, n \geq r > n_1 \wedge c \leq n_1\Big\}.$$

In order to improve the linearity between geographical and genetic dissimilarities, it is often advantageous to use log-transformed geographical distances as is done for example in Vekemans and Hardy (2004); Rousset (1997); Hausdorf and Hennig (2020). Zero geographical distances can occur if two individuals are observed in the same location. Therefore, following Hausdorf and Hennig (2020), the following transformation is considered:

$$f(d_x(\mathbf{z}_r, \mathbf{z}_c)) = \ln(d_x(\mathbf{z}_r, \mathbf{z}_c) + F_x^{-1}(0.25)), \tag{2.2}$$

where F_x is the empirical cdf of all geographical distances in the dataset, and $F_x^{-1}(\alpha)$ is the corresponding α-quantile. In our comparative study we will consider both untransformed and log-transformed $\mathbf{D}_x$. In order to keep notation light, all methods will be described using untransformed geographical distances.

Although the shared allele distance is used here, the discussed methods are based on models for dissimilarities that do not rely on the specific dissimilarity. The methods can therefore also be applied to other dissimilarities. In conspecificity testing, sometimes data come at population level with genetic distances between populations rather than individuals, e.g., in Clarke et al. (2002) and one example in Hausdorf and Hennig (2020).

3 Methods

The methodologies presented here use the information on the relationship between the distances in $\mathbf{D}_y$ and $\mathbf{D}_x$ with the aim of testing a conspecificity presumption encoded in $\mathbf{D}_g$. In the presence of isolation by distance behaviour (positive association between genetic and geographical dissimilarities), two groups of individuals belonging to the same species might display a certain degree of genetic structure that is explained by their geographical separation. If the genetic dissimilarities are too large to be compatible with the geographical separation between the two groups, this will constitute evidence for lineage separation, i.e., for distinctness. As Hausdorf and Hennig (2020) write, "*it is often difficult to assess whether observed differences between allopatric meta-populations would be sufficient to prevent the fusion of these meta-populations upon contact.*" In these situations, non-spatial models (to whom the putative grouping is often ascribed in practical applications) may be biased, and IBD patterns should be taken into account (Meirmans 2012).

The computation of dissimilarities implies information loss: the complex biological mechanisms (e.g., dispersal, see Cayuela et al. 2018) that act on the allele frequencies of the two investigated putative species have an indirect effect on the relationship between genetic and geographical dissimilarities, which can be nonlinear (Hutchison and Templeton 1999). The methods discussed here do not attempt to model such evolutionary processes, but rather work at the dissimilarity level, where the information from the P loci is summarized. The conspecificity null hypothesis is operationalised by these methods as having the same trend in the relation between genetic and geographical dissimilarities within groups and between groups. For the alternative hypothesis, genetic dissimilarities would be expected to be larger between groups than within groups when adjusted for geographical distances. The methods are based on linear regression and correlation, i.e., they are based on a linearity assumption. Note however that it can normally be expected with enough data that a zero correlation or regression slope can also be rejected if the relation is nonlinear but monotonic. Therefore the methods can be used also to detect nonlinear monotonic deviations from the null hypothesis, even though linearity would be the ideal condition. Incidentally, for Euclidean data, Székely et al. (2007) even show that indepen-

dence is equivalent to a "distance correlation", closely related to what is considered here, being zero. Nonlinearity occurs in some real datasets for example because of saturation effects, i.e., genetic distances reaching the maximum possible value; the Supplementary Material shows examples. Furthermore, the methods treated here do not require the triangle inequality, and monotonic transformations of dissimilarities can also be used.

All methods considered here are heuristic with weak (if any) theoretical justification for the given situation. Some of the methods have been explicitly criticized for the violation of model assumptions of existing theory, see below. To our knowledge, however, there are no alternatives with a stronger foundation, and the methods that are already published are used in practice. In general, statistical model assumptions can be seen as idealizations and are rarely (if at all) fulfilled in practice. This means that the violation of certain assumptions and the lack of theory applying to the specific situation does not automatically make a method invalid. The relevant question is whether certain issues with the data (such as violation of standard assumptions) have the potential to cause a method to produce seriously misleading results. For this reason an empirical investigation as done here should be worthwhile; we also give some remarks regarding the validity of the heuristics. More theoretical investigation is desirable, but it will probably be hard and is not the topic of this paper. Note in particular that our aim is not to advertise all involved methods or a specific one, but rather to explore strengths and weaknesses of all methods, and certainly all of them need to be applied cautiously.

In the following, the statistical methods involved in this comparative study are described.

3.1 Regression on dissimilarities with jackknife testing

Hausdorf and Hennig (2020) proposed to regress the genetic dissimilarities on the log-transformed geographical distances trying to clarify whether the genetic structure found between the two candidate species can be compatible with their IBD behaviour. To this end, a regression line based on the within-group dissimilarities (red and black observations in Fig. 1) is compared with a regression line based on all dissimilarities. The null hypothesis of conspecificity is rejected if the between-groups dissimilarities (green in Fig. 1) are systematically too large compared to what would be expected from the regression computed on the within-group dissimilarities. Dependence between dissimilarities is taken into account by running the test using a jackknife scheme that treats the individuals rather than the dissimilarities as observational units.

This approach is complicated by the fact that the test just mentioned relies on a single regression line being appropriate for the within-group dissimilarities in both groups. Hausdorf and Hennig (2020) propose a test protocol where it is first tested whether this is the case (H_{01}). Then, depending on the result, either a null hypothesis of a joint regression for all dissimilarities is tested (H_{02}, corresponding to conspecificity), or, in case that H_{01} is rejected, it is tested whether the between-groups distances are in line with at least one of the group-wise regressions of the within-group dissimilarities (H_{03}; in case that this is rejected, it is taken as evidence against

conspecificity, whereas non-rejection is an ambiguous result that would need closer biological investigation).

The first of the three tests focuses on the relationship between genetic and geographical dissimilarities within the two groups, assuming the following linear relationship:

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = \begin{cases} a_1 + b_1\{d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d_x}^W\} + e(\mathbf{z}_r, \mathbf{z}_c) & \text{with } c < r \leq n_1 \\ a_2 + b_2\{d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d_x}^W\} + e(\mathbf{z}_r, \mathbf{z}_c) & \text{with } n_1 < c < r \end{cases}. \quad (3.1)$$

a_1, b_1, a_2 and b_2 are estimated via least squares, and

$$\bar{d_x}^W = \frac{1}{|W|} \sum_{r,c \in W} d_x(\mathbf{z}_r, \mathbf{z}_c)$$

is the mean within-group geographical distance taken over both candidate species. The errors *e* in (3.1) are assumed to have zero mean, but not to be independent. Only the genetic random variation of individuals is assumed to be independent, but not dissimilarities involving the same individual.

The first test tests $H_{01}: \ a_1 = a_2$ and $b_1 = b_2$. It is tested against the two-sided alternative that $a_1 - a_2 \neq 0$ *or* $b_1 - b_2 \neq 0$. Both of these are tested and combined using Bonferroni, i.e., multiplying the minimum of the two p-values by 2.

In order to deal with the dependence between dissimilarities, Hausdorf and Hennig (2020) use non-parametric jackknife (already suggested by Clarke et al. (2002)) to obtain a measure of the variability of the estimates. Jackknifing (Efron and Tibshirani 1993, ch.11) here consists in computing as many OLS estimates as the number of individuals involved in a given regression model (e.g., n_1 for group 1) by fitting it on the n_1 datasets obtained by removing one individual at a time. In this particular setup, the removal of one individual implies the removal of all the dissimilarities related to it, so each jackknife replicate of the OLS estimates for group 1 is based on $(n_1 - 1)(n_1 - 2)/2$ data points instead of $n_1(n_1 - 1)/2$.

In jackknifing, so-called pseudovalues $u_i, \ i = 1, \ldots, n$, for a statistic U computed on data $\mathbf{X}$ with *n* observations are computed as $u_i = nU(\mathbf{X}) - (n-1)U(\mathbf{X}_{(i)})$ where $\mathbf{X}_{(i)}$ has the i^{th} observation left out. The variability of the difference between parameter estimates is quantified by pooling the within-group jackknife estimates of standard error (Efron and Tibshirani 1993, ch.11) in order to run a Welch's t-test (Welch 1947).

This principle is applied here to both the difference between intercepts and to the difference between slopes of the two within-group regressions, where the null hypothesis for Welch's t-test is that the expected difference is zero, see Hausdorf and Hennig (2020) for more details. Jackknife testing is a heuristic idea that has a theoretical justification only in specific situations (Shao and Wu 1989), the assumptions of which are not fulfilled here. Therefore, its characteristics have to be explored experimentally in all but the simplest situations, which is done here in Sect. 4. Jackknifing individuals treats the individuals rather than the dissimilarities as independent units of the analysis, and can be expected to lead to a larger jackknife standard error,

and therefore more conservative tests than jackknifing dissimilarities. This means in particular that the dependence between dissimilarities belonging to the same individual will not specifically invalidate the jackknife test (the t-test is not run on the dissimilarities but on the jackknife pseudovalues as originally proposed by Tukey, see Miller (1974)) beyond the fact that it is not covered by the existing theory.

If H_{01} is not rejected, a unique regression is fitted on all the within-group dissimilarities, regardless of the membership, because the IBD behaviour of the two candidate species looks compatible. In this situation, hypothesis H_{02} is tested. The following ordinary least squares model is fitted:

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = a_* + b_*(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d_x}^W) + e(\mathbf{z}_r, \mathbf{z}_c), \tag{3.2}$$

where $\mathbb{E}(e(\mathbf{z}_r, \mathbf{z}_c)) = 0$ and $r, c \in W$. This fit will be compared with the following model, which is based on all the dissimilarities in the dataset (within and between-group), regardless of the grouping:

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = a + b(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d_x}^W) + e(\mathbf{z}_r, \mathbf{z}_c), \tag{3.3}$$

where $c < r \leq n$ and $\mathbb{E}(e(\mathbf{z}_r, \mathbf{z}_c)) = 0$. Define $\bar{d_x}^B = \frac{1}{|B|}\sum_{r,c\in B} d_x(\mathbf{z}_r, \mathbf{z}_c)$, the average between-group geographical distance. $H_{02}: \ a = a^*$ and $b = b^*$ is then tested against the one-sided alternative

$$a + b(\bar{d_x}^B - \bar{d_x}^W) > a_* + b_*(\bar{d_x}^B - \bar{d_x}^W), \tag{3.4}$$

i.e., genetic dissimilarities predicted at $\bar{d_x}^B$ by all dissimilarities combined are systematically larger than predicted by within-group dissimilarities only. The statistic on which jackknife testing is based is

$$\hat{a} + \hat{b}(\bar{d_x}^B - \bar{d_x}^W) - \hat{a_*} - \hat{b_*}(\bar{d_x}^B - \bar{d_x}^W), \tag{3.5}$$

where $\hat{a}$, $\hat{b}$, $\hat{a_*}$ and $\hat{b_*}$ are the corresponding OLS estimates.

If H_{01} is rejected, the IBD behaviour of the two candidate species cannot be described by a unique model and model (3.1) is adopted. In this situation, H_{03} is tested, that is, the compatibility of IBD behaviour and genetic structure is checked for each group separately. Two models similar to (3.3) are set up, each one comprising within-group dissimilarities from one of the groups only, together with the between-group dissimilarities, and two jackknife tests are run on statistics analogous to (3.5). $H_{03}: \ a_j = a_j^*$ and $b_j = b_j^*$ for at least one of $j = 1, 2$ is tested, where a_j, b_j refer to regressions based on dissimilarities within group j only, and a_j^*, b_j^* refer to regressions based on all dissimilarities involving a member of group j. The alternative is defined by analogy to (3.4). The test rejects H_{03} if the maximum of the p-values for the two tests regarding groups $j = 1, 2$ is too small. Note that jackknifing these tests involves two different kinds of pseudovalues. For example consider the test regarding group 1. Members of group 1 are involved in dissimilarities between other members of group 1 and group 2, whereas members of group 2 are only involved in dissimi-

larities to members of group 1. This is accounted for by the computation of the jackknife estimate of the standard error to be used for the t-test, see Hausdorf and Hennig (2020) and the documentation of the `prabclus R`-implementation (Hausdorf and Hennig 2019) for details.

A rejection to the test for either H_{02} or H_{03} constitutes evidence against the null hypothesis of conspecificity, suggesting that the relationship between genetic and geographical dissimilarities displayed by the two meta-populations cannot explain their genetic differences and they might thus represent two separated lineages.

3.2 The partial Mantel test

The null hypothesis of the simple Mantel test states that "*the distances among objects in matrix* $\mathbf{D}_y$ *are not (linearly or monotonically) related to the corresponding distances in* $\mathbf{D}_x$" (Legendre and Legendre 2012, p.600). The original test statistic by Mantel (1967) was a cross-product of the vectors of dissimilarities,

$$\sum_{c<r\leq n} d_y(\mathbf{z}_r, \mathbf{z}_c) \cdot d_x(\mathbf{z}_r, \mathbf{z}_c)\,,$$

the standardized version of which corresponds to the sample correlation coefficient between the vectors of dissimilarities:

$$r(\mathbf{D}_y, \mathbf{D}_x) = \frac{\sum_{c<r\leq n}(d_y(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_y)(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_x)}{\sqrt{\sum_{c<r\leq n}(d_y(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_y)^2 \sum_{c<r\leq n}(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_x)^2}}, \quad (3.6)$$

where $\bar{d}_y = \frac{1}{w}\sum_{c<r\leq n} d_y(\mathbf{z}_r, \mathbf{z}_c)$ is the overall average genetic dissimilarity and $\bar{d}_x = \frac{1}{w}\sum_{c<r\leq n} d_x(\mathbf{z}_r, \mathbf{z}_c)$ is the overall average geographical distance, with $w = n(n-1)/2$.

Partial Mantel tests were proposed by Smouse et al. (1986) and are based on a partial correlation coefficient here defined as

$$r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x) = \frac{r(\mathbf{D}_y, \mathbf{D}_g) - r(\mathbf{D}_y, \mathbf{D}_x)r(\mathbf{D}_g, \mathbf{D}_x)}{\sqrt{(1 - r(\mathbf{D}_y, \mathbf{D}_x)^2)(1 - r(\mathbf{D}_g, \mathbf{D}_x)^2)}}\,. \quad (3.7)$$

(3.7) quantifies the correlation between the genetic dissimilarities and the grouping distances after having accounted for the geographical distances. Medrano et al. (2014) tested the null hypothesis that $\rho(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x) = 0$, where ρ is the true underlying partial correlation in the population in order to ascribe the genetic structure found in two subgroups of trumpet daffodils to their lineage separation. The rejection of such a hypothesis led them to maintain that the IBD behaviour displayed by the groups was not sufficient to explain the genetic dissimilarity found between the groups and that these should therefore not be considered conspecific. In Fig. 1 the null hypothesis means that conditionally on geographical distances between-groups

genetic dissimilarities (i.e., green) are not systematically larger than within-group ones (i.e., red or black).

Hypothesis testing is usually carried out by means of permutations. Legendre (2000) carried out empirical comparisons of four permutation strategies for partial Mantel tests. His first strategy, the one used in this study, consists in permuting just one of the three dissimilarity matrices and recomputing the partial correlation coefficient a large number of times. The default number of permutations in the `ecodist` package by Goslee and Urban (2007), which was used in this study, is 1000. For each of these 1000 iterations, rows and corresponding columns in matrix $\mathbf{D}_y$ are permuted to yield $\mathbf{D}_y^*$, which implies the modification of $r(\mathbf{D}_y^*, \mathbf{D}_g)$ and $r(\mathbf{D}_y^*, \mathbf{D}_x)$ to be included in (3.7). If the two groups are separated species, the partial correlation between genetic and group dissimilarities should be positive (larger genetic dissimilarity between groups). Therefore a one-sided test is carried out, and the associated p-value is equal to the share of $r(\mathbf{D}_y^*, \mathbf{D}_g|\mathbf{D}_x)$ permutation replicates that are at least as large as the original value $r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x)$. Legendre (2000) remarked that this permutation strategy may lead to inflated type-I error if outlying dissimilarity values are present in the data, whereas skewness in the dissimilarities distribution should not represent an issue. Another difficulty is that the entries in $\mathbf{D}_g$ are binary, and Pearson's correlation is originally defined for continuous entries. It has been applied as point-biserial correlation also to binary vectors, and the assessment of significance by the permutation principle accounting for dependence of entries of $\mathbf{D}_g$ in the same row or column is arguably not invalidated by binary data. Mantel tests can be run with other correlation measures, see Dietz (1983), but it is not obvious why this should lead to improvements.

3.2.1 Testing with jackknife

Significance in partial Mantel tests is typically assessed via permutations. This, however, might introduce a distortion. Permuting $\mathbf{D}_y$ while keeping $\mathbf{D}_g, \mathbf{D}_x$ fixed generates data for which $\rho(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x) = 0$ as prescribed by the null hypothesis. However, on top of that, the permuted $\mathbf{D}_y$ will be independent of both $\mathbf{D}_g$ and $\mathbf{D}_x$, which may be inappropriate in a real situation. Other permutation schemes as listed in Legendre (2000) also come with potentially unrealistic implicit structural assumptions. Partial Mantel tests have been controversially discussed with mixed empirical results in various situations, see, e.g., Guillot and Rousset (2013); Legendre et al. (2015).

The potential distortion from permutation can be prevented by jackknifing the partial correlation (3.7), leaving one individual out at a time, and then generate pseudovalues and run a t-test as explained in Sect. 3.1.

3.2.2 Testing with bootstrap

Another option to assess the variability of the partial correlation coefficient $r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x)$ is by resampling n individuals with replacement, generating nonparametric bootstrap samples. This idea was discouraged in Clarke et al. (2002) and Hausdorf and Hennig (2020) because, whenever two identical individuals are sam-

pled more than once, the associated dissimilarities will be equal to zero, generating bootstrap samples that in most cases tend to display a larger proportion of zero dissimilarities with respect to the original data. However, to date, no systematic study has demonstrated the performance of nonparametric bootstrap for species delimitation tasks.

A seminal reference for this technique is Efron and Tibshirani (1993). We use the *bias-corrected* (BC) bootstrap confidence intervals here, defined and motivated in (Efron and Tibshirani (1993), ch.22.5). The null hypothesis of conspecificity is rejected in the lower bound of a $(1-2\alpha)$ bootstrap confidence interval (α-quantile of the bootstrap distribution adjusted for bias correction) for the partial correlation (3.7) is larger than 0.

3.3 The linear mixed effects model

Another approach to model the dependence between dissimilarities involving the same individual is via introducing individual random effects into a regression between geographical and genetic dissimilarities.

Clarke et al. (2002) proposed such a model. They were working with population-level genetic and geographical data. After centering the geographical distances to remove correlation between the intercept and slope estimates, they extended the linear regression between genetic and (log-transformed) geographical distances by introducing one random effect for each of the two populations on which the dissimilarity value was based. With the notation defined above and considering an individual-level analysis, it is possible to specify their model as

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = a + b\left(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_x\right) + \tau_r + \tau_c + \epsilon(\mathbf{z}_r, \mathbf{z}_c) \quad \text{with}\, n \geq r > c\,, \tag{3.8}$$

where a is a constant term, τ_i ($i = r$ or $i = c$) is a random effect representing the average deviation of d_y values involving population i from that expected from its d_x distances to the other populations and τ_i and $\epsilon(\mathbf{z}_r, \mathbf{z}_c)$ are assumed to be independent with $\epsilon(\mathbf{z}_r, \mathbf{z}_c)$ i.i.d. normally distributed. This specification assumes that dependence between two dissimilarities involving the same population can be expressed by an additive random value. Technically this allows for dissimilarities smaller than zero, and does not take into account dependence that involves more than two pairs of populations, as exists for distances at least due to the triangle inequality. The model can therefore not be fully correct for a regression between distances, but given that all models are idealisations and simplifications, the model can still be suitable if it allows for inference with good performance characteristics.

The authors fitted the model via restricted maximum likelihood (REML). It has gained popularity in the landscape genetics literature (Peterman and Pope 2021), being used to assess the effect of landscape variables on gene flow (Van Strien et al. 2012) and for landscape model selection (Shirk et al. 2018). It can be fitted using the `mlpe_rga` function from the `ResistanceGA R` package (Peterman 2018), based on the `lme4` package (Bates et al. 2015).

In order to apply model (3.8) to species delimitation, a fixed effect associated with the grouping distance $\mathbf{D}_g$ needs to be incorporated:

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = b_0 + b_1 d_x(\mathbf{z}_r, \mathbf{z}_c) + b_2 d_g(\mathbf{z}_r, \mathbf{z}_c) + \tau_r + \tau_c + \epsilon(\mathbf{z}_r, \mathbf{z}_c) \quad \text{with } n \geq r > c\,. \tag{3.9}$$

b_2 here is an intercept update for between group genetic dissimilarities, and the null hypothesis $b_2 = 0$ corresponds to conspecificity, which is tested against the one-sided alternative $b_2 > 0$. In Fig. 1, b_2 would be the amount by which the green between-groups dissimilarities are on average higher than the red and black within-group dissimilarities. This is similar to H_{02} in Sect. 3.1, assuming implicitly that there is no difference between the within-group regressions for group 1 and group 2. Even if there is such a difference, it can be seen as relevant to whether the dissimilarities between groups are systematically larger than a model defined on the aggregated within-group dissimilarities.

Note that different from the approaches in Sects. 3.1 and 3.2, (3.9) provides a *generative* model for dissimilarities, but we will not use it as such because it does not use information about the underlying genetic dissimilarities, and also, as argued above, it cannot be fully correct for these.

The test can be based on profile likelihood-based confidence intervals (CI) (Venzon and Moolgavkar 1988; Royston 2007). The null hypothesis is rejected if the lower boundary of the $(1 - 2\alpha)$ profile likelihood-based CI is larger than zero. These CIs are obtained in R via the `confint` function applied on the `mlpe_rga` output. Profile-likelihood-based CIs are connected to likelihood ratio tests. Therefore, since inference involves a fixed effect, model (3.9) should not be fitted with REML (West et al. 2022).

A related approach was used by Yang (2004) for estimating and testing for isolation by distance. Instead of introducing random effects explicitly, several standard correlation patterns for the $\epsilon(\mathbf{z}_r, \mathbf{z}_c)$ as available in the SAS PROC MIXED (SAS 2000) were used to model the dependence in the dissimilarities. This can be expected to be inferior to (3.9), because it does not use the information which dissimilarities belong to the same individual.

3.3.1 A linear regression model ignoring dependence

In our study we will also compare a straight linear regression model without the random effects that ignores the dependence between dissimilarities:

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = b_0^* + b_1^* d_x(\mathbf{z}_r, \mathbf{z}_c) + b_2^* d_g(\mathbf{z}_r, \mathbf{z}_c) + \epsilon(\mathbf{z}_r, \mathbf{z}_c) \quad \text{with } n \geq r > c\,, \tag{3.10}$$

assuming $\epsilon(\mathbf{z}_r, \mathbf{z}_c)$ i.i.d. normally distributed with zero mean. The null hypothesis is once more $b_2^* = 0$, tested against $b_2^* > 0$ with the standard regression t-test. A similar approach has been taken for distances in Spriggs et al. (2018). Note that if (3.10) held indeed, the null hypothesis $\rho(\mathbf{D}_y, \mathbf{D}_g | \mathbf{D}_x) = 0$ of the partial Mantel test would be equivalent to $b_2^* = 0$ (Legendre 2000).

4 Simulations

The simulators employed in this study simulate genetic data for individuals with geographical locations rather than dissimilarities, see Fig. 2. They are based on models for the evolutionary processes that lie behind the modification of the alleles in the loci sampled from the individuals' DNA. These algorithms simulate the life cycle of individuals that inhabit an artificial map and, generation by generation, exchange their genetic material through migration schemes that give rise to different IBD behaviours. The genetic make-up of the output individuals is the result of this complex set of factors, which will indirectly impact the relationship between genetic and geographical distances.

The choice of parameter values for complex simulations like these is a challenging task. While it is crucial to inform this choice via the best available knowledge from real data (Adrion et al. 2020), retrieving relevant information for the specific landscape model of interest can be hard. As an indication, Pope et al. (2015) found that almost one third of the publicly available datasets they surveyed could not be recreated, while 40% did not report geographical information; on top of this, only a share of the reproducible datasets would contain co-dominant markers compatible with the computation of shared allele distances. We mostly focus on simple archetypical situations that allow for a good understanding of the key factors that influence the methods' performances. In Sect. 4.3 we present a simulation that was designed based on two groups present in the real dataset presented in Sect. 5.

All tests are run at level $\alpha = 0.05$. In the following, the two simulators used in this comparative study and the corresponding results are described.

4.1 Simulations based on the SLiM simulator

As explained in its user manual (Haller and Messer 2016), SLiM (Haller and Messer 2023) is a forward-in-time simulation package for constructing genetically explicit individual-based evolutionary models. Its default settings include non overlapping generations, diploid individuals and offspring generated by recombination of parental chromosomes with the addition of new mutations. Within each species, an arbitrary number of subpopulations can be simulated, connected by any pattern of migration. Mutations at specific base positions in the genomes are explicitly modelled, also as nucleotide sequences. SLiM provides support for continuous space, either in one, two or three dimensions, and this can help simulate dispersal (mate choice with spatial kernel or nearest-neighbor search) and spatial competition. Importantly, SLiM allows the simulation of more than one species in a single SLiM model, opening the door to ecological interactions and coevolutionary dynamics. Virtually any feature of the simulated evolutionary scenario can be controlled via the integrated Eidos scripting language, which was created specifically for SLiM.

The short summary above hints at how its numerous simulation possibilities could be exploited to investigate the type I error rate and power of the methodologies explained in the previous sections. A two-group continuous space simulation, with individuals that compete for foraging areas (resulting in a more likely reproduction of more isolated individuals), choose mates among their nearest neighbors and gen-

erate offspring in their surroundings, will lead to observations isolated by distance. Instead, absence of competition and less parent-dependent offspring positioning will lead to quasi-panmictic results, i.e., to the lack of association between genetic and geographical distances. The individuals from the two groups might inhabit the same geographical area or might be segregated into two disjoint areas. As regards conspecificity and distinctness, we used five different scenarios:

- A simulation with one species and one subpopulation sampled with an artificial random split will return a naive conspecificity scenario, which can be used as baseline, referred to as *random split (rs)*.
- a simulation with only one species and two subpopulations, which descend from the same parent population and are able to exchange migrants, may yield a scenario consistent with the null hypothesis of conspecificity (referred to as *conspecifity*, with two versions explained below), because the two groups are related by their history and their individuals are expected to display a similar genetic make-up;
- a simulation exploiting SLiM's multispecies engine (introduced in chapter 19 of the manual), with two distinct species that cannot interact, will generate a scenario consistent with the alternative hypothesis of distinctness (referred to as *distinctness*, with two versions as explained below), since their genetic information is expected to be completely unrelated.

The details about SLiM's assumptions and key parameters for the simulations carried out in this study are reported below. The code can be found in the Supplementary Material.

The simulation was initialized on a two-dimensional map and with an explicit nucleotide sequence 1000 base pairs long: this means that a total of 1000 diploid loci (technically, two genomes with 1000 positions) were simulated, where four different alleles can be found - corresponding to the four nucleobases: adenine, cytosine, guanine and thymine. The initial nucleotide sequence was random and the recombination rate was set to the default low value of 10^{-8} (loci were not independent). Mutations were handled according to the Jukes-Cantor mutational model (Jukes and Cantor 1969) by specifying a matrix containing the mutation rates from one nucleobase to the other: a unique rate applied to all transitions among nucleobases and in these simulations it was set to 0.0025, a value that is larger than the default: too low values would lead to too few mutations and thus less genetic structure, given the timescale of the simulation. The map was always a square 200×200 units wide, but in any case species were not allowed to get out of the central 100×100 area. As far as mating is concerned, in all scenarios individuals would randomly pick a mate among their three closest neighbors, selected within a circular area of radius 3.

On top of these shared parameter options, the following settings varied according to the scenario:

- In the *random split* scenario, throughout the simulation, all individuals inhabited the central 100×100 area. In the *conspecificity* scenario, the parent population inhabited the central square area, but, depending on the version, the two "chil-

dren" subpopulations would continue to share the same wide area (*overlapping conspecific (oc)*) or start to migrate to two disjoint areas of the map (*separated conspecific (sc)*). By the time of the last simulated generation, the first subpopulation would inhabit the area between point (50, 50) and point (100, 100), whereas the second group would inhabit the area between point (100, 100) and point (150, 150), both included in the original wide central square area. Also in the *distinctness* scenario there were two versions: in the *overlapping distinct (od)* one, both species would inhabit the central 100×100 area, whereas in the *separated distinct (sd)* one, they would inhabit, since the very first generation, the area between point (50, 50) and point (100, 100) and the area between point (100, 100) and point (150, 150), respectively.

- In all simulations, SLiM would output the data regarding the individuals only at the last generation. In the *random split* scenario, one hundred generations were simulated. In the *conspecificity* scenarios, the parent population would be simulated for the first 100 generations and be removed afterwards; the two children subpopulations would originate from the parent one at generation 90 and then be simulated up to generation 150. In the *distinctness* scenarios, both species would be simulated for 50 generations.
- The only subpopulation existing in the *random split* scenario and the parent population in the *conspecificity* scenarios were made up of 400 individuals. The children subpopulations in the *conspecificity* scenarios and the two species in the *distinctness* scenarios were made up of 200 individuals each. Note that, in each SLiM cycle, individuals are born, mate and die, but the default behaviour of the software is to keep the number of simulated individuals steady throughout the generations.
- In order to generate diverse data richness situations, regardless of the scenario, the number of loci available for the computation of the shared allele distance was either 4, 40 or 89 (randomly selected) and the total number of individuals sampled was either 12, 30 or 90. The whole simulation study was carried out either with equally sized groups (e.g., 6 versus 6 individuals) or with one group being twice as large as the other (e.g., 4 versus 8). In the *random split* scenario there was just one big group from which individuals were drawn, and these individuals were then randomly assigned to the two groups used to test conspecificity. Given the membership, the drawing of individuals was random, except for the *separated conspecific* scenario, when it was constrained to those individuals inhabiting the subpopulation-specific geographical area of the map: indeed, given the migration process involved in that scenario, it could well happen that some individuals at the last generation were still positioned in the area specific to the other group, typically because one of their parents belonged there.
- As far as spatial competition is concerned, it was modeled through the effect that interactions between individuals had on their probability to reproduce. Each individual experienced an interaction strength that was the sum of all its interactions with individuals in the neighbourhood. In particular, a Gaussian kernel was used to translate the distance between two individuals in the strength of the interaction between them: when trying to enforce a strong IBD behaviour in the individuals, this Gaussian distribution would have mean 2.5 and standard deviation 5 and the

interactions with individuals out of the circular area of radius 15 would be set to zero; when trying to mimic no IBD species, the distribution would have mean 0.5 and standard deviation 1 and the circular area with positive-valued interactions would have radius 3. With the first parameter settings, given a certain Euclidean distance between two individuals, the strength of the interaction would be assigned a larger value: the stronger the total interaction felt by an individual, the lower its probability to reproduce, leading to the formation of isolated subgroups and hence to restricted gene flow. With narrower Gaussian kernels, instead, the total interaction strength on each individual would tend to be smaller and thus there would be less incentive to dispersal, resulting in a more panmictic-like behaviour.

- In the *conspecificity* scenarios, the two children subpopulations were allowed to exchange migrants. A migration rate of 20% means that when creating the offspring for, say, the first subpopulation, 20% of the parents (with some stochastic variability) were picked from individuals belonging to the second subpopulation. In the *overlapping conspecific* scenario, the two subpopulations would exchange parents at a rate randomly oscillating between 40 and 50% till the last generation. In the *separated conspecific* scenario, the migration rate would start off at 20% and then linearly decrease to reach zero in the last generation, at a pace that is consistent with the progressive separation of the geographical areas.
- As far as offspring generation is concerned, it occurred at every simulation cycle after the choice of the two mating parents: its position was shifted from that of the first parent according to a draw from a zero-mean Gaussian kernel with standard deviation 1 in case of strong IBD behaviour and 9 in case of quasi-panmictic behaviour. Thus, with strong spatial competition, the emerging isolated groups would tend to be preserved because offspring were more likely to emerge in a narrow neighbourhood of their parents. In the *separated conspecific* scenario, the location parameter of the Gaussian distribution involved in this process was modified according to the group: for the first group, that would end up in the square area between point (50, 50) and point (100, 100), the parameter was set to -0.5, whereas it was equal to 0.5 for the second group. Also in this respect, this is consistent with the gradual process of separation that affected the groups since the 100^{th} generation.

In the Supplementary Material, example distance-distance plots similar to Fig. 1 from the scenarios are shown, with equal or unequal group sizes, both for quasi-panmictic groups and for isolated by distance groups.

4.1.1 Results from SLiM

Five scenarios were simulated with SLiM: *rs* (one group, random split), *oc* (two groups, same parent population, same inhabited area), sc (two groups, same parent population, disjoint inhabited areas), *sd* (two species living in disjoint areas) and *od* (two species inhabiting the same area). Across the scenarios sorted in this way, the rejection rate from species delimitation methods is expected to be non-decreasing, since we transition from a clear conspecificity setup (*rs*) to a clear distinctness setup

(*od*); if species inhabit the same area, both conspecificity and distinctness can be more easily diagnosed, so that we expect a lower rejection rate for *oc* than for *sc*, but a higher one for *od* than for *sd*. On top of these scenarios, other varying parameters (all shared by both groups) were the IBD behaviour and the number of loci available for the computation of $\mathbf{D}_y$ (out of the 1000 loci simulated). In half of the cases the two groups were equally sized and in the other half $n_2 = 2n_1$. The combination of all these factors generated 36 scenarios and in each of them the techniques described in Sect. 3 were applied both with untransformed and log-transformed geographical distances. One hundred replicates of each of these combinations were generated and the number of rejections was recorded for all the methods. This information is visualized in power plots, one for equal and one for unequal group sizes. In these plots, HH20 denotes the protocol by Hausdorf and Hennig (2020); PMantel denotes PMT with permutations; MJack denotes PMT with jackknife; BC denotes PMT with bias-corrected bootstrap; MXD denotes the mixed effects model (3.9); LM is the multiple regression ignoring dependence.

In Fig. 3, rejection rates from the scenarios with equally sized groups are reported (methods' abbreviations are explained in the caption). Results with 40 available loci are only shown in the Supplementary Material, same regarding Sect. 4.2. The expected non-decreasing trend in the rejection rates was confirmed, with some minor exceptions between the *rs* and the *oc* (same parent, same area) scenario. In the *rs* scenario, all methods (surprisingly, model (3.10), too) displayed a type I error rate close to the significance level 5%, although the bias-corrected bootstrap with untransformed $\mathbf{D}_x$ rejected the null hypothesis of conspecificity too often in some setups. In general, the jackknife-based methods (HH20, MJack) showed type I error rates very close to zero, whereas all other methods had them slightly above the significance threshold. In the second scenario, especially with large samples, these methods showed rejection rates close to 10%.

In *sc* (same parent, disjoint areas) with quasi-panmictic groups, all rejection rates registered a strong increase: especially with many individuals and loci, all methods seemed to suggest that the two groups should be considered distinct, despite having originated from the same parent population. This was probably due to the combination of geographical separation and absence of IBD behaviour: the genetic structure that was formed because of the decreasing migration rate could not be explained by geographical distance as individuals in the same group tended to be quasi-panmictic. Indeed, when species were isolated by distance and with sufficient genetic information (89 available loci), the rejection rates in the *sc* scenario all fell below the significance level. Recalling the remark in the Introduction about different levels of biological differentiation, it can be controversial whether the groups generated with this particular SLiM script should be considered conspecific. Most methods concluded they are not, which is important information for biologists using these tests.

Under the multispecies setups, all methods displayed a rejection rate close to 1. The jackknife-based methods, though, tended to display lower power than the other methods, particularly with small sample sizes. This trend was common to all scenarios: PMantel, MXD, LM and BC always showed rejection rates larger than HH20 and MJack. In this respect, it is worth noting that MJack, representing a compromise between HH20 and PMantel, displayed satisfactory type I error rate and larger power

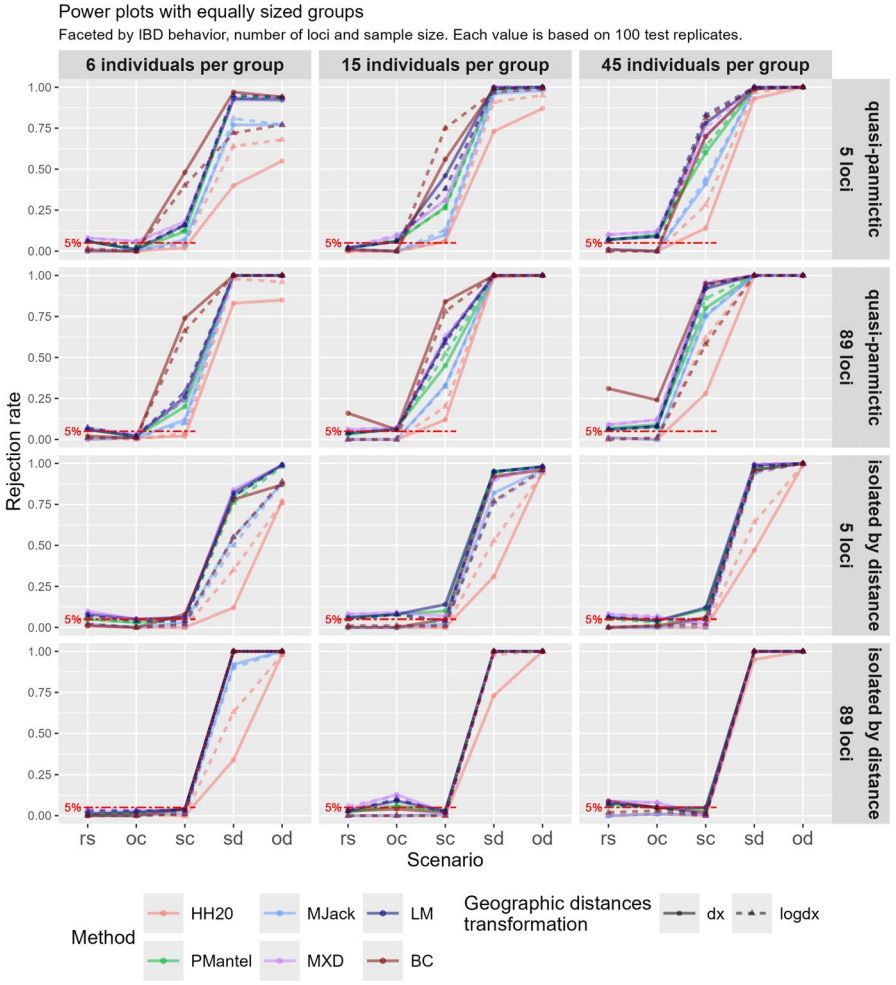

Fig. 3 Power plot based on **SLiM** data simulated **with equally sized groups**. Panels are done by number of individuals per group (rows) and a combination of IBD behavior and number of available loci (columns). Each rejection rate is based on 100 simulations with the same parameter settings. Circles and solid lines refer to analyses with untransformed geographical distances, whereas triangles and dashed lines refer to analyses with log-transformed ones. The horizontal dashed red line is superimposed to help assess type I error rates

than HH20 in all setups. Now, if the rejection of the null hypothesis in *rs* is seen as a Bernoulli random variable with success probability equal to 0.05, 10 rejections out of 100 would represent a significant result at 5% level. PMantel did not consistently display such significant figures in the simulations, but in an ad hoc simulation under the null hypothesis, with 15 individuals per group and 40 loci (not shown here), this test rejected 70 times out of 1000 repetitions (one-sided p-value $= 0.0023$ against the null hypothesis that the rejection probability is 0.05). By replacing the permutation-based significance assessment with a jackknife-based one, MJack achieved more power

than HH20 while keeping its same low type I error rate. This might support the idea that permutations introduce a distortion in the distribution of geographical distances.

Also in Fig. 4, with unequal group sizes, the rejection rates were mostly non-decreasing when going from *rs* to *od*. The most important difference regards the type I error rate of LM: when one group was twice as large as the other, the rejection rate of LM in the *rs* scenario often was above the significance level, sometimes strongly so, especially with the largest sample sizes.

Regardless of the other parameter options, the transformation of the geographical distances did not seem to have a relevant impact on the methods' performance. The only exception is HH20, whose power increased with log-transformed geographical dissimilarities.

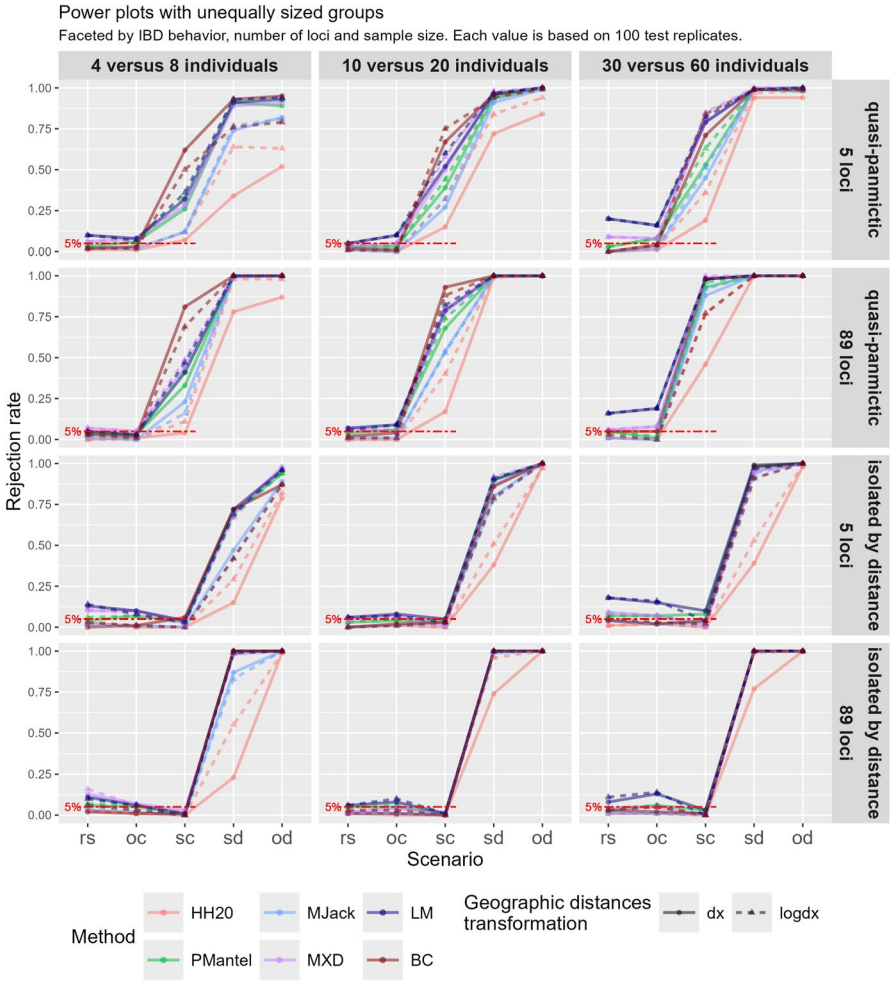

Fig. 4 Power plot based on **SLiM** data simulated **with unequally sized groups**. See the caption to Fig. 3 for further description

4.2 Simulations based on the GSpace simulator

GSpace is a coalescent-based simulator designed to model genetic variation in spatially structured populations under isolation by distance. It reconstructs the genealogical history of sampled individuals while accounting for recombination and migration across a lattice-based landscape. The lattice structure allows for flexible spatial modeling, where each vertex on the lattice can represent a local population or the position of individual organisms in a continuous habitat approximation.

The simulation proceeds backward in time: starting from the sampled individuals, their genealogies are traced until the most recent common ancestor (MRCA) is reached. At each generation, migration is modeled using two-dimensional dispersal distributions (e.g., uniform, geometric, or Zeta), determining how individuals move between locations. Once the MRCA is identified, mutations are introduced forward in time along the branches of the coalescence tree. Further details can be found in Leblois et al. (2009).

As far as the specific settings for the simulations carried out in this study are concerned, a 200×200 vertices map was created, with diploid individuals always inhabiting the central 60×60 area. A first group would inhabit the area between vertex (70, 70) and vertex (90, 90), whereas a second group would inhabit the square between vertices (110, 110) and (130, 130). Their coordinates were drawn uniformly within the allowed range. A sequence of di-allelic loci was simulated, with both the mutation rate per generation per locus and the recombination rate per generation between loci equal to 0.005 (ten times the default) for all simulated loci. The so-called K-allele mutation model was used, according to which the initial allelic state is changed into one of the other possible states - in this case the only other allele allowed in the di-allelic setting. In order to mimic a continuous habitat, in each vertex there could be up to one individual.

In addition to this, the following parameters varied according to the scenario:

- In order to obtain a baseline scenario where the conspecificity hypothesis was trivially true, the two groups were simulated within the same software execution, so that the algorithm would reconstruct a unique genealogy common to all individuals (referred to as *1* in the results plots). In all other cases, *2s*, *2o*, *2n*, see below), the two geographically separated groups were generated during two distinct software executions and collected in the same dataset. Note that the baseline scenario in SLiM had the two groups inhabiting the same area, unlike what is done here with GSpace.
- The two groups would obviously share the same allele pools (the same two allele options for all loci) when generated within the same software execution, whereas they could share both alleles (referred to as *2s*, "s" for "same allele pool"), one allele (*2o*) or no allele (*2n*) otherwise. Of course, when no allele was shared, the genetic dissimilarities between individuals from different groups would always take value one.
- IBD behaviour was controlled via the choice of the univariate dispersal distribution: GSpace assumes that dispersal occurs independently in each dimension. In order to yield panmictic groups, a uniform dispersal was used, according to which

the probability of moving t steps in one direction is $\frac{m}{d_{max}}$, where m is the total migration rate, equal to 0.5, and d_{max} is the maximum distance reachable in any migration event, set to 200 (the largest possible value) in all scenarios. As regards IBD species, a Zeta (or truncated discrete Pareto) dispersal distribution was used, assigning value $\frac{m}{2|t|^{\kappa}}$ to the probability to move t steps in one direction, with $\kappa = 5$ being the shape parameter.

- The total number of simulated individuals was either 12, 30 or 90, whereas the genetic sequence was either 4, 40 or 100 loci long. As with SLiM, all scenarios were investigated both with equally sized groups and with one group twice as large as the other.

The script on which these simulations were based and distance-distance plots from all simulated scenarios can be found in the Supplementary Material.

4.2.1 Results from GSpace

The combination of the parameter settings illustrated above led to a total of four scenarios: the simulation involving a unique software execution represented a conspecificity setup, whereas the situation where the allele pools of the two groups shared no allele constituted a distinctness one. The other two setups were included as intermediate situations, with the two groups not sharing ancestors while still showing similarities in their genetic information. Recall that in all these scenarios the geographical separation was always the same: the two groups inhabited two disjoint areas of the map. The four scenarios can be sorted as follows: one execution (*1*), two executions and same alleles (*2 s*), two executions and one allele in common (*2o*), two executions and no allele in common (*2n*). In this order, the rejection rate is expected to be non-decreasing.

On top of these scenarios, other varying parameters were the number of individuals per group, the number of loci available for the analysis and the IBD behaviour (absent with Uniform dispersal distribution and strong with Zeta dispersal distribution). Each of these parameter (and scenario) combinations was simulated one hundred times and the number of rejections was recorded.

In Fig. 5 the results related to situations in which the two groups are equally sized are reported. It is apparent how the ranking of the methods in terms of performance was similar to that observed in SLiM, apart from the odd behaviour of BC, which displayed very large type I error rates especially with fewer loci and fewer individuals. This is due to resampling with replacement, which led to the emergence of too many zero dissimilarities in the bootstrap samples. A closer look at the related distance-distance plots (not shown here) suggests that these zero dissimilarities were outlying with respect to the bulk of the data, leading to a biased distribution of bootstrap replicates of the partial correlation coefficient. Indeed, accounting for geographical separation, in those situations there appeared to be a large positive association between genetic and grouping distance, as zero distances could only be found when the grouping indicator was equal to zero. Consequently, bootstrap replicates had an upward bias that often caused a rejection of the null hypothesis. This phenomenon was not evident with SLiM data because its scenario *rs* did not involve geographically sepa-

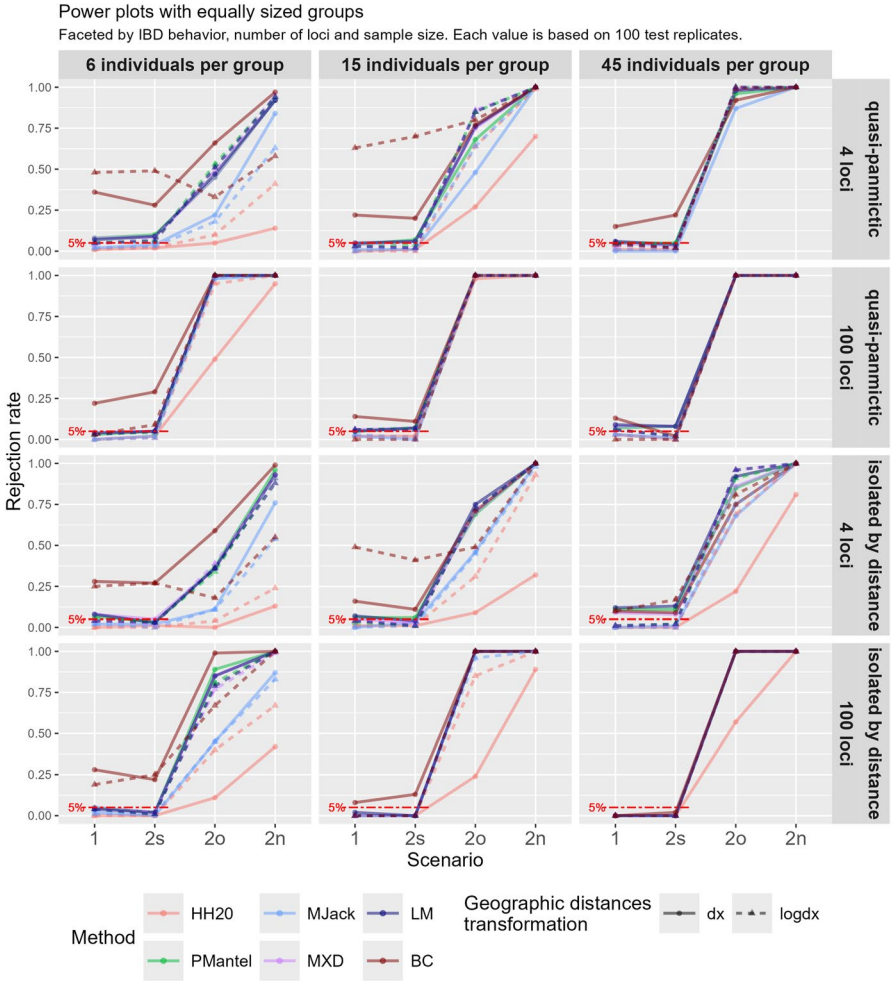

Fig. 5 Power plot based on **GSpace** data simulated **with equally sized groups**. Panels by combination of IBD behaviour and number of loci (columns) and number of individuals per group (rows). See the caption to Fig. 3 for further description

rated groups, unlike GSpace. These findings represent empirical evidence that partial Mantel testing with bootstrap can generate unreliable results.

The jackknife-based methods showed lower power with respect to other methods, and once again the logarithmic transformation of $\mathbf{D}_x$ helped HH20 to be more powerful. In scenario *2o*, given the number of loci and the total sample size, the rejection rates were lower with IBD species than with quasi-panmictic ones. Indeed, with species isolated by distance, geographical separation was sometimes enough to explain the genetic discrepancies between the two groups, and jackknife-based methods were more sensitive to this data feature than other methods.

Figure 6 shows that, when the two groups had unequal sizes, the rejection rates were similar to those with equally sized groups. A slight decrease of some rejection rates across all scenarios can be spotted in the column with IBD species and four loci.

4.3 A simulation inspired by real data

In order to incorporate a simulation using realistic parameters backed up by real data, we considered the Cassioides and Tyndarus subgroups of the brassy ringlets data analyzed and discussed in more detail in Sect. 5 and shown in the last plot in Fig. 1.

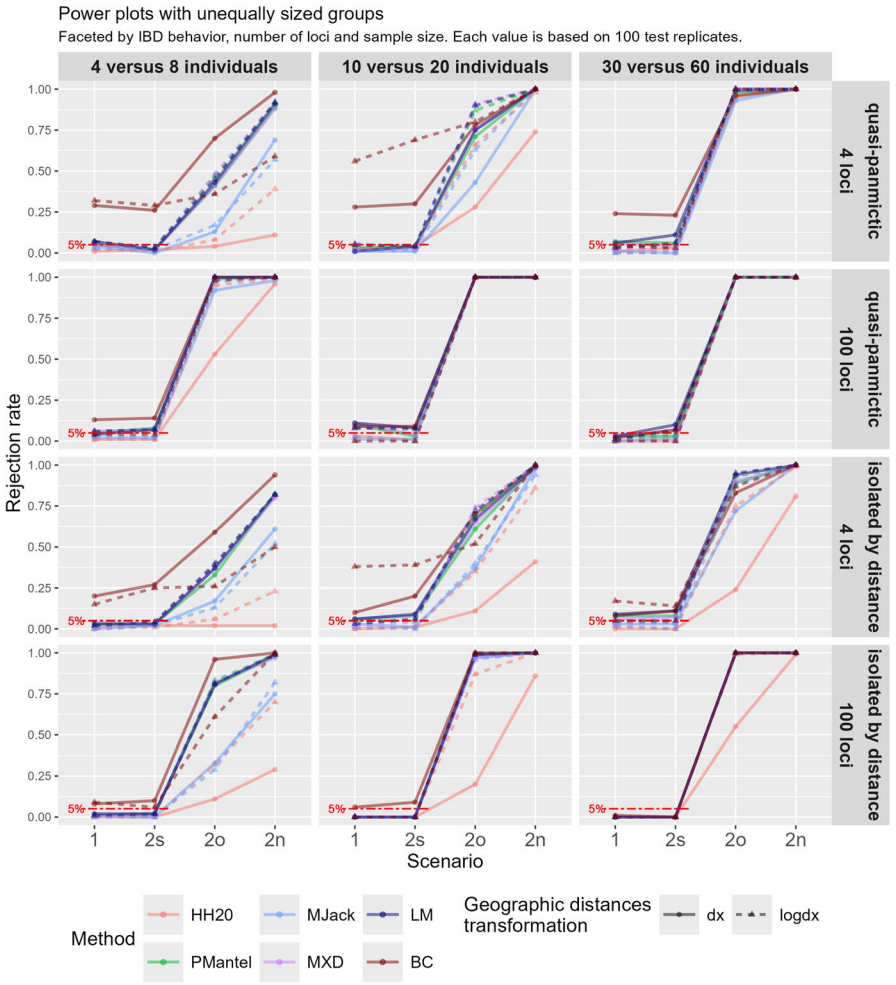

Fig. 6 Power plot based on **GSpace** data simulated **with unequal group sizes**. Panels by combination of IBD behaviour and number of loci (columns) and number of individuals per group (rows). See the caption to Fig. 3 for further description

For this task, GSpace was preferred to the SLiM simulator because the latter does not allow to fix the geographic positions of the individuals in advance and involves many more parameters to be set.

Geographical coordinates of the individuals (available in the dataset) were converted to integers and mapped on the vertices of the GSpace lattice. For each of these two subgroup of brassy ringlets, 32 parameter combinations were compared (narrowed down from earlier experiments with 300 combinations) in order to find the parameter settings that best mimicked the original trend in the dissimilarities. For details on how the best parameters were found, consider the Supplementary Material, which also has the GSpace script. As a result, for both subgroups, a Zeta distribution with $\kappa = 1$, eight unique allelic states, the KAM mutation model, a recombination rate of 5^{-7}, and the mutation rate of 10^{-6} were used.

As shown in Sect. 5, the evidence is very strong that the Cassioides and Tyndarus subgroups constitute two different species. In a first simulation, individuals from two groups were simulated using the parameter settings above and based on the original geographical positions, but during two distinct software executions. 100 such datasets were simulated. We assume that in this situation the subgroups are indeed different species, so that a rejection of the null hypothesis is correct. Estimating the rejection probability therefore amounts to measuring the power of the tests.

In a second simulation, we simulated a single group of 36 individuals (as many as in both subgroups combined) with the above parameters. We then splitted individuals into two subgroups according to their original membership (Cassioides vs Tyndarus) in the real data. This simulates a single species, so that a rejection of the null hypothesis constitutes a type I error.

The results of both simulations are given in Table 1. Largely in line with the earlier simulation results, HH20, MJack, and BC achieved a type I error probability below 0.05, whereas the other three methods rejected clearly too often in the type I error simulation, which makes their good power useless; MXD was worse here than in the other simulations. Once more, MJack achieved a clearly higher power than HH20.

4.4 Discussion

The individual-based spatially-explicit simulations carried out via SLiM and GSpace allowed to study the type I error rate and power of the techniques described in the methods section. A consistent ranking in the overall performance of the methodologies could be observed, with jackknife-based methods being more conservative and less powerful and the other techniques being more powerful, but at the cost of occasional type I error rates above the significance level.

By combining jackknife significance assessment and the usage of partial correlation coefficients, MJack managed to achieve better power than HH20, while show-

Table 1 Empirical rejection rates of the methodologies discussed in Sect. 3 for the first (power) and second (type I error) simulation described in Sect. 4.3 based on 100 datasets each

Method	HH20	PMantel	MJack	MXD	LM	BC
Power	0.84	0.96	0.94	1	0.98	0.94
Type I error	0.03	0.11	0.03	0.18	0.22	0.04

ing type I error rates consistently below the significance level, unlike PMantel. The bias-corrected bootstrap-based PMT showed a too large type I error rate in the most challenging data setups. No method apart from HH20 clearly benefited from the logarithmic transformation of the geographical distances. Despite the trend in the dissimilarities often being convex, the assumption of linearity did not seem to have a severe impact on performance.

Despite wrongly assuming that all dissimilarities in the dataset are independent, OLS estimation in model (3.10) displayed good type I error rate and power in several situations. In particular, it often performed better than the random effects model MXD, avoiding certain increased type I error probabilities of the latter. However, with SLiM data, the type I error rate of LM was consistently above the significance level when unequally sized groups were being compared. This is in line with the expectation that ignoring the dependence between distances should lead to an underestimation of variability.

The performance of HH20 was summarized by a single rejection rate in the power plots, but recall that this method is hierarchical. In our study, a rejection corresponded to cases when either H_{02} or H_{03} was rejected, and thus, non-rejections included also the inconclusive results that can arise when testing H_{03}. Although in all simulations both groups always had the same IBD behaviour, some rejections of H_{01} occurred, so that H_{03} rather than H_{02} was tested. The possibility to take into consideration unequal IBD behaviour is unique to HH20, but here may have led to a degradation of its performance. More precisely, lower rejection rates may have been recorded here in spite of the fact that, in some unclear cases, a practitioner could have concluded that two distinct species were being compared.

5 Analysis of brassy ringlets data

Gratton et al. (2016) discussed the biological delimitation of a taxon of butterflies (brassy ringlets; *Erebia tyndarus* complex, Lepidoptera) endemic to Southern Europe, the Altai Republic and the Rocky Mountains. They studied the morphological, genetic and geographic information of 45 individuals netted during the summer of 2012 across the Italian Appennines, the Alps and the Pyrenees. Four subgroups of this clade were represented in the sample, namely *E. Tyndarus*, *E. Nivalis*, *E. Calcaria* and *E. Cassioides*, with the latter possibly divisible in three populations according to the area of collection. After selecting a subset of 389 diploid loci, they applied k-means clustering on the principal components obtained from the genetic data, Bayesian model-based clustering using the `STRUCTURE` software (Pritchard et al. 2000) and coalescent-based Bayes factor delimitation (Leaché et al. 2014), integrating their results by examining the isolation by distance behaviour and morphological differentiation of the individuals in each putative cluster. The study in Gratton et al. (2016) did not only back up the distinction between the four groups mentioned above from a genetic point of view, but also supported further differentiation within the *Cassioides* group among the Eastern and Orobian Alps population, the Southern and Central Appennines population and the population inhabiting Northern Appennines, Pyrenees and Western Alps.

Hausdorf and Hennig (2020) applied their testing protocol to these data in order to replicate and deepen the IBD investigation carried out by Gratton et al. (2016). With the exclusion of the *Calcaria* group, which could not be examined due to the small sample size (only 3 specimens), the distinction among the groups was confirmed. The classification within the *Cassioides* group, instead, was slightly amended: HH20 suggested that there was no evidence of distinctness between the Southern and Central Appennines population and the population inhabiting Northern Appennines, Pyrenees and Western Alps, whereas the genetic dissimilarity between these populations taken together and the Eastern and Orobian Alps population was too large to be explained by isolation by distance.

In Table 2, the results from all methods involved in this study are reported. See Fig. 1 for the dissimilarity data on which three of these comparisons are based. The only non-rejections of the null hypothesis of conspecificity occurred for *E. Cassioides*: W_Alps + Pyrenees + N_Apennines **vs** Central + S_Apennines (second panel in Fig. 1) by the bias-corrected bootstrap-based Partial Mantel test (the CI contains 0) and the larger one of the two group-wise p-values for H_{03} by HH20.

In all other cases, all methods agreed upon the distinctness results, confirming the conclusions shared by Gratton et al. (2016) and Hausdorf and Hennig (2020) - including also the distinction between the Western and Appennines populations versus the Eastern and Orobian populations (last row in Table 2). According to the evidence collected in this study, not only the *E. Tyndarus*, *E. Nivalis* and *E. Cassioides* groups should be considered distinct: also the three subgroups identified within the *E. Cassioides* group, namely the Eastern and Orobian Alps population, the Southern and Central Appennines population and the population inhabiting Northern Appennines, Pyrenees and Western Alps, display a genetic structure that cannot be explained by their geographic separation.

6 Conclusions

We investigated methods that model the relationship between genetic and geographical dissimilarities in biological species in order to perform species delimitation. These techniques check whether the genetic structure existing between two putative species is compatible with the way genetic dissimilarities within each group increase with the geographical separation of the individuals. The type I error rate and power of these methods were compared by means of individual-based simulations carried out with the simulators GSpace and SLiM. Results showed that the method of Hausdorf and Hennig (2020) (HH20) has a very conservative type I error rate and lower power than Partial Mantel tests (PMTs) as applied by Medrano et al. (2014), which in turn had type I error rates slightly above the significance level in some setups. Testing PMTs with jackknife instead of permutations fixed this behaviour, ensuring more power than HH20 while keeping the type I error rate still close to zero. This method can therefore be seen as the best in the simulated setups. Testing PMTs with bias-corrected bootstrap confidence intervals, instead, often led to inflated type I error rates. The linear mixed effects model displayed a performance similar to PMT. A linear regression without random effects (LM), i.e., wrongly assuming independence

Table 2 Results from all methods compared here on the brassy ringlets data

Groups compared	H_{01}	H_{02}	H_{03}	PMantel	MJack	MXD	LM	BC
E. Tyndarus vs *E. Nivalis*	0.074	$<10^{-5}$	n.a	0.001	$<10^{-29}$	(0.296, 0.306)	$<10^{-113}$	(0.983, 0.995)
E. Nivalis vs *E. Cassioides*	0.094	$<10^{-4}$	n.a	0.001	$<10^{-55}$	(0.378, 0.389)	~ 0	(0.972, 0.988)
E. Tyndarus vs *E. Cassioides*	$<10^{-9}$	n.a	both $<10^{-24}$	0.001	$<10^{-67}$	(0.345, 0.352)	~ 0	(0.977, 0.986)
E. Cassioides: W_Alps + Pyrenees + N_Apennines vs Orobian + E_Alps	0.487	0.004	n.a	0.001	$<10^{-5}$	(0.025, 0.036)	$<10^{-17}$	(0.493, 0.745)
E. Cassioides: W_Alps + Pyrenees + N_Apennines vs Central + S_Apennines	$<10^{-4}$	n.a	0.098; 0.004	0.002	0.015	(0.030, 0.045)	$<10^{-5}$	(− 0.148, 0.552)
E. Cassioides: Central + S_Apennines vs Orobian + E_Alps	0.144	0.009	n.a	0.001	$<10^{-4}$	(0.041, 0.066)	$<10^{-15}$	(0.477, 0.484)
E. Cassioides: W_Alps + Pyrenees + ALL Apennines vs Orobian + E_Alps	1	$<10^{-4}$	n.a	0.001	$<10^{-8}$	(0.020, 0.029)	$<10^{-25}$	(0.329, 0.618)

For the three tests in HH20, *p*-values are reported (two of them are given for H_{03} when discordant); *p*-values are reported for PMantel, MJack and LM, too; for MXD and BC, confidence intervals are reported for the b_2^* regression coefficient and the partial correlation coefficient, respectively: if their lower boundary is larger than zero, the null hypothesis is rejected

among the dissimilarities, showed inflated type I error rates only with unequally sized groups simulated with SLiM, and performed surprisingly well otherwise.

The ranking in the overall performance of the methods was consistent over both simulators, and the log-transformation of the geographical dissimilarities did not seem to have a considerable impact on methods other than HH20, where it improved matters. The impact of transformations will in general depend on the used dissimilarities.

Due to the extremely large amount of possibilities defined by parameter choices of the simulators, many more potentially interesting situations could be simulated, and, as is always the case with such simulations, generalizability of results cannot be guaranteed beyond the simulated scenarios. Some scenarios worth exploring could involve comparisons between groups with different IBD behaviour, size and separation of the inhabited areas, but also simulations with independent loci, different settings for the time scale and migration rates, etc. The number of generations can be an important parameter in the distinctness scenarios simulated with SLiM, but it is hard to investigate systematically. A too large number of generations can mean that shared allele distances between co-specific individuals might become so large that they could be indistinguishable from distances between species, and the effect of changing this may also interact with many other parameter choices. Moreover, the investigation could be extended to population-based simulations, using genetic dissimilarity measures like F_{st} (Weir and Cockerham 1984) or the chord distance (Cavalli-Sforza and Edwards 1967). The individual-based methods compared in this study often assume independence between observations in the same group, something that is not biologically grounded and is indeed avoided in population-based studies. Independence between individuals is actually questionable even for the data simulated from GSpace and SLiM, where models account for contact between individuals within species. We do however think that it is realistic that statistical methods assuming independence are applied to real data that are not in fact truly independent, and such a situation is actually simulated.

In addition, another simulation-based investigation could be set up in order to compare the performance of these methods in scenarios where there is no putative grouping known in advance. To this end, an automated unsupervised routine for species delimitation may be conceived that uses the tests investigated here to decide whether groups that come out of a cluster analysis produced by methods such as `STRUCTURE` (Pritchard et al. 2000) should be merged. The `conStruct` approach (Bradburd et al. 2018) does something similar, but is only applicable when the number of loci is much larger than n, whereas, as shown in this study, most methods investigated here worked fairly good even with very small n and P.

Applying the methods treated here to other problems of regression between dissimilarities such as relating similarity between languages or dialects to geographical distance (Bella et al. 2021) may also be of interest.

Regarding the simulation in Sect. 4.3, despite using the best parameters we could find, there were still systematic differences between simulated and real data, so there may be further scope to improve the realism of the simulations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-026-00669-6.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

Data Availability The code for generating the data used in the simulations is provided in the Supplementary Material. The brassy ringlets data were published by the authors of Gratton et al. (2016) under https://doi:10.5061/dryad.3n5c9.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Informed consent For this kind of study informed consent is not required.

References

Adriaensen F, Chardon J, De Blust G, Swinnen E, Villalba S, Gulinck H, Matthysen E (2003) The application of 'least-cost' modelling as a functional landscape model. Landsc Urban Plan 64(4):233–247. https://doi.org/10.1016/S0169-2046(02)00242-6

Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, Gower G, Kyriazis CC, Ragsdale AP, Tsambos G, Baumdicker F, Carlson J, Cartwright RA, Durvasula A, Gronau I, Kim BY, McKenzie P, Messer PW, Noskova E, Ortega-Del Vecchyo D, Racimo F, Struck TJ, Gravel S, Gutenkunst RN, Lohmueller KE, Ralph PL, Schrider DR, Siepel A, Kelleher J, Kern, AD (2020). A community-maintained standard library of population genetic models. eLife 9:e54967. https://doi.org/10.7554/eLife.54967

Balkenhol N, Cushman SA, Storfer A, Waits LP (2015) Introduction to landscape genetics - concepts, methods, applications. John Wiley & Sons, USA

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

Bella G, Batsuren K, Giunchiglia F (2021). A database and visualization of the similarity of contemporary lexicons. In: Ekstein K, Partl F, Konopik M (eds), In: Proceedings of the 24th International Conference on Text, Speech, and Dialogue, Olomouc, Czech Republik, pp. 95–104. Springer Nature, Switzerland

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368(6470):455–457

Bradburd GS, Coop GM, Ralph PL (2018) Inferring continuous and discrete population genetic structure across space. Genetics 210(1):33–52. https://doi.org/10.1534/genetics.118.301333 (https://academic.oup.com/genetics/article-pdf/210/1/33/49481005/genetics0033.pdf)

Burbrink FT, Ruane S (2021) Contemporary philosophy and methods for studying speciation and delimiting species. Ichthyol Herpetol 109(3):874–894. https://doi.org/10.1643/h2020073 (https://meridian.allenpress.com/copeia/article-pdf/109/3/874/2926176/i2766-1520-109-3-874.pdf)

Carstens BC, Pelletier TA, Reid NM, Satler JD (2013) How to fail at species delimitation. Mol Ecol 22(17):4369–4383
Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis. models and estimation procedures. Am J Hum Genet 19:233–257
Cayuela H, Rougemont Q, Prunier JG, Moore J, Clobert J, Besnard A, Bernatchez L (2018) Demographic and genetic approaches to study dispersal in wild animal populations: a methodological review. Mol Ecol 27(20):3976–4010
Clarke RT, Rothery P, Raybould AF (2002) Confidence limits for regression relationships between distance matrices: estimating gene flow with distance. J Agric Biol Environ Stat 7(3):361–372
De Queiroz K (2007) Species concepts and species delimitation. Syst Biol 56(6):879–886. https://doi.org/10.1080/10635150701701083 (https://academic.oup.com/sysbio/article-pdf/56/6/879/24203468/56-6-879.pdf)
Dietz EJ (1983) Permutation tests for association between two distance matrices. Syst Zool 32(1):21–26
Edwards DL, Knowles LL (2014) Species detection and individual assignment in species delimitation: can integrative data increase efficacy? Proc Royal Soc B Biol Sci 281(1777):20132765
Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall, New York London
Frantz AC, Cellina S, Krier A, Schley L, Burke T (2009) Using spatial bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? J Appl Ecol 46(2):493–505. https://doi.org/10.1111/j.1365-2664.2008.01606.x
Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. J Stat Softw 22:1–19
Gratton P, Trucchi E, Trasatti A, Riccarducci G, Marta S, Allegrucci G, Cesaroni D, Sbordoni V (2016) Testing classical species properties with contemporary data: how "bad species" in the brassy ringlets (erebia tyndarus complex, lepidoptera) turned good. Syst Biol 65(2):292–303
Guillot G, Rousset F (2013) Dismantling the mantel tests. Methods Ecol Evol 4(4):336–344. https://doi.org/10.1111/2041-210x.12018
Haller, B.C., and Messer, P.W. (2016). SLiM: An Evolutionary Simulation Framework. http://benhaller.com/slim/SLiM_Manual.pdf
Haller BC, Messer PW (2023) Slim 4: multispecies eco-evolutionary modeling. Am Nat 201(5):E127–E139
Hausdorf B (2011), 04. Progress toward a general species concept. Evolution 65(4): 923–931. https://doi.org/10.1111/j.1558-5646.2011.01231.x. https://academic.oup.com/evolut/article-pdf/65/4/923/47936342/evolut0923.pdf
Hausdorf B, Hennig, C (2019). R-Package 'prabclus', version 2.3-2. https://cran.r-project.org/web/packages/prabclus/
Hausdorf B, Hennig C (2020) Species delimitation and geography. Mol Ecol Resour 20(4):950–960. https://doi.org/10.1111/1755-0998.13184
Hutchison DW, Templeton AR (1999) Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of gene flow and drift on the distribution of genetic variability. Evolution 53(6):1898–1914
Ishida Y (2009) Sewall wright and gustave malécot on isolation by distance. Philos Sci 76(5):784–796
Jukes TH, Cantor, CR (1969). Evolution of protein molecules, In Mammalian Protein Metabolism, ed. Munro, H.N., 21–132. Academic Press. https://doi.org/10.1016/B978-1-4832-3211-9.50009-7
Kimura M, Weiss GH (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49(4):561–576
Leaché AD, Fujita MK, Minin VN, Bouckaert RR (2014) Species delimitation using genome-wide snp data. Syst Biol 63(4):534–542
Leblois R, Estoup A, Rousset F (2009) Ibdsim: a computer program to simulate genotypic data under isolation by distance. Mol Ecol Resour 9(1):107–109
Legendre P (2000) Comparison of permutation methods for the partial correlation and partial mantel tests. J Stat Comput Simul 67(1):37–73
Legendre P, Fortin MJ, Borcard D, Peres-Neto P (2015) Should the mantel test be used in spatial analysis? Methods Ecol Evol 6(11):1239–1247
Legendre P, Legendre L (2012). Numerical ecology. Developments in environmental modelling (3rd English ed.) pp 24. Elsevier, Amsterdam
Mantel N (1967) The detection of disease clustering and a generalized regression approach. Can Res 27:209–220
McRae BH (2006) Isolation by resistance. Evolution 60(8):1551–1561

McRae BH, Dickson BG, Keitt TH, Shah VB (2008) Using circuit theory to model connectivity in ecology, evolution and conservation. Ecology (Durham) 89(10):2712–2724
Medrano M, López-Perea E, Herrera CM (2014) Population genetics methods applied to a species delimitation problem: endemic trumpet daffodils (narcissus section pseudonarcissi) from the southern iberian peninsula. Int J Plant Sci 175(5):501–517
Meirmans PG (2012) The trouble with isolation by distance. Mol Ecol 21(12):2839–2846. https://doi.org/10.1111/j.1365-294X.2012.05578.x
Miller RG (1974) The jackknife-a review. Biometrika 61(1):1–15. https://doi.org/10.1093/biomet/61.1.1
Peterman WE (2018) Resistancega: an r package for the optimization of resistance surfaces using genetic algorithms. Methods Ecol Evol 9(6):1638–1647. https://doi.org/10.1111/2041-210X.12984
Peterman WE, Pope NS (2021) The use and misuse of regression models in landscape genetic analyses. Mol Ecol 30(1):37–47. https://doi.org/10.1111/mec.15716
Pope LC, Liggins L, Keyse J, Carvalho SB, Riginos C (2015) Not the time or the place: the missing spatio-temporal link in publicly available genetic data. Mol Ecol 24(15):3802–3809. https://doi.org/10.1111/mec.13254
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics (Austin) 155(2):945–959
RannalaB, Yang Z (2020) Species delimitation. In: Phylogenetics in the genomic era (eds) Scornavacca C, Delsuc F, Galtier N 5.5:1–5.5:18. (self published). https://inria.hal.science/PGE/
Raxworthy CJ, Ingram CM, Rabibisoa N, Pearson RG (2007) Applications of ecological niche modeling for species delimitation: a review and empirical evaluation using day geckos (Phelsuma) from Madagascar. Syst Biol 56(6):907–923. https://doi.org/10.1080/10635150701775111
Rissler LJ, Apodaca JJ (2007) Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (Aneides flavipunctatus). Syst Biol 56(6):924–942. https://doi.org/10.1080/10635150701703063
Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145(4):1219–1228. https://doi.org/10.1093/genetics/145.4.1219
Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for windows and linux. Mol Ecol Resour 8(1):103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x
Royston P (2007) Profile likelihood for estimation and confidence intervals. Stand Genomic Sci 7(3):376–387
SAS (2000). Statistical Analysis System User's Guide. Version 8. Cary, NC. SAS Institute Inc.
Scapini F, Aloia A, Bouslama MF, Chelazzi L, Colombini I, ElGtari M, Fallaci M, Marchetti GM (2002) Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, talitrus saltator and talorchestia brito, from an exposed mediterranean beach. Behav Ecol Sociobiol 51:403–414
Shao J, Wu CFJ (1989) A general theory for jackknife variance estimation. Ann Stat 17(3):1176–1197. https://doi.org/10.1214/aos/1176347263
Shirk AJ, Landguth EL, Cushman SA (2018) A comparison of regression methods for model selection in individual-based landscape genetic analysis. Mol Ecol Resour 18(1):55–67. https://doi.org/10.1111/1755-0998.12709
Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47(1):264–279
Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the mantel test of matrix correspondence. Syst Zool 35(4):627–632
Spriggs EL, Eaton DAR, Sweeney PW, Schlutius C, Edwards EJ, Donoghue MJ (2018) Restriction-site-associated DNA sequencing reveals a cryptic viburnum species on the north American coastal plain. Syst Biol 68(2):187–203. https://doi.org/10.1093/sysbio/syy084
Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP (2010) Landscape genetics: where are we now? Mol Ecol 19(17):3496–3514
Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. Ann Stat 35(6):2769–2794. https://doi.org/10.1214/009053607000000505
Van Strien MJ, Keller D, Holderegger R (2012) A new analytical approach to landscape genetic modelling: least-cost transect analysis and linear mixed models. Mol Ecol 21(16):4010–4023
Vekemans X, Hardy OJ (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. Mol Ecol 13(4):921–935. https://doi.org/10.1046/j.1365-294X.2004.02076.x
Venzon DJ, Moolgavkar SH (1988) A method for computing profile-likelihood-based confidence intervals. J Roy Stat Soc: Ser C (Appl Stat) 37(1):87–94

Virgoulay T, Rousset F, Leblois R (2021) Gspace: an exact coalescence simulator of recombining genomes under isolation by distance. Bioinformatics 37(20):3673–3675
Waits LP, Storfer A (2015) Basics of population genetics: quantifying neutral and adaptive genetic variation for landscape genetic studies. John Wiley & Sons, USA
Weir BS, Cockerham CC (1984) Estimating f-statistics for the analysis of population structure. Evolution 38(6):1358–1370
Welch BL (1947) The generalization of "Student's" problem when several different population variances are involved. Biometrika 34(1–2):28–35. https://doi.org/10.1093/biomet/34.1-2.28
West BT, Welch KB, Galecki AT (2022) Linear mixed models: a practical guide using statistical software. CRC Press, USA
Yang R (2004) A likelihood-based approach to estimating and testing for isolation by distance. Evolution 58(8):1839–1845. https://doi.org/10.1111/j.0014-3820.2004.tb00466.x

Authors and Affiliations

Gabriele d'Angella[1] · Christian Hennig[1]

✉ Christian Hennig
christian.hennig@unibo.it

Gabriele d'Angella
gabriele.dangella2@unibo.it

[1] Alma Mater Studiorum - University of Bologna, Via Belle Arti, 41, 40126 Bologna, BO, Italy