ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
SCIENZE STATISTICHE

Ciclo 37

Settore Concorsuale: 13/D1 - STATISTICA

Settore Scientifico Disciplinare: SECS-S/01 - STATISTICA

STATISTICAL DELIMITATION OF BIOLOGICAL SPECIES BASED ON GENETIC AND SPATIAL DATA

Presentata da: Gabriele d'Angella

Coordinatore Dottorato
Prof. Angela Montanari

Supervisore
Prof. Christian Hennig

Esame finale anno 2025

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 data set, is key to the conservation of biodiversity. In the presence of spatial patterns of genetic differentiation, delimitation methods based on genetic data might overestimate the number of species in a dataset.

This work tackles this problem in two settings. When individuals are divided into two putative groups, methods that model the relationship between genetic and geographic dissimilarity are used to test whether the two groups belong to the same species. Existing approaches based on partial Mantel testing and regression on distances are explored and new ones are proposed. A modelling challenge is connected to the fact that dissimilarities are not independent. All methodologies are compared through an extensive simulation study involving SLiM and GSpace, two different software packages that can simulate spatially-explicit genetic data at an individual level. A proposed version of the partial Mantel test that uses jackknife instead of permutations is found to provide fairly good power while controlling for the type I error rate in all simulated scenarios.

In a setting where no putative grouping is available, existing model-based clustering algorithms (sNMF and TESS3) are integrated with distance-based approaches for the estimation of the number of species in the dataset. Further considered approaches use null models to calibrate tests for the presence of more than one species in the dataset. In particular, a weighted null model is developed that can capture spatial patterns of genetic differentiation. When calibrated with this null model, a test statistic proposed to adapt the ΔK method to TESS3 is found to display promising type I error and power properties on SLiM data.

Table of contents

List of figures

List of tables

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 Z. Yang 2020). Species delimitation is crucial for the preservation of biodiversity (D. Liu 2024) and has applications in several areas, such as ecology and medicine (Burbrink and Ruane 2021). It relies on species conceptualization, which determines what key traits are relevant for the identification of a species (De Queiroz 1998, 2007; Hausdorf 2011). The empirical data employed to delimit species can be molecular (see, e.g., Rannala and Z. 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).

The use of 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 geographic information when delimiting species can lead to the overestimation of the genetic structure in the data (Frantz et al. 2009). This is more likely to happen in the presence of spatial patterns of genetic differentiation, such as isolation by distance (IBD; Wright 1943): clustering methods might wrongly assign individuals to different species given that their genetic dissimilarity increases with geographic separation (Bradburd, Coop, and Ralph 2018), violating the assumption of random mating within the population.

This project deals with the delimitation of species using molecular and geographic data in the presence of spatial patterns of genetic differentiation. Chapter 1 considers scenarios in which individuals are divided into two (putative) groups and the goal is to test whether they belong to the same species. Chapter 2, instead, concerns settings where no grouping information on the individuals is available and the objective is to clarify how many species are represented in the dataset.

A way to include spatial information in molecular species delimitation routines is to study the relationship between genetic and geographic dissimilarity. In Chapter 1, it is checked whether genetic discrepancies between individuals from different groups

are compatible with the way genetic dissimilarities evolve with geographic separation within the groups. More precisely, a grouping distance is computed that takes value 1 when two individuals belong to different putative groups and 0 otherwise. If genetic dissimilaries are associated with grouping distances even after accounting for geographic distances, this constitutes evidence that the putative groups might belong to distinct species. Existing approaches to test this include permutation-based partial Mantel testing (Medrano et al. 2014; Smouse et al. 1986) and regression (Hausdorf and Hennig 2020), and inference strategies are pursued that take into account the dependence between dissimilarities. In Chapter 1, jackknife and bootstrap-based partial Mantel tests are introduced, together with an extension to the linear mixed model by Clarke et al. (2002). A multiple linear regression that assumes dissimilarities to be independent is also examined. All these methodologies are systematically compared in terms of type I error and power through a broad simulation study involving two recent spatially explicit genetic simulators, SLiM (Haller and Messer 2023) and GSpace (Virgoulay et al. 2021b). These are used to generate datasets with one or more species displaying various patterns of isolation by distance. The performance of the methods is assessed in relation with several factors, such as the assumption of linearity, the way dependence is accounted for and the relative sizes of the putative groups.

In Chapter 2, approaches to estimate the number of species in the dataset are considered that take into account the presence of spatial patterns of genetic differentiation. First, two strategies are proposed in which existing model-based clustering algorithms are integrated with methods based on the relationship between genetic and geographic distance. In one of them, groups identified by sNMF (Frichot, Mathieu, et al. 2014) are iteratively tested for merging using the techniques from Chapter 1. The other strategy revolves around TESS3 (Caye et al. 2018), which instead takes both geographic and genetic information into account when clustering individuals. Its output is used to fit a regression on dissimilarities developed to help estimate the number of species in the dataset. In the second part of Chapter 2, null models are conceived that attempt to reproduce all features of the observed data that cannot be interpreted as clustering information. This is done in order to construct tests that detect the presence of more than one species in the dataset, but also to calibrate an adaptation of the ΔK statistic by Evanno et al. (2005), which allows to estimate the number of species. All these approaches are evaluated through extensive SLiM simulations containing from 1 to 6 species.

The methodologies presented here can thus integrate delimitation studies based on other data sources, such as morphological or behavioural information. Moreover, the investigation of the relationship between dissimilarities may be relevant also for other applications.

Chapter 1

Approaches with known groups

Species delimitation is of paramount importance in systematics and has practical implications for conservation and management (e.g., Pedraza-Marrón et al. 2019). In the attempt to delimit species using genetic and geographic data, the presence of spatial patterns of genetic differentiation can be confounding (Bradburd, Coop, and Ralph 2018; Frantz et al. 2009). That is, it can be hard to assess whether the genetic differences displayed by two populations inhabiting separate areas are consistent with species distinctness or can be explained by geographic separation (Hausdorf and Hennig 2020). To tackle this issue, species delimitation routines have to control for the presence of spatial patterns of genetic differentiation.

A way to do this is to study the relationship between genetic and geographic dissimilarities. Consider a setup with two putative groups of individuals to be tested for cospecificity. Inference is based on checking whether the relationship between genetic and geographic dissimilarity differs between pairs of individuals in the same group and pairs in different groups. Each of the three panels in Figure 1.1 shows genetic and geographic dissimilarities of two groups for which a test for conspecificity is of interest (see section 1.6.1 for details). 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 geographic distance, also within groups. In the first plot, genetic ("shared allele") distances between groups seem slightly higher on average than genetic distances within groups, but also the geographic 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 geographic distances only (in which case there is no reason to consider the two groups as different species), even less so in the second plot. In the third 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.

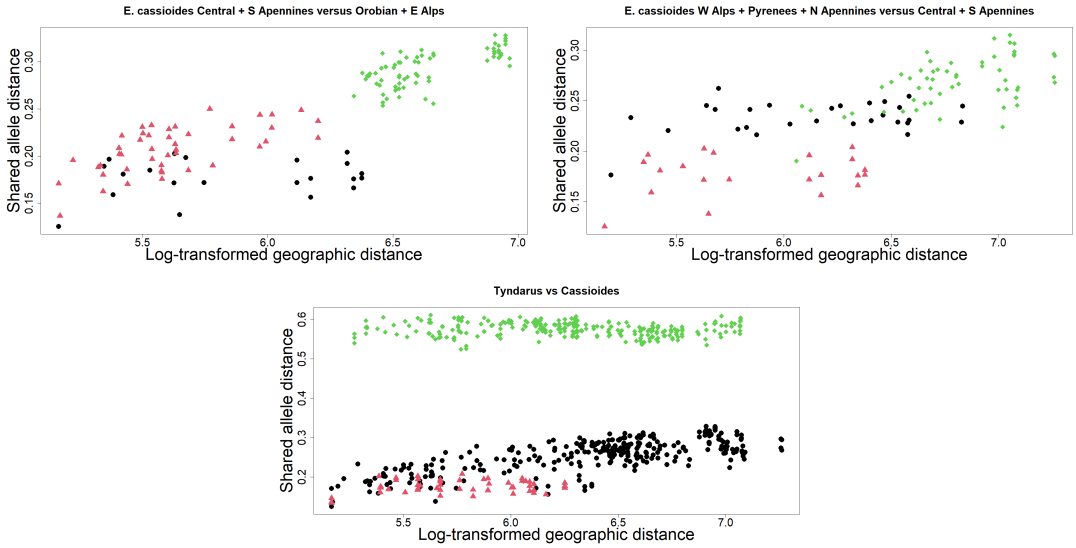

Fig. 1.1 Log-transformed geodesic distances vs. shared allele distances for three pairs of groups from the brassy ringlets data (left side *E. Cassioides central + S. Apennines* vs. *Orobian + E. Alps*; middle *E. Cassioides W. Alps, Pyrenees + N. Apennines* vs. *central + S. Apennines*; right side *E. Tyndarus* vs. *E. Cassioides*), see Section 1.6.1, 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.

The impact of grouping on the genetic dissimilarity can be quantified controlling for the effect of geographic 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 geographic 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 to 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 belonging to the same individual. This approach is here adapted to the species delimitation problem by testing for an effect of the grouping dissimilarity. As further new approaches, alternative versions of the partial Mantel test that use jackknife or bootstrap instead of permutations are considered. All these techniques take into account the dependence between dissimilarities involving the same individual. 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 geographic and grouping dissimilarities as explanatory variables. A similar model was used by Spriggs et al. (2018) to integrate a rich molecular species delimitation analysis.

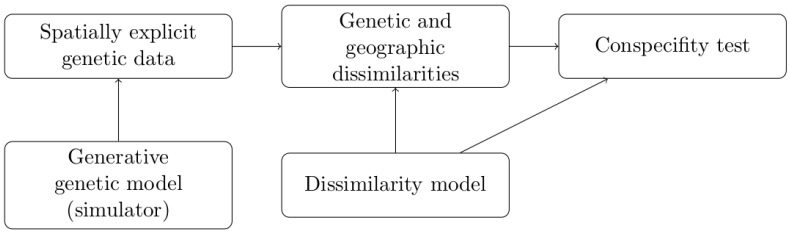

Fig. 1.2 The conspecificity tests treated here are computed on genetic and geographic dissimilarities assuming models for general dissimilarities (i.e., not taking into account how exactly the dissimilarities came about). The dissimilarities are computed from the originally observed spatially explicit genetic data, which in this study are simulated from generative genetic models.

In this chapter, type I error rate and power of the aforementioned methods are systematically compared based on data from two distinct spatially-explicit genetic simulators, GSpace (Virgoulay et al. 2021b) and SLiM (Haller and Messer 2023).

A distinctive feature of this analysis is that the genetic dissimilarities on which inference is based are much simpler than the data from which they are computed, see Section 1.2. 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 Figure 1.2. 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. In this regard, note that also in other fields dissimilarities are used in a regression framework, but usually they are employed as covariates and not as response variables, or viceversa (e.g., Vera 2022).

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 formalize a key aspect of the information in the data.

This chapter is organized as follows. Section 1.1 provides and overview on research problems concerning species delimitation and the study of the relationship between measures of genetic relatedness and geographic separation. Some key definitions are provided, and the contribution of this work is positioned in the literature. The data and the distances employed in this project are introduced in section 1.2, while dissimilarity-based methods are discussed in section 1.3. In section 1.4, type I error and power properties of these methods are compared through a broad simulation study based on

data from SLiM and GSpace. The analysis of the relationship between genetic and geographic distances poses modelling issues related to the assumption of linearity and to dependence. These are discussed in section 1.5, where the three strategies to assess significance in partial Mantel tests, permutations, jackknife and bootstrap, are further compared. Section 1.6 contains an application on real data, while section 1.7 digresses on some probabilistic properties of the genetic distance employed in this work, the shared allele distance.

1.1 Literature review

Species are one of the fundamental units of biology (Hausdorf 2011). However, they exist at a higher level of organization than the humans observing them. This makes it hard for humans to perceive entire species by simply looking at them, as they do with genes and cells, "which is why biologists have symposia devoted to the topic of species delimitation" (De Queiroz 2007). Species delimitation consists in determining from empirical data which groups of individual organisms constitute different populations of a single species and which constitute different species (Rannala and Z. Yang 2020). This differs from species conceptualization, which concerns species definitions and clarifies the properties on which these definitions are based. There exist various species concepts (see, e.g., De Queiroz 1998): for instance, "the biological species concept defines a species taxon as a group of organisms that can successfully interbreed and produce fertile offspring" (Ereshefsky 2007). Some of these concepts can be partially incompatible: De Queiroz (2007) proposed a unified species concept, according to which a species is defined as a separately evolving metapopulation lineage. In his view, all other species concepts become secondary properties, operational criteria that serve the species delimitation task. The more criteria suggest that two candidate species should be considered distinct, the stronger the indication of lineage separation. When no criterion supports distinctness, there should be agreement about conspecificity. Therefore, once this reconciliation is adopted, disagreement about species delimitation results will not be based on the conceptualization of the species, but rather on incomplete lineage separation (some criteria support conspecificity and some do not), on the relevance of the available data and on the reliability of the methodologies used to infer species membership. Consistently with this framework, the techniques treated in this work only attempt to formalize one aspect of species distinctness, and are not deemed to be sufficient to delimit species in themselves.

Burbrink and Ruane (2021) maintain that "researchers are far from agreeing upon a set of criteria to delimit species". The task gets thornier when lineage separation is

incomplete, as no single framework has been proposed to model speciation using the six major parameters in evolutionary biology, namely mutation, genetic drift, selection, migration, recombination and population demography. When two candidate species lie in the gray zone of speciation, gene flow, isolation by distance, introgression and other factors can make the species delimitation task complicated. Integration of genetic data with all available data types, especially geographical information, is advised to tackle these difficulties.

Waits and Storfer (2015) review several definitions that are central to this work. Molecular species delimitation is based on the information extracted from the DNA: by comparing the DNA sequences at the same location (called locus) in the genome of two individuals, the genetic variation between them is quantified. The variants of the genes residing at a particular locus are called alleles, and loci containing two alleles are called diploid. Loci are said to be adaptive when they are affected by selection, which modifies allele frequencies according to what is genetically favorable in a specific environmental condition. Neutral loci, instead, are influenced mainly by gene flow and drift. Gene flow is the result of individuals dispersing and reproducing, whereas "genetic drift is the change in allele frequencies due to random sampling effects as alleles are passed on from one generation to the next" (Waits and Storfer 2015). In order to obtain genetic data from nuclear DNA, there exist a co-dominant and a dominant approach. Dominant loci methods, such as amplified fragment length polymorphism (AFLP), do not allow to identify both alleles in a diploid locus, so they yield presence-absence data. Co-dominant ones, instead, return genotypes where both alleles are visible: examples are microsatellites and single-nucleotide polymorphisms (SNPs). All these methods translate in different costs of acquisition, proportions of surveyed genome, genetic variability and thus suitability to answer specific biological questions. Lastly, a critical distinction is that between genetic data analyses based on individuals, e.g. single animals or plants, and those based on populations, i.e., communities of individuals found at the same geographic location.

This work focuses on individual-level analyses of neutral co-dominant loci. The interest is in genetic structure, namely in the distribution of the genetic variation among individuals. Populations from a species are said to be panmictic when mating is completely random and all genetic combinations are equally likely. In these populations there is virtually infinite gene flow and thus they display no genetic structure. Panmictic populations represent an ideal benchmark, as several factors - such as geographic isolation - can be expected to impede random mating. Genetic structure is quantified by means of genetic dissimilarities. Reviews and comparative studies on these measures - some of which are mentioned below - can be found, e.g., in Meirmans and Hedrick

(2011), Whitlock (2011), P. Legendre and L. Legendre (2012, ch. 7), Shirk et al. (2018) and Sentinella et al. (2022).

Among the various sources of information that can help in the species delimitation task, geography is one of the most studied. Indeed, in the analysis of genetic data, it is key to account for its spatial structure (e.g., Battey et al. 2020): an example is positive spatial autocorrelation, i.e., when nearby individuals are more genetically similar than distant ones (P. Legendre and L. Legendre 2012, ch. 1). Here lies the focus of landscape genetics, a discipline that combines population genetics and landscape ecology (Balkenhol et al. 2015). Various approaches to account for spatial structure in landscape genetics are reviewed in Wagner and Fortin (2015). One of these is link-level analysis, in which the genetic structure is quantified by means of dissimilarity measures and their association with geographic dissimilarities is investigated. Techniques explored in this project work with genetic and geographic dissimilarities.

The umbrella term "landscape distance" (Shirk et al. 2018) refers to the various ways to quantify the geographic dissimilarity between two locations. Several options exist, each associated with a specific construct on how genetic structure varies geographically. The simplest theory on this is isolation by distance (IBD; Ishida 2009; Wright 1943), according to which genetic dissimilarity is positively related to Euclidean (or geodesic) distance. IBD assumes movement to happen in straight lines, which of course can be limiting in some setups. This is why some authors maintain that IBD should be considered a null model against which to test for more realistic hypotheses (MacDonald et al. 2020), such as isolation by resistance (IBR; McRae 2006). IBR attempts to more explicitly take into account the landscape features when evaluating the resistance to gene flow between two locations on the map. Maps become resistance surfaces and there exist two approaches to translate these into landscape distances: the least-cost path approach (Adriaensen et al. 2003) assumes that individuals are able to optimize and tend to move along the path that minimizes the resistance, whereas the circuit-based framework (McRae et al. 2008) tries to aggregate all possible pathways connecting two locations. More recently, least-cost transect analysis was proposed by Van Strien, Keller, and Holderegger (2012). See also Cayuela et al. (2018) on this. Isolation by environment (IBE; Wang and Bradburd 2014) has more to do with selection and adaptive loci than with gene flow and neutral ones: it prescribes that species will display larger genetic divergence the more their habitats differ, independently of geographic separation. This work will focus on isolation by distance: Euclidean distance will be used in simulations, while geodesic distance will be used in real data analysis. However, in principle the techniques discussed here are compatible with any landscape distance.

The relationship between genetic and geographic dissimilarities has been studied for decades. In his ecological conception of IBD, Wright (1943) investigated how local

genetic differentiation decreases as the number of individuals whose gametes may come together (the neighborhood size) grows. Based on this, Malécot's genetic version of IBD studied the connection between migration and geographic distance, maintaining that "genetic relatedness of individuals decreases as the distance between them increases" (Ishida 2009). In these studies, typically a specific migration model was assumed, such as the island model or the stepping stone model (Kimura and Weiss 1964). The former assumes that "there is a mainland of infinite size exchanging migrants with an island of finite size". In the latter, migration only occurs among nearest neighbor populations (Waits and Storfer 2015). The goal of these models was to go beyond the starting simplistic Hardy-Weinberg assumption (see, e.g., Hedrick 2009, ch. 2) and to quantify evolutionary processes, such as dispersal, gene flow and random drift. Building on Malécot's remarks (Kimura 1955), Kimura and Weiss (1964) showed that the correlation of gene frequencies decreases exponentially with geographic distance when a one-dimensional setup (e.g. a river, coastal line or mountain ridge) is considered and that the decrease becomes faster as the dimensionality grows. Nei (1972) introduced a measure of genetic distance, D, based on the identity of genes between populations and found it to be linearly related to distance in a one-dimensional stepping-stone model, but non-linearly related with it in a two-dimensional stepping-stone model. Under the same models, Slatkin (1993) found empirical evidence that the logarithm of F_{st} (Weir and Cockerham 1984; Wright 1949) and the logarithm of the geographic distance display an approximately linear relationship. Nei (1973) also proposed G_{st}, an extension of F_{st} that accounts for the presence of more than two alleles at a given locus. Rousset (1997) showed that the ratio $\frac{F_{st}}{1-F_{st}}$ displays a linear relationship with geographic distance in one-dimensional setups and a log-linear relationship in two-dimensional ones.

All these dissimilarity measures refer to population-level analyses. Starting from the 90s, many individual-based genetic dissimilarity measures were introduced. Bowcock et al. (1994) proposed a measure based on the proportion of shared alleles, which will be used in this study. With the aim of examining fine-scale genetic structure in plant populations, Loiselle et al. (1995) introduced an estimator of coancestry based on the correlation of the frequencies of homologous alleles at a given locus of two individuals. Rousset (2000) extended his ideas for F_{st} to an individual-based analogue, the $\hat{a}$ coefficient. Vekemans and Hardy (2004) proposed the "*Sp*" statistic, which - at a given spatial scale, with dispersal-drift equilibrium and excluding selection pressures - is deemed to correctly estimate the neighborhood size. The same estimation was also the goal of Watts et al. (2007), who developed the $\hat{e}$ estimator starting from the proposal by Loiselle et al. (1995). Kinship, relatedness and fraternity coefficients were studied at the individual level, too (see the user manual to the SPAGeDi software (Hardy and Vekemans 2002) for a review). Also Rousset's $\hat{a}$ and kinship coefficients were found to

display an approximately linear relationship with log-transformed geographic distances in two-dimensional setups, at least within some distance ranges.

Albeit not exhaustive, this review shows how most of these studies on the relationship between genetic relatedness and geographic separation focused on the validation of migration models or the estimation of demographic parameters. Main issues included how to efficiently measure genetic relatedness, which transformations to apply and whether the resulting relationship was linear or not. More recently, the focus of some studies was rather on the choice of the landscape distance, and the goal was to clarify what environmental factors were best at explaining the genetic structure in a species (Jaquiéry et al. 2011; MacDonald et al. 2020; Peterman and N. S. Pope 2021). The impact of these environmental factors was often measured controlling for the geographic (e.g., Euclidean) distance.

The general idea in this project is that grouping information can be treated as an environmental factor and its impact on genetic structure can be assessed controlling for geographic separation. As an example, two putative groups might be identified on the basis of morphological traits (Gratton et al. 2016). A grouping distance can then be defined that takes value 1 if two individuals come from different groups and 0 otherwise. If genetic distances are positively associated with this grouping distance even after accounting for the geographic distance - thus for patterns like isolation by distance - this supports the delimitation of the two groups. Instead, if no significant association is spotted, this means that any genetic structure spotted between the putative groups is compatible with what IBD can predict. As explained above, this kind of analysis only operationalizes one aspect of the heterogeneity in the data and does not suffice to decide about species distinctness. Species delimitation considerations may follow once other operational criteria are explored for the putative groups considered (De Queiroz 2007).

Fairly recent species delimitation studies where tests of this kind were carried out are Medrano et al. (2014), Spriggs et al. (2018) and Hausdorf and Hennig (2020). The methodologies in these papers will be explored in this work, and new ones will be introduced. When delimiting species, neglecting geographic information can lead to the overestimation of the genetic structure in the data (Frantz et al. 2009). That is, in the presence of IBD patterns, geographically separate groups might display genetic differences that may be wrongly interpreted as species distinctness. By controlling for geographic separation when assessing the explanatory power of grouping information, the techniques discussed in this work attempt to tackle this issue, ensuring more reliable species delimitation results.

1.2 Data

Spatially explicit genetic data consists of individuals carrying information about their location and genetic make-up. In this chapter, methods will be applied on n individuals from two groups, with known membership. Hence, two columns in our data-frame 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. Individual-level co-dominant 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 (see, e.g., 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$^{\text{th}}$ individual.

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

$$d_x(\mathbf{z}_i, \mathbf{z}_j) = \sqrt{\left(z_i^{(x)} - z_j^{(x)}\right)^2 + \left(z_i^{(y)} - z_j^{(y)}\right)^2}. \tag{1.1}$$

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

$$d_y(\mathbf{z}_i, \mathbf{z}_j) = 1 - \frac{1}{2P} \sum_{p=1}^{P} \left| Z_i^p \cap Z_j^p \right| \cdot \left[1 + \mathbb{1}\left(|Z_i^p| + |Z_j^p| = 2 \right) \right], \tag{1.2}$$

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 Appendix B), but missing values can cause a violation of the triangle inequality. In this study, this dissimilarity was computed with functions from the `R` package `prabclus` (Hausdorf and Hennig 2019). 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. Some statistical properties of the shared allele distance are discussed in section 1.7.

The grouping distance is $d_g(\mathbf{z}_i, \mathbf{z}_j) = \mathbb{1}(z_i^c \neq z_j^c)$. For instance, given the toy data frame

$$\mathbf{Z} = \begin{pmatrix} z_1^{(x)} & z_1^{(y)} & 1 & \{A\} & \{B\} & \{A,C\} \\ z_2^{(x)} & z_2^{(y)} & 1 & \{A,B\} & \{A,B\} & \{B,C\} \\ z_3^{(x)} & z_3^{(y)} & 2 & \{C\} & \{A\} & \{C\} \end{pmatrix},$$

we get:

$$\begin{array}{lll} d_y(\mathbf{z}_1,\mathbf{z}_2) = \frac{1}{2} & d_y(\mathbf{z}_1,\mathbf{z}_3) = \frac{5}{6} & d_y(\mathbf{z}_2,\mathbf{z}_3) = \frac{2}{3}; \\ d_g(\mathbf{z}_1,\mathbf{z}_2) = 0 & d_g(\mathbf{z}_1,\mathbf{z}_3) = 1 & d_g(\mathbf{z}_2,\mathbf{z}_3) = 1. \end{array}$$

Let $n_g = |\{i : z_i^c = g\}|$ be the number of individuals belonging to group $g = 1, 2$. These represent the two candidate (or putative) species to be tested for conspecificity. In practice, this grouping information may be based on, say, morphological (Gratton et al. 2016), ecological (Raxworthy et al. 2007; Rissler and Apodaca 2007), behavioural (Scapini et al. 2002) grounds or can simply represent the researcher's hypothesis. The number of geographic 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 $n \times n$ block matrices $\mathbf{D}_y$ and $\mathbf{D}_g$ store the genetic and grouping dissimilarities.

In the literature, a logarithmic transformation of the geographic distances is sometimes applied to achieve a more linear relationship between genetic and geographic distances, as done for example in Rousset (1997), Vekemans and Hardy (2004) and Hausdorf and Hennig (2020). Zero geographical distances can occur if two individuals are observed in the same location. Therefore, based on Hausdorf and Hennig (2020), the following transformation is considered:

$$f(d_x(\mathbf{z}_r, \mathbf{z}_c)) = \ln(d_x(\mathbf{z}_r, \mathbf{z}_c) + F_x^{-1}(0.25)), \tag{1.3}$$

where

$$\begin{aligned} F_x^{-1}(q) &= \inf\{d : F_x(d) \geq q\}, \quad \text{with} \\ F_x(d) &= \frac{1}{w} \sum_{c<r\leq n} \mathbb{1}(d_x(\mathbf{z}_r, \mathbf{z}_c) \leq d) \end{aligned}$$

being the empirical cumulative distribution function of all geographic distances in the dataset, and $w = n(n-1)/2$. Both untransformed and log-transformed $\mathbf{D}_x$ will be considered in this work. In order to keep notation light, all methods will be described using untransformed geographic distances.

Although the shared allele distance and the Euclidean (or geodesic) distance are used in this project, the discussed methods are based on models for dissimilarities that do not rely on these specific dissimilarities. The methods can therefore also be applied to other dissimilarities.

1.3 Methods

The methodologies presented in this chapter leverage the information on the relationship between the distances in $\mathbf{D}_y$ and $\mathbf{D}_x$ with the aim of confirming or falsifying a conspecificity presumption encoded in $\mathbf{D}_g$. In the presence of isolation by distance behaviour (positive association between genetic and Euclidean distance), two groups of individuals belonging to the same species might display a certain degree of genetic structure that is explained by their geographic separation. If the genetic dissimilarities are too large to be compatible with the geographic 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 metapopulations would be sufficient to prevent the fusion of these metapopulations 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 geographic dissimilarities, which can be non-linear (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 geographic 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 geographic distances. The methods are based on linear regression and correlation, i.e., they assume linearity. 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. Incidentally, for Euclidean data, Székely et al. (2007) even show that independence is equivalent to a "distance correlation", closely related to what is considered here, being zero. Furthermore, the methods treated here do not require the triangle inequality, and monotonic transformations of dissimilarities can also be used.

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

1.3.1 Regression on dissimilarities with jackknife testing

Hausdorf and Hennig (2020) propose to regress the genetic dissimilarities on the log-transformed geographic 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 Figure 1.1) is compared with a regression line based on all dissimilarities. The null hypothesis of conspecificity is rejected if the between-groups dissimilarities (green in Figure 1.1) 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).

To begin, let us define the sets of all index pairs referring to within-group and between-group dissimilarities, respectively:

$$W = \left\{ r,c \leq n \,|\, c < r \leq n_1 \vee n_1 < c < r \right\},\ B = \left\{ r,c \leq n \,|\, r > n_1 \wedge c \leq n_1 \right\}.$$

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

$$d_y(\mathbf{z}_r,\mathbf{z}_c) = \begin{cases} a_1 + b_1\{d_x(\mathbf{z}_r,\mathbf{z}_c) - \bar{d}_x^{\,W}\} + e(\mathbf{z}_r,\mathbf{z}_c) & \text{with}\, c < r \leq n_1 \\ a_2 + b_2\{d_x(\mathbf{z}_r,\mathbf{z}_c) - \bar{d}_x^{\,W}\} + e(\mathbf{z}_r,\mathbf{z}_c) & \text{with}\, n_1 < c < r \end{cases}. \tag{1.4}$$

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 geographic distance taken over both candidate species. The errors e in (1.4) are assumed to have zero mean, but not to be independent. Only the genetic random variation of individuals is assumed to be independent (which is a limitation with respect to population-level studies), 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 $u_i,\ 1=1,\ldots,n$ 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). Jackknife testing is a heuristic idea that has a theoretical justification in specific situations (Shao and Tu 2012), of which the assumptions are not fulfilled here. Therefore, its characteristics have to be explored experimentally in all but the simplest situations, as is done in this study. It 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. We write, for $r,c \neq i$:

$$d_y(\mathbf{z}_r,\mathbf{z}_c) = \begin{cases} a_{1(i)} + b_{1(i)}\{d_x(\mathbf{z}_r,\mathbf{z}_c) - \bar{d}_x^{\,W}\} + e(\mathbf{z}_r,\mathbf{z}_c) & \text{with } c < r \leq n_1 \qquad \text{for } i = 1,\ldots,n_1 \\ a_{2(i)} + b_{2(i)}\{d_x(\mathbf{z}_r,\mathbf{z}_c) - \bar{d}_x^{\,W}\} + e(\mathbf{z}_r,\mathbf{z}_c) & \text{with } n_1 < c < r \quad \text{for } i = n_1+1,\ldots,n. \end{cases} \tag{1.5}$$

In (1.5), n_1+n_2 regressions are defined, from which jackknife replicates of the coefficients are obtained via OLS. As explained above, pseudovalues $\hat{a}^{\psi}_{1(i)}$ and $\hat{b}^{\psi}_{1(i)}$ are then computed for group 1 and pseudovalues $\hat{a}^{\psi}_{2(i)}$ and $\hat{b}^{\psi}_{2(i)}$ are obtained for group 2, using OLS estimates of the coefficients in (1.4). Average pseudovalues and jackknife estimates of the standard errors will be used to construct the test statistic. For instance, the standard error estimate for the intercept coefficient in the first group is

$$se^2_{a_1^{\psi}} = \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n_1} \left(\hat{a}^{\psi}_{1(i)} - \bar{a}^{\psi}_1\right)^2,$$

where $\bar{a}^{\psi}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{a}^{\psi}_{1(i)}$ is the average pseudovalue and $\hat{a}^{\psi}_{1(i)} = \hat{a}_1 n_1 - (n_1-1)\hat{a}_{1(i)}$ is the i^{th} pseudovalue, with $i=1,\ldots,n_1$. Similarly, the other average pseudovalues $\bar{a}^{\psi}_2$, $\bar{b}^{\psi}_1$ and $\bar{b}^{\psi}_2$ and standard error estimates $se^2_{a_2^{\psi}}$, $se^2_{b_1^{\psi}}$ and $se^2_{b_2^{\psi}}$ are obtained.

The p-value for H_{01} is then computed as the minimum between 1 and

$$2 \cdot \min\left(2 \cdot \mathbb{P}\left(t_{\nu_{a^{\psi}}} > \frac{\left|\bar{a}^{\psi}_1 - \bar{a}^{\psi}_2\right|}{\sqrt{se^2_{a_1^{\psi}} + se^2_{a_2^{\psi}}}}\right), 2 \cdot \mathbb{P}\left(t_{\nu_{b^{\psi}}} > \frac{\left|\bar{b}^{\psi}_1 - \bar{b}^{\psi}_2\right|}{\sqrt{se^2_{b_1^{\psi}} + se^2_{b_2^{\psi}}}}\right)\right), \tag{1.6}$$

where t_ν is a Student-t random variable with ν degrees of freedom and the Welch–Satterthwaite formula prescribes

$$\nu_{a^\psi} = \frac{\left(se^2_{a_1^\psi} + se^2_{a_2^\psi}\right)^2}{\frac{se^4_{a_1^\psi}}{n_1-1} + \frac{se^4_{a_2^\psi}}{n_2-1}},$$
$$\nu_{b^\psi} = \frac{\left(se^2_{b_1^\psi} + se^2_{b_2^\psi}\right)^2}{\frac{se^4_{b_1^\psi}}{n_1-1} + \frac{se^4_{b_2^\psi}}{n_2-1}}.$$

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:

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = a_* + b_*(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d_x}^W) + e(\mathbf{z}_r, \mathbf{z}_c) \tag{1.7}$$

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:

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = a + b(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d_x}^W) + e(\mathbf{z}_r, \mathbf{z}_c) \tag{1.8}$$

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

$$a + b(\bar{d_x}^B - \bar{d_x}^W) > a_* + b_*(\bar{d_x}^B - \bar{d_x}^W), \tag{1.9}$$

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

$$\hat{a} + \hat{b}(\bar{d_x}^B - \bar{d_x}^W) - \hat{a_*} - \hat{b_*}(\bar{d_x}^B - \bar{d_x}^W), \tag{1.10}$$

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

For $r,c \neq i$, the following batch of regression models is fitted:

$$d_y(\mathbf{z}_r,\mathbf{z}_c) = \begin{cases} a_{*(i)} + b_{*(i)}(d_x(\mathbf{z}_r,\mathbf{z}_c) - \bar{d}_x^{\,W}) + e(\mathbf{z}_r,\mathbf{z}_c) & \text{with } r,c \in W \\ a_{(i)} + b_{(i)}(d_y(\mathbf{z}_r,\mathbf{z}_c) - \bar{d}_x^{\,W}) + e(\mathbf{z}_r,\mathbf{z}_c) & \text{with } r > c \end{cases} \quad \text{for } i = 1,\dots,n\,. \tag{1.11}$$

In (1.11), $2n$ regression models are fitted and n replicates of $\hat{a_{*(i)}}$, $\hat{b_{*(i)}}$, $\hat{a_{(i)}}$ and $\hat{b_{(i)}}$ are generated. Pseudovalues $\hat{a}^{\Psi}_{*i}$, $\hat{b}^{\Psi}_{*i}$, $\hat{a}^{\Psi}_{i}$ and $\hat{b}^{\Psi}_{i}$ are obtained, for $i = 1,\dots,n$, and used to compute n jackknife pseudovalues for (1.10):

$$T^{\Psi}_{2i} = \hat{a}^{\Psi}_{i} - \hat{a}^{\Psi}_{*i} + (\hat{b}^{\Psi}_{i} - \hat{b}^{\Psi}_{*i})(\bar{d}_x^{\,B} - \bar{d}_x^{\,W}) \quad \text{for } i = 1,\dots,n\,.$$

The jackknife estimate of the standard error of (1.10) is then

$$se_{T^{\Psi}_2} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(T^{\Psi}_{2i} - \bar{T}^{\Psi}_{2})^2}\,,$$

where $\bar{T}^{\Psi}_{2} = \frac{1}{n}\sum_{i=1}^{n} T^{\Psi}_{2i}$. H_{02} is thus tested, with p-value equal to $\mathbb{P}\left(t_{n-1} > \frac{\bar{T}^{\Psi}_{2}}{se_{T^{\Psi}_2}}\right)$.

If H_{01} is rejected, the IBD behaviour of the two candidate species cannot be described by a unique model and model (1.4) is adopted. In this situation the compatibility of IBD behaviour and genetic structure is checked for each group separately. $H_{03}:\ a_g = a^*_g$ and $b_g = b^*_g$ for at least one of $g = 1,2$ is tested, where a_g, b_g refer to regressions based on dissimilarities within group g only, and a^*_g, b^*_g refer to regressions based on all dissimilarities involving a member of group g. We define:

$$\begin{aligned} B_1 &= B \cup \left\{ r,c \,|\, c < r \leq n_1 \right\} \\ B_2 &= B \cup \left\{ r,c \,|\, n_1 < c < r \right\} \\ \bar{d}_x^{\,(1)} &= \frac{2}{n_1(n_1-1)} \sum_{c<r\leq n_1} d_x(\mathbf{z}_r,\mathbf{z}_c) \\ \bar{d}_x^{\,(2)} &= \frac{2}{n_2(n_2-1)} \sum_{n_1<c<r} d_x(\mathbf{z}_r,\mathbf{z}_c) \end{aligned}$$

where $\bar{d}_x^{\,(1)}$ and $\bar{d}_x^{\,(2)}$ are the average within-group geographic dissimilarities for group 1 and 2, respectively. B_1 is the set of index pairs referring to either between-group dissimilarities or dissimilarities within group 1. B_2 analogously for group 2. Based on these two partially overlapping sets, the following linear models are fitted:

$$d_y(\mathbf{z}_r,\mathbf{z}_c) = \begin{cases} a^*_1 + b^*_1(d_x(\mathbf{z}_r,\mathbf{z}_c) - \bar{d}_x^{\,(1)}) + e(\mathbf{z}_r,\mathbf{z}_c) & \text{with } r,c \in B_1 \\ a^*_2 + b^*_2(d_x(\mathbf{z}_r,\mathbf{z}_c) - \bar{d}_x^{\,(2)}) + e(\mathbf{z}_r,\mathbf{z}_c) & \text{with } r,c \in B_2 \end{cases} \tag{1.12}$$

where $\mathbb{E}(e(\mathbf{z}_r, \mathbf{z}_c)) = 0$. For each group, model (1.4) and (1.12) are compared and, similarly to the previous test, the comparison is based on their prediction at $\bar{d}_x{}^B$. So, H_{03} is tested against the alternative:

$$a_g^* + b_g^*(\bar{d}_x{}^B - \bar{d}_x{}^{(g)}) > a_g + b_g(\bar{d}_x{}^B - \bar{d}_x{}^{(g)})$$

and the test rejects if the maximum of the p-values for the two tests regarding groups $g = 1, 2$ is too small.

Let us here focus on the test based on the set B_1 for brevity: the one based on set B_2 will be easily reproducible for symmetry. From (1.4) and (1.12), OLS estimates $\hat{a_1}$, $\hat{b_1}$, $\hat{a}_1^*$ and $\hat{b}_1^*$ are obtained. The goal is to come up with a measure of uncertainty for

$$T_{31} = \hat{a}_1^* - \hat{a_1} + \left(\hat{b}_1^* - \hat{b_1}\right)\left(\bar{d}_x{}^B - \bar{d}_x{}^{(1)}\right). \tag{1.13}$$

When excluding the elements in B_1, some jackknife iterations will impact only the estimates of $\hat{a}_1^*$ and $\hat{b}_1^*$ (when an individual from the second group is dropped) whereas all the other iterations will impact also $\hat{a_1}$ and $\hat{b_1}$ (when an individual from group 1 is dropped). To handle this with the notation, let us clarify that $\hat{a}_{1(i)} = \hat{a}_1$ and $\hat{b}_{1(i)} = \hat{b}_1$ if $i = n_1 + 1, \ldots, n_1 + n_2$: there is no jackknife replicate of the intercept and slope estimates if the individual that is excluded in the iteration does not belong to the first group, so the original estimates are considered. In order to get n jackknife replicates of (1.13), we fit the following batch of linear regressions:

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = a_{1(i)}^* + b_{1(i)}^*\left(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_x{}^{(1)}\right) + e(\mathbf{z}_r, \mathbf{z}_c) \quad \text{with } r, c \in B_1 \wedge r, c \neq i \quad i = 1, \ldots, n. \tag{1.14}$$

The n replicates $a_{1(i)}^*$ and $b_{1(i)}^*$ from (1.14), together with the n_1 replicates obtained in (1.5) and the full-sample estimates $\hat{a_1}$ and $\hat{b_1}$, allow us to compute the n jackknife replicates of (1.13) as follows:

$$T_{31(i)} = \hat{a}_{1(i)}^* - \hat{a}_{1(i)} + \left(\hat{b}_{1(i)}^* - \hat{b}_{1(i)}\right)\left(\bar{d}_x{}^B - \bar{d}_x{}^{(1)}\right). \tag{1.15}$$

We are now able to compute the n pseudovalues

$$T_{31(i)}^{\psi} = nT_{31} - (n-1)T_{31(i)}, \tag{1.16}$$

whose mean will constitute the numerator of the test statistic. Note how we first computed jackknife replicates of the test statistic and afterwards obtained pseudovalues with the usual formula, whereas, when testing H_{02}, we first attained pseudovalues of the intercept and slopes estimates and used them to directly get pseudovalues of the

test statistic. These two ways of obtaining pseudovalues are equivalent; in the same spirit, one can also compute the jackknife estimate of the standard error using jackknife replicates instead of pseudovalues, taking care of the multiplying factor (Efron and Tibshirani 1993, ch. 11).

The computation of the standard error estimate here presents one additional step with respect to the procedure for testing H_{02}, because of the unequal impact that the removal of the individuals has on the values in (1.15): when $i = 1, \ldots, n_1$, both pairs of estimates are modified, whereas, when the individual belongs to the second group, only the estimates in (1.14) are updated. To take this into account, we pool the variance of the pseudovalues generated in the first scenario with the variance of those generated under the second scenario:

$$V_{31.1} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (T^{\psi}_{31(i)} - \bar{T}^{\psi}_{31.1})^2$$
$$V_{31.2} = \frac{1}{n_2 - 1} \sum_{i=n_1+1}^{n} (T^{\psi}_{31(i)} - \bar{T}^{\psi}_{31.2})^2,$$

where $\bar{T}^{\psi}_{31.1} = \frac{1}{n_1} \sum_{i=1}^{n_1} T^{\psi}_{31(i)}$ and $\bar{T}^{\psi}_{31.2} = \frac{1}{n_2} \sum_{i=n_1+1}^{n} T^{\psi}_{31(i)}$. The jackknife estimate of the standard error is then:

$$se^2_{T_{31}} = \frac{n_1 V_{31.1} + n_2 V_{31.2}}{n^2}.$$

The computation of the degrees of freedom relies again on the Welch-Satterthwaite approximation:

$$\nu^{\psi}_{31} = \frac{se^4_{T_{31}}}{\frac{1}{n_1 - 1}(\frac{n_1 V_{31.1}}{n})^2 + \frac{1}{n_2 - 1}(\frac{n_2 V_{31.2}}{n})^2}.$$

The p-value is then computed as $\mathbb{P}\left(t_{\nu^{\psi}_{31}} > \frac{\sqrt{n}\bar{T}^{\psi}_{31}}{se_{T_{31}}}\right)$, where $\bar{T}^{\psi}_{31} = \frac{1}{n} \sum_{i=1}^{n} T^{\psi}_{31(i)}$. The same procedure is followed based on the set B_2 to obtain the second p-value and the largest of the two is compared with the desired significance level. However, situations in which the two tests disagree (one p-value is smaller than the significance threshold and the other is not) require data-specific considerations, as the result of this testing protocol is inconclusive.

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 geographic dissimilarities displayed by the two metapopulations cannot explain their genetic differences and they might thus represent two separate lineages.

1.3.2 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 $\mathbf{D}_x$" (P. Legendre and L. 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:

$$r(\mathbf{D}_y, \mathbf{D}_x) = \frac{\sum_{c<r\leq n}(d_y(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_y)(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_x)}{\sqrt{\sum_{c<r\leq n}(d_y(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_y)^2 \sum_{c<r\leq n}(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_x)^2}}, \tag{1.17}$$

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 geographic distance.

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

$$r(\mathbf{D}_y, \mathbf{D}_g | \mathbf{D}_x) = \frac{r(\mathbf{D}_y, \mathbf{D}_g) - r(\mathbf{D}_y, \mathbf{D}_x) r(\mathbf{D}_g, \mathbf{D}_x)}{\sqrt{(1 - r(\mathbf{D}_y, \mathbf{D}_x)^2)(1 - r(\mathbf{D}_g, \mathbf{D}_x)^2)}}\,. \tag{1.18}$$

(1.18) quantifies the correlation between the genetic dissimilarities and the grouping distances after having accounted for the geographic distances. A partial Spearman correlation could be employed to test for monotonic relationships (Q. Liu et al. 2018). Medrano et al. (2014) tested the null hypothesis that the true population value of (1.18) is zero in order to ascribe the genetic structure found in two subgroups of trumpet daffodils to their lineage separation. The rejection of such 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 Figure 1.1 the null hypothesis means that conditionally on geographic distances between-groups genetic dissimilarities (i.e., green) are not systematically larger or smaller than within-group ones (i.e., red or black).

Hypothesis testing is usually carried out by means of permutations, although there exists an asymptotically normal transformation of (1.17) (P. Legendre and L. Legendre 2012, p. 600). P. 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 rows and corresponding columns of $\mathbf{D}_y$ and recomputing the partial

correlation coefficient a large number of times. The permutable units for the test are indeed the n individuals. 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 (1.18). If the two groups are from distinct species, the partial correlation between genetic and grouping 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 $r(\mathbf{D}_y, \mathbf{D}_g | \mathbf{D}_x)$. P. 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.

There has been a long debate in the literature concerning Mantel tests, see Diniz-Filho et al. (2013) for a review. P. Legendre and Fortin (2010) mathematically show that tests on the correlation between raw data vectors do not correspond to tests on the correlation between the dissimilarities computed from that raw data. A brief history of the controversies on Mantel tests is reported in the appendix to P. Legendre, Fortin, et al. (2015), who claim that Mantel tests should only be used when research hypotheses concern relationships between dissimilarities and not between the raw data vectors. Guillot and Rousset (2013) maintain that the permutation procedure traditionally employed in Mantel testing in most cases leads to inflated type I error rates - see next paragraph. Bradburd, Ralph, and Coop (2013) discuss the limitations of partial Mantel tests in landscape genetics applications. If one is interested in relationships between raw data vectors, a node-level analysis (Wagner and Fortin 2015) can be preferable and techniques that have shown better performance with respect to the Mantel test (P. Legendre, Fortin, et al. 2015) are described by P. Legendre and L. Legendre (2012, ch. 14), but also mentioned by Diniz-Filho et al. (2013). Moreover, strongly non-linear and heteroskedastic relationship between genetic and geographic dissimilarities can hinder Mantel test performance: examining the Mantel correlogram (Borcard and P. Legendre 2012) may help grasp how the genetic dissimilarities evolve with landscape distance. Goslee and Urban (2007) proposed a piecewise version of the correlogram that can help handle these assumptions' violation.

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 the true population value of (1.18) is zero as prescribed by the null hypothesis, but 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 P. Legendre (2000) also come with potentially unrealistic implicit structural assumptions.

The potential distortion from permutation can be prevented by jackknifing the partial correlation (1.18). By leaving one individual out at a time, n jackknife replicates of the partial correlation coefficient are obtained:

$$r(\mathbf{D}_{y(i)},\mathbf{D}_{g(i)}|\mathbf{D}_{x(i)}) = \frac{r(\mathbf{D}_{y(i)},\mathbf{D}_{g(i)}) - r(\mathbf{D}_{y(i)},\mathbf{D}_{x(i)})r(\mathbf{D}_{g(i)},\mathbf{D}_{x(i)})}{\sqrt{(1-r(\mathbf{D}_{y(i)},\mathbf{D}_{x(i)})^2)(1-r(\mathbf{D}_{g(i)},\mathbf{D}_{x(i)})^2)}}, \tag{1.19}$$

where $\mathbf{D}_{\cdot(i)}$ is the version of the dissimilarity matrix $\mathbf{D}_{\cdot}$ in which the row and corresponding column involving the i^{th} individual in the dataset are excluded. Then, n pseudovalues are computed as

$$r_i^{\psi} = n\, r(\mathbf{D}_y,\mathbf{D}_g|\mathbf{D}_x) - (n-1)\, r(\mathbf{D}_{y(i)},\mathbf{D}_{g(i)}|\mathbf{D}_{x(i)})$$

and used to obtain a jackknife estimate of the standard error of the partial correlation coefficient:

$$se_r = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(r_1^{\psi} - \bar{r}^{\psi})^2},$$

where $\bar{r}^{\psi} = \frac{1}{n}\sum_{i=1}^{n} r_i^{\psi}$. The p-value is obtained as $\mathbb{P}\left(t_{n-1} > \frac{\bar{r}^{\psi}}{se_r}\right)$.

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 sampled 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.

Bias-corrected (BC) bootstrap confidence intervals were used here, as defined and motivated in Efron and Tibshirani (1993, ch. 14.3 and 22.5). Let $r_b(\mathbf{D}_y,\mathbf{D}_g|\mathbf{D}_x)$, with $b = 1,\ldots,B$, be bootstrap replicates of the partial correlation coefficients and $F_r(d) = \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\left(r_b(\mathbf{D}_y,\mathbf{D}_g|\mathbf{D}_x) \leq d\right)$ the empirical distribution function thereof. The lower boundary r_{lo} of bias-corrected bootstrap confidence intervals is picked from the distribution of bootstrap replicates according to how far the original value of the partial correlation falls from its median bootstrap replicate. The null hypothesis of the PMT is rejected at α significance level if $r_{lo} = F_r^{-1}\left(\Phi(2\cdot\hat{\zeta} - \Phi^{-1}(1-\alpha))\right) > 0$,

where $\hat{\zeta} = \Phi^{-1}\left(F_r(r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x))\right)$ and Φ is the cumulative standard normal distribution function. This type of confidence interval was used by Mason and Mimmack (1992) for a simple correlation coefficient.

Section 1.5.3 will expand on these three ways of partial Mantel testing.

1.3.3 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 geographic and genetic dissimilarities.

Clarke et al. (2002) proposed such a model. They were working with population-level genetic and geographic data. After centering the geographic distances to remove correlation between the intercept and slope estimates, they extended the linear regression between genetic and (log-transformed) geographic 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

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = a + b\left(d_x(\mathbf{z}_r, \mathbf{z}_c) - \bar{d}_x\right) + \tau_r + \tau_c + \varepsilon(\mathbf{z}_r, \mathbf{z}_c) \quad \text{with} \, n \geq r > c\,, \tag{1.20}$$

where a is a constant term, τ_r is a random effect representing the average deviation of d_y values involving individual r from that expected from its d_x distances to the other individuals and τ_r and $\varepsilon(\mathbf{z}_r, \mathbf{z}_c)$ are assumed to be independent with $\varepsilon(\mathbf{z}_r, \mathbf{z}_c)$ i.i.d. normally distributed. This specification assumes that dependence between two dissimilarities involving the same individual 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 individuals, 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 idealizations 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 N. S. Pope 2021), being used to assess the effect of landscape variables on gene flow (Van Strien, Keller, and Holderegger 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 (1.20) for species delimitation, a fixed effect associated with the grouping distance $\mathbf{D}_g$ needs to be incorporated:

$$d_y(\mathbf{z}_r,\mathbf{z}_c) = b_0 + b_1 d_x(\mathbf{z}_r,\mathbf{z}_c) + b_2 d_g(\mathbf{z}_r,\mathbf{z}_c) + \tau_r + \tau_c + \varepsilon(\mathbf{z}_r,\mathbf{z}_c) \quad \text{with} n \geq r > c\,. \tag{1.21}$$

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 Figure 1.1 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 Section 1.3.1, 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 unlike the approaches in Sections 1.3.1 and 1.3.2, (1.21) 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) (Royston 2007; Venzon and Moolgavkar 1988). 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, model (1.21) should not be fitted with REML (West et al. 2022). Anyway, given the small number of fixed effects included in the model, estimates from ML and REML should not differ much (Verbeke and Molenberghs 2009, section 5.3.5).

A related approach was used by R. Yang (2004) for estimating and testing for isolation by distance. Instead of introducing random effects explicitly, several standard correlation patterns for the $\varepsilon(\mathbf{z}_r,\mathbf{z}_c)$ as available in the SAS PROC MIXED (SAS Institute 2001) were used to model the dependence in the dissimilarities. This can be expected to be inferior to (1.21), because it does not use the information which dissimilarities belong to the same individual.

A linear regression model ignoring dependence This study also features a straight linear regression model without the random effects that ignores the dependence between dissimilarities:

$$d_y(\mathbf{z}_r,\mathbf{z}_c) = b_0^* + b_1^* d_x(\mathbf{z}_r,\mathbf{z}_c) + b_2^* d_g(\mathbf{z}_r,\mathbf{z}_c) + \varepsilon(\mathbf{z}_r,\mathbf{z}_c) \quad \text{with} n \geq r > c\,, \tag{1.22}$$

assuming $\varepsilon(\mathbf{z}_r, \mathbf{z}_c)$ 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 (1.22) held indeed, the null hypothesis of the partial Mantel test in section 1.3.2 would be equivalent to $b_2^* = 0$ (P. Legendre 2000).

Type I error and power properties of all these methodologies will be assessed by means of simulations, which are introduced in the following section. All tests will be run at level $\alpha = 0.05$.

1.4 Simulations

The relationship between genetic and geographic dissimilarities in real data can be influenced by a plethora of factors, including sampling scale (Anderson et al. 2010), introgression, i.e., gene flow between distinct species (Bamberger et al. 2021), missingness in the genetic data (Séré et al. 2017), habitat configuration and heterogeneity (Van Strien, Holderegger, and Van Heck 2015), to name just a few. By enabling researchers to control some of these factors, simulations have proved crucial in the related literature (Epperson et al. 2010). Landguth et al. (2015) highlight the importance of simulations, particularly individual-based and spatially explicit, in landscape genetics.

More than 20 software packages for simulating genome-wide data are listed in Bourgeois and Warren (2021). In population genetics, simulation algorithms can be mainly divided in backward-time and forward-time approaches (Yuan et al. 2012). The former are related to the coalescent process (Kingman 1982) and reconstruct the history of alleles observed at the present generation up to the most recent common ancestor; allele mutations are then applied on the reconstructed genealogy. Since the history of lineages whose offsrping did not survive up to the present generation are ignored in the process, these algorithms are often quite efficient (Carvajal-Rodríguez 2008). A widely used state-of-the-art software package for coalescent simulations is msprime (Baumdicker et al. 2021), which is not spatially-explicit. Packages that include a coalescent component and take into account the geographic position of populations are, e.g., SPLATCHE3 (Currat et al. 2019), quantinemo2 (Neuenschwander et al. 2018) and GSpace (Virgoulay et al. 2021b). Forward-time approaches tend to be less computationally efficient because they simulate the evolutionary history of the species, generation after generation, from an ancestral population to the present. Thanks to this, they allow more flexible modeling of demographic scenarios related to mutation, recombination and selection. Among this kind of packages we find geonomics (Terasaki

Hart et al. 2021), CDmetaPOP2 (Day et al. 2023) and SLiM (Haller and Messer 2023), all spatially-explicit. Attempts to combine the computational efficiency of coalescent simulations with the flexibility of forward-time ones also exist (Kelleher et al. 2018).

The software packages employed in this study, SLiM and GSpace, simulate genetic data for individuals with geographic locations rather than dissimilarities, see Figure 1.2. 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 patterns. 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 geographic distances.

In the following, the two simulators used in this comparative study are described and the performance of the methods in section 1.3 on simulated datasets is discussed.

1.4.1 Simulations based on the SLiM simulator

As explained in its user manual (Haller and Messer 2022), 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. Individuals can be hermaphroditic or sexual and mating need not be biparental. 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 SLiM manual describes more than 100 scripts (called "recipes") for simulating particular evolutionary scenarios. 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 section 1.3. 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 generate 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 geographic distances. In this project, the term "quasi-panmictic" is used for convenience to refer to species simulated withouth IBD pattern, i.e., to scenarios where the genetic distance between co-specific individuals does not grow with geographic separation. As explained in section 1.1, though, the assumption of panmixia can of course be violated for several other reasons. The individuals from the two groups might inhabit the same geographic area or might be segregated into two disjoint areas. As regards the conspecificity and distinctness 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;
- 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, 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, may generate a scenario consistent with the alternative hypothesis of distinctness, 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 for the three scenarios above can be found in Appendix C.1.

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 *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 sub-scenarios, the two "children" subpopulations would continue to share the same wide area (*overlapping conspecific*) or start to migrate to two disjoint areas of the map (*separate conspecific*). 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 sub-scenarios: in the *overlapping distinct* one, both species would inhabit the central 100×100 area, whereas in the *separate distinct* 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 *split* scenario, 100 generations were simulated. In the *conspecificity* scenario, 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* scenario, both species would be simulated for 50 generations.

- The only subpopulation existing in the *split* scenario and the parent population in the *conspecificity* scenario were made up of 400 individuals. The children subpopulations in the *conspecificity* scenario and the two species in the *distinctness* scenario 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 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 *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 *separate conspecific* scenario, when it was constrained to those individuals inhabiting the subpopulation-specific geographic 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 quasi-panmictic 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 scenario, 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 *separate 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 geographic 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 *separate conspecific* sub-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 Appendix A.1, example distance-distance plots similar to Figure 1.1 from these five scenarios are shown, with equal or unequal group sizes, both for quasi-panmictic groups and for isolated by distance groups. Section 1.4.4 further expands on these parameter choices for simulations.

Results from SLiM As explained above, five scenarios were simulated with SLiM: a *split* scenario (one group, random split), an *overlapping conspecific* scenario (two groups, same parent population, same inhabited area), a *separate conspecific* scenario (two groups, same parent population, disjoint inhabited areas), a *separate distinct* scenario (two species living in disjoint areas) and an *overlapping distinct* scenario (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 (the *split* strategy) to a clear distinctness setup (the multispecies simulation). 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 Section 1.3 were applied both with untransformed and log-transformed geographic distances. 100 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

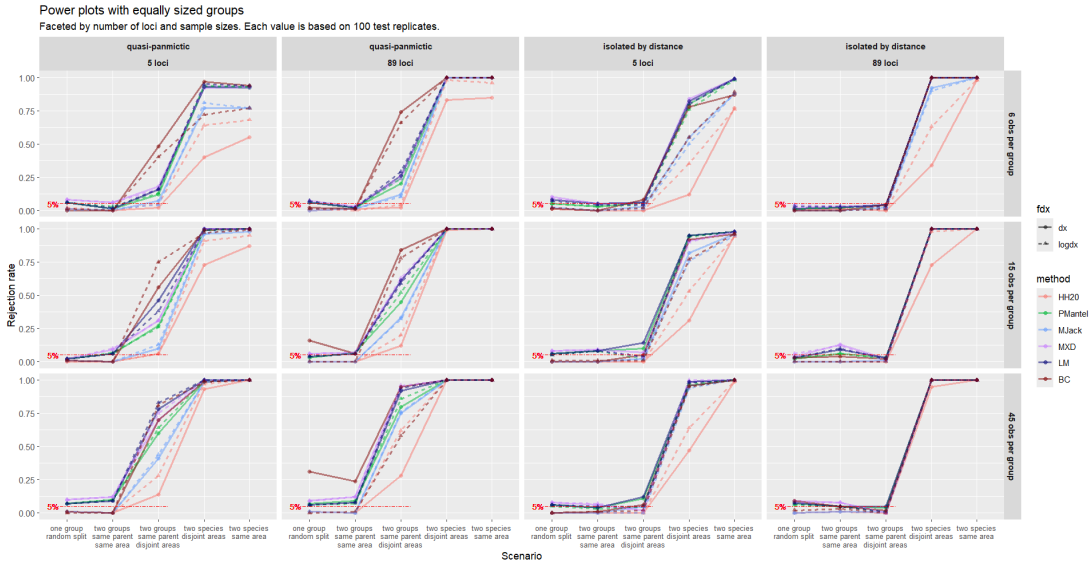

Fig. 1.3 Power plot based on **SLiM** data simulated **with equally sized groups**. Panels by number of individuals per group (rows) and a combination of IBD behavior and number of available loci (columns). Each rejection rate is based on 100 simulations with the same parameter settings. Circles and solid lines refer to analyses with untransformed geographic 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.

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 (1.21); LM is the multiple regression ignoring dependence.

In Figure 1.3, rejection rates from the scenarios with equally sized groups are reported. Results with 40 available loci are only shown in Appendix A.1 (see Figures A.9, A.10, A.11 and A.12), same regarding the GSpace simulations in Section 1.4.2 (see Figures A.13, A.14, A.15 and A.16). The expected non-decreasing trend in the rejection rates was confirmed, with some minor exceptions between the *split* and the *overlapping conspecific* (same parent, same area) scenario. In the *split* scenario, all methods (surprisingly, model (1.22), 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 the *separate conspecific* (same parent, disjoint areas) scenario 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 geographic separation and absence of IBD behaviour: the genetic structure that was formed because of the decreasing migration rate could not be explained by geographic 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 *separate conspecific* scenario all fell below the significance level. Recalling the remarks in the previous sections 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 than HH20 in all setups. Now, if the rejection of the null hypothesis in the *split* scenario 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 geographic distances - see section 1.5.3.

Also in Figure 1.4, with unequal group sizes, the rejection rates were mostly non-decreasing when going from the *split* scenario to the *overlapping distinct* scenario. 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 *split* scenario often lay above the significance level, sometimes strongly so, especially with the largest sample sizes. This confirms that neglecting the dependence in the dissimilarities can lead to type I error inflation for these testing procedures, as further discussed in section 1.5.2.

Regardless of the other parameter options, the transformation of the geographic distances did not seem to have a relevant impact on the methods' performance. The

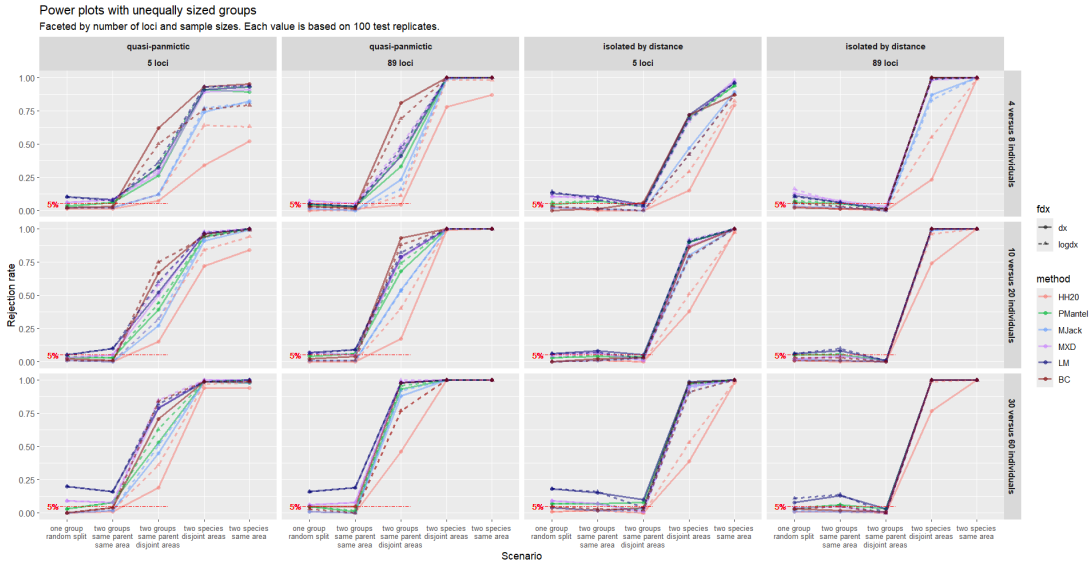

Fig. 1.4 Power plot based on **SLiM** data simulated **with unequally sized groups**. See the caption to Figure 1.3 for further description.

only exception is HH20, whose power increased with log-transformed geographic dissimilarities.

1.4.2 Simulations based on the GSpace simulator

As clarified in its user manual (Virgoulay et al. 2021a), GSpace (Virgoulay et al. 2021b) is a "a backward generation by generation exact coalescence algorithm with recombination" that allows to simulate allelic data on a lattice under isolation by distance. Each deme on the lattice can represent the locality in population-level simulations or the position of single individuals on the map, approximating continuous habitats (but see the remarks in Battey et al. (2020) on this). Given the sampled individuals, the history of neutral genes is generated going backward in time and then mutations are simulated from the most recent common ancestor starting from the top of the coalescence tree down to the branches. At each generation, migration is handled by means of two-dimensional dispersal distributions, to be chosen among uniform, geometric, Zeta, etc. 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 demes map was created, with diploid individuals always inhabiting the central 60×60 area. A first group would inhabit the area between point $(70, 70)$ and point $(90, 90)$, whereas a second group would inhabit the square between points $(110, 110)$ 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 deme 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. In all other cases, the two geographically separated groups were generated during two distinct software executions and collected in the same dataset.

- The two groups would obviously share the same allele pool (the same two allelic stata for all loci) when generated within the same software execution, whereas they could share both alleles, one allele or no allele 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 quasi-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.

Distance-distance plots from all simulated scenarios and the related script can be found in Appendix A.1 and C.2, respectively.

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

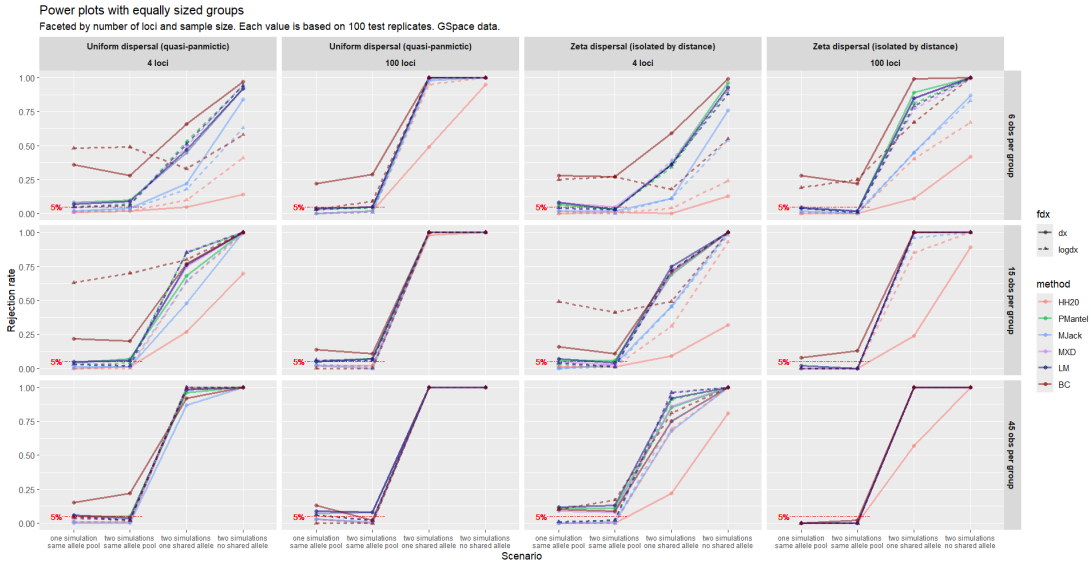

Fig. 1.5 Power plot based on **GSpace** data simulated **with equally sized groups**. Panels by combination of IBD behaviour and number of loci (columns) and number of individuals per group (rows). See the caption to Figure 1.3 for further description.

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 geographic 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 (*split*), two executions and same alleles (*same allele pool*), two executions and one allele in common (*overlapping allele pools*), two executions and no allele in common (*disjoint allele pools*). 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 100 times and the number of rejections was recorded.

In Figure 1.5 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. The type I error inflation of BC, occasionally spotted on SLiM datasets, was here exharcerbated, especially with fewer loci and fewer individuals. As it will be further discussed in section 1.5.3, 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 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.

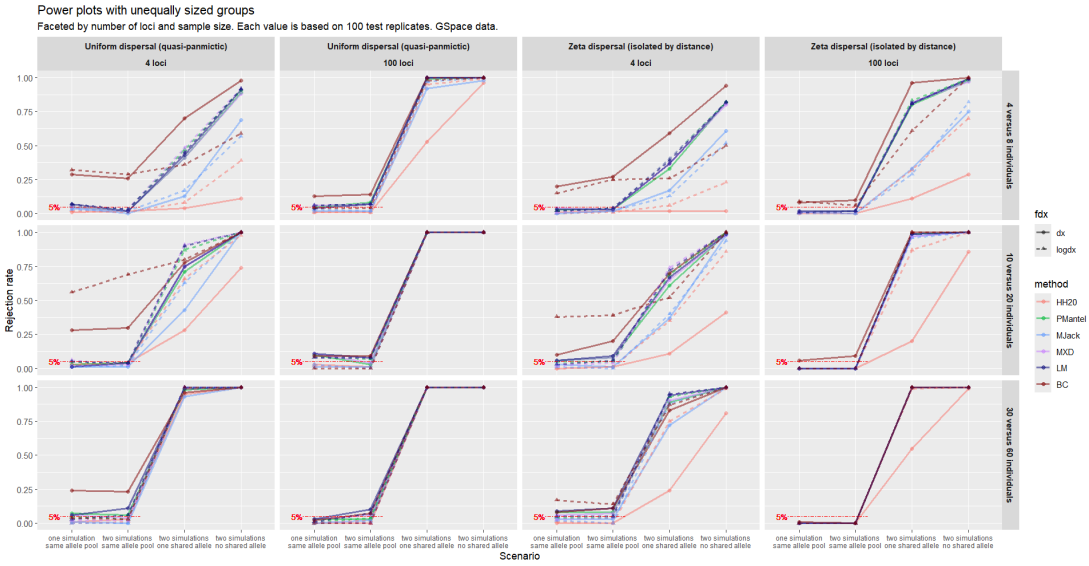

Fig. 1.6 Power plot based on **GSpace** data simulated **with unequal group sizes**. Panels by combination of IBD behaviour and number of loci (columns) and number of individuals per group (rows). See the caption to Figure 1.3 for further description.

Indeed, accounting for geographic 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 as evident with SLiM data because its *split* scenario did not involve geographically separate 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 the *overlapping allele pools* scenario, 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, geographic 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 1.6 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.

1.4.3 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 showing type I error rates consistently beneath the significance level, unlike PMantel. The bias-corrected bootstrap-based PMT showed too large type I error rate in the most challenging data setups. Additional insights on these three ways of assessing significance in partial Mantel tests can be found in section 1.5.3.

Despite wrongly assuming that all dissimilarities in the dataset are independent, OLS estimation in model (1.22) 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 can invalidate the inference on their relationship. More on this in section 1.5.2.

No method apart from HH20 clearly benefited from the logarithmic transformation of the geographic distances. Its performance 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. The investigation of scenarios where groups show different IBD patterns is left for future work.

The role of the assumption of linearity, made by all methodologies compared here, will be further explored in section 1.5, where also the issue of the dependence in the dissimilarities will be discussed. Moreover, additional insights on the use of permutations, jackknife and bootstrap in partial Mantel tests will be presented.

In the remainder of this section, parameter choices for the simulations above are discussed more in depth.

1.4.4 Parameter choices for simulations

The choice of parameter values for complex simulations like those employed in this study 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, L. C. 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 geographic 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. Although both GSpace and SLiM model evolutionary processes and return geographic coordinates and genotypic data, here they were used to study the performance of inferential methods that act on genetic and geographic dissimilarities - thus on a different analytic level. Therefore, the choice of simulation parameters was mainly guided by dissimilarity-based considerations, in connection with the hypotheses of interest. By visual inspection of distance-distance plots and Mantel correlograms, parameter values were selected in the light of their association with key data features, such as the enforcement of IBD behavior within the species, range separation in the dissimilarities within and between the groups, geographic segregation. Similar data visualizations were contrasted with those from compatible real data, i.e., with datasets where spatially-explicit co-dominant markers were available, such as those in Figures 1.1, 1.14 and 1.8. The estimation of some of the parameters described in this section from real data is left for future work.

In the following, after the description of correlograms, insights from the parameter explorations carried out for SLiM and GSpace are discussed.

Mantel correlograms Correlograms display measures of spatial correlation against geographic distance classes (Borcard, Gillet, and Legendre 2011, ch. 7) and can be used as diagnostics plots to detect isolation by distance in our setup. In Mantel correlograms (Borcard and P. Legendre 2012), the Pearson correlation between genetic dissimilarity and distance class indicators is plotted. Geographic distances are divided into distance classes, usually using Sturges' rule - which gives $\log_2(w)+1$ classes - where w is the number of dissimilarities. Given a distance class, two individuals are assigned class indicator 0 if they belong to the same distance class and 1 otherwise. If individuals closer in space tend to be more genetically similar than those geographically separate (a pattern like isolation by distance), we expect these correlations to be larger for the first distance classes and smaller for the last distance classes. Indeed, in that case, for the first distance classes, belonging to the same class (indicator equal to 0) means being close in space, so that individuals from distinct classes (indicator equal to 1) tend to

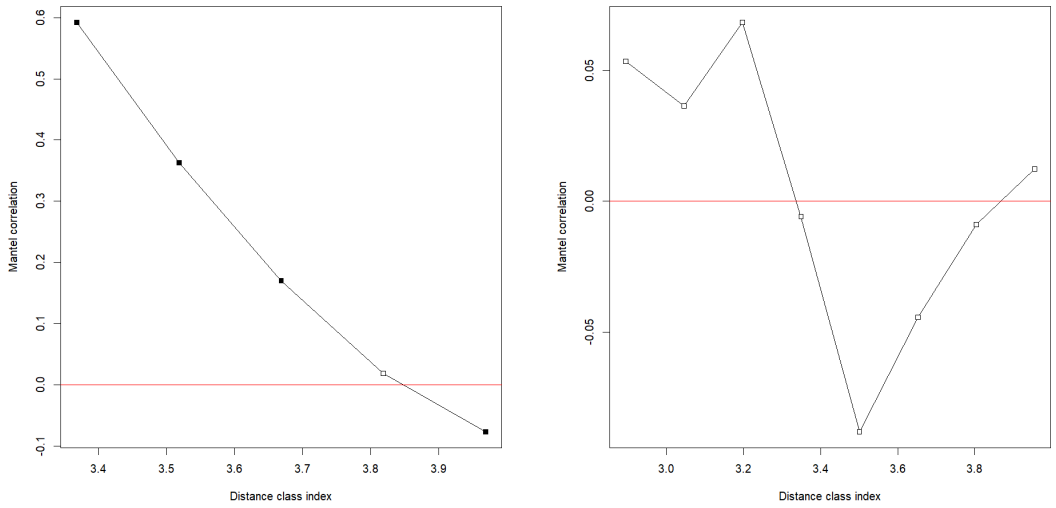

Fig. 1.7 Mantel correlograms for two exemplary SLiM datasets from the split scenario, with a total of 80 individuals and 40 loci. Based on log-transformed geographic distances. Species with IBD behaviour on the left, with quasi-panmictic behaviour on the right. Plot generated with the `vegan` package (Oksanen et al. 2022).

display larger genetic dissimilarities. Viceversa, under IBD, in the last distance classes an indicator equal to 1 will be associated with smaller genetic dissimilarities, leading to negative correlations. For each distance class, a simple Mantel test is carried out and Mantel correlation coefficients significantly different from zero are depicted in black on the chart. The path drawn by the significant coefficients informs the assessment of the spatial structure found in the genetic dissimilarities.

In Figure 1.7, Mantel correlograms are shown for two exemplary SLiM datasets from the split scenario. When parameter settings for IBD behaviour are used, a clear negative trend appears, with significant Mantel tests for most of the distance classes. With quasi-panmictic individuals, instead, the range of the Mantel correlations is narrower and no significant association between genetic dissimilarities and distance class indicators is spotted.

According to François and Waits (2015), also PCA applied on genotypic data can be used to detect the presence of IBD patterns. Under an equilibrium IBD model, a "horseshoe" shape in the scatterplot of the (first two) principal components should appear.

SLiM With respect to di-allelic models, nucleotide-based models in SLiM helped obtain datasets with a wider range of genetic dissimilarities, where the separation between within-group and between-groups dissimilarities, if present, was more pronounced. In these models, as explained in section 1.4.1, mutations were governed by the Jukes-

Cantor model with a unique transition rate equal to 0.0025, larger than the default. Too low rates would prevent sufficient genetic discrepancies between individuals to build up, resulting in very low values of the shared allele distance. Values larger than 0.0025, instead, would imply way longer execution times. Similarly to what happened with GSpace, mutations could mask IBD patterns also in SLiM: in IBD scenarios, given the time-scale of the simulation, large mutation rates would induce constant genetic dissimilarities, with very low variance. The impact of mutations clearly changes with the time-scale of the simulation: in this study the related parameters were chosen in such a way so to control execution time, avoid the masking of IBD patterns and minimize the number of monomorphic loci.

The recombination rate is associated with the independence of loci. A value equal to 0.5 enforces independent loci, but requires longer execution time. In IBD scenarios, this rate would induce clearer linearity in the increase of genetic dissimilarities up to the level of geographic separation where they plateau; with the default rate, a more convex shape is spotted in this area of the distance-distance plot. In quasi-panmictic scenarios, independent loci led to more concentrated genetic dissimilarities - see section 1.7.2.

The number of simulated individuals obviously interacted with IBD patterns, because of the constrained habitat. More individuals fill the map better, facilitating gene flow: this would dampen IBD patterns in scenarios where spatial parameters were set to enforce them, whereas panmictic-like behaviour would be intensified. Given the number of *simulated* individuals, a larger number of *sampled* individuals would make IBD patterns more apparent, also on correlograms.

A sufficiently large number of generations before data output was needed in order for mutations to build up, as explained above. With too short simulations, IBD patterns would not be visible. While, in scenarios with one species, longer simulations would not result in any dramatic change in IBD patterns and shape of genetic distance distribution, in scenarios with two species it was crucial to keep the number of generations as low as 50. Simulations with IBD behavior and 200 generations, for instance, would yield datasets in which the shared allele distances between co-specific individuals would plateau at the same level observed for between-groups dissimilarities. In quasi-panmictic scenarios with same time-scale, the range of genetic dissimilarities would be very similar within and between the groups. This is connected to the fact that the four nucleobases were directly used as allelic stata for the diploid loci in these abstract simulations, i.e., no allelic status was exclusive to some species. Therefore, similarly to what happened in the *same allele pool* scenario with GSpace, even unrelated species would never end up displaying maximum shared allele distance values, on average. Datasets where this happened were thus considered incompatible with the alternative hypothesis of distinctness for the sake of this simulation study, because the positive

shift in genetic dissimilarities that can be expected to appear when individuals from distinct species are compared would be masked. Another possible workaround, which is left for future work, is to specify species-dependent matrices of mutation rates in the Jukes-Cantor model, so that the genetic material in the two species might converge to different allele frequencies in the long run. However, keeping the number of generations low ensured that the desired structure in the genetic dissimilarities was obtained while minimizing the execution time of the simulations.

As far as parameters related to IBD patterns are concerned, as explained in section 1.4.1, the most relevant ones were the mean and standard deviation of the Gaussian kernel used to translate geographic distance into interaction strength (spatial competition) and the standard deviation of the Gaussian kernel used for offspring positioning. The choice of their values in the IBD and quasi-panmictic scenarios was based on a thorough parameter exploration. Experiments showed how offspring positioning was more impactful than spatial competition in enforcing IBD patterns in the species, while the number of nearest neighbors from which to choose a mate and the maximum distance to find one were less relevant. In IBD scenarios, larger values of these two parameters weakened the positive association between genetic and geographic dissimilarities, generating less evident plateaus. This translated in milder slopes in the correlograms.

GSpace Since the most recent common ancestor is found for every individual in the sample, individuals from the same GSpace execution can hardly be seen as belonging to distinct species. As a consequence, in scenarios not under the null hypothesis of conspecificity, the two groups were generated during different software executions. The allele pools in the two executions would either consist of the same two allelic stata (e.g., “A” and “B”), or share only one of them (e.g., “B” and “C” in the second group) or be completely disjoint (e.g., “C” and “D” in the second group). Theoretically, a common ancestor can be compatible with distinctness, since two separate lineages might originate from it due to a range of evolutionary triggers. However, in order to mimic such a phenomenon, we would have to simulate the process of speciation over time, which is not possible in GSpace - since one cannot explicitly manage the evolution generation by generation as in purely forward-time simulators.

GSpace can simulate allelic and DNA sequence data, i.e., supports both allelic and nucleotidic mutation models. In this study, an allelic model was used, which allows for whatever number of allelic stata (different alleles in loci). A larger number of possible allele options implies larger genetic dissimilarities on average, as will be discussed in sections 1.5.1 and 1.7.1. A longer sequence of loci available for the computation of the genetic distance, instead, leads to a more concentrated distribution thereof.

The IBD behaviour in GSpace is obviously influenced by dispersal options - such as the shape parameter κ for the Pareto distribution, but also by total emigration rate, edge effects and maximum dispersal distance. When κ equals 1, individuals are quasi-panmictic, whereas values larger than 5 induce strong isolation by distance patterns. Smaller total emigration rates make the execution slower and reduce the effect of the Pareto shape parameter. To limit the impact of edge effects, map boundaries were set far from the area where individuals are positioned. Maximum dispersal distance, instead, was set to the maximum so to more directly regulate IBD behaviour via dispersal distribution choices.

On top of these parameters, also the mutation rate has a non negligible impact on isolation by distance. A rate of 50% translates in a narrower range of genetic dissimilarities, probably because mutations tend to modify the genetic make-up more than dispersal distributions can tell. With very low rates, such as 5^{-6}, more and more zero genetic dissimilarities occur, the reason being that certain individuals never change their allelic stata and stay identical to their original genetic make-up - which is random but shared by all individuals generated under the same software execution. The rate chosen for the simulations thus constituted a nice compromise in order for the dispersal options to be the main drivers of the desired dissimilarity trends.

In GSpace, simulating one individual in each deme (geographic location) allows to approximate settings where individuals from a population inhabit a continuous habitat. However, original applications of GSpace and IBDSim (Leblois et al. 2009) are related to population-level analyses, where a community of individuals is generated at each deme. When more than one individual is simulated at a deme but only one of them is sampled and used for the analysis, data generation takes much more time and the impact of dispersal parameters is dampened. All in all, this simulation setting seems incompatible with the goal of reproducing continuous dispersal across the map.

1.5 Modelling issues

This section offers additional insights on the dissimilarity-based methods introduced in section 1.3, whose type I error and power properties were examined on simulated data in section 1.4.

Section 1.5.1 expands on the assumption of linearity in the relationship between genetic and geographic distances, with insights from real and simulated data. An ad hoc simulator is introduced to show in a simplified setup how the relationship between shared allele distance and geographic distance evolves depending on the number of loci, allelic stata and on the strength of IBD patterns. The behaviour of the single-locus

shared allele distance - defined in (1.26) - is also studied in relation to more abstract quantities than the geographic distance, such as a measure of the difference between the allele frequencies in the two compared loci.

Section 1.5.2 discusses how dissimilarity-based methods are affected by the dependence in the dissimilarities. The role of cluster imbalance (ratio between n_1 and n_2) on type I error rates is further illustrated, also using Gaussian data. This stresses that the relevance of these results goes beyond the specific biogeographic application of this work, as the regression between distances may be of interest in its own right.

Section 1.5.3 further explores the use of permutations, jackknife and bootstrap in partial Mantel tests. By visualizing the distribution of the partial correlation replicates they generate under the null hypothesis of conspecificity (as it was operazionalized in previous sections), some type I error and power properties of these techniques are clarified.

1.5.1 The assumption of linearity

Methods in section 1.3 assume that the genetic dissimilarities vary linearly with the (log-transformed) geographic distances. As the literature cited in section 1.1 illustrates, an extensive list of factors can affect the relationship between genetic and geographic dissimilarities, making the assumption of linearity problematic. This list includes:

- whether the analysis is carried out at population level, where communities of individuals are compared, or at the level of the single specimen, as happens in this study;
- the type of the genetic marker extracted, such as AFLPs, microsatellites, SNPs, etc. (e.g., Gaudeul et al. 2004).
- the specific measure of genetic dissimilarity employed - e.g., see the manual to the SPAGeDi software (Hardy and Vekemans 2002);
- the choice and dimensionality of the geographic distance, which is connected to the adopted landscape model (Shirk et al. 2018). In section 1.4, patterns of isolation by distance were analyzed using a Euclidean distance based on abstract two-dimensional coordinates. A geodesic distance could be used in real data applications, where one-dimensional setups (e.g., a river or a mountain ridge) and three-dimensional setups like oceans can also be of interest. In isolation by resistance frameworks, more sophisticated landscape dissimilarities are usually employed. In principle, methods described in this work are compatible with any of these alternative choices.

- any transformation applied on either d_y or d_x (or both), which can be based on ratios (e.g., Rousset 1997), logarithms (e.g., Vekemans and Hardy 2004) or square roots (e.g., P. Legendre, Fortin, et al. 2015), etc.

- other factors, such as those mentioned at the beginning of section 1.4.

The shared allele distance takes value in the interval $(0, 1)$, thus linear predictions can in principle fall outside of the admitted range, especially if large genetic differences are frequent within a species. In these scenarios, the tests described in section 1.3 might show low power, as genetic discrepancies between groups can hardly be higher than what is observed within them (Hausdorf and Hennig 2020). More in general, the combination of strong IBD behaviour and sufficient geographic separation can jeopardize similar tests, since even maximum genetic dissimilarities between distinct species will be explainable according to linear predictions. The natural upper boundary of the shared allele distance can induce a convex trend in its relationship with geographic distance, but similar trends arise also when different evolutionary factors act at different spatial scales (Hutchison and Templeton 1999).

In this section, examples from real and simulated data are reported where these characteristics of the relationship between the shared allele distance and the (log-transformed) Euclidean distance are highlighted. Despite their simplicity, linear models can still capture the association between genetic and grouping dissimilarities, controlling for geographic separation - as the simulation study in section 1.4 showed. Indeed, null hypotheses of zero linear correlation or regression slope can be rejected even if the relationship is nonlinear but monotonic. Implementations of the methods in section 1.3 based on polynomial regression or rank correlation coefficients are left for future work.

Insights from real data Geo-referenced SNP data from 21 Albinaria virginea specimens sampled in Crete in 2018 was obtained thanks to Prof. Bernhard Hausdorf (Bamberger et al. 2021). These individuals come from 9 different localities, thus may also be analyzed at population-level. The relationship between shared allele distance and both untransformed and log-transformed geodesic distances is shown in Figure 1.8, where we spot a positive association. However, individuals from the same locality - obviously at zero geographic distance - display lower genetic dissimilarities than the overall linear trend can predict. This is only partially mitigated when log-transformed geodesic distances are used. Anyway, as far as individuals with positive geographic distance are concerned, the OLS fit works fairly good.

As regards the data from Gratton et al. (2016), described in section 1.6.1, the related dissimilarities are reported in Figures 1.1 and 1.14. The linearity of the within-group relationship between genetic and log-geographic dissimilarities can be assessed by

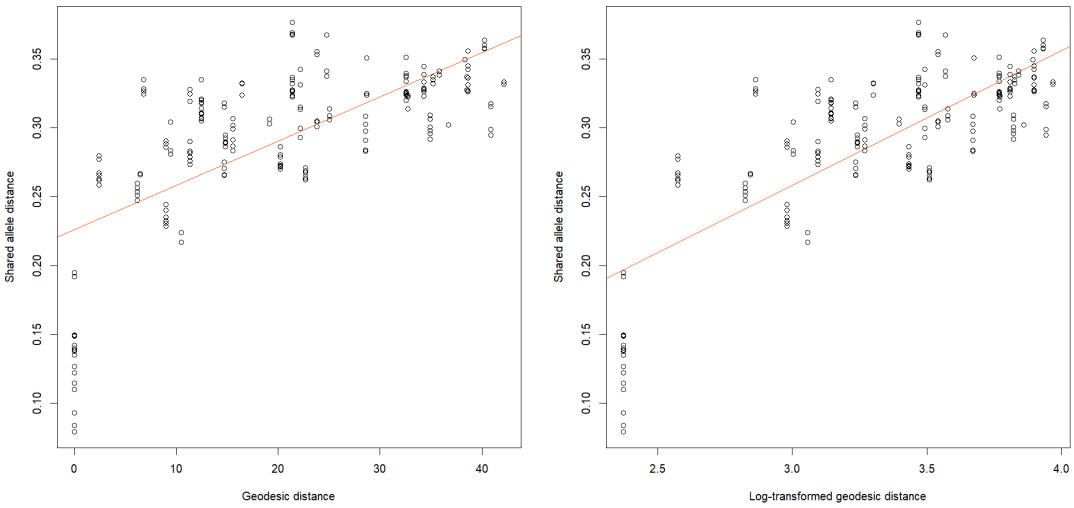

Fig. 1.8 Distance-distance plot based on 21 Albinaria virginea specimens (Bamberger et al. 2021). The applied log-transformation is that in (1.3).

looking at black and red points. While specimens from the Nivalis subgroup were only sampled at two different locations, there is enough data from the Tyndarus and Cassioides subgroups. The latter shows a fairly linear trend, although also here individuals observed at the same geographic location are slightly outlying. The Cassioides group is quite heterogeneous, resulting in a heteroskedastic trend, which however does not depart remarkably from linearity. Heteroskedasticity is partly mitigated when subpopulations of the Cassioides group are considered in isolation.

Insights from simulations As it can be seen in many of the distance-distance plots in Appendix A.1, isolated by distance species simulated with SLiM and GSpace often showed a nonlinear trend in the relationship between genetic and geographic dissimilarities, which was not mitigated by the log-transformation of the latter. The observed convex trend resembled the fourth relationship hypothesized by Hutchison and Templeton (1999). Nevertheless, as explained above, the violation of this assumption apparently did not deteriorate the performance of the species delimitation methods in section 1.3 in terms of type I error rate and power.

An ad hoc simulator can be constructed whose parameters more directly influence the relationship between genetic and geographic distances. Without making reference to any evolutionary process, this simulator models the allele frequencies in the loci of the individuals based on their geographic location. A single parameter, ι, drives spatial patterns of differentiation. Thanks to these features, this algorithm helps study in a simplified setup how factors like the number of loci, the number of possible allelic stata

and IBD patterns are related to the (lack of) linearity of the relationship between d_y and d_x.

This simulator generates two-dimensional coordinates and P diploid loci by associating to each locus a spatial linear gradient. Assuming that M allelic stata exist for every locus, the probability to observe each of them at locus p will depend on the position of the individual relative to the gradient associated with that locus. By imposing various allele frequency distributions along the gradients, a range of different IBD behaviours can be recreated.

More precisely, the algorithm to simulate n individuals and P loci, all with M allelic stata, is the following:

i. uniformly sample n two-dimensional coordinates on a user-specified rectangular area;

ii. set P gradients with random directions: the p^{th} gradient can be visualized as a line, passing through the center of the two-dimensional map, along which the probability to observe the M allelic stata in the p^{th} locus varies. M equidistant points are found on each of these lines;

iii. for $p = 1,\ldots,P$, the i^{th} individual is assigned to the closest of the M points identified on the p^{th} line, according to the Euclidean distance. In Figure A.17, this is illustrated for six loci. This is formalized with an assignment indicator $c_i^p \in \{1,\ldots,M\}$;

iv. for $m = 1,\ldots,M$ the probability vector $\mathbf{q}^m = \{q_{m'}^m\}_{m'}$ is defined, where

$$q_{m'}^m = \left(\sum_{m'=1}^{M} \frac{1}{\iota^{|m'-m|}} \right)^{-1} \frac{1}{\iota^{|m'-m|}}, \tag{1.23}$$

which satisfies $\sum_{m'} q_{m'}^m = 1, \forall m$. Note that, for $m = 1,\ldots,M$, the vector $\mathbf{q}^m$ is the same across the P loci: the position of an individual with respect to the gradient informs the distribution of allelic stata along that gradient in the same way for all loci. Value of the user-defined parameter ι larger than 1 enforce stronger IBD patterns in the simulated data. Indeed, given that $c_i^p = m$, the larger ι, the more likely it is for individual i to display the m^{th} allelic status at locus p. Consequently, genetic discrepancies between individuals will grow with their geographic separation, depending on the number of loci and allelic stata;

v. the two allelic stata at the p^{th} diploid locus of the i^{th} individual are drawn according to the multinomial distribution specified in (1.27), using the probability vector $\mathbf{q}^{c_i^p}$, for $i = 1,\ldots,n$ and $p = 1,\ldots,P$.

In Figures A.18, A.19 and A.20, distance-distance plots based on data simulated with this algorithm are reported. As anticipated, $\iota > 1$ induces a positive association between the distances, while $\iota = 1$ yields datasets with no IBD structure. Larger values of P generate more concentrated trends, whereas more allelic stata increase the average value of genetic dissimilarities, ceteris paribus. Remarkably, it is evident how dissimilarities preserve a linear trend as long as the upper boundary of the shared allele distance is far from their bulk: when $M = \iota = 10$ the combination of strong spatial patterns of differentiation and the overall high level of genetic diversity generates extremely convex trends, that can be barely mitigated via the logarithmic transformation of the geographic dissimilarities. On the contrary, in di-allelic simulations ($M = 2$), this transformation applied on a linear trend sometimes even induces concavity in the relationship between d_y and d_x. On a side note, when $\iota = 10$ and $P = 5$, all $2P+1$ values the shared allele distance can take - as prescribed in (1.31) - are observed.

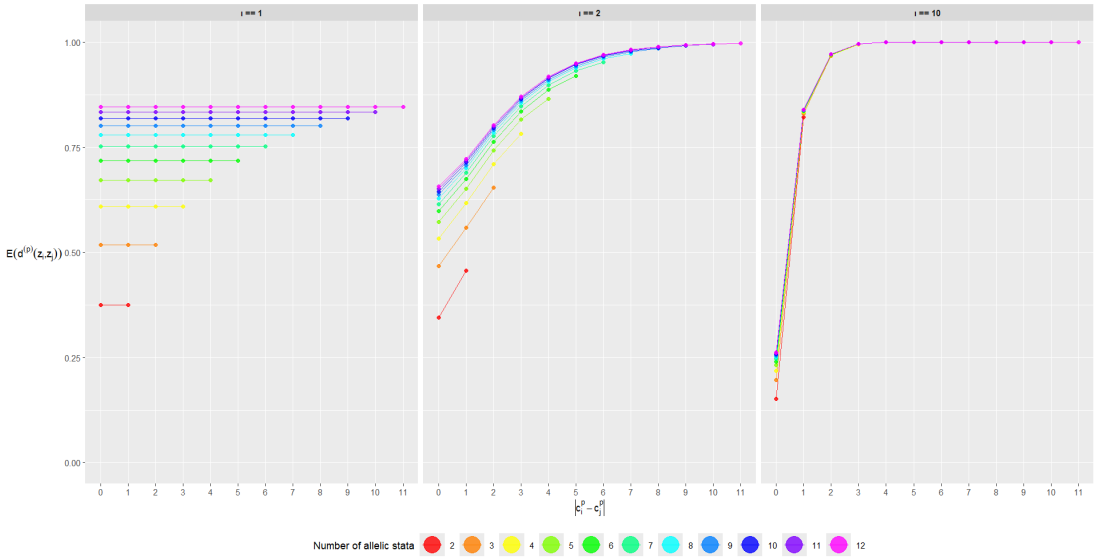

Fig. 1.9 Single-locus expected shared allele distance between two individuals as a function of the absolute value of the lag between the categories c_i^p and c_j^p to which they were assigned for locus p. Panels by value of ι, colours by number of possible allelic stata.

To further characterize how these factors drive the behaviour of d_y, the generative model underlying the algorithm above can be used to study the trend in the expectation of the single-locus shared allele distance, obtained in (1.34). This is simply the shared allele distance computed with the information in one locus only. In Figure 1.9, the value of $\mathbb{E}\left(d^{(p)}(\mathbf{z}_i, \mathbf{z}_j)\right)$ is reported as a function of $|c_i^p - c_j^p|$, the absolute value of the difference between the labels of the categories to which individuals i and individual j are assigned. The range of this lag is $\{0, \ldots, M-1\}$. Given the lag, the differences between vectors $\mathbf{q}_i^{c_i^p}$ and $\mathbf{q}_j^{c_j^p}$, on which $d^{(p)}$ is based, vary according to the number M

of allelic stata and the value of parameter ι. As before, we observe that larger M given ι are associated with larger genetic discrepancies between individuals: larger allele pools make it easier for individuals to have different genetic make-up at any locus. If M is fixed, larger values of ι induce a positive association between lag and expected single-locus genetic distance, and when the latter approaches its upper boundary, convexity kicks in. Note that larger ι's also yield more similar individuals than $\iota = 1$ at zero lag, because the probability to display the same allelic state is larger.

Since M equidistant points are found on the p^{th} linear gradient, the Euclidean distance between them depends on M itself and on the height and width of the rectangular area on which individuals are uniformly positioned - see Figure A.17. Hence, the connection between the lags in Figure 1.9 and the geographic distance between individuals is indirect. Nonetheless, this visualization offers a general picture of how the single-locus shared allele distance grows with geographic separation assuming that genetic variability follows a linear gradient. Results suggest that linearity only applies for low values of M and ι. However, since the overall shared allele distance is an average of P such distances, also from Figures A.18, A.19 and A.20 one can see that a linear prediction will still constitute an acceptable approximation in a fairly wide range of scenarios.

At a higher level of abstraction, a similar investigation can be carried out by comparing the value of $\mathbb{E}\left(d^{(p)}(\mathbf{z}_i, \mathbf{z}_j)\right)$ with a measure of probability mismatch between the vectors $\mathbf{q}_i^{(p)}$ and $\mathbf{q}_j^{(p)}$. Note that now $\mathbf{q}^{(p)}$ is used to indicate a generic vector of allele probabilities at locus p and superscripts without brackets are used for powers. Indeed, this investigation is not related to the simulator described above, and concerns the way the shared allele distance between two loci is expected to vary depending on the differences in their allele distributions. An option to measure this probability mismatch is the Manhattan distance $d_q(\mathbf{q}_i^{(p)}, \mathbf{q}_j^{(p)}) = \sum_{m=1}^{M} |q_{im}^{(p)} - q_{jm}^{(p)}|$, similarly to what will be done for the admixture dissimilarities in (2.6). In Figure A.21, the outcome of a Monte Carlo experiment with 10000 replicates is displayed for $M \in \{2, 4, 10\}$. Each row in $\mathbf{q}_i^{(p)}$ was generated in two steps: first, M values were drawn uniformly from the unit interval; then, these values were used as parameters in a Dirichlet draw, so to generate M probability values that sum to 1. More in detail, given $\{u_m \sim \text{Unif}(0,1)\}_{m=1,\dots,M}$, $\mathbf{q}_i^{(p)}$ is sampled from a distribution with pdf:

$$f(q_1, \dots, q_M, u_1, \dots, u_M) = \frac{1}{B(\{u_m\}_{m=1,\dots,M})} \prod_{m=1}^{M} q_m^{u_m}, \tag{1.24}$$

where each q_m lies between 0 and 1, $\sum_{m=1}^{M} q_m = 1$ and the normalizing constant is the beta function:

$$B(\{u_m\}_{m=1,\dots,M}) = \frac{\prod_{m=1}^{M} \Gamma(u_m)}{\Gamma(\sum_{m=1}^{M} u_m)},$$

with Γ being the gamma function, that extends the factorial function to complex numbers. The same is done for $\mathbf{q}_j^{(p)}$, and $d_q(\mathbf{q}_i^{(p)}, \mathbf{q}_j^{(p)})$ is computed. Then, the probability that the shared allele distance between a locus sampled with allele distribution $\mathbf{q}_i^{(p)}$ and a locus sampled with allele distribution $\mathbf{q}_j^{(p)}$ takes value 0.5 or 1 is obtained from (1.29) and (1.30), respectively. As before, the value of the expected single-locus shared allele distance follows from (1.34).

The charts in Figure A.21 shows that the relationship between these two quantities highly depends on M, but is always monotonically non-decreasing. $\mathbb{E}\left(d^{(p)}\right)$ appears to be bounded from below by $d_q(\mathbf{q}_i^{(p)}, \mathbf{q}_j^{(p)})/2$, and an upper boundary that changes with M can also be spotted. Increasing values of M tend to generate more values of d_q close to its theoretical upper boundary, equal to 2. The variability of $\mathbb{E}\left(d^{(p)}\right)$, instead, decreases with d_q, but this decrease gets milder as M grows.

The relationship between $d_q(\mathbf{q}_i^{(p)}, \mathbf{q}_j^{(p)})$ and geographic distance can be governed by any kind of spatial pattern, thus it is hard to relate the trend observed in Figure A.21 with that between genetic and geographic distance. However, this experiment illustrates how the shared allele distance is expected to vary as the unrelatedness between two genetic make-ups grows, regardless of what drives it.

Too far to be distinct As anticipated at the beginning of this section, dissimilarity-based methods that assume linearity are at risk of losing power if the separation between two IBD groups is too large relative to their within group spatial pattern of genetic differentiation. Indeed, if the slope in the regression of genetic dissimilarities over geographic ones is positive, there will be a value of d_x such that even maximum shared allele distances will be explained by isolation by distance. This poses an identifiability issue, since if maximum values of d_y can be explained already within a species, it is not possible to tell apart one species from two species.

To illustrate this, additional SLiM simulations from the *separate distinct* scenario reported in the Figure 1.4 were carried out, where the two groups have 30 and 60 individuals each and 89 loci are available for the computation of the shared allele distance (bottom-right corner of the chart). In Figure 1.10, the location labelled "B" is the same as in section 1.4.1, while, for the sake of this experiment, 100 datasets where the second group is located either in "C", "D", "E" or "F" were genereted. One can see how the raising geographic separation of the groups annihilates the power of all methods, with the exception of the bootstrap-based partial Mantel test, whose

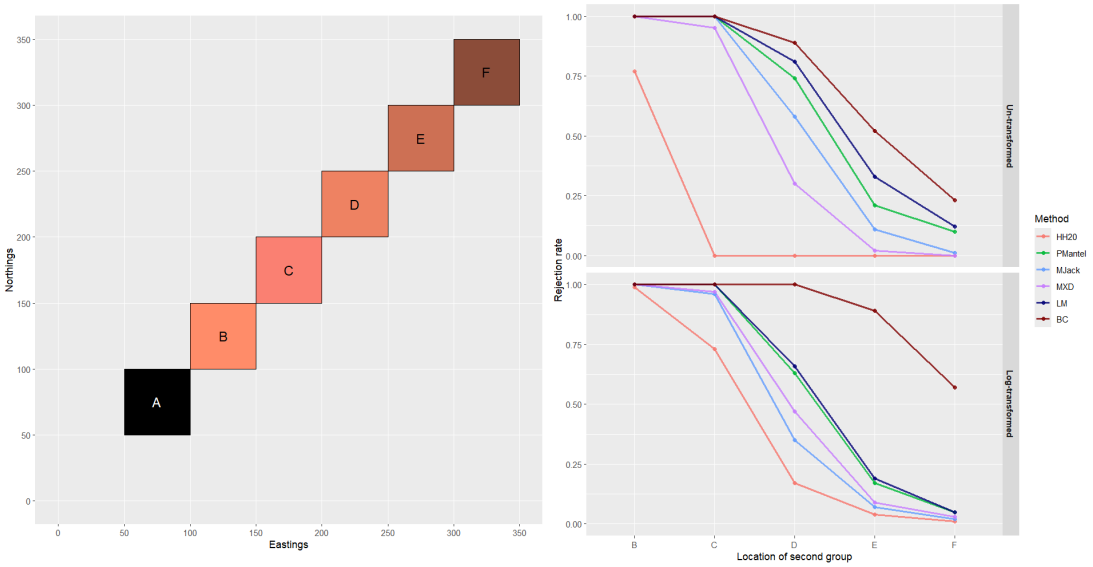

Fig. 1.10 SLiM simulation on the effect of group separation. Left: map showing the location of the first group (A) and five possible locations of the second group: in Figure 1.4, the second group in the *separate distinct* scenario (two species, disjoint areas) was located in B. For the sake of this analysis, 100 datasets with IBD species, large sample sizes and 89 loci for each of the second group locations from C to F were generated. Right: empirical rejection rates of the methods described in section 1.3, with panels by transformation of geographic distances.

already mentioned flaws will be further described in section 1.5.3. In general, the decrease is faster when d_x is log-transformed, because this widens the range of within-group geographic dissimilarities with respect to between-groups ones, leading to larger estimates for the slope and lower (even negative) intercept updates for between-group dissimilarities. The ranking of methods stays the same, although MJack deteriorates faster than MXD when a log-transformation of geographic distances is applied.

The same analysis on scenarios with quasi-panmictic groups (not reported) did not result in a power loss, which is to be expected given that there is no relationship in the dissimilarities within the groups - thus isolation by distance can never explain genetic discrepancies between the groups. This experiment stresses how species delimitation tests based on dissimilarities are sensitive to the scale of the study and have to be interpreted in the light of other factors, such as the level of saturation in the within-group dissimilarities. For instance, microsatellite data tend to generate large shared allele distances already within the groups, unlike SNP data. Furthermore, Figure 1.10 highlights what difficulties can arise when engineering a simulation study like the one in section 1.4, since parameter choices have to yield datasets whose geographic scale is compatible with the spatial patterns of differentiation induced in the species. Only on this condition, the analysis of type I error rate and power of the inferential tools will be reliable.

1.5.2 Dependence

As explained previously in this chapter, in dissimilarity-based methods the assumption of independence can be hardly defended. The $n-1$ dissimilarities involving individual i, with $i = 1,\ldots,n$, will be dependent by construction, because they are all based on the information from the i^{th} statistical unit in the sample. A way to illustrate this is to construct an ad hoc experiment with Gaussian data. Suppose three IID observations are sampled from a P-dimensional Multivariate Normal distribution a large number of times. Three Euclidean distances can be computed between them, and any pair of them will have one observation in common. In a simulation where 1000 such pairs were generated with P ranging from 1 to 20, they showed a Pearson correlation between 0.2 and 0.3, regardless of the value of P. This obviously did not happen when four observations were sampled and the correlation between pairs of L2 distances not sharing any observation was computed: in this case the paired distances were unrelated.

The linear regression in (1.22) ignores this association and assumes that the w dissimilarities can be seen as independent units. On the contrary, the jackknife takes into consideration the relationship between the statistical unit i and its related dissimilarities, dropping them altogether for the computation of pseudovalues. The linear mixed model in (1.21) resorts to random intercepts to describe how genetic dissimilarities involving the i^{th} unit tend to deviate from their average at a given geographic distance. As pointed out in section 1.3.3, this can only model the dependence structure due to pairwise relationships, while more complex interactions between more than two individuals are not captured. The triangle inequality, for instance, establishes that $d_y(\mathbf{z}_i,\mathbf{z}_j) \leq d_y(\mathbf{z}_i,\mathbf{z}_k) + d_y(\mathbf{z}_j,\mathbf{z}_k)$, $\forall k \neq i,j$ - see the proof in Appendix B. In principle, this upper boundary is not guaranteed by the linear mixed model. Experiments on SLiM data (not reported) show that there is no significant relationship between each $d_y(\mathbf{z}_i,\mathbf{z}_j)$ and the sums $d_y(\mathbf{z}_i,\mathbf{z}_k) + d_y(\mathbf{z}_j,\mathbf{z}_k)$, $\forall k \neq i,j$, that bound it from above, although a positive association can be spotted especially in quasi-panmictic scenarios. All things considered, the generative model in section 1.3.3, which can even yield negative dissimilarities, is not fully compatible with the nature of dissimilarities.

Permutations and bootstrap, being non-parametric, circumvent these issues. However, when used for assessing the significance in partial Mantel tests (section 1.3.2), they might introduce a distortion in the distribution of the partial correlation coefficient under the null hypothesis that its true value in the population is zero. Not only permuting rows and columns of $\mathbf{D}_y$ leads to zero partial correlation, it also removes any association between $\mathbf{D}_y$ and both $\mathbf{D}_g$ and $\mathbf{D}_x$, which might cause type I error inflation. As regards the bootstrap, by resampling individuals with replacement, it generates datasets under the null hypothesis that often have more zero dissimilarities than the original ones.

The methodologies in 1.3, each with its own limitations, attempt to take the dependence in the dissimilarities into account by making a diverse set of assumptions. Despite the limitations, results in section 1.4 showed that only bootstrap-based PMTs were consistently off in terms of type I error - see section 1.5.3. The LM in (1.22), which neglects dependence, displayed inflated type I error rates only in SLiM simulations with unequally-sized groups. The remainder of this section delves into the drivers of this behaviour.

Recall that, although they deal with the dependence structure in the dissimilarities, these methodologies assume that the n individuals in the dataset are independent, which is obviously irrealistic when individuals belong to the same species. This is a limitation of individual-level analyses with respect to population-level ones, whose investigation is left for future work.

Experiments with Gaussian data The behaviour of the linear regression model neglecting dependence can be investigated using Gaussian data in order to clarify whether results displayed in section 1.4 are reproduced out of that specific data scenario. In this experiment, for each individual $i = 1, \dots, n$, vectors $\mathbf{X}_i = \{X_{ip}\}_p$, $\mathbf{Y}_i = \{Y_{ip}\}_p \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_P)$ are sampled and used as raw data for the computation of distances. Given individuals i and j, Euclidean distances $d_X(i,j) = \sqrt{\sum_{p=1}^{P}(X_{ip} - X_{jp})^2}$ and $d_Y(i,j) = \sqrt{\sum_{p=1}^{P}(Y_{ip} - Y_{jp})^2}$ are obtained, which will be respectively used as d_x and d_y were used in previous sections. Grouping information is then introduced by allocating the n individuals to two groups, whose sizes are either $n_1 = n_2$ or $n_2 = 2n_1$. With data labelled accordingly, a grouping distance d_g is computed as explained in section 1.2.

A simple linear regression can be fitted:

$$d_Y(i,j) = b_0^\dagger + b_1^\dagger d_X(i,j) + \varepsilon(i,j) \quad i \neq j\,, \tag{1.25}$$

where $\varepsilon(i,j)$ is assumed to be IID normally distributed with zero mean. The null hypothesis of interest is that $b_1^\dagger = 0$, tested against $b_1^\dagger > 0$ using standard regression t-tests. The presence of positive association between d_Y and d_X can be also assessed by testing the null hypothesis that their Pearson correlation is not larger than zero. This can be done either via a simple Mantel test, as available in packages `vegan` (Oksanen et al. 2022) or `ecodist` (Goslee and Urban 2007), or via jackknife resampling, similarly to what is done in section 1.3.2 for the partial correlation coefficient.

Testing for the presence of positive correlation between d_Y and d_g while controlling for d_X is also of interest. Hence, the model in (1.22) can be fitted replacing d_y with d_Y and d_x with d_X, and the null hypothesis that $b_2^* = 0$ can be tested against $b_2^* > 0$. This

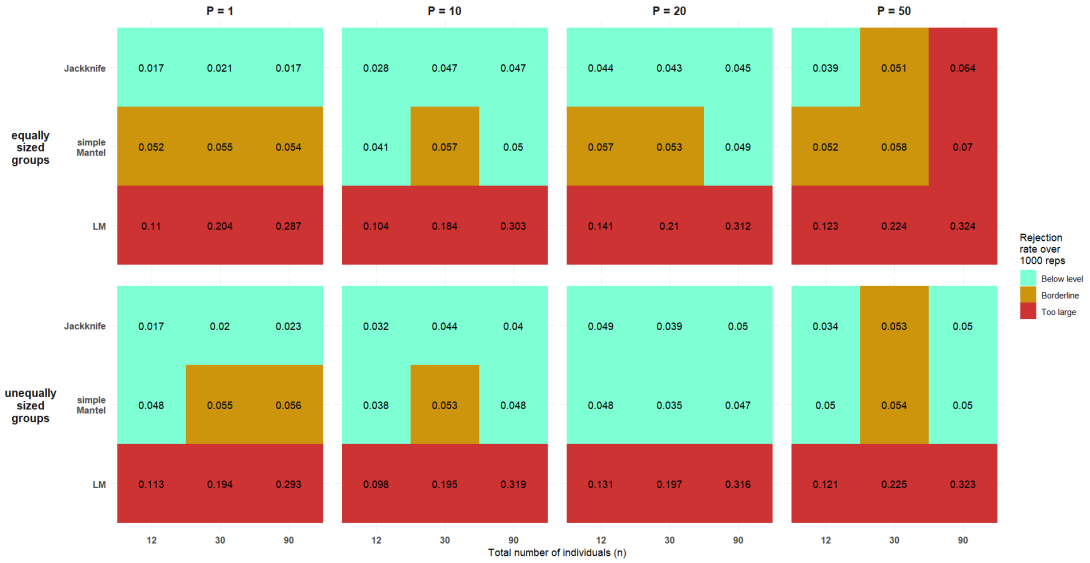

Fig. 1.11 Rejection rate out of 1000 tests on distances d_Y and d_X based on n P-variate normally distributed coordinates $\{\mathbf{X}_i\}_{i=1,\dots,n}$ and $\{\mathbf{Y}_i\}_{i=1,\dots,n}$. Jackknife and simple Mantel are tests on Pearson correlation, LM tests $b_1^\dagger = 0$ in (1.25). In the top panel, each group is made up of $n/2$ individuals, whereas in the bottom panel $n_1 = n/3$ and $n_2 = 2n_1$. Vertical panels by value of P. Cells are coloured in green for values equal to or lower than 0.05, in gold for values between $(0.05, 0.062)$ and in red for values above 0.062. The threshold 0.062 is chosen because $\mathbb{P}(B \geq 62) < 0.05$, where B is distributed as a Binomial RV with size 1000 and success probability 0.05.

can be compared with the permutation-based and jackknife-based partial Mantel tests described in section 1.3.2.

The significance level adopted for all these tests is 5%.

In Figure 1.11, results for jackknife and simple Mantel tests on $r(d_Y, d_X)$ and standard regression t-tests on the OLS estimate $\hat{b}_1^\dagger$ are reported for settings with equally and unequally sized groups, three total sample sizes n and four values of P. Each combination of these settings was used to generate 1000 datasets and empirical rejection rates of the three tests were recorded.

As explained above, the data in $\{\mathbf{X}_i\}_{i=1,\dots,n}$ and $\{\mathbf{Y}_i\}_{i=1,\dots,n}$ was generated independently, therefore d_Y and d_X can be expected to show no association. Despite this, standard linear regression t-tests rejected the hypothesis that $b_1^\dagger = 0$ too often with respect to the adopted significance level, with rates that rose with n. This is not surprising, since LM assumes IID errors, neglecting the dependence structure in the dissimilarities. On the contrary, jackknife and simple Mantel tests that take this dependence into account showed empirical rejection rates close to 5% - with the exception of the scenario with $n_1 = n_2 = 45$ and $P = 50$.

In Figure 1.12, results for MJack and PMantel on $r(d_Y, d_g | d_X)$ and standard regression t-tests on the OLS estimate $\hat{b}_2^*$ are reported for the same scenarios of Figure 1.11.

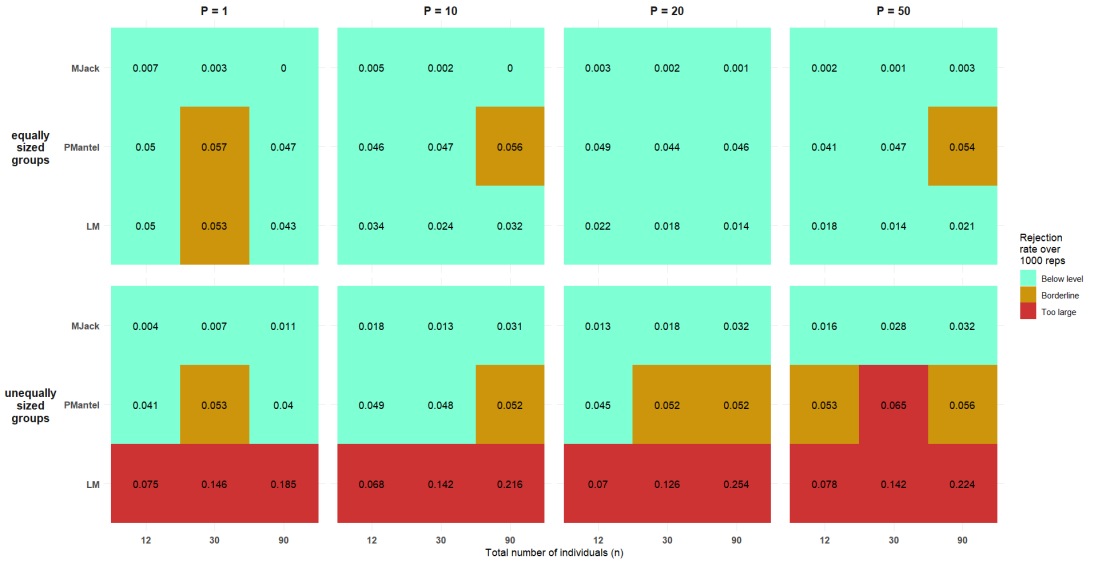

Fig. 1.12 Rejection rate out of one thousand tests on distances d_Y, d_X and d_g based on n P-variate normally distributed coordinates $\{\mathbf{X}_i\}_{i=1,\dots,n}$ and $\{\mathbf{Y}_i\}_{i=1,\dots,n}$, grouped as explained in the text. Method names as in section 1.4. See Figure 1.11 for further description.

The data generating process in this section is also compatible with the hypothesis that d_Y and d_g are not associated even after controlling for d_X, thus these empirical rejection rates help assess the type I error rate of the compared tests.

Consistently with results in section 1.4, rejection rates of jackknife-based partial Mantel tests were very low in all scenarios, whereas those from permutation-based PMTs were often borderline (significantly above 5% in one case). Also here, LM showed nice type I error properties when the groups were equally sized, whereas type I error inflation occurred with unequally sized groups. Rejection rates rose with n, showing that neglecting the dependence in the dissimilarities leads to the underestimation of the standard error of $\hat{b}_2^*$. The value of P did not seem to have a systematic impact on the tests.

The fact that cluster imbalance drives the type I error of the LM in (1.22) is thus confirmed for Gaussian data, too. This behaviour also applied to SLiM simulations, while it was not as apparent on GSpace data. Additional experiments on GSpace data (not reported) showed that this was not due to the fact that the two groups were geographically segregated under the *split* scenario, unlike in SLiM's *split* scenario. Moreover, the specific ratio n_1/n_2 was also not impactful in this sense.

Future work will have to clarify whether the effect of cluster imbalance is related to the computation of the grouping distance. For instance, consider an experiment where a Bernoulli RV with success probability π is used to assign individuals to one of the two groups. Adopting the notation from section 1.2, three group membership

labels z_i^c, z_j^c and z_k^c can be generated a large number of times, similarly to what was done with Gaussian triplets at the beginning of this section. When $\pi = 0.5$, individuals have the same probability to belong to the two groups, so in the long run the groups will be equally sized. When generating z_i^c, z_j^c and z_k^c 1000 times with $\pi = 0.5$, grouping distances $\mathbb{1}(z_i^c \neq z_j^c)$ and $\mathbb{1}(z_i^c \neq z_k^c)$ showed no association, despite having individual i in common. This did not happen with $\pi \neq 0.5$. Is this connected with how the type I error rate of LM in Figure 1.12 changed with cluster imbalance?

1.5.3 On partial Mantel correlation replicates

In section 1.3.2, the significance of the observed value of $r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x)$ was assessed using either permutations, jackknife or bootstrap. Permutations represent the traditional tool to do this in (partial) Mantel tests (Mantel 1967; Smouse et al. 1986), and imply the resampling without replacement of rows and columns of the only $\mathbf{D}_y$. For the jackknife, instead, one subsets matrices $\mathbf{D}_y$, $\mathbf{D}_g$ and $\mathbf{D}_x$ by removing rows and columns referring to the individual dropped each time. The bootstrap applies resampling with replacement to all these three matrices. By adopting different strategies, these techniques generate a distribution of $r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x)$ under the null hypothesis that its true value in the population is zero. The information from this distribution is then used for inference: in this work, quantiles of the permutation distribution were compared with the original coefficient, bootstrap replicates served the construction of a confidence interval, and a t-test was carried out based on jackknife pseudovalues. Here, the features of the distribution of $r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x)$ obtained through these methods are dissected in order to comment on some type I error and power properties they showed in section 1.4.

As exemplified in Figures A.22 and A.23 for SLiM scenarios with equally and unequally sized groups, the differences between these methods can already be seen from the way they replicate distance-distance plots. Jackknife replicates are of course rather similar to the original ones, they just contain $n-1$ less dissimilarities. On one hand, this brings about the smallest impact on the dependence structure originally observed in the data. Any IBD pattern or clustering of genetic distances will be preserved. On the other hand, this limits the number or replicates that can be obtained to the size of the sample, n. In contrast, up to $n!$ replicates of the test statistic can be obtained via permutations. This constraint might explain in part the lack of power of MJack and constitutes the main reason why it is often impractical to compare the observed correlation value with the quantiles of the distribution of jackknife pseudovalues: for instance, if $n < 21$, at 5% significance level, a rejection of the null hypothesis could occur only if the original partial correlation were to be larger than all its pseudovalues.

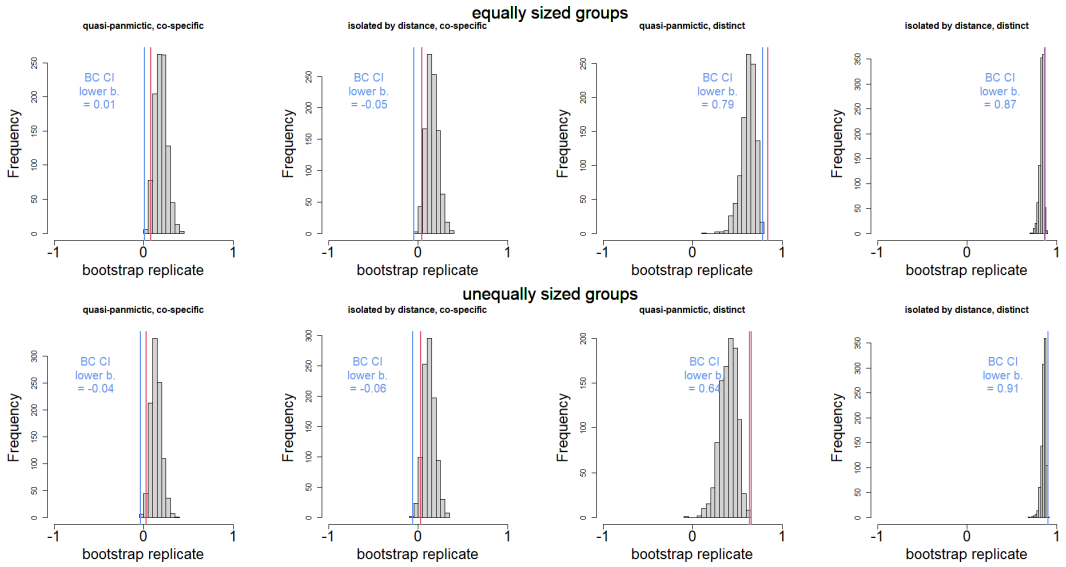

Fig. 1.13 Empirical distribution of 1000 bootstrap replicates of the partial correlation for eight SLiM datasets with $n = 30$ and $P = 89$. Panels by ratio of group sizes, IBD behaviour and conspecificity scenario. A vertical red line is drawn at the value of the original partial correlation coefficient, while a blue line indicates the selected lower boundary of the BC confidence interval. When this blue line lies on the right of value 0, the bootstrap-based PMT rejects the null hypothesis of conspecificity.

Bootstrap dataset replicates also look quite similar to their original counterparts. However, the generation of cloned individuals implies that fewer unique values in $\mathbf{D}_y$ and $\mathbf{D}_x$ are usually spotted and, much more importantly, zero distances appear even if they were not present in the original dataset. When quasi-panmictic species are considered, these zero distances often constitute outlying data points on the distance-distance plots (see the first and third plot on the last row of Figures A.22 and A.23). On top of this, they are always associated with a zero grouping distance, because clones belong to the same group. These outlying distances will thus cause the average within-group genetic dissimilarity to be substantially lower than its between-groups counterpart. In conspecificity scenarios, this gap will often lead to values of $r(\mathbf{D}_y, \mathbf{D}_g | \mathbf{D}_x)$ larger than the original one. As a result, most of the distribution of bootstrap replicates of the partial correlation will fall on the right of value 0. This might lead to the selection of a confidence interval lower boundary larger than zero, producing a rejection of the null hypothesis of conspecificity - as is the case in the first plot of Figure 1.13.

More in detail, since bias-corrected bootstrap CIs adjust the percentile to be selected as lower boundary according to how far the observed partial correlation falls with respect to its median bootstrap replicate, the issue caused by zero distances will be partly mitigated by the choice of percentiles below 5%. Indeed, evidence from the simulations in section 1.4 (not reported) suggests that this was the reason why bias-

corrected CIs were superior to standard percentile ones (Efron and Tibshirani 1993, ch. 13) in terms of type I error rate. Clearly, no bias correction can help if the whole empirical distribution of correlation replicates consists of positive values, and a rejection will follow necessarily. Recall that the lower boundary of bias-corrected bootstrap is chosen to be a percentile (of the empirical distribution of bootstrap replicates) below 5% if the original correlation is lower than its median bootstrap replicate, while a percentile larger than 5% is chosen otherwise. As it can be seen from the distinctness scenarios reported in Figure 1.13, this could even lead to the selection of CI lower boundaries above the median bootstrap replicate.

Despite the distortion due to the introduction of an inflated number of zero distances in the bootstrap dataset replicates, a cursory inspection of the histograms in Figure 1.13 suggests that a test with better type I error properties might be constructed by avoiding confidence intervals. That is, an alternative way to assess significance using the bootstrap could be to compare the original partial correlation with the quantiles of its bootstrap replicates, as it is done with permutations, instead of looking at the lower boundary of the bias-corrected CIs. A systematic assessment of this approach is left for future work.

As regards permutations, in Figures A.22 and A.23 these display the most striking differences with respect to the original plots, as both the association between $\mathbf{D}_y$ and $\mathbf{D}_x$ and that between $\mathbf{D}_y$ and $\mathbf{D}_g$ are disrupted. Namely, in scenarios with IBD co-specific individuals, any positive correlation between genetic and geographic distances is cancelled. Moreover, between-groups dissimilarities (colored in green) cease to be clustered. In particular, the fact that any IBD behaviour is lost through the permutation process enforces a null model that is stricter than what the null hypothesis prescribes, i.e., that the true value of $r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x)$ in the population is zero. In Figures A.24 and A.25 it is shown how not only the permutation distribution of $r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x)$ is centered at zero, but also the distribution of $r(\mathbf{D}_y, \mathbf{D}_x)$ and $r(\mathbf{D}_y, \mathbf{D}_g)$ is. This can happen regardless of the IBD behaviour and the conspecificity setup, and when the original value of these simple correlations is larger than zero, its permutation distribution appears to be slightly skewed on the right. These charts help visualize how permutations act on the dependence structure in the data by erasing any correlation the genetic distances might have with geographic or grouping distances considered separately. This is not implied by the use of the jackknife and might partly explain why the type I error rate of permutation-based PMTs was often borderline - also in section 1.5.2. On a side note, the empirical distribution of permutation replicates for these datasets looks bell-shaped with sufficiently large n and P. This was also the case for the jackknife (not shown), where significance is assessed via a t-test for which pseudovalues are assumed to be asymptotically normal.

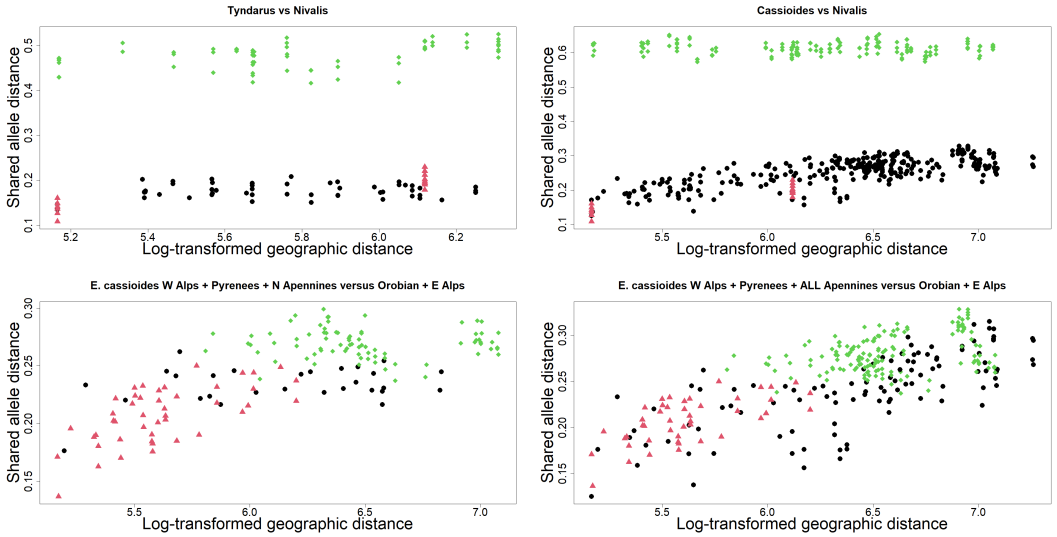

Fig. 1.14 Log-transformed geodesic distances vs. shared allele distances for the four pairs of groups indicated in Table 1.1 but not shown in Figure 1.1 from the brassy ringlets data. The black circles (first group) and red triangles (second group) show distances between pairs of individuals belonging to the same group. The green diamonds show the distances between two individuals belonging to different groups.

1.6 Real data analysis

In this section, real data from Gratton et al. (2016), later examined also by Hausdorf and Hennig (2020), is analyzed with the methods introduced in section 1.3.

1.6.1 Brassy ringlets

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

Table 1.1 Results from all methods compared here on the brassy ringlets data. 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.

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	$< 10^{-8}$	$(0.020, 0.029)$	$< 10^{-25}$	$(0.329, 0.618)$

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 this 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 1.1, the results from all methods involved in this study are reported. See Figures 1.1 and 1.14 for the dissimilarity data on which 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 (middle panel in Figure 1.1) 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 1.1). 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.

1.7 Properties of the shared allele distance

As a measure of similarity, Bowcock et al. (1994) looked at the proportion of alleles two individuals share at a given locus, then averaged this proportion across all P loci. By subtracting this quantity to 1, the shared allele distance is obtained, which was used in this study also because of its intuitive interpretation. Building on its definition, this section deepens some features of the shared allele distance.

In section 1.7.1, some probabilistic properties of the shared allele distance in simplified setups are outlined. The occurrence of specific alleles in loci is modeled with a multinomial law and combinatorics arguments are employed to determine the probability mass function of the shared allele distance. In section 1.7.2, central limit theorems are applied to describe the asymptotic behaviour of this distance as the number of loci or the number of individuals diverges. Results are compared with simulations, also from the SLiM software.

1.7.1 Probabilistic description

By looking at (1.2), we notice that:

$$\begin{aligned} d_y(\mathbf{z}_i,\mathbf{z}_j) &= \frac{1}{P}\sum_{p=1}^{P}\left\{1-\frac{1}{2}|Z_i^p \cap Z_j^p|\left[1+\mathbb{1}\left(|Z_i^p|+|Z_j^p|=2\right)\right]\right\} \\ &= \frac{1}{P}\sum_{p=1}^{P} d_y^{(p)}(\mathbf{z}_i,\mathbf{z}_j) \end{aligned} \tag{1.26}$$

where $d_y^{(p)}(\mathbf{z}_i,\mathbf{z}_j)$ is the shared allele distance (shald, for short) between individual i and individual j based only on the p^{th} locus. So, the overall shared allele distance is the mean of the P single-locus shared allele distances. The single-locus shald can only

take value 1 when there is no allele in common between the loci being compared, value 0.5 in case of partial overlap (no matter the ordering) or 0 in case of complete match. Throughout section 1.7.1, let us assume that the probability of any of these three values does not change across the P loci:

$$p_0 = \mathbb{P}\left(d_y^{(p)}(\mathbf{z}_i, \mathbf{z}_j) = 0\right)$$
$$p_{.5} = \mathbb{P}\left(d_y^{(p)}(\mathbf{z}_i, \mathbf{z}_j) = 0.5\right)$$
$$p_1 = \mathbb{P}\left(d_y^{(p)}(\mathbf{z}_i, \mathbf{z}_j) = 1\right)$$

with $p = 1, \ldots, P$. The value of $d_y^{(p)}(\mathbf{z}_i, \mathbf{z}_j)$ is based on the comparison of the genetic make-up of individuals i and j at locus p. Assuming that M different allelic stata can be observed at each diploid locus, their occurrence can be modeled as a random experiment. Suppose that each allelic status m, with $m = 1, \ldots, M$, has probability q_m of being observed. Since these loci can be homozygous (i.e., same allelic status appears twice in a locus), they can be modeled using two independent extractions with repetitions based on the same vector $\mathbf{q}_i = (q_{i1}, \ldots, q_{iM})$, with $\sum_{m=1}^{M} q_{im} = 1$. Assuming that the same allele distribution applies to the extraction of all P loci from individual i, we can say that

$$L_i^{(p)} \sim \text{Multinomial}(n = 2, \mathbf{q}_i) \quad \forall p\,. \tag{1.27}$$

That is, before we observe the genetic information carried by individual i, each of its P loci, $L_i^{(p)}$, is a Multinomial random variable with two trials and vector of probabilities $\mathbf{q}_i$. If we denote by $H_{im}^{(p)}$ the number of times allele m is extracted for individual i at locus p, the probability that the observed value of locus $L_i^{(p)}$ is Z_i^p can be expressed as:

$$\begin{aligned}\mathbb{P}(L_i^{(p)} = Z_i^p) &= \mathbb{P}(H_{i1} = h_{i1}, \ldots, H_{iM} = h_{iM}) \\ &= \begin{cases} \frac{2}{h_{i1}! \cdots h_{iM}!} q_{i1}^{h_{i1}} \cdots q_{iM}^{h_{iM}} & \text{if} \quad \sum_{m=1}^{M} h_{im} = 2 \\ 0 & \text{otherwise}\,, \end{cases}\end{aligned}$$

where we dropped the superscript referring to the p^{th} locus to keep notation light. Clearly, also the loci from individual j can be assumed to follow a Multinomial distribution, with a (possibly) different vector of probabilities, $\mathbf{q}_j$.

Given the probability to get any locus starting from the allele frequencies, it is possible to dissect the situations that will lead to the each of the three values the single-locus shald can take. Once again, in all loci from individual i, allelic stata are assumed to have the same probabilities of occurrence, stored in $\mathbf{q}_i$. Let us start from $d_y^{(p)}(\mathbf{z}_i, \mathbf{z}_j) = 0$:

- the probability to get an homozygous locus containing the m^{th} allele twice is equal to q_{im}^2 for individual i: the probability that individual j will have that same locus is q_{jm}^2. As a consequence, the probability of minimum distance between homozygous loci is:
$$\sum_{m=1}^{M} q_{im}^2 q_{jm}^2;$$
- the probability to get an heterozygous locus, made up of allele m and allele m', with $m \neq m'$, is $2q_{im}q_{im'}$ for individual i and $2q_{jm}q_{jm'}$ for individual j, so the probability of null shald between heterozygous loci is:
$$\sum_{m=1}^{M} \sum_{m<m'} (2q_{im}q_{im'})(2q_{jm}q_{jm'}).$$

Overall, when comparing individual i and individual j at a given locus p, we obtain:

$$p_0 = \sum_{m=1}^{M} \left[q_{im}^2 q_{jm}^2 + \sum_{m<m'} 4q_{im}q_{im'}q_{jm}q_{jm'} \right]. \tag{1.28}$$

The same reasoning can be applied to retrieve the cases that will lead to a single-locus shald equal to 0.5:

- a homozygous locus can lead to such value only if the other locus is heterozygous and contains the relevant allele, event that has probability:
$$\sum_{m=1}^{M} q_{im}^2 \sum_{m' \neq m} q_{jm}q_{jm'} = \sum_{m=1}^{M} q_{im}^2 q_{jm}(1-q_{jm});$$
- heterozygous loci with allele m and m' will display partial overlap with the two relevant homozygous loci (one for m and one for m') and with heterozygous loci containing one of these two alleles, but not both. Given an heterozygous locus for individual i (whose probability is $2q_{im}q_{im'}$), the probability that the corresponding locus for individual j will share only one allele is:
$$q_{jm}^2 + q_{jm'}^2 + \sum_{o \neq m,m'} (2q_{jm}q_{jo} + 2q_{jm'}q_{jo}) = q_{jm}^2 + q_{jm'}^2 + 2(q_{jm} + q_{jm'})(1 - q_{jm} - q_{jm'}).$$

Table 1.2 Nine possible combinations of two single-locus shald's and corresponding value of their sum S_2.

$d_y^{(1)}$	0	0	0	0.5	0.5	0.5	1	1	1
$d_y^{(2)}$	0	0.5	1	0	0.5	1	0	0.5	1
S_2	0	0.5	1	0.5	1	1.5	1	1.5	2

As a result, after some simple algebra, we obtain:

$$p_{.5} = \sum_{m=1}^{M} \left[2q_{im}^2 q_{jm}(1-q_{jm}) + 2q_{im} \sum_{m<m'} q_{im'}(2q_{jm} + 2q_{jm'} - q_{jm}^2 - q_{jm'}^2 - 4q_{jm}q_{jm'}) \right] \tag{1.29}$$

and the shortcut to calculate the remaining probability is:

$$p_1 = 1 - p_0 - p_{.5}\,. \tag{1.30}$$

In this way, starting from some assumptions on the distribution of the alleles, the probability that the comparison between two individuals at a given locus yields each of the three possible values of the shald was reconstructed. The interest is in the distribution of the overall shared allele distance, which is the mean of the P single-locus shald values.

We now prove by induction that the sum of P single-locus shald's can take $2P+1$ distinct values:

- recall that the single-locus shald ($P = 1$) can take three values, 0, 0.5 and 1;
- when $P = 2$, we expect to find five distinct values. Let us denote the sum of P single-locus shald values with:

$$S_P = \sum_{p=1}^{P} d_y^{(p)}.$$

  Each comparison of loci from two individuals can return three shald values, so there will be 9 possible cases, reported in Table 1.2, leading to $S_2 \in \{0, 0.5, 1, 1.5, 2\}$. A trend can be spotted: the sequence of distinct values for the sum of P single-locus shald values goes from 0 to P with steps equal to $\frac{1}{2}$;
- for $b \in \mathbb{N}^+$, let us assume that the number of distinct values when $P = b$ is indeed $2b+1$. We need to prove that, when $P = b+1$, this count grows to $2b+3$. We noticed that S_b takes value in $\{0, 0.5, \ldots, b-0.5, b\}$. Now, three things can happen with the additional single-locus shald value: either it is equal to 0 and

we get again the same sequence of values from 0 to b, or it takes value 0.5 and thus the resulting value of S_{b+1} belongs to the set $\{0.5, 1, \ldots, b, b+0.5\}$ (with the only new distinct value being $b+0.5$) or it takes value 1, leading to a set of possible values for S_{b+1} equal to $\{1, 1.5, \ldots, b+0.5, b+1\}$ (where the novelty is $b+1$). We navigated all possible cases and the only new distinct values that S_{b+1} can take with respect to S_b were $b+0.5$ and $b+1$: so we get a count of distinct values equal to $2b+1+2 = 2b+3$, as expected.

Therefore, we proved that:

$$\left| \left\{ s : \sum_{p=1}^{P} d_y^{(p)} = s \,,\, d_y^{(p)} \in \{0, 0.5, 1\} \right\} \right| = 2P+1 \,. \tag{1.31}$$

Thus, when averaging P single-locus shald values there will be 3^P possible distance value combinations that will lead to $2P+1$ distinct overall shald values, whose probability needs to be determined.

It is crucial to recall that the single-locus shald is here assumed to be independent and identically distributed $\forall p$: allele frequencies do not vary across loci and all loci are independent. In biology, independence among loci is related to the concept of linkage equilibrium (Waits and Storfer 2015). In this simplified setup, the multinomial distribution can be exploited once again: in each of the P trials, one value of the single-locus distance is extracted according to the vector of probabilities $\mathbf{p} = (p_0, p_{.5}, p_1)$. By denoting with Δ_0 the number of zero distances among the P ones and by δ_0 its observed value (and likewise for $\delta_{.5}$ and δ_1), it is possible to specify:

$$\begin{aligned} \mathbb{P}(S_P = s) &= \mathbb{P}(\Delta_0 = \delta_0, \Delta_{.5} = \delta_{.5}, \Delta_1 = \delta_1) \\ &= \begin{cases} \frac{P!}{\delta_0! \delta_{.5}! \delta_1!} p_0^{\delta_0} p_{.5}^{\delta_{.5}} p_1^{\delta_1} & \text{if} \quad \delta_0 + \delta_{.5} + \delta_1 = P \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

This does not suffice to describe the probability of the $2P+1$ values of the overall shald, because different outcomes of this multinomial experiment will lead to the same distance value. We need to aggregate the related probabilities, because each shald value is associated with the union of these events. It can be beneficial to visualize this problem by means of the layers of *Pascal's pyramid*, which are closely related to trinomial expansions. The trinomial we are expanding, $(p_0 + p_{.5} + p_1)^P$, leads to $\frac{P}{2}(P+1)$ terms which we can arrange as in Figure 1.15, where the case $P = 3$ is considered: by putting $p_{.5}$ on top, we get that the values of d_y grow from left to right and that the natural arrangement of the expansion terms in columns helps us retrieve the addends to be aggregated in order to obtain the probability of each shald value.

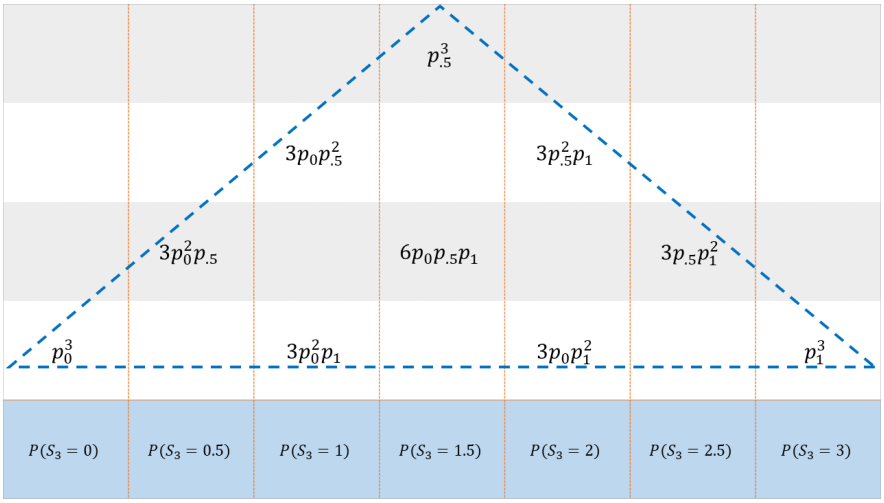

Fig. 1.15 Terms from the trinomial expansion with $P = 3$ arranged in a triangle. By summing the values in each column we get the probabilities specified at the bottom.

Having grasped this, we need a rule to identify the addends for aggregation. To this aim, we define the following set:

$$\mathscr{S}_{s,P} = \left\{ \boldsymbol{\delta}^\intercal = (\delta_0, \delta_{.5}, \delta_1) : \boldsymbol{\delta}^\intercal \mathbf{t} = s \wedge \delta_0 + \delta_{.5} + \delta_1 = P \right\} \tag{1.32}$$

where $\mathbf{t}^\intercal = (0, 0.5, 1)$ and $s \in \{0, 0.5, \ldots, P\}$. $\mathscr{S}_{s,P}$ is the set of all vectors of observed counts $\boldsymbol{\delta}^\intercal$ that sum to P and whose resulting S_P is equal to s, which is one of the possible $2P+1$ values the measure can take.

Thanks to this definition, we can now compute the probability that the sum of P single-locus shald values equals some value s:

$$\mathbb{P}(S_P = s) = \sum_{\boldsymbol{\delta} \in \mathscr{S}_{s,P}} \frac{P!}{\delta_0! \delta_{.5}! \delta_1!} p_0^{\delta_0} p_{.5}^{\delta_{.5}} p_1^{\delta_1} . \tag{1.33}$$

Of course, this is also the probability that the overall shared allele distance takes value s/P, which is the genetic distance of our interest.

1.7.2 Asymptotic results

As it is apparent from (1.26), the shared allele distance is the average of P discrete random variables (RVs) that take value 0 with probability $p_0^{(p)}$, value 0.5 with probability $p_{.5}^{(p)}$ and value 1 with probability $p_1^{(p)}$. To begin, we can assume that these P RVs are independent and identically distributed. Identity in distribution, instead, implies that the vectors of allele probabilities $\mathbf{q}_i^{(p)}$ and $\mathbf{q}_j^{(p)}$ are the same for every locus. By adapting the notation to these assumptions, we can write $p_0^{(p)} = p_0$, $p_{.5}^{(p)} = p_{.5}$ and $p_1^{(p)} = p_1$ for

$p = 1, \ldots, P$. In this section, to ease notation, the subscript y of single-locus shared allele distances is dropped. If we denote by $d^{(1)}$ the first single-locus distance (distributed as all the others), we can compute its expected value and variance:

$$\begin{aligned} \mathbb{E}\left(d^{(1)}\right) &= \frac{1}{2}\mathbb{P}\left(d^{(1)} = 0.5\right) + \mathbb{P}\left(d^{(1)} = 1\right) = \frac{p_{.5}}{2} + p_1\,, \\ \mathbb{V}\left(d^{(1)}\right) &= \mathbb{E}\left((d^{(1)})^2\right) - \mathbb{E}\left(d^{(1)}\right)^2 = p_1 + \frac{p_{.5}}{4} - p_1^2 - \frac{p_{.5}^2}{4} - p_{.5}p_1\,. \end{aligned} \tag{1.34}$$

The expected value is bounded between 0 (when $p_0 = 1$) and 1 (when $p_1 = 1$), whereas the variance is bounded between 0 (when all the probability mass falls on one of the three possible values) and 0.25 when $p_0 = p_1 = 0.5$.

Now, given the IID sequence $\{d^{(p)}\}_{p=1,\ldots,P}$, we will have

$$\begin{aligned} \mathbb{E}\left(\sum_p d^{(p)}\right) &= \sum_p \mathbb{E}\left(d^{(p)}\right) = P\mathbb{E}\left(d^{(1)}\right)\,, \\ \mathbb{V}\left(\sum_p d^{(p)}\right) &= \sum_p \mathbb{V}\left(d^{(p)}\right) = P\mathbb{V}\left(d^{(1)}\right) \end{aligned}$$

and the classical central limit theorem will apply (Billingsley 1995, p. 357):

$$\frac{\sum_p d^{(p)} - P\mathbb{E}\left(d^{(1)}\right)}{\sqrt{P\mathbb{V}\left(d^{(1)}\right)}} \Rightarrow \mathcal{N}(0,1)\,. \tag{1.35}$$

By the delta method, this amounts to claim that the overall shared allele distance

$$d_y = \frac{1}{P}\sum_p d^{(p)} \Rightarrow \mathcal{N}\left(\mathbb{E}\left(d^{(1)}\right), \sqrt{\frac{\mathbb{V}\left(d^{(1)}\right)}{P}}\right)\,. \tag{1.36}$$

That is, under the IID assumption, when the number of loci P goes to infinity, the distribution of the shared allele distance converges to a Normal distribution with mean equal to the single-locus shald expected value and variance equal to the single-locus shald variance divided by P.

The hypothesis of identity in distribution can be relaxed. Suppose we have P different vectors $(p_0^{(p)}, p_{.5}^{(p)}, p_1^{(p)})$, resulting in P different values of $\mathbb{E}\left(d^{(p)}\right)$, which we assume to be known. We can subtract these expected values to each distance in our sequence $\{d^{(p)}\}_{p=1,\ldots,P}$ so to get a new sequence, say, $\{X_p\}_{p=1,\ldots,P}$, such that $\mathbb{E}(X_p) = 0$ for every $p = 1, \ldots, P$ and the whole sequence is bounded by 1. Note that centering does not alter the variance of the elements in the sequence, thus $\mathbb{V}(X_p) = \mathbb{V}\left(d^{(p)}\right)$ for

every p. We assume that the variance $\mathbb{V}\left(\sum_p X_p\right) = \sum_p \mathbb{V}\left(X_p\right) = \sum_p \mathbb{E}\left(|X_p|^2\right)$ goes to infinity as P grows. In this situation, the Lyapunov's condition is verified for $\delta = 1$ (Billingsley 1995, p. 362):

$$\sum_{p=1}^{P} \frac{\mathbb{E}\left(|X_p|^3\right)}{\left(\sum_p \mathbb{E}\left(|X_p|^2\right)\right)^{3/2}} \leq \frac{\sum_{p=1}^{P} \mathbb{E}\left(|X_p|^2\right)}{\left(\sum_p \mathbb{E}\left(|X_p|^2\right)\right)^{3/2}} = \frac{1}{\sqrt{\sum_p \mathbb{E}\left(|X_p|^2\right)}} \to 0\,. \tag{1.37}$$

This amounts to say that

$$\frac{\sum_p X_p}{\sqrt{\sum_p \mathbb{E}\left(|X_p|^2\right)}} = \frac{\sum_p d^{(p)} - \sum_p \mathbb{E}\left(d^{(p)}\right)}{\sqrt{\sum_p \mathbb{V}\left(d^{(p)}\right)}} \Rightarrow \mathcal{N}(0,1)\,, \tag{1.38}$$

which boils down to (1.35) if there is identity in distribution.

It is crucial to remark that here we are considering two unrelated individuals whose P loci are being compared, thus it is the number of loci that diverges. If a very large number of non degenerate single-locus shared allele distances is available to compare the genetic make-up of two unrelated individuals, we expect the overall shared allele distance to be approximately normally distributed.

In practice, this is not the usual situation when working with dissimilarity data. What we observe is a set of n individuals and we compute the $w = n(n-1)/2$ genetic dissimilarities among them, which, by construction, will not be independent - see section 1.5.2. Let us consider the asymptotic distribution of these w dissimilarities.

If the number of individuals n is even, we can allocate them into $n-1$ groups where each individual appears only once and is coupled with another individual: by doing so, the $n/2$ dissimilarities between paired individuals in each group will be unrelated. By repeating this until all dissimilarities in the dataset have been assigned, the w dissimilarities can be partitioned into $n-1$ groups of size $n/2$. Although dissimilarities within each group will be independent, every group will share some information with any other group because each individual will be present in more than one group. If n is odd, n groups of $(n-1)/2$ unrelated dissimilarities can be formed. This way of partitioning the w dissimilarities was already hinted at in Bohonak (2002). We can

visualize this with $n = 7$, where we spot 7 groups of 3 dissimilarities,

$$\begin{pmatrix} d_y(\mathbf{z}_1,\mathbf{z}_2) & & & & & & \\ d_y(\mathbf{z}_1,\mathbf{z}_3) & d_y(\mathbf{z}_2,\mathbf{z}_3) & & & & & \\ d_y(\mathbf{z}_1,\mathbf{z}_4) & d_y(\mathbf{z}_2,\mathbf{z}_4) & d_y(\mathbf{z}_3,\mathbf{z}_4) & & & & \\ d_y(\mathbf{z}_1,\mathbf{z}_5) & d_y(\mathbf{z}_2,\mathbf{z}_5) & d_y(\mathbf{z}_3,\mathbf{z}_5) & d_y(\mathbf{z}_4,\mathbf{z}_5) & & & \\ d_y(\mathbf{z}_1,\mathbf{z}_6) & d_y(\mathbf{z}_2,\mathbf{z}_6) & d_y(\mathbf{z}_3,\mathbf{z}_6) & d_y(\mathbf{z}_4,\mathbf{z}_6) & d_y(\mathbf{z}_5,\mathbf{z}_6) & & \\ d_y(\mathbf{z}_1,\mathbf{z}_7) & d_y(\mathbf{z}_2,\mathbf{z}_7) & d_y(\mathbf{z}_3,\mathbf{z}_7) & d_y(\mathbf{z}_4,\mathbf{z}_7) & d_y(\mathbf{z}_5,\mathbf{z}_7) & d_y(\mathbf{z}_6,\mathbf{z}_7) & \end{pmatrix},$$

and with $n = 8$, where we can see 7 groups made up of 4 dissimilarities:

$$\begin{pmatrix} d_y(\mathbf{z}_1,\mathbf{z}_2) & & & & & & \\ d_y(\mathbf{z}_1,\mathbf{z}_3) & d_y(\mathbf{z}_2,\mathbf{z}_3) & & & & & \\ d_y(\mathbf{z}_1,\mathbf{z}_4) & d_y(\mathbf{z}_2,\mathbf{z}_4) & d_y(\mathbf{z}_3,\mathbf{z}_4) & & & & \\ d_y(\mathbf{z}_1,\mathbf{z}_5) & d_y(\mathbf{z}_2,\mathbf{z}_5) & d_y(\mathbf{z}_3,\mathbf{z}_5) & d_y(\mathbf{z}_4,\mathbf{z}_5) & & & \\ d_y(\mathbf{z}_1,\mathbf{z}_6) & d_y(\mathbf{z}_2,\mathbf{z}_6) & d_y(\mathbf{z}_3,\mathbf{z}_6) & d_y(\mathbf{z}_4,\mathbf{z}_6) & d_y(\mathbf{z}_5,\mathbf{z}_6) & & \\ d_y(\mathbf{z}_1,\mathbf{z}_7) & d_y(\mathbf{z}_2,\mathbf{z}_7) & d_y(\mathbf{z}_3,\mathbf{z}_7) & d_y(\mathbf{z}_4,\mathbf{z}_7) & d_y(\mathbf{z}_5,\mathbf{z}_7) & d_y(\mathbf{z}_6,\mathbf{z}_7) & \\ d_y(\mathbf{z}_1,\mathbf{z}_8) & d_y(\mathbf{z}_2,\mathbf{z}_8) & d_y(\mathbf{z}_3,\mathbf{z}_8) & d_y(\mathbf{z}_4,\mathbf{z}_8) & d_y(\mathbf{z}_5,\mathbf{z}_8) & d_y(\mathbf{z}_6,\mathbf{z}_8) & d_y(\mathbf{z}_7,\mathbf{z}_8) \end{pmatrix}.$$

Note that the groupings displayed here are arbitrary as other allocations of the dissimilarities into groups are possible that fulfill the unrelatedness requirement explained above. Since we are now studying the asymptotic behaviour for $n \to +\infty$ (while P is fixed), the odd versus even distinction is not crucial. We will consider the array we need when $n = 7$ (odd): let us arrange the 21 dissimilarities into 7 rows of 3 elements each:

$$\underbrace{\begin{matrix} d_y(\mathbf{z}_1,\mathbf{z}_2) & d_y(\mathbf{z}_3,\mathbf{z}_7) & d_y(\mathbf{z}_5,\mathbf{z}_6) \\ d_y(\mathbf{z}_1,\mathbf{z}_3) & d_y(\mathbf{z}_4,\mathbf{z}_6) & d_y(\mathbf{z}_5,\mathbf{z}_7) \\ d_y(\mathbf{z}_1,\mathbf{z}_4) & d_y(\mathbf{z}_2,\mathbf{z}_5) & d_y(\mathbf{z}_3,\mathbf{z}_6) \\ d_y(\mathbf{z}_1,\mathbf{z}_5) & d_y(\mathbf{z}_2,\mathbf{z}_6) & d_y(\mathbf{z}_4,\mathbf{z}_7) \\ d_y(\mathbf{z}_1,\mathbf{z}_6) & d_y(\mathbf{z}_3,\mathbf{z}_4) & d_y(\mathbf{z}_2,\mathbf{z}_7) \\ d_y(\mathbf{z}_1,\mathbf{z}_7) & d_y(\mathbf{z}_2,\mathbf{z}_4) & d_y(\mathbf{z}_3,\mathbf{z}_5) \\ d_y(\mathbf{z}_2,\mathbf{z}_3) & d_y(\mathbf{z}_4,\mathbf{z}_5) & d_y(\mathbf{z}_6,\mathbf{z}_7) \end{matrix}}_{7\times 3} = \begin{matrix} d_{1,1} & d_{1,2} & d_{1,3} \\ d_{2,1} & d_{2,2} & d_{2,3} \\ d_{3,1} & d_{3,2} & d_{3,3} \\ d_{4,1} & d_{4,2} & d_{4,3} \\ d_{5,1} & d_{5,2} & d_{5,3} \\ d_{6,1} & d_{6,2} & d_{6,3} \\ d_{7,1} & d_{7,2} & d_{7,3} \end{matrix}$$

where each row is a collection of shared allele distances based on disjoint pairs of individuals. Following (Billingsley 1995, p. 359), we can assume without loss of generality that these w elements have zero mean: indeed, we can fill the array with the

centered version of the dissimilarities,

$$\underbrace{\begin{matrix} X_{1,1} & \cdots & X_{1,(n-1)/2} \\ X_{2,1} & \cdots & X_{2,(n-1)/2} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,(n-1)/2} \end{matrix}}_{n \times \frac{n-1}{2}}, \tag{1.39}$$

where $X_{m,h} = d_{m,h} - \mathbb{E}\left(d_{m,h}\right)$, for $m = 1, \ldots, n$ and $h = 1, \ldots, (n-1)/2$. Each $X_{m,h}$ has zero mean, finite variance and is bounded by 1. We can now define

$$\begin{aligned} \sigma^2_{m,h} &= \mathbb{E}\left(X^2_{m,h}\right) \\ s^2_m &= \sum_{h=1}^{(n-1)/2} \sigma^2_{m,h}, \end{aligned}$$

where the latter is the sum of the variances of the elements in the m^{th} row and is assumed to be positive for large m. The number of rows diverges when the number of individuals n diverges, so in Lyapunov's condition with $\delta = 1$ we can write

$$\lim_{n \to +\infty} \sum_{h=1}^{(n-1)/2} \frac{\mathbb{E}\left(|X_{m,h}|^3\right)}{s^3_m} \leq \lim_{n \to +\infty} \frac{\sum_{h=1}^{(n-1)/2} \mathbb{E}\left(X^2_{m,h}\right)}{s^3_m} = \lim_{n \to +\infty} \frac{1}{s_m} = 0, \tag{1.40}$$

since the number of positive addends in s_m is $(n-1)/2$. This proves that:

$$\frac{1}{s_m} \sum_{h=1}^{(n-1)/2} X_{m,h} = \frac{\sum_{h=1}^{(n-1)/2} d_{m,h} - \sum_{h=1}^{(n-1)/2} \mathbb{E}\left(d_{m,h}\right)}{\sqrt{\sum_{h=1}^{(n-1)/2} \mathbb{V}\left(d_{m,h}\right)}} \Rightarrow \mathcal{N}(0,1), \tag{1.41}$$

identifying an asymptotic distribution for the average genetic dissimilarity even in real datasets where dissimilarities display dependence because of the way they are computed. Note that we assumed nothing about the number of loci P on which each $d_{m,h}$ is based, although these dissimilarities cannot be degenerate, i.e., have zero variance.

In Figure 1.16 we can visualize how closely the empirical distribution of the shared allele distances based on n individuals and P IID loci resembles the asymptotic distribution described in (1.36). This data was generated by randomly imputing diploid loci to the individuals, with four equally likely allelic stata, thus there exists no relationship at all among the individuals and the only dependence embedded in the genetic dissimilarities is due to their computation. We can see how, with small P, the kernel density estimation based on the w dissimilarities is wiggly because there are few values the

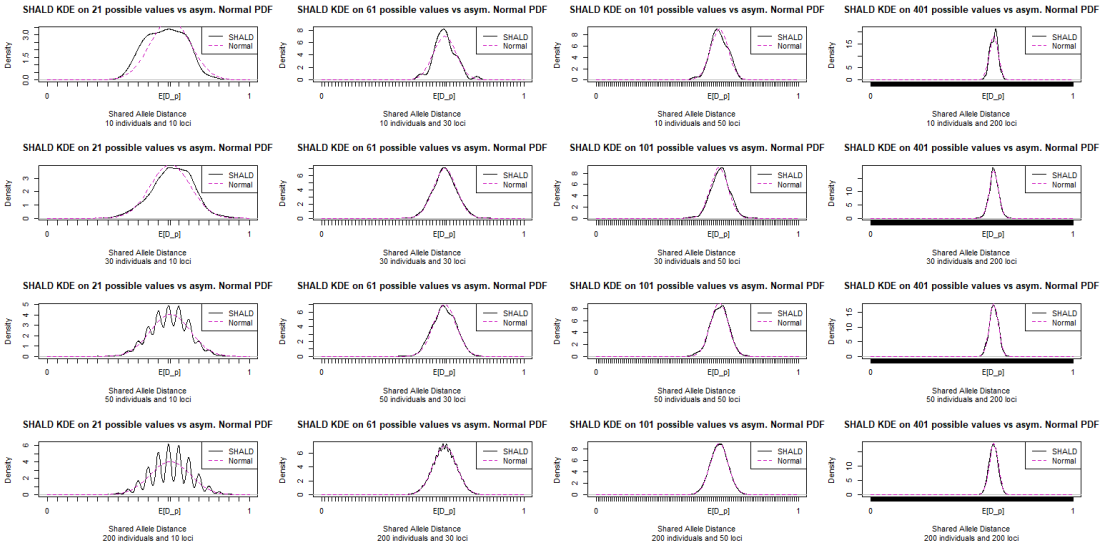

Fig. 1.16 Graphical comparison between KDE based on w dissimilarities and P independent and identically distributed loci (four equally likely allelic stata) and the associated asymptotic Normal distribution. Faceted by number of individuals (rows) and number of loci (columns).

shared allele distance can take, but already with $n = P = 30$, the two curves overlap almost completely. As expected, the variability of the dissimilarities decreases with P and they get more and more concentrated around the average single-locus shared allele distance, $\mathbb{E}\left(d^{(1)}\right)$.

With independent but not identically distributed loci, the approximation works similarly (not shown here): the asymptotic Normal distribution will have mean and variance equal to the average mean and variance of the single-locus shared allele distances, respectively.

The kernel density estimate based on data simulated from the *split* scenario with 90 individuals in SLiM (section 1.4.1) helps us visualize the impact of the additional dependence factors among observations: in such scenario, individuals are all related in that they belong to the same population and share ancestors; also, loci were not simulated as independent, thus the asymptotic normality of the shared allele distance as the number of loci increases is not guaranteed. From Figure 1.17, it is clear that the assumptions on which (1.38) is based are not fulfilled. The mean of the asymptotic Normal density is obtained as the mean (over the number of loci, P) of the average (over the number of dissimilarities, $w = 4005$) single-locus shared allele distance; similarly, the variance of that Normal curve is the average (over loci) of the sample variance (over w) of the single-locus shared allele distances: these would be the predicted parameters of the asymptotic Normal approximation if there was no dependence among individuals or among loci in the dataset, something that is rarely to be seen in real data.

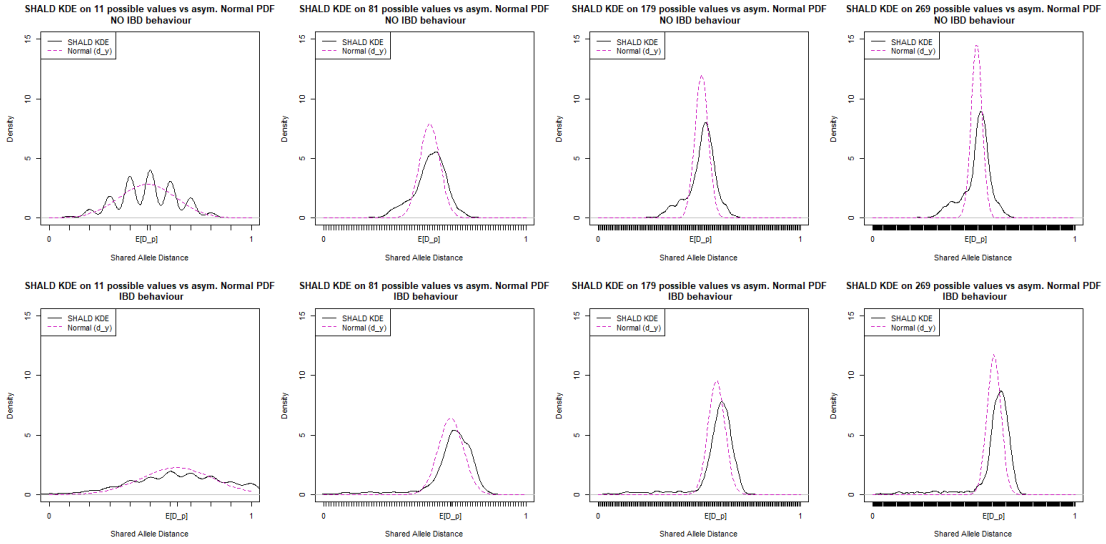

Fig. 1.17 Graphical comparison between KDE based on 4005 dissimilarities and P **dependent loci** simulated with SLiM (*split* scenario) and the associated asymptotic Normal distribution assuming independence. Mean and variance of the Normal proxy are averages over their single-locus sample estimates. Panels by IBD behaviour (rows) and number of loci (columns).

The approximation looks poorer when the data are generated without any isolation by distance behaviour, which might be due to the larger share of extremely low values recorded for IBD simulations: larger variance estimates for the normal proxy return wider curves that come closer to the real data bulk, still neglecting its skewness, though. Under no IBD scenarios, real data display a less Gaussian density, which makes the proxy look poorer overall. In order to assess the impact of the dependence among loci, the same plot is reported in Figure 1.18 for a SLiM simulation where the recombination rate was set to 0.5 and thus all loci are independent. We can see how the approximation looks much more satisfactory when there is no IBD behaviour, as can be expected. When individuals display an IBD behaviour, instead, the kernel density estimate does not get close to the theoretical asymptotic distribution and the mismatch resembles the one observed with dependent loci.

Lastly, in order to visualize how good is the proxy in (1.41), the kernel density estimate of the average shared allele distance from one hundred SliM simulations was plotted in Figure 1.19 for six combinations of IBD behaviour and sample size. We can see how the asymptotic approximation gets closer to the empirical distribution as n grows, as expected. The impact of the IBD behaviour does not look remarkable here, although the curves look more similar for IBD species - a difference which goes in the same direction as in Figure 1.17.

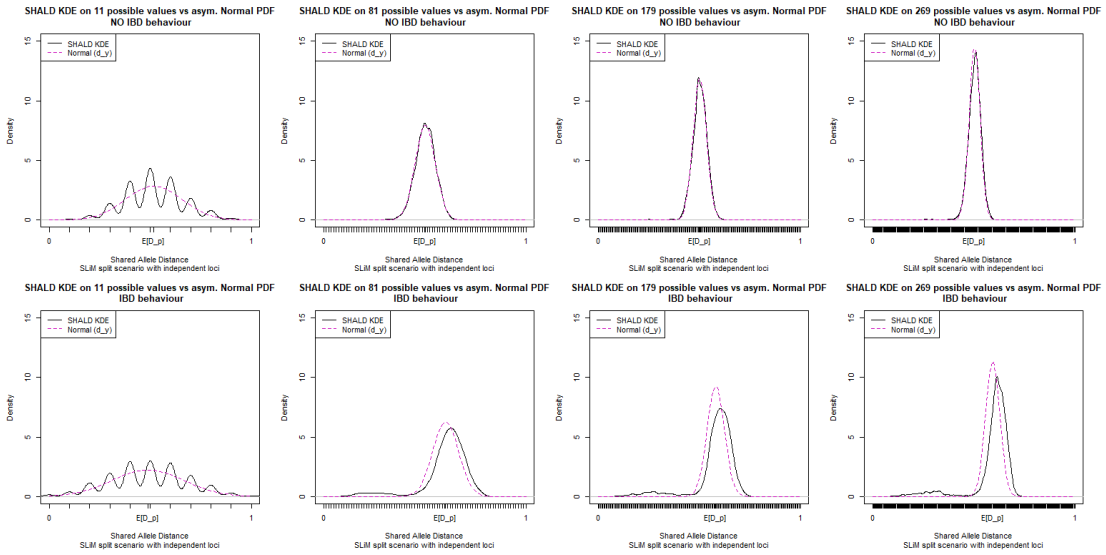

Fig. 1.18 Graphical comparison between KDE based on 4005 dissimilarities and P **independent loci** simulated with SLiM (*split* scenario) and the associated asymptotic Normal distribution assuming independence. Mean and variance of the Normal proxy are averages over their single-locus sample estimates. Panels by IBD behaviour (rows) and number of loci (columns).

1.8 Closing remarks

This chapter considered the problem of testing with genetic and geographic data whether two putative groups of individuals belong to the same species. To this aim, species delimitation methods that model the relationship between genetic and geographic dissimilarities at individual level were investigated. 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 geographic 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. This occasional anti-conservativeness might be partly explained by the fact that permuting the genetic distances not only makes them independent of grouping distances controlling for geographic distances - as the null hypothesis requires - but also disrupts any association between genetic and geographic distances and genetic and grouping distances considered separately. 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. Testing PMTs with bias-corrected

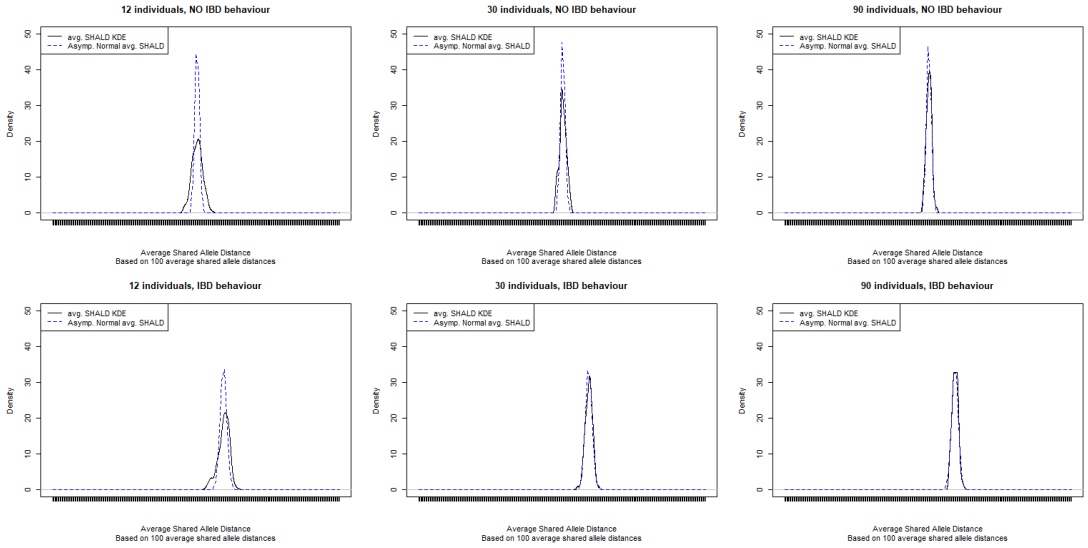

Fig. 1.19 Graphical comparison between KDE based on 100 average shald's (with $P = 134$ loci) based on data simulated with SLiM (*split* scenario) and the associated asymptotic Normal distribution. The mean of the Normal proxy is the grand mean per scenario, whereas the variance is the average of the 100 variances from that scenario, divided by 100. Faceted by IBD behaviour (rows) and number of individuals per group (columns).

bootstrap confidence intervals, instead, often led to inflated type I error rates. This was due to the generation of too many zero dissimilarities in the bootstrap replicates of the datasets, which caused the bootstrap distribution of partial correlation coefficients to be biased. The extension of the linear mixed effects model by Clarke et al. (2002) displayed a performance similar to PMT. A linear regression without random effects (LM), i.e., wrongly assuming independence among the dissimilarities, showed inflated type I error rates only with unequally sized groups simulated with SLiM, and performed surprisingly well in other scenarios. However, the effect of unequal group sizes on the type I error of LM was confirmed also on Gaussian data, suggesting that neglecting the dependence in the dissimilarities can lead to liberal tests in most cases. The ranking in the overall performance of the methods was consistent over both simulators, and the log-transformation of the geographic dissimilarities did not seem to have a considerable impact on methods other than HH20, where it improved matters.

Although the relationship between genetic and geographic dissimilarities was often found to be convex, both on real and simulated data, this did not seem to substantially affect the methodologies. The assumption of linearity was also studied using an ad hoc spatially-explicit genetic simulator, which assumes that allele frequencies in the loci vary along geographic linear gradients. As far as the shared allele distance is concerned, data from this simulator showed that the trend in the dissimilarities can be expected

to be linear only when the number of possible allelic stata is small and IBD patterns are weak. SLiM simulations also illustrated how dissimilarity-based methods can face identifiability issues if there is strong genetic variability already within the putative groups, as Hausdorf and Hennig (2020) suggested.

The shared allele dissimilarity was proved to fulfill the triangle inequality when no missing loci are present in the data. In the ideal scenario of independent and identically distributed loci, the shared allele distance between two unrelated individuals was proved to be asymptotically normal.

Several aspects of this study lend themselves to future investigations. Due to the extremely large amount of possibilities defined by parameter choices of the simulators, many more potentially interesting situations could be simulated. As an example, scenarios worth exploring could involve comparisons between groups with different IBD behaviour or groups covering areas of different size. Some simulation parameters may be informed via estimates from real data. Moreover, the investigation could be extended to other genetic and landscape distances and the analysis could be carried out at population-level. Techniques that do not assume linearity, such as polynomial regressions or PMTs based on rank correlations, might also be studied. Lastly, applying the methods treated here to other problems of regression between dissimilarities such as relating similarity between languages or dialects to geographic distance (Bella et al. 2021) may also be of interest.

Chapter 2

Approaches with unknown groups

When no putative grouping is available, methods able to identify co-specific groups in the dataset are required. The diversity in the dataset is summarized in genetic clusters (or discrete populations), which do not necessarily correspond to distinct species. These clusters can then inform species delimitation considerations according to the operational criteria adopted for a specific study, in line with De Queiroz (2007).

As explained in Chapter 1, in this work individuals are deemed to be co-specific when they display genetic discrepancies compatible with their geographic separation. This operational criterion only tackles one aspect of species delimitation, a complex task for which a range of criteria have to be considered together (De Queiroz 1998). Nevertheless, to ease the discussion, evidence from the methodologies presented here will be used to draw conclusions on species membership, i.e., to delimit the analyzed individuals into distinct species. See also the remarks in section 1.1 on this.

This chapter tackles two problems: a) clarifying whether a spatially-explicit genetic dataset contains more than one species and b) estimating the number of species. Since the presence of spatial patterns of genetic differentiation can lead to the overestimation of the number of groups in the data (Bradburd, Coop, and Ralph 2018), unsupervised species delimitation methodologies have to take these patterns into account. To this aim, in section 2.3.1 a routine is proposed that integrates sNMF (Frichot, Mathieu, et al. 2014), an existing clustering algorithm for genetic data, with the distance-based methods introduced in section 1.3. In section 2.3.2, TESS3 (Caye et al. 2018) is used, which instead takes both geographic and genetic information into account when clustering individuals. Its output is exploited by a distance-based method here developed for the estimation of the number of species in the dataset.

Through a broad simulation study based on SLiM, a critical limitation of both these approaches is highlighted. To overcome this limitation, null models specific to this application are proposed. These are used to calibrate tests based on TESS3 output both

for the detection of the presence of more than one species in the dataset and for the estimation of the number of species.

Throughout this chapter, the notation established in section 1.2 will be preserved, unless otherwise stated. The data consists of n individuals whose locations and genetic make-ups are known, unlike the grouping information stored in $\{z_1^c, \ldots, z_n^c\}$. Also here locations consist of two coordinates and genetic information is expressed in P diploid loci.

In the following, after a brief review of the literature in section 2.1, additional SLiM simulations are presented in section 2.2. They will be used to compare all the methodologies proposed in the chapter. Section 2.3 describes the integrated methods that build on sNMF and TESS3 solutions, whereas section 2.4 discusses null model calibration.

2.1 Literature review

Rannala and Z. Yang (2020) review the history of molecular species delimitation, with reference to how the use of genetic information integrated taxonomic practices mainly based on morphological characters. The approach of taxonomists to cluster analysis is contrasted with that of numerical ecologists in P. Legendre and L. Legendre (2012, ch. 8), where techniques based on measures of resemblance between individuals are surveyed. Clustering spatially-explicit genetic datasets is at the heart of landscape genetics (Manel, Schwartz, et al. 2003), where it is connected to assignment methods (Manel, Gaggiotti, and Waples 2005). In supervised settings where predefined populations are available, assignment methods are used to classify individuals accordingly. In unsupervised settings, instead, these populations are constructed from the data - usually by assuming that they fulfill some criteria, such as linkage equilibrium. The role of the assumptions in clustering methods guides the review in François and Waits (2015), where a distinction is introduced between exploratory data analysis and model-based clustering. In that review, algorithms that do not use geographic information, such as sNMF (Frichot, Mathieu, et al. 2014) and STRUCTURE (Pritchard et al. 2000), and methods that exploit it, such as GENELAND (Guillot, Mortier, and Estoup 2005) and TESS (C. Chen et al. 2007), are briefly discussed, together with their built-in criterion for the choice of K, the number of clusters to be identified. The latest version of TESS, introduced by Caye et al. (2018), replaces its Bayesian algorithm with least squares optimization.

Another recent model-based clustering method for spatially-explicit data is conStruct (Bradburd, Coop, and Ralph 2018), which attempts to explicitly include patterns of isolation by distance in the assignment of the individuals to the groups. Indeed, TESS3

embeds geographic information by introducing spatial priors on group membership, but then assumes constant allele frequencies across the species' range. In conStruct, instead, spatial patterns of genetic differentiation are modeled also within the groups. In its Bayesian framework, estimation is based on Hamiltonian Monte Carlo, with potentially longer computational times than TESS3. Additionally, TESS3 can provide clustering configurations for datasets where $n > P$, unlike conStruct.

Bradburd, Coop, and Ralph (2018) also report several references to existing methodologies, and discuss their criterion to select K in the light of related literature - see Verity and Nichols (2016). The problem of the choice of K is indeed well known in molecular ecology, a field where one of the most used criteria adopted for the task is the ΔK statistic by Evanno et al. (2005) (Janes et al. 2017). In statistics, null models have been used to tackle this kind of problems - see, e.g., Tibshirani et al. (2001). The formulation of null models can help guide assignment methods in the delimitation of distinct species, since spatial patterns of differentiation can induce within-species heterogeneity. Clustering methods can identify genetic clusters existing within the species level due to these patterns, but the goal in this chapter is to determine the value of K that corresponds to the number of species in the dataset. This will require the development of data-influenced null models tailored to the task. On a general note, see Jain and Dubes (1988, ch. 4) for the role of null models in cluster analysis and Gordon (1996) for the relevance of data-influenced null models.

In this chapter, strategies to delimit species in the presence of IBD patterns are investigated in an unsupervised setting. To this aim, the performance of the proposed clustering methodologies is assessed based on SLiM datasets from a wide range of scenarios. In the first part of the chapter, two new clustering routines are conceived that build on the output of sNMF (Frichot, Mathieu, et al. 2014) and TESS3 (Caye et al. 2018), respectively. Since several SLiM datasets considered in this work have $n > P$, and also because execution time is a critical parameter in simulation studies, TESS3 will be used as spatially-explicit clustering method instead of conStruct. The comparison of their performance on datasets where they can both be applied is left for future work. In the second part of the chapter, test statistics computed from TESS3 output are calibrated by means of null models in order to detect the number of clusters while controlling for spatial patterns of differentiation.

2.2 Additional SLiM simulations

In this chapter, datasets generated via the SLiM software (Haller and Messer 2023) will be used to compare the performance of the methods discussed. They are described here

for convenience, as many following sections will make reference to them. The SLiM software was introduced in section 1.4.1, whereas related code is in Appendix C.1.

Eight data scenarios were conceived, some of which were already used in Chapter 1. For each of them, a version with quasi-panmictic species and one with species showing the same IBD behaviour were generated. Additionally, three sample size options were explored. Each combination of scenario, sample size and IBD behaviour was simulated 100 times, returning geographic positions of the individuals and 1000 diploid loci with four possible allelic stata (the four nucleobases: adenine, cytosine, guanine and thymine). As in Chapter 1, a random selection of 134 of these loci would then be recorded, and divided in a subset with 5, one with 40 and one with 89 loci, which allowed to investigate the impact of data richness on methods' performance.

In Figure 2.1, the locations of the individuals from exemplary datasets are shown for each combination of scenario and IBD behaviour. Eight different SLiM scripts were used to create:

- a scenario with only one species covering the whole map;
- a scenario with two species inhabiting the whole map, one having twice the number of members of the other;
- a scenario with two unequally sized species inhabiting separate areas of the map;
- a scenario with two distinct species inhabiting the whole map, but where one of them has two subpopulations, generated as in the *overlapping conspecific* scenario in section 1.4.1. Each of these subpopulations is made up of as many members as those from the other species;
- a scenario with three equally sized species inhabiting the whole map;
- a scenario with three equally sized species segregated in three disjoint areas of the map;
- a scenario with six species inhabiting the whole map, where three of them have a number of individuals twice as big as the number of the individuals from the other three species;
- a scenario with six species as the previous one, but where now each species inhabits its exclusive niche on the map.

Scenarios with three or fewer species were simulated with a total number of individuals equal either to 12, 30 or 90, while for the two scenarios with six species there were either 36, 90 or 270 individuals. In Table 2.1, the number of individuals belonging to

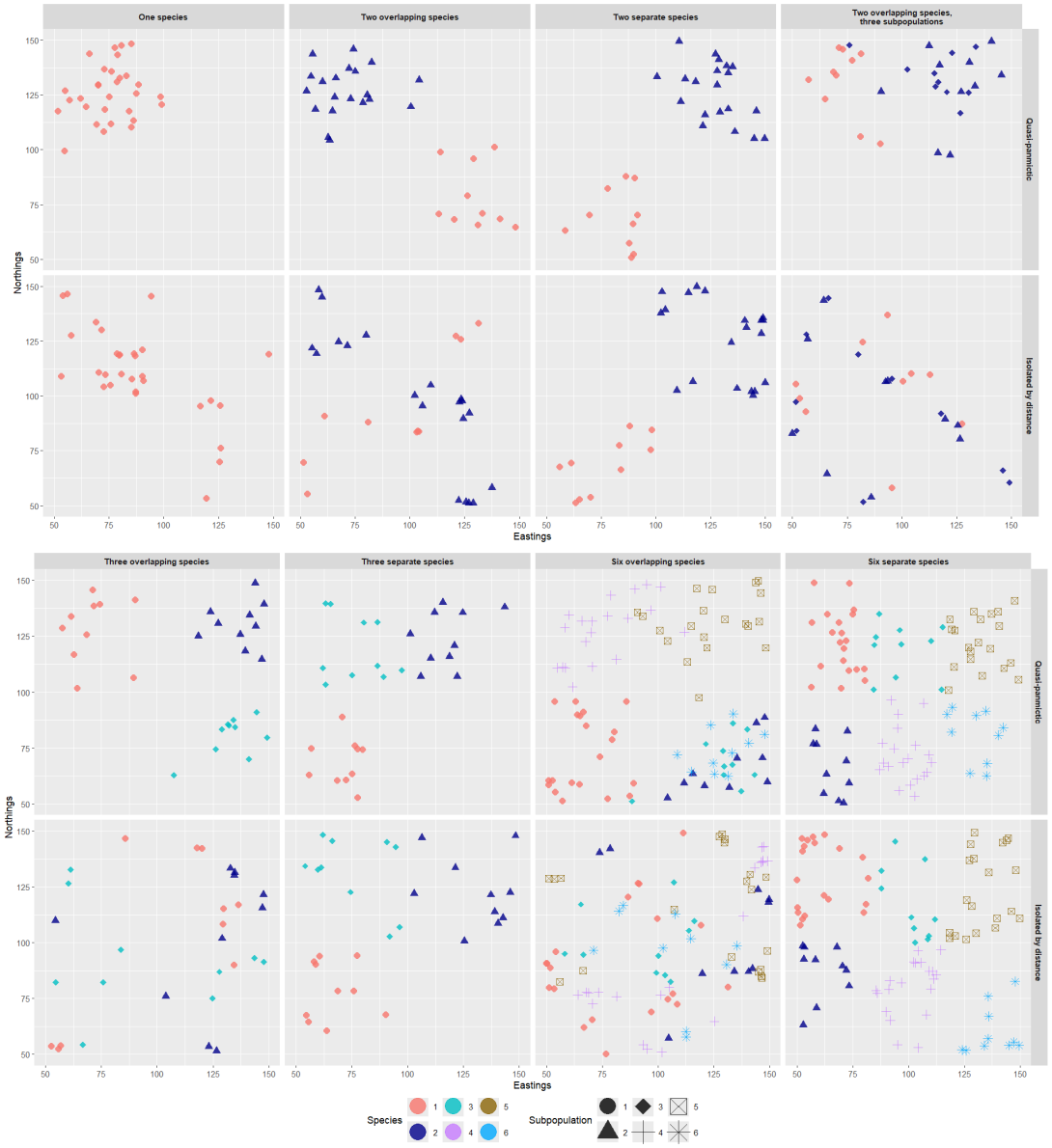

Fig. 2.1 Maps of the geographic positions of individuals from exemplary SLiM datasets with average sample sizes. Panels by scenario and IBD behaviour. Colours map species membership, while shapes map subpopulations: the latter only differ from species membership in the scenario where one of the two species has two subpopulations. All visual cues are only comparable within a scenario, as the individuals from different scenarios are not related.

each species is recapitulated for setups with average sample sizes. For instance, in the last scenario there were three species with 10 individuals each and three species with 20 individuals each, for a total sample size of 90.

As it can be seen on Figure 2.1, individuals from quasi-panmictic species tended to be geographically close, even in scenarios where they were allowed to populate the whole map. On the contrary, individuals from species isolated by distance could

Table 2.1 Number of species members by scenario (setups with average sample size).

Species	1	2	3	4	5	6
One species	30	0	0	0	0	0
Two overlapping species	10	20	0	0	0	0
Two separate species	10	20	0	0	0	0
Two overlapping species, three subpopulations	10	20	0	0	0	0
Three overlapping species	10	10	10	0	0	0
Three separate species	10	10	10	0	0	0
Six overlapping species	20	10	10	20	20	10
Six separate species	20	10	10	20	20	10

well be separated by very large geographic distances, consistently with the mating and reproductive behaviour described in section 1.4.1. As a result, in absence of spatial patterns of differentiation, the amount of geographic segregation in scenarios with overlapping species often ended up being not that different from that observed with separate species, because distinct species would occasionally populate disjoint areas of the map. This was most evident in scenarios with 2 species and less impactful in scenarios with 6 species. Similarly to Chapter 1, also in these simulations all species from a dataset exhibited the same IBD behaviour.

In the following sections, this data will be used to examine the performance of routines whose goal is to identify the number of species in the data and, possibly, to allocate individuals to them.

2.3 Integrating assignment methods with distance-based approaches

In this section, the problem of assigning n individuals to K groups, with K unknown, is tackled by means of two new approaches. These two-step approaches first generate a clustering solution using existing software and then resort to distance-based methods to integrate and possibly improve the solution. The first of the two strategies uses sNMF (Frichot, Mathieu, et al. 2014) to cluster individuals based on genetic information only and then relies upon distance-based methods from section 1.3 to test groups for merging in an iterative routine. The second strategy generates the preliminary clustering configuration using TESS3 (Caye et al. 2018), which instead exploits spatial information. Then, a jaccknife-based iterative routine that builds on TESS3 output is executed to find the number of groups K whose associated TESS3 configuration is most supported. These routines are evaluated based on the SLiM datasets presented in section

2.2. A critical limitation affecting both these proposed methodologies is highlighted, motivating section 2.4.

2.3.1 Merging routine based on sNMF configuration

The sNMF function from the `R` package `LEA` (Frichot and François 2015) provides statistical estimates of ancestry proportions using multilocus genotype data. In admixture models, the alleles that make up the genetic material of the individuals are supposed to be proportionally inherited from K ancestral populations (François and Waits 2015). Ancestry coefficients quantify to what extent each individual's genetic material originated from these populations: in practice, individuals in the dataset are assigned to K groups according to their largest estimated ancestry proportion.

Consider a genotypic matrix $\mathbf{X}$, whose entries contain the number of instances of each allele for all loci and all individuals. As far as simulated data is concerned, in this chapter we deal with diploid individuals and nucleotidic models: each locus contains two alleles and there are four possible allelic stata ("A", "C", "G", "T"). Thus, matrix $\mathbf{X}$ will have n rows and $4P$ columns, where P is the number of loci. For instance, if the first locus of individual i is "CG", the first four elements on the i^{th} row of $\mathbf{X}$ will be "0110", whereas with locus "TT" we would have "0002". Technically, if any locus in the genetic dataset displays less than four different allelic stata (e.g., no guanine ever spotted at a given locus), $\mathbf{X}$ will contain fewer than $4P$ columns. The same thing happens if some loci contain redundant information: examples are monomorphic loci that never change across all individuals in a dataset or loci where one nucleobase is always present and we only have to record which one of the other three occurs. For the sake of notation, let us assume that $\mathbf{X}$ contains $4P$ columns (no missing allelic stata or redundant loci).

Denote by $\widetilde{\mathbf{X}} = (\widetilde{X}_{il})$ the expanded version of $\mathbf{X}$, containing three times its number of columns: its entries are either zero or one and indicate for each allele count in $\mathbf{X}$ (0, 1 or 2 in our setup) whether an individual has that many instances of a given allele in a given locus.

By using an allelic status indicator m, equal to 1 for adenine, 2 for cytosine, 3 for guanine and 4 for thymine, the probability that individual i displays c instances of the m^{th} allelic status at locus p can be expressed as:

$$\mathbb{P}(\widetilde{X}_{il} = c) = \sum_{k=1}^{K} Q_{ik} G_{lk}\,, \quad c = 0,1,2\,, \tag{2.1}$$

where Q_{ik} is the share of genetic information that individual i inherited from the ancestral population k, G_{lk} - with $l = 3(\widetilde{p}+m)+c-2$ - quantifies how often c instances of the m^{th} allelic status are observed at locus p in population k and $\widetilde{p} = 4(p-1)$.

Estimates of Q_{ik} and G_{lk} for every i, k and l are found by minimizing the least-squares criterion:

$$||\widetilde{\mathbf{X}} - \mathbf{Q}\mathbf{G}^{\intercal}||_F^2, \tag{2.2}$$

where $\mathbf{Q} = (Q_{ik})$ is the $n \times K$ matrix of admixture proportions, $\mathbf{G} = (G_{lk})$ is the $12P \times K$ matrix of allele count frequencies and $||\cdot||_F$ denotes the Frobenius norm. These two matrices are subject to the following constraints:

$$\begin{aligned} Q_{ik}, G_{lk} > 0 \quad & \forall i,k,l\,; \\ \sum_{k=1}^{K} Q_{ik} = 1 \quad & \forall i\,; \\ \sum_{c=0,1,2} G_{lk} = 1 \quad & \forall p,m\,. \end{aligned} \tag{2.3}$$

Under these constraint, the optimization in (2.2) is solved via sparse nonnegative matrix factorization (hence, the acronym). The details of the optimization can be found in Frichot, Mathieu, et al. (2014).

In order to select a value for K, the authors suggest to adopt a cross-validation scheme. A share (e.g., 5%) of all entries in $\widetilde{\mathbf{X}}$ is masked (artificially set as missing) and their value is predicted with (2.1). Note that this exclusion is not row-wise, but rather cell-wise. The optimization in (2.2) is carried out using the masked version of $\widetilde{\mathbf{X}}$ and the resulting matrices $\mathbf{Q}$ and $\mathbf{G}$ are employed to predict the probability of each masked genotype. The actual values of the masked entries that were excluded from the optimization are then used to compute a cross entropy criterion equal to the average negative log-probability of the masked allele counts. The best K corresponds to the lowest value of the cross-validation cross entropy.

Unlike STRUCTURE (Pritchard et al. 2000) and ADMIXTURE (Alexander et al. 2009), the sNMF framework does not assume Hardy-Weinberg equilibrium (see, e.g., Hedrick 2009, ch. 2) in the ancestral populations, yet delivers estimates of the ancestry coefficients comparable to them in absence of strong inbreeding (Frichot, Mathieu, et al. 2014). STRUCTURE is known to be at risk of overestimating K when there is spatial structure in the data (Frantz et al. 2009). Therefore, when using these algorithms to inform species delimitation considerations, it can be beneficial to integrate them with the approaches based on genetic and geographic dissimilarities described in section 1.3. The advantage of using sNMF lies in the fact that it is implemented in the `R` package `LEA` (Frichot and François 2015) and is estimated to be from 10 to 30 times faster than

ADMIXTURE (Frichot, Mathieu, et al. 2014) - which was in turn published as an optimized version of STRUCTURE.

Clustering routine A procedure is described here that delimits species using the sNMF algorithm and distance-based techniques from section 1.3 in tandem. The configuration selected via the cross-validation approach by sNMF serves as starting point for a merging routine involving a distance-based method. Pairs of groups are sorted according to a measure of cluster dissimilarity and sequentially tested for merging: if their genetic differences are compatible with their geographic separation according to a distance-based method (e.g., HH20), that is, it the test does not reject the null hypothesis of conspecificity, the two groups are lumped together. The cluster similarity ranking is then updated and the process continues until all individuals are grouped together or no test rejection occurs for any of the pairs still standing.

In the simulation study on the SLiM datasets presented in section 2.2, the sNMF algorithm was run five times for each investigated K. The maximum K explored with the algorithm changed according to the sample size: it was equal to 6 when $n = 12$, to 10 when $n \in \{30, 36\}$ and to 20 when $n \in \{90, 270\}$. The selected K at this stage was the one leading to the smallest average cross-entropy across the five runs. The estimates for $\mathbf{Q}$ and $\mathbf{G}$ were taken from the solution with smallest cross-entropy out of the five executions with the selected K. The information in these matrices was then used to carry out distance-based tests for merging. Obviously, no merging routine was required if the selected K was equal to 1.

The k^{th} column of matrix $\mathbf{G}$ stores allele count frequencies specific to group k, with $k = 1, \ldots, K$. It constitutes a summary of the genetic make-up observed in that group, thus can be used to quantify the overall dissimilarity between groups. One of the simplest ways to compute such a dissimilarity between group k and group k', with $k \neq k'$, is by using the Manhattan distance:

$$\sum_{l=1}^{12P} |G_{lk} - G_{lk'}| \,. \tag{2.4}$$

Formula (2.4) was adopted in this study to quantify group dissimilarity. Another option, not explored here, could employ a linkage criterion based on the shared allele distance between individuals (Kaufman and Rousseeuw 1990, ch. 5.5).

In the analysis of SLiM datasets, given the estimates for $\mathbf{Q}$ and $\mathbf{G}$ obtained according to the aforementioned criterion, the following routine began:

i. Individual i was assigned to the group it was most admixed with, i.e., finding k corresponding to the largest entry in $\mathbf{Q}_i$, the i^{th} row of matrix $\mathbf{Q}$.

ii. If groups with one or two individuals were present, they were merged with their most similar group according to (2.4). The first of these tiny groups to be merged was the one displaying the smallest group dissimilarity from any other group. This was done because distance-based methods cannot be applied when one of the two groups being compared has less than three members. However, this meant that no such group could ever be included in the final solution. This limitation may be tackled in future work. As long as the scenarios in section 2.2 are concerned, it was not too impactful since none of them involved species with these few individuals. On the other hand, in scenarios with $n = 12$, it could easily happen that more than one merging decision was made in this way, neglecting geographic information.

iii. Once all groups contained at least three individuals, all combinations of these groups in couples were ranked according to their group dissimilarity (2.4). From the least dissimilar to the most dissimilar, pairs of groups were tested for merging applying one of the tests described in section 1.3 - using log-transformed geographic distances: if a distance-based test failed to reject the hypothesis of conspecificity between the groups, they were lumped together. This led to the re-computation of the related entries in matrix $\mathbf{G}$: weighted averages of the allele count proportions were computed, based on the size of the two merged groups.

iv. Upon merging, a new ranking of group dissimilarities was produced and the search for the next pair of groups to be merged began.

v. The routine stopped when no more merging was suggested or when all groups were merged together.

Results Figure 2.2 reports the number of times any K was selected by the sNMF cross-validation procedure and by five versions of the clustering routine described above. Each version uses a different distance-based method in the merging phase of the routine: all techniques from section 1.3 except bootstrap-based partial Mantel tests were employed. The count related to the sNMF cross-entropy criterion is based on all of the one hundred dataset replicates for each combination of SLiM scenario, sample size and IBD behaviour. The count related to the five methods, instead, is based on datasets for which the initial configuration selected by sNMF had more than one group. Hence, it varies for each setup. In all cases, genetic datasets with 40 loci were analyzed. All tests were run at 5% significance level.

It is sensible to begin by looking at the effectiveness of the cross-entropy criterion to indicate the correct value of K. Although the modal choice seemed to correspond with the correct K in most non IBD scenarios with average sample sizes (top middle

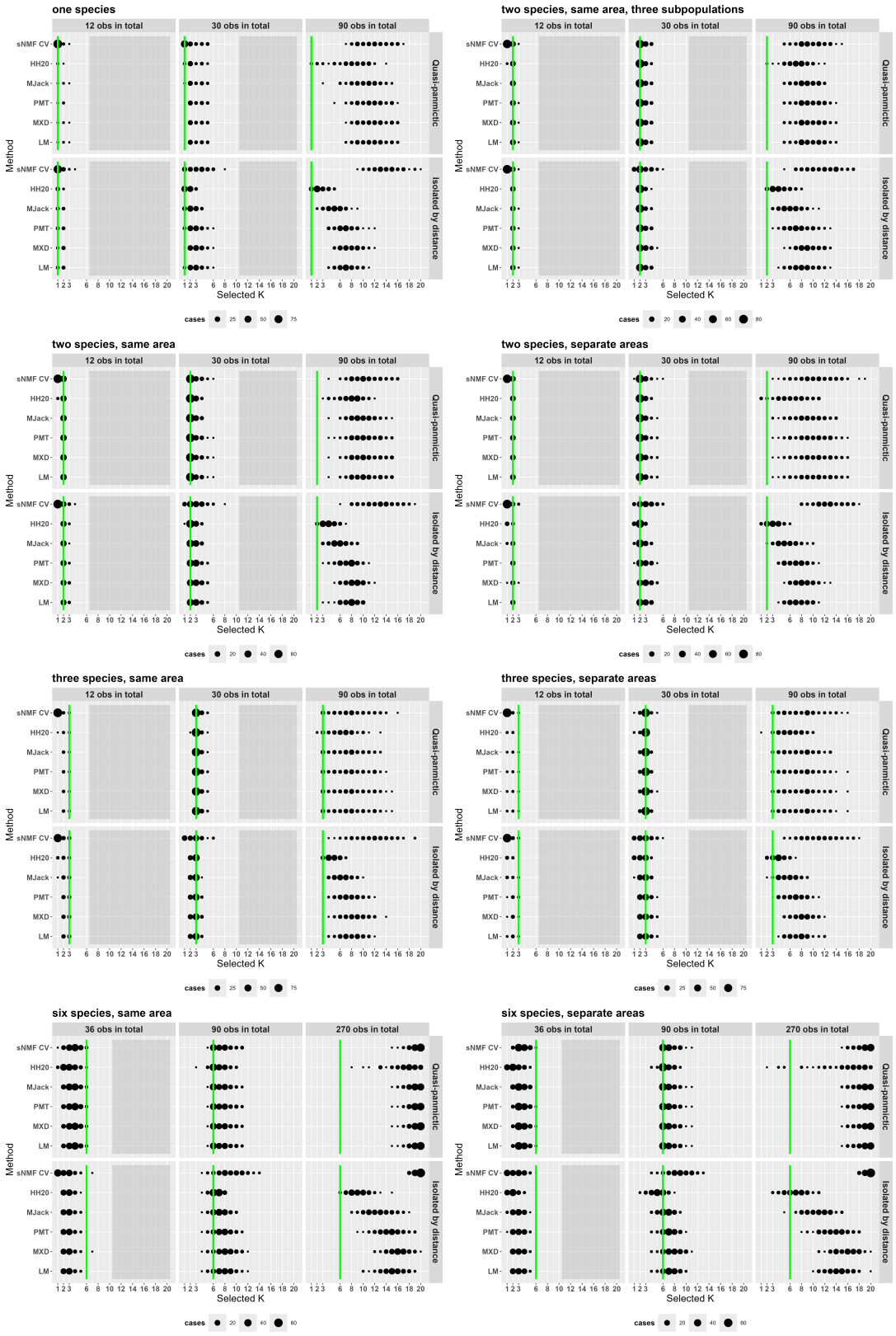

Fig. 2.2 Counts of times each K was selected by each version of the clustering routine (circle size proportional to count). The sNMF cross-entropy criterion was used on all 100 datasets from each setup (scenario, sample size and IBD behaviour), while the merging procedures based on five different distance-based methods were applied only when sNMF CV selected $K > 1$. The green line indicates the true K and the shaded area refers to Ks that were not explored by sNMF given the sample size. The same acronyms explained in section 1.4.1 are used to refer to the five versions of the merging procedure described in this section.

plot), in general this criterion was mainly influenced by sample size. Indeed, regardless of the IBD behaviour, K was underestimated in setups with small sample sizes and overestimated in those with large sample sizes. More in detail, isolation by distance interacted with large sample sizes leading to even larger estimated Ks, as it is expected given that sNMF does not leverage geographic information in its optimization. Isolation by distance patterns existing within the species called for additional groups to improve the fit of the solution.

As far as the merging routine is concerned, differences among the dissimilarity-based methods were more apparent in setups with IBD behaviour and larger sample sizes, and resembled those already observed in Chapter 1. As a premise, it is necessary to mention that groups with 3 individuals were sometimes hard to handle for MXD, where they interfered with the construction of profile confidence intervals. This resulted in additional merging decisions for this method. As regards HH20, these small groups often forced the preliminary test on equality of the within-group dissimilarity trends to be skipped, so that the hypothesis H_{02} was tested a priori.

Jackknife-based tests, such as those in HH20 and MJack, usually led to fewer rejections of the null hypothesis of conspecificity, resulting in more frequent cluster agglomerations. Permutation-based PMTs and LM showed rather similar figures, while the least prone to merge were MXD tests (despite the aforementioned bug). In setups with quasi-panmictic species, the only method consistently distinguishable from the others was HH20, whose empirical distribution of the chosen Ks was often shifted to the left or at least skewed in that direction with respect to the other methodologies.

The impact of geographic segregation can be assessed by comparing setups where distinct species inhabited the same area with those where they inhabited separate areas. While keeping the IBD behaviour constant, geographic segregation did not really influence the choice of K in a systematic way, especially if the cross-entropy criterion is considered. However, if one looks at its impact on jackknife-based methods in IBD setups, geographic segregation did lead to excessive merging. This was probably due to the identifiability issues existing in dissimilarity-based methods - discussed in section 1.5.1: when a positive relationship in the dissimilarities is present within the species, any sufficient geographic separation between them can lead to lack of evidence for rejecting the conspecificity hypothesis. This is indeed consistent with what observed in Chapter 1.

All in all, by looking at the green reference line in Figure 2.2, it might seem that the merging routine with HH20 represented the best strategy for choosing K in this simulation study. However, as argued above, this performance can be mostly explained as the result of the compensation of two biases: the overestimation of K using the

cross-entropy criterion in the initialization phase and the marked conservativeness of HH20 in the merging phase.

Furthermore, when merging techniques are considered, although it can be argued that the overestimation of K is partly due to the maximum explored K being larger for larger sample sizes, setups where this setting was not binding reveal an additional insight. In the last two scenarios, the maximum K was 20 both with average and with large sample sizes. There, one can see that, despite the initial Ks selected via cross-entropy being similar, scenarios with larger sample sizes led to fewer merging decisions across all methods. This happened because dissimilarity-based methods tested hypotheses that were data-driven since the groups involved had been selected via the optimization in sNMF. In other words, the hypothesis of conspecificity was tested for pairs of groups that had been previously separated *because of* their genetic discrepancies. This two-step procedure made it easier for tests to detect evidence of distinctness and possibly invalidated the related p-values, leading to liberal tests. The larger the sample size, the easier for sNMF to find differences in the groups, the harder for dissimilarity-based techniques to support merging decisions.

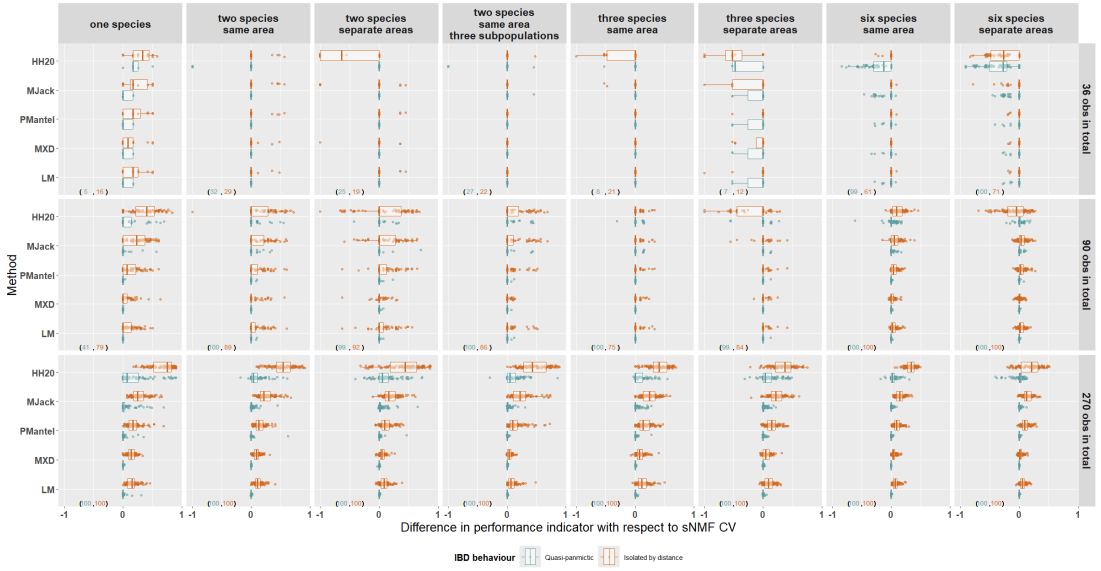

Fig. 2.3 Difference between performance indicator based on the final configuration reached by each dissimilarity-based method and same indicator based on the configuration selected by sNMF CV from which the merging routine started. Panels by scenario and sample size, groups by IBD behaviour. Given the IBD behaviour, the brackets in the bottom left corner report the number of cases for that scenario where the sNMF CV did not return $K = 1$ and a merging routine was applied.

To complete the picture, Figure 2.3 helps visualize to what extent each version of the merging routine improved the configuration selected via sNMF CV. It shows the difference between the value of a performance indicator obtained on clustering after

merging and its value obtained on the initial configuration. In the scenario with one species only, the performance indicator is the number of individuals assigned to the largest group divided by the total number of individuals in the dataset. In all other scenarios, it is the adjusted Rand index (Hubert and Arabie 1985). Since no merging phase took place when the cross-entropy criterion suggested to select $K = 1$, the number of replicates analyzed for each combination of scenario, sample size and IBD behaviour varies. This piece of information is reported in the bottom left corner of each panel.

Once again, the combination of the overestimation of K by the sNMF CV and the conservativeness of jacknife-based methods made them look as the best performing combinations for this task. In Figure 2.3, however, it is easier to see how this feature proved to be a two-edged sword in scenarios with more species: HH20, but also MJack, often worsened the initial solution, especially with geographic segregation. Nonetheless, MJack seems to be the strategy with the best trade-off between improvements and deteriorations, together with permutation-based PMTs. This is consistent with the findings from Chapter 1.

By looking at the panels in the one species scenario, we also observe how the cross-entropy criterion went from overestimating K five times out of 100 with small sample sizes to 41 times with average sample sizes when no spatial pattern of differentiation was present in the species. With IBD behaviour, instead, overestimations were more frequent, respectively 16 and 71 times. This confirms that clustering methods like sNMF that do not take spatial information into account tend to identify more groups in a genetic dataset than there actually are, because IBD patterns generate structure below the species level.

2.3.2 Admixture dissimilarity tests based on TESS3 configuration

TESS3 (Caye et al. 2018) can be seen as an extension of sNMF in which the estimation considers geographic information on top of genetic information. This program estimates matrices $\mathbf{Q}$ and $\mathbf{G}$ by minimizing:

$$||\widetilde{\mathbf{X}} - \mathbf{Q}\mathbf{G}^{\intercal}||_F^2 + \frac{1}{2}\sum_{i,j=1}^{n} \omega_{ij}||\mathbf{Q}_i - \mathbf{Q}_j||^2, \tag{2.5}$$

where:

- $\omega_{ij} = \exp(-d_x(\mathbf{z}_i, \mathbf{z}_j)^2/(\rho\bar{d_x})^2)$ is the weight associated with individuals i and j.
- $d_x(\mathbf{z}_i, \mathbf{z}_j)$ is the geographic distance defined in (1.1).
- ρ is a tuning parameter, whose default value is 0.05.

- $\bar{d}_x$ is the average geographic distance across all pairs of individuals.

This optimization is subject to the same constraints in (2.3). The inclusion of the second addend in (2.5) with respect to (2.2) enforces a geographically-informed regularity condition, such that individuals close in space tend to display similar admixture coefficients. Note that this does not model co-specific individuals that can be located far away, as it happens in many IBD datasets shown in Figure 2.1. The authors discuss two numerical methods to carry out the optimization in (2.5), whose details can be found in Caye et al. (2018).

Also in this case a cross-validation procedure is advised in order to choose K (Caye et al. 2018). However, the built-in cross-entropy criterion in TESS3 was found to severely overestimate K for the SLiM datasets presented in section 2.2. As it can be seen in Figure 2.4, excluding SLiM scenarios where only 5 loci were available, the most common selected value for K according to this criterion was the maximum explored one, regardless of the true number of species in the dataset. This performance looks poorer than that observed for sNMF, despite the inclusion of geographic information in TESS3 optimization. In the following, a testing procedure based on TESS3 output is described that can inform the choice of K without relying on its cross-entropy criterion.

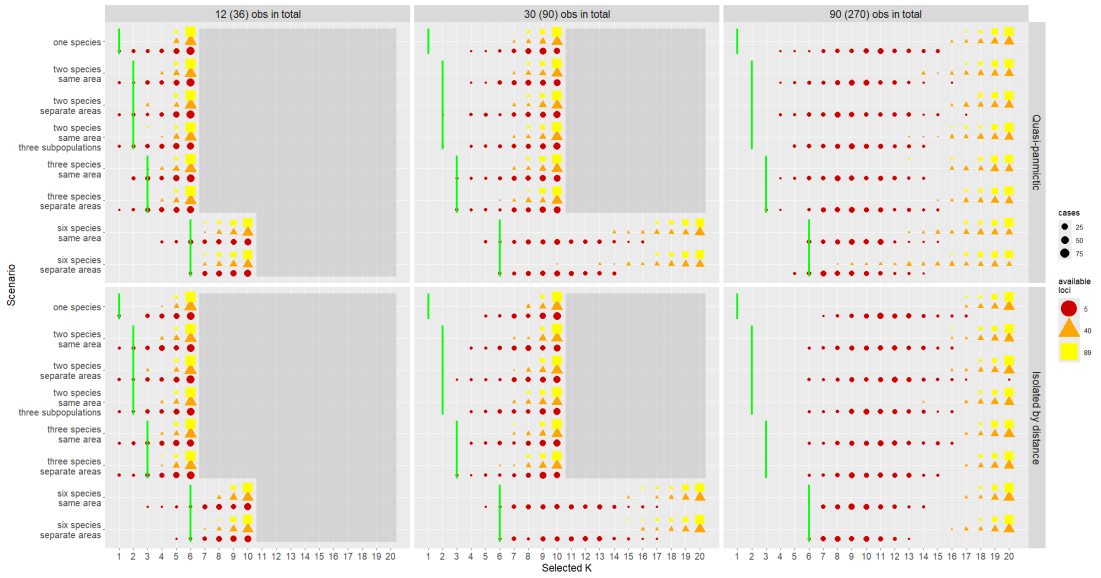

Fig. 2.4 Counts of times each K was selected by the cross-entropy criterion in TESS3 (symbol size proportional to count). Panels by sample size and IBD behaviour, color and shape by number of loci. The green line indicates the true K for the specific scenario; a grey rectangle covers unexplored values of K given the sample size. Cross-entropy based on 20 runs with 5% masked genotypes.

Jackknife-based tests for the choice of K As explained above, in admixture models the estimates in $\mathbf{Q}$ can be used to allocate individuals to the groups. The admixture coefficients specific to each individual can be compared to those of other individuals by means of admixture dissimilarities: these quantify how differently any pair of individuals inherited genetic material from the assumed ancestral populations. Given an initial TESS3 configuration with K groups, one can compute admixture dissimilarities and compare them with those based on a solution with $K+1$ groups. This comparison can happen in a regression framework, where genetic dissimilarities are regressed over geographic distances and these two admixture dissimilarities. If the solution with the additional group is able to explain some variation in the genetic dissimilarities even after controlling for spatial information and for the information embedded in the solution with K groups, this can constitute an indication to prefer it. Otherwise, there is no evidence that an additional group is needed and the solution with K groups can be selected.

To formalize this procedure, denote with $\mathbf{Q}_i^{(K)}$ the $1 \times K$ vector of admixture proportions for individual i estimated by a TESS3 execution with K groups. Define the admixture dissimilarity between individuals i and j based on a TESS3 solution with K groups as:

$$d_{q,K}(\mathbf{z}_i, \mathbf{z}_j) = \sum_{k=1}^{K} \left| Q_{ik}^{(K)} - Q_{jk}^{(K)} \right|, \tag{2.6}$$

where $Q_{ik}^{(K)}$ quantifies how much the i^{th} individual is admixed with the k^{th} ancestral population given that K of them were assumed. Note that, unlike with the cluster dissimilarity in (2.4), here the focus is on individuals and therefore the estimates in $\mathbf{Q}$ are used. The Manhattan distance is chosen also in this application because it treats all deviations between admixture coefficients equally. The distance in (2.6) takes its minimum value 0 when $\mathbf{Q}_i^{(K)} = \mathbf{Q}_j^{(K)}$ and its maximum value 2 if the two vectors are completely different. Its empirical distribution as K grows becomes more and more concentrated on the upper boundary and skewed to the left - see Figure A.26.

Note that admixture dissimilarities can be seen as a generalization of the grouping distance d_g defined in section 1.2. If one replaces the largest value in $\mathbf{Q}_i^{(K)}$ with 1 and sets all other values to zero, a "crisp" version of the admixture coefficients vector is obtained. (2.6) will then equal 1 if two individuals are allocated to different groups and 0 otherwise. By doing this in a setup with $K > 2$ groups, all between-group dissimilarities are treated in the same way, regardless of the groups where the specific individuals belong. The admixture dissimilarity, instead, more accurately quantifies the differences in the estimated ancestry of the individuals.

The following regression on dissimilarities can be fitted via least squares:

$$d_y(\mathbf{z}_r, \mathbf{z}_c) = \beta_0 + \beta_1 d_x(\mathbf{z}_r, \mathbf{z}_c) + \beta_2 d_{q,K}(\mathbf{z}_r, \mathbf{z}_c) + \beta_3 d_{q,K+1}(\mathbf{z}_r, \mathbf{z}_c) + e(\mathbf{z}_r, \mathbf{z}_c), \quad (2.7)$$

with $c < r \leq n$ and $\mathbb{E}(e(\mathbf{z}_r, \mathbf{z}_c)) = 0$. Since the admixture dissimilarities can be seen as a generalization of the grouping distance, the intuition behind the choice of the null hypothesis to be tested can be explained with reference to it. Indeed, if the solution with $K+1$ groups is to be preferred, this means that, with K groups, some individuals from distinct species have been assigned to the same group. The genetic dissimilarities between them are presumably larger than those these individuals have with other representatives of their species. However, the grouping distance will take value 0 for all of them. With an additional group, some of these individuals from distinct species that were lumped together will be possibly assigned to different groups: thanks to this, the grouping distance will gain explanatory power in terms of the genetic dissimilarities, since now many more genetic dissimilarities between individuals from distinct species will pair with grouping distances equal to 1. As argued in Chapter 1, tests on the grouping distances are indeed one-tailed, since in this framework the fact that genetic dissimilarities are larger when the grouping distance is equal to 1, controlling for geographic distance, is evidence for distinctness. Likewise, genetic dissimilarities can be expected to be larger when admixture dissimilarities are larger, given the geographic separation. As a consequence, we are interested in testing the null hypothesis that $\beta_3 \leq 0$.

Similarly to section 1.3.1, errors are not independent in (2.7), thus standard inference tools are not available. Therefore, we can again resort to jackknife. When $K = 1$, $\{d_{q,K}(\mathbf{z}_r, \mathbf{z}_c)\}_{c<r\leq n}$ consists in a vector of zeroes and is thus not included in the regression. With $K = 1$, it is tested whether including the admixture dissimilarity based on two groups in the regression adds explanatory power to geographic distance alone. Starting from such test with $K = 1$, the routine to determine K consists in the following steps: i) fit the model in (2.7) and jackknife test the null hypothesis that $\beta_3 \leq 0$; ii) upon failure to reject this null hypothesis, choose K as solution, otherwise set $K = K + 1$ and repeat. In this way, a value of K can be determined without having to choose in advance a maximum value for it to be investigated.

Results This routine to choose K was applied on all SLiM datasets described in section 2.2, and exploring various data richness setups: only 5 loci, 40 loci or 89 loci. Figure 2.5 displays the number of times any value of K was selected for a given dataset following the procedure described above. Log-transformed geographic distances were employed.

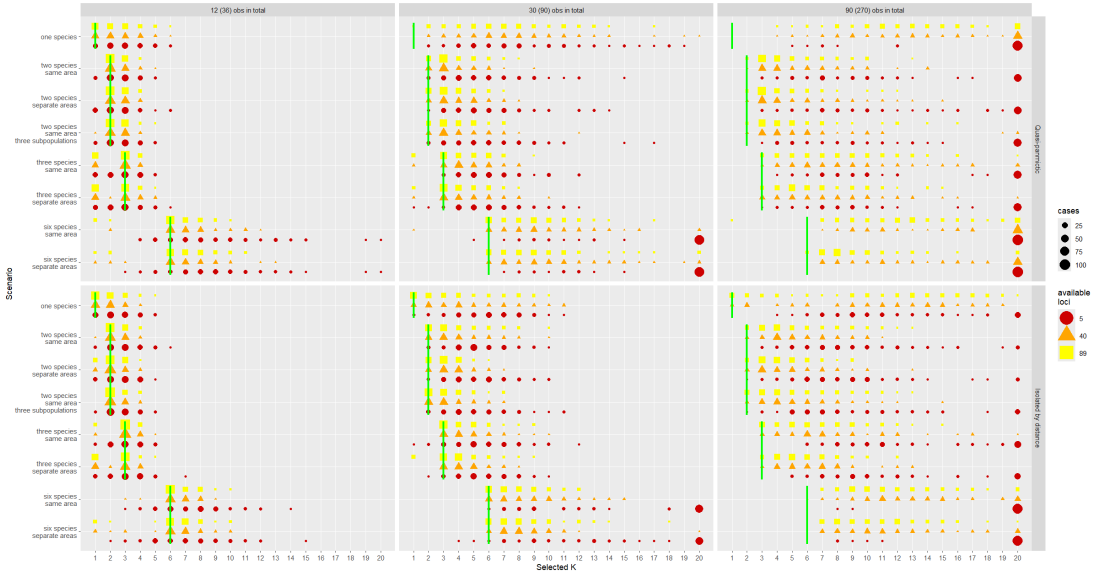

Fig. 2.5 Counts of times each K was selected by the admixture dissimilarity test (symbol size proportional to count). Panels by sample size and IBD behaviour, color and shape by number of loci. The green line indicates the true K for the specific scenario.

From a cursory inspection, it seems that the correct K was determined most of the times when datasets had small sample sizes, but the modal solution was systematically larger than the correct one in setups with bigger datasets. Besides that, since the maximum K explored was 20 for computational reasons, some censoring in the selection of K can be observed, especially in setups with large sample sizes or with 6 species: it could be that the strategy would have selected even larger Ks in those setups, had the data analyses not stopped at 20.

As far as patterns of isolation by distance are concerned, they interacted with geographic segregation, impacting the results on various levels. On one hand, when quasi-panmictic species were geographically clustered, this helped TESS3 to identify them. This is because the optimization in (2.5) favours similar admixture coefficients for close-by individuals, hence it is particularly suited for setups where homogeneous species inhabit separate areas. On the other hand, with species being isolated by distance, geographic separation could translate in lack of evidence that additional groups were needed for better fit. This resembles the behaviour of the merging routines in the previous section, and happened especially with small sample sizes, where we observe more underestimations of K when species inhabited separate areas.

In addition, the large amount of solutions with $K = 1$ in scenarios with 3 or 6 species can be due to the nature of the testing strategy. Since solutions with subsequent Ks are compared, it could happen that with, say, 3 species being represented in the dataset, there was not enough evidence to reject the hypothesis that $\beta_3 \leq 0$ in model

(2.7) specified with $K = 1$ (thus, comparing $K = 1$ vs $K = 2$). A possible remedy could be to fit a regression including all $\{d_{q,K}(\mathbf{z}_r, \mathbf{z}_c)\}_{c<r\leq n}$'s for $K = 2, \ldots, K_{\max}$, where $K_{\max}$ is some maximum explored value for K: the selected K could then be the largest such that the regression coefficient associated with $\{d_{q,K}(\mathbf{z}_r, \mathbf{z}_c)\}_{c<r\leq n}$ is significantly larger than zero. While possibly circumventing the aforementioned issue, this strategy would re-introduce the need to pick a maximum value for K to be explored in advance. Moreover, preliminary experiments (not reported) suggested that the performance of such strategy might be inferior to what is applied here.

Above all, the performance of this testing strategy for choosing K seems to be haunted by the same inferential issues highlighted for the merging routine described above. In (2.7), the genetic and geographic dissimilarities used to test for the need of an additional group are based on the same data fed to TESS3 in order to cluster the individuals in the dataset. Since admixture coefficients, and thus the dissimilarities based on them, are the product of TESS3 optimization, a test on $\beta_3 \leq 0$ at significance level α will be prone to display a type I error rate larger than α. Indeed, it will be often the case that an additional group better accommodates the heterogeneity in the dataset and this will be more apparent with larger sample sizes. The fact that the additional information provided by a longer sequence of available loci was associated with better results can be explained in terms of data dimensionality: the search of differentiated groups gets thornier when more variables are employed, and tests on their discrepancies lose power. This inefficiency mitigated the bias described above, leading to fewer rejections and to more solutions with smaller K. Despite the two clustering strategies not being comparable, it can be expected that the selection bias described here is more marked using TESS3 than using sNMF, because the latter does not optimize its cluster configurations according to geographic separation. The behaviour of these two experimental routines illustrates how two-step strategies of this kind, where differences between groups are tested based on the same information used to find the groups in the first place, have to be avoided or the described selection bias offset.

The following subsection provides an illustration of how the bias of the OLS estimated $\hat{\beta}_3$ of the coefficient in (2.7) interacted with the main data features in the SLiM scenario with two species inhabiting the same area.

A note on the bias of this test An experiment can be constructed to show that when model (2.7) is fitted using admixture dissimilarities based on TESS3 execution, $\hat{\beta}_3$ tends to be overestimated. This can be done with data from the SLiM scenario with two species that cover the whole geographic map.

In Figure 2.6, the value of $\hat{\beta}_3$ in the regression (2.7) fitted for $K = 1$ is reported for three different ways to obtain the matrix $\mathbf{Q}^{(2)}$ from which $d_{q,2}$ is calculated. Consistently

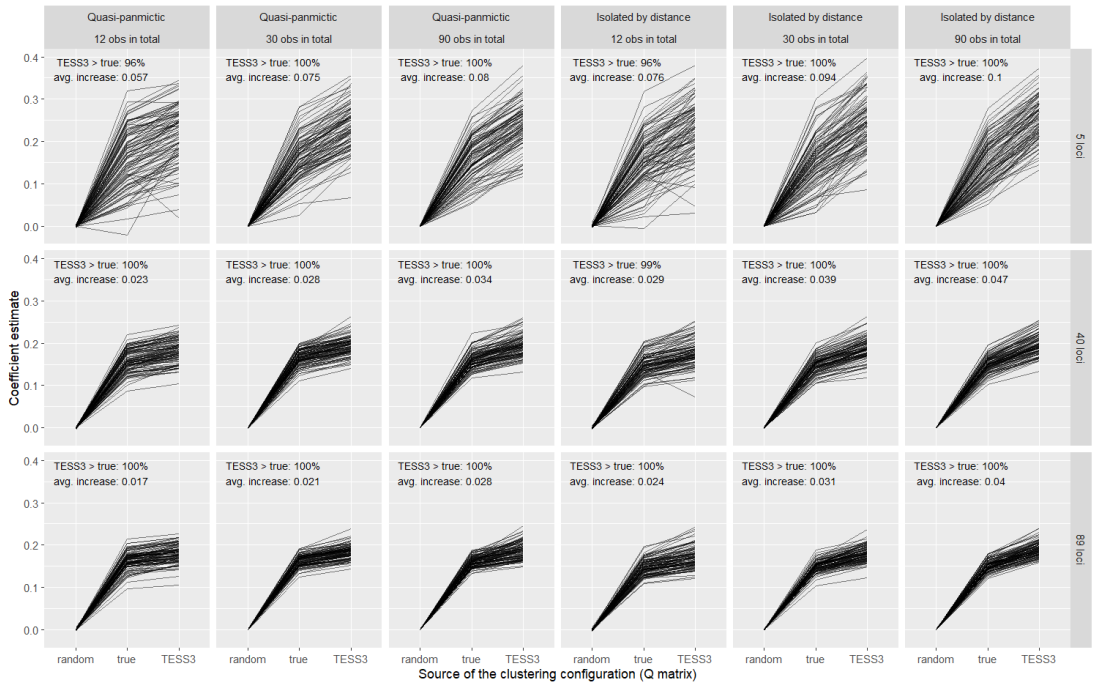

Fig. 2.6 Estimates of $\hat{\beta}_3$ based on regression (2.7) for the comparison between $K = 1$ and $K = 2$ obtained in three different ways, described in the text. Values that stem from the same dataset are connected by a line. All datasets are from the SLiM scenario with two overlapping species. Panels by IBD behaviour, sample size and number of available loci. The superimposed text in each panel reports the percentage of datasets for which TESS3 estimate was larger than that with true membership and the average positive shift between these two quantities across the one hundred datasets in that panel.

with previous charts, each panel reports results based on 100 SLiM datasets for each combination of IBD behaviour, total sample size and number of available loci. On the horizontal axis:

- the category labelled "random" refers to coefficient estimates obtained by generating 1000 replicates of $\mathbf{Q}^{(2)}$ in a random fashion. For $i = 1,\ldots,n$, $Q_{i1}^{(2)}$ is sampled from the Uniform distribution on the unit interval and then $Q_{i2}^{(2)}$ is set to be $1 - Q_{i1}^{(2)}$. The average value of $\hat{\beta}_3$ across the 1000 estimates obtained in this way is reported for every dataset;

- the category labelled "true" reports, for each dataset, the value of $\hat{\beta}_3$ associated with a choice of $\mathbf{Q}^{(2)}$ informed by true membership. If individual i belongs to the first species, $\mathbf{Q}_i^{(2)} = \{1,0\}$, while $\mathbf{Q}_i^{(2)} = \{0,1\}$ otherwise;

- for the category labelled "TESS3", the values of $\hat{\beta}_3$ originated from TESS3 solution with $K = 2$ optimized over 20 runs. Thus, the values in $\mathbf{Q}^{(2)}$ contained

admixture coefficients estimated by the algorithm as done for the analysis discussed in the previous subsection.

The three values obtained in this way for the same dataset are connected by a line in Figure 2.6, easing the comparison. As expected, random fuzzy configurations of the individuals in the two groups led to an estimate for the regression coefficient on average equal to zero, while true membership was associated with systematically larger values - apart from a couple of exceptions in setups with 5 loci. The estimates based on TESS3 were in turn larger than almost all corresponding estimates based on true "crisp" membership, that is, TESS3 was found to overestimate $\hat{\beta}_3$ in most of the cases.

This experiment also shows how the way this bias varies with parameters like n and P is consistent with findings in Figure 2.5. In Figure 2.6, the average increase in the estimated coefficient going from "true" to "TESS3" configurations was larger when the dataset contained more individuals and smaller when the number of loci increased. The variability in the estimates was more marked when fewer loci were available for the assessment of genetic differences between individuals, and with small n this could even result in some "TESS3" estimates being smaller than their "true" counterpart.

The use of "true" membership in this experiment reproduces the testing setting seen in Chapter 1, where the investigated dissimilarity-based methods would test hypothesis not driven by the data. In that setup, the grouping information employed in the testing procedure could be seen as fixed, while here it is derived by a preliminary analysis carried out in TESS3. The fact that a preliminary analysis informs the specification of the null hypothesis makes it adaptive, and testing adaptive null hypotheses invalidates the subsequent inference (Fithian et al. 2014). Estimates based on TESS3 being systematically larger than those based on fixed grouping suggest that the test in (2.7) tends to be liberal.

In the following, possible remedies to the limitations of the testing procedures described in this section are discussed.

2.4 Calibrating test statistics for the choice of K

This section considers new strategies to test whether more than one species is represented in a dataset and to estimate the number of species. Null models are constructed in the attempt to offset the bias observed in statistics like $\hat{\beta}_3$ in the regression (2.7), but also to calibrate measures of clustering quality that are expected to decrease as the number of clusters grows. In section 2.4.1, a weighted null model based on TESS3 is proposed and challenges in its parametrization are discussed. In section 2.4.2, three statistics are considered that can be used to test whether the true K equals 1 or to estimate the best

K. Then, a selection of SLiM datasets from section 2.2 is used to compare the type I error and power properties of these statistics when calibrated with three versions of the weighted null model or using a simpler uniform null model. Lastly, some alternative strategies to null models are mentioned in section 2.4.3.

The role of calibration Given the requested number of ancestral populations K, TESS3 will return a clustering configuration with K groups even if the dataset contains only one species. Model-based clustering algorithms like this assume that K ancestral populations are to be found, and will describe the genotypes as a mixture of the contributions of these populations - regardless of species delimitation considerations. Discrepancies between the K co-specific groups may be explained by spatial patterns of differentiation, if present. However, even with panmictic species displaying no geographic segregation, TESS3 might capture noise patterns and the resulting groups will still look different, since their differences informed their identification.

Therefore, the objective is to discriminate clustering solutions where distinct species are being delimited from those where the assignment is guided by within-species heterogeneity or just noise. To this aim, the value of statistics like $\hat{\beta}_3$ in section 2.3.2 obtained on the original dataset can be compared with their distribution under appropriate null models (calibration). Null models attempt to capture all features of the data that cannot be interpreted as clustering information - see Huang et al. (2015) for a review. Ideally, for species delimitation as tackled in this project, datasets simulated from the null model should have same number of individuals, number of loci, geographic locations and IBD patterns of the original dataset. Indeed, as exemplified by SLiM scenarios where distinct species cover the same area of the map, any strictly geographic grouping is not necessarily associated with species distinctness. Moreover, mimicking IBD patterns in the data is crucial because a dataset containing only one species might display heterogeneity due to spatial patterns: data from a null model neglecting this may then look significantly different from the original dataset, leading to the selection of configurations with $K > 1$ in a setup where the goal is to conclude that $K = 1$. On the other hand, the recovery of these data features should not cause any relevant clustering information to enter the null model: in this case any test relying upon it is likely to have little power when the data truly contains more than one species.

Null models can be used to construct tests for the presence of more than one species in the dataset ($K = 1$ versus $K > 1$), but also to estimate the number of species. As explained in Hennig and Lin (2015), while null models try to replicate all data features that do not constitute clustering information, the clustering alternative is implied by the choice of the test statistic employed. The value that this quantity assumes on null datasets has to be systematically different from that it takes on datasets where there

truly is more than one species. Statistics like $\hat{\beta}_3$ in the comparison between $K = 1$ and $K = 2$ or the partial Mantel correlation defined in (1.18) are associated with a clustering alternative expressed in terms of dissimilarities: admixture or grouping dissimilarities are positively associated with genetic distances even after controlling for geographic dissimilarities, meaning that the heterogeneity in the dataset cannot be explained only in terms of isolation by distance. The presence of more than one species can also be detected by looking at how TESS3 solutions evolve as K increases, as explained below.

The ΔK method In molecular ecology, a widely used statistic for determining K is the ΔK proposed by Evanno et al. (2005). This statistic is connected with the search for elbows on charts where some clustering error measure is plotted against K. In TESS3, this measure is the root mean squared error:

$$\text{RMSE}_K = \sqrt{\left(\frac{\sum_{i=1}^{n} \sum_{l=1}^{12P} \left(\widetilde{X}_{il} - \sum_{k=1}^{K} Q_{ik} G_{lk} \right)^2}{12Pn} \right)}, \tag{2.8}$$

that quantifies how accurate the estimation of allele counts for each individual and each locus is when K ancestral populations are assumed. Unlike the cross-validation cross-entropy criterion, which is supposed to hit its minimum at the best K, (2.8) tends to decrease as K grows. Theoretically, given an indermediate configuration with $K < n$, it is always possible to introduce an additional ancestral population and split the individuals in one of the K starting groups so to better describe their genotypes, up to the point where $K = n$ and each column in $\mathbf{G}$ reports the allele count frequencies observed for each individual (zero RMSE). In practice, due to the numerical optimization in TESS3, a random TESS3 run with $K + 1$ can actually display larger RMSE than one with K groups, but on average a decreasing trend emerges upon the execution of a sufficiently large number of runs for each K. On top of this, often TESS3 solutions with $K = n$ do not return a $\mathbf{Q}$ matrix where each row contains only one positive value, i.e. they do not necessarily suggest a one-to-one relationship between ancestral populations and individuals. This is consistent with the idea that admixture models assign alleles to the ancestral populations, not individuals (François and Waits 2015). In these models, it is possible to assume that the number of ancestral populations from which the individuals in the dataset derived their genetic material is $K > n$, and TESS3 can return a configuration accordingly. However, experiments on SLiM scenarios (not reported) showed how, on average, the RMSE for TESS3 solutions with $K > n$ tends to be larger than the minimum hit at $K = n$ and becomes considerably instable.

Given that measures like the RMSE are expected to decrease with K, the rationale of the elbow search is that adding one more group to the solution brings about stronger improvements when the best K is yet to be reached, while, once it is overcome, additional groups are not that useful anymore. The change in the rate of improvement will result in a flattened trend, generating an elbow. This heuristic dates back at least to Thorndike (1953). In the field of molecular ecology, it was translated in mathematical terms by Evanno et al. (2005). They tailored it to the built-in error measure in the STRUCTURE program (Pritchard et al. 2000), the posterior probability of the data for a given K, but it can be applied on (2.8), too. The visual inspection of the elbow plot is replaced by the maximization of a function of the second order rate of change of the average RMSE:

$$\begin{aligned} \text{RMSE}'_K &= \overline{\text{RMSE}}_K - \overline{\text{RMSE}}_{K-1}\,, \\ \Delta K &= \frac{|\text{RMSE}'_{K+1} - \text{RMSE}'_K|}{s(\text{RMSE}_K)}\,, \end{aligned} \tag{2.9}$$

where $\overline{\text{RMSE}}_K$ and $s(\text{RMSE}_K)$ are the mean and standard deviation of the RMSE across all TESS3 runs executed for a given K. The chosen K is the one that maximizes ΔK. Evanno et al. (2005) motivated the division by $s(\text{RMSE}_K)$ by observing that the variance of the posterior probability in STRUCTURE tends to increase with K. However, this is not the case for the root mean squared error in TESS3, whose variability only soars for $K > n$.

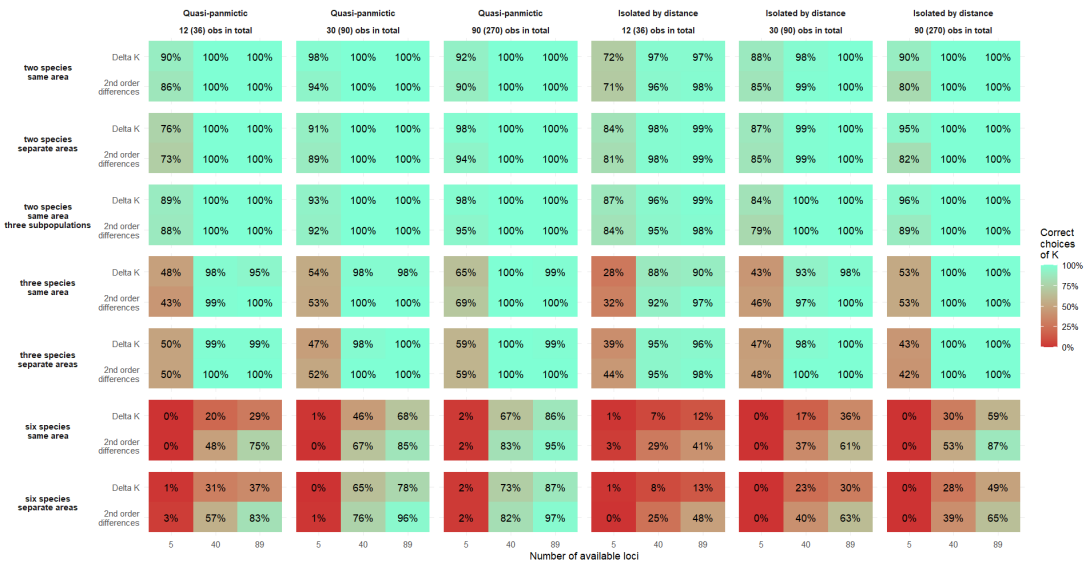

Fig. 2.7 Proportion of datasets for which the heuristics to choose K selected the number of species. Each tile of the heatmap refers to a combination of heuristic rule and number of available loci. Panels by scenario, IBD pattern and sample size. The number of TESS3 runs for each value of K was 20. Maximum K per scenario chosen as in section 2.3.1.

Figure 2.7 helps assess the efficacy of the ΔK rule in recovering the number of species in the SLiM datasets. In order to check whether dividing by $s(\text{RMSE}_K)$ does any good when using this statistic with the RMSE from TESS3, the Figure also shows how accurate were the choices when $\Delta K \times s(\text{RMSE}_K)$ was maximized. As apparent from (2.9), this quantity amounts to the absolute value of the second order differences (SODs) in the $\overline{\text{RMSE}}_K$ trend. In scenarios with two or three species, these heuristic rules showed almost perfect results, provided that at least 40 loci were available. In these SLiM scenarios, second order differences did slightly worse than ΔK. However, the overall performance strongly deteriorated in scenarios with six species, where dividing SODs by $s(\text{RMSE}_K)$ was found to have a negative impact on accuracy. Maximizing absolute second order differences led to success rates equal or larger than those obtained using ΔK in 100 out of 126 setups shown in Figure 2.7. In addition, when ΔK's rate was lower, its gap from SODs' rate was on average four times larger than that observed in the cases where ΔK's rates were better. This performance can be imputed to the low variability sometimes displayed by RMSE values across TESS3 runs for a given K: in scenarios with six species, TESS3 configurations with $K < 6$ often recorded very small values of $s(\text{RMSE}_K)$, generating a peak in the ΔK trend. This peak would frequently stand out with respect to the ΔK value at $K = 6$. All in all, these simulations suggest that it might be better to rely upon absolute SODs than to ΔK when using TESS3 to determine the best value for K.

As far as other data features are concerned, larger sample sizes and longer available genetic sequences helped improve the accuracy of the methods, in line with the conclusions in Evanno et al. (2005). The presence of within-species IBD patterns, instead, was associated with worse results.

Regardless of the specific criterion that is maximized to determine K, elbow rules can never suggest that the best K to be chosen is 1. On SLiM scenarios with one species, the empirical distribution of the Ks selected via the ΔK rule had mode on $K = 2$ and was skewed on the right. Somewhat unexpectedly, the presence of IBD patterns did not seem to have an impact. With the aim of testing whether a dataset is homogeneous, i.e., contains only one species, the trend in the SODs can be contrasted with that observed under a null model. The absence of deviations from the SODs trend based on null data can constitute evidence that none of the configurations returned by TESS3 for $K = 2, \ldots, K_{\max}$ is sufficiently different from what could be expected by partitioning homogeneous data.

In the following, possible strategies to construct null models are discussed. They will be used to calibrate statistics like $\hat{\beta}_3$ from model (2.7), the partial Mantel correlation defined in (1.18) and second order differences in the trend of TESS3 RMSE values.

2.4.1 Null models

As delineated above, in order to construct tests of homogeneity ($K = 1$ versus $K > 1$) or routines to determine the best value of K, it is required: a) to conceive a null model that imitates all the features of the dataset in hand that cannot be interpreted as clustering information and b) to choose a test statistic sensitive to the presence of the clustering structure of interest. If TESS3 is used to cluster individuals, it is crucial to apply it both on the original and the null data, a minimum requirement for the comparability of the resulting test statistics.

This section deals with the challenges in the development of null models and with their interaction with the selection of the test statistic to use.

Uniform null model One of the simplest ways to construct a null model in this context is to only use the information about the area inhabited by all individuals and the genotypes observed in the dataset. For dataset like those output by SLiM, which feature generic eastings and northings on a two-dimensional map, individuals under this null model can be uniformly sampled from the rectangular area covered by the original ones. Their genetic make-up can be re-created via a random resampling with replacement of the diploid loci observed in the dataset. This null model discards information about any geographic clustering and generates completely panmictic datasets, with no structure whatsoever. It only reproduces the original number of individuals and loci, and merely preserves the range of the original geographical coordinates.

By neglecting information on geographic structure and spatial patterns of differentiation, this null model can be expected to lead to liberal tests of homogeneity. For instance, genetic discrepancies between geographically segregated groups from a species with IBD behaviour are likely to constitute a deviation from what is expected to be observed under this uniform null model.

Weighted null model A more data-influenced null model can be conceived that is based on resampling loci at the original locations of the individuals in the dataset. Since the geographic position of individuals is not necessarily associated with species membership, it can be exactly reproduced in null datasets. Based on these positions, spatial weights can be computed for each pair of individuals and used to inform a resampling of the genetic make-up at each position. An appropriate quantification of the weights may help reconstruct the spatial patterns of differentiation in the datasets, allowing to recover all data features a null model needs for our aims.

Given the locations of the n individuals, the i^{th} row in the original genotypic matrix $\mathbf{X}$ can be resampled based on spatially-weighted allele count frequencies:

$$\boldsymbol{\xi}_i = \frac{\boldsymbol{\omega}_i \mathbf{X}}{2\boldsymbol{\omega}_i \mathbf{1}_n}, \tag{2.10}$$

where $\boldsymbol{\omega}_i = \{\omega_{ij}\}_{j=1,\ldots,n}$ is the $1 \times n$ vector of weights related to individual i and $\mathbf{1}_n$ is a $n \times 1$ vector of 1s. These weights are the same used in (2.5) for TESS3 optimization, provided that $\rho = 0.05$. Assuming once again that $\mathbf{X}$ has $4P$ columns, each of the P sub-vectors in $\boldsymbol{\xi}_i$ will contain four elements that sum to one. These can be used as event probabilities in a multinomial experiment. For example, the vector $\{\xi_{3,5}, \xi_{3,6}, \xi_{3,7}, \xi_{3,8}\}$ is used to resample the second locus of the third individual in the dataset. In general, the draw of the two alleles for individual i at locus p is modelled as a multinomial random variable with probability mass function equal to

$$P(X_{i,\widetilde{p}+1} = x_{i,\widetilde{p}+1}, \ldots, X_{i,\widetilde{p}+4} = x_{i,\widetilde{p}+4}) = \frac{2}{x_{i,\widetilde{p}+1}! \cdots x_{i,\widetilde{p}+4}!} \xi_{i,\widetilde{p}+1}^{x_{i,\widetilde{p}+1}} \cdots \xi_{i,\widetilde{p}+4}^{x_{i,\widetilde{p}+4}} \tag{2.11}$$

when $x_{i,\widetilde{p}+1} + x_{i,\widetilde{p}+2} + x_{i,\widetilde{p}+3} + x_{i,\widetilde{p}+4} = 2$ and to 0 otherwise. Recall that $\widetilde{p} = 4(p-1)$. This drawing mechanism assumes that loci are independent.

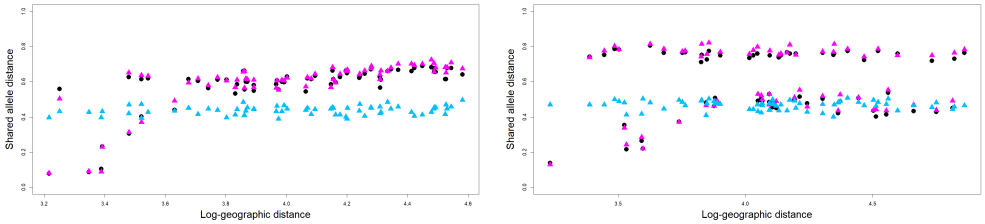

Fig. 2.8 Distance-distance plots from a SLiM dataset with one species (left) and a dataset with two overlapping species (right). Both datasets have $n = 12$ and $P = 134$. Black circles are distances based on original data, blue triangles are from a weighted null model with $\rho = 5$ and magenta triangles are from a weighted null model with $\rho = 0.05$.

The choice of ρ in the weights has a remarkable impact on the null datasets generated with this model. In Figure 2.8, dissimilarities from dataset replicates obtained with this null model are superimposed to the original dissimilarities from two exemplary SLiM datasets - taken from the first two scenarios described in section 2.2:

- when the parameter $\rho = 0.05$ (its default value in TESS3 optimization), the resulting dataset is very similar to the original one, therefore it possesses the same IBD pattern and variability of the genetic dissimilarities. However, this comes at the cost of absorbing clustering information, since also in the null dataset it is possible to spot an upward shift in the genetic dissimilarities that perfectly

matches species membership (plot on the right). As a consequence, tests based on similar null datasets will hardly lead to the rejection of the homogeneity hypothesis, as null datasets will be too similar to the original one.

- when ρ is set to 5, the dissimilarities look more homogeneous, but they also follow a flat trend and there is less genetic variability. Calibrating test statistics with null datasets like this where IBD patterns are not reproduced might yield similar results to when uniform null models are employed.

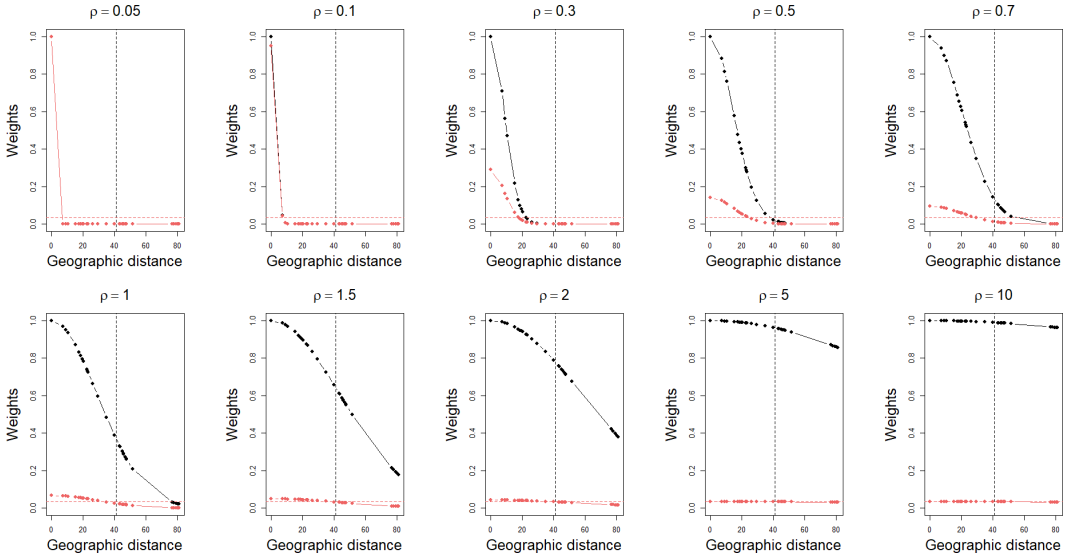

Fig. 2.9 Weights $\boldsymbol{\omega}_i = \{\omega_{ij}\}_{j=1,\dots,n}$ (black) and their scaled version $\frac{\boldsymbol{\omega}_i}{2\boldsymbol{\omega}_i \mathbf{1}_n}$ (red) against geographic distance, for a random individual i from a SLiM dataset with $n = 30$ individuals. Panels by value of parameter ρ. ω_{ij} is defined in (2.5), scaled weights in (2.10). The vertical black line indicates the average geographic distance $\bar{d}_x$, whereas the horizontal line indicates the uniform weight $1/n$.

The factor ρ tweaks the fraction of the overall average geographic distance that is used to normalize d_x in the computation of the weights. Figure 2.9 shows how the weights used to resample the loci of a random individual i in a SLiM dataset with $n = 30$ vary with ρ. For instance, an individual found one $\bar{d}_x$ away and another individual two $\bar{d}_x$'s away from individual i are given weights approximately equal to 0.37 and 0.02, respectively, when $\rho = 1$; when $\rho = 2$, instead, these weights would be approximately equal to 0.78 and 0.37, respectively. Thus, the closer of the two individuals goes from having a weight ~ 18 times larger than that of the further individual to only having twice its weight, approximately. With values as low as 0.05, only the individuals closest to the i^{th} location will have a non-negligible impact on the re-sampling of its genetic material. More precisely, since ω_{ii} is included in $\boldsymbol{\omega}_i$, too, the original genetic make-up at that location will have the largest weight in its resampling, leading to null datasets very close

the original one. Setting $\omega_{ii} = 0$, however, does not solve the problem because now the i^{th} location is filled with genetic information mostly based on its closest neighbour. As ρ increases, more and more individuals are assigned a positive weight and scaled weights converge to the reciprocal of the sample size n. In practice, with values as large as 5, all individuals contribute almost equally to the resampling of the genetic material at the i^{th} location. This smooths genetic discrepancies and causes the loss of spatial patterns of differentiation.

Since the choice of ρ is connected to this trade-off, there might be an optimal value between those considered in Figure 2.8 that hits a good balance between recovering key IBD patterns and preserving the conspecificity requirement for null datasets. In order to inform the choice of ρ in a data-driven way, a possibility is to explore for what values of ρ some key data features are reproduced. The slope coefficient estimate in the simple linear regression between genetic and log-transformed geographic dissimilarities can be used as proxy to describe the IBD pattern that the null model needs to reproduce. In the same spirit of Hennig and Hausdorf (2004), the relationship between ρ and this coefficient can be investigated by simulating datasets from the weighted null model over a set of ρ values. The resulting trend can be captured by, e.g., linear interpolation. That is, points with coordinates consisting of each ρ value and its associated slope estimate can be connected by straight lines. The resulting polygonal chain can be used to map any value of ρ within the covered interval to an estimate of the slope. The value of ρ associated to the slope parameter estimate on null data that is closest to that observed on the original dataset can then be selected. Null datasets generated with the selected ρ parameter are used for test statistic calibration.

On SLiM datasets from scenarios with one, two or three species inhabiting the same area, the association between the parameter ρ used in the null model and the resulting slope tended to be non-linear. As shown in Figure 2.10, at $\rho = 0.05$ slope estimates usually resemble the value they take on the original data. An increase follows for ρs in the interval $(0.3, 0.5)$ and then estimates fall down, often below the starting point. The dissimilarity plots in Figure A.27 show how this is the result of factors that intervene in sequence: at first, genetic dissimilarities between individuals at short geographic distance are smoothed, causing an increase in the correlation between dissimilarities - which usually hits a maximum between $\rho = 0.3$ and $\rho = 0.5$. Then, the reduction in the variability of genetic dissimilarities kicks in, followed by the progressive flattening of the trend due to large ρ values enforcing quasi-panmictic behaviour.

In Figure A.28, the standard deviation of genetic dissimilarities for SLiM scenarios is reported, both for original data and for null datasets with varying value of ρ. Larger values of ρ tend to yield more homogeneous datasets, with diminished genetic variability. The impact of the parameter is mitigated by clustering structure and by the absence of

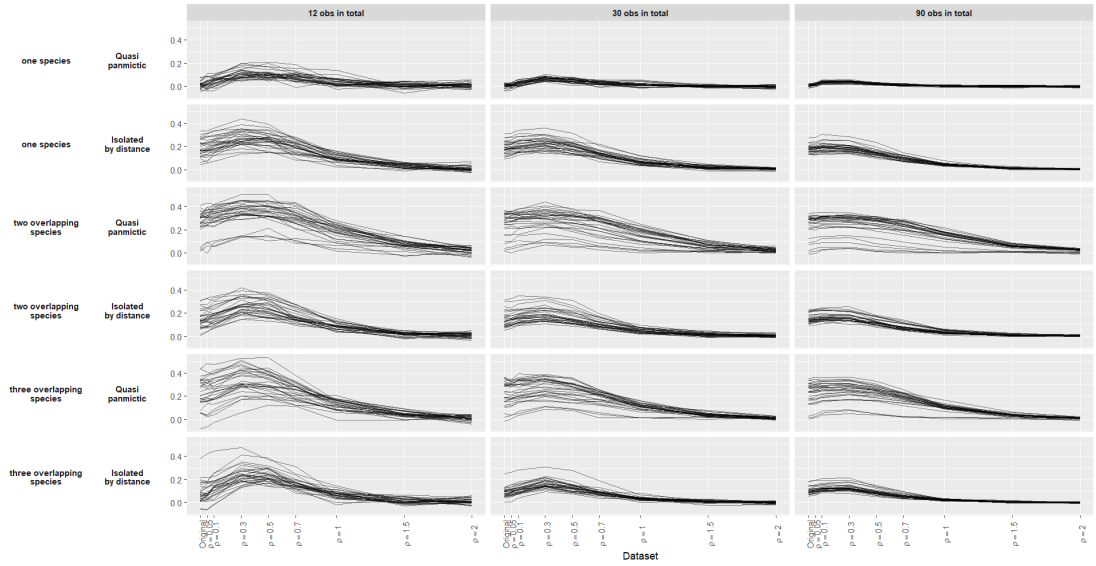

Fig. 2.10 Slope estimates from the simple linear regression between genetic and log-transformed geographic distances on original SLiM data or on data simulated with the weighted null model according to various values of the parameter ρ. Connected estimates pertain to the same SLiM dataset. Panels by scenario, IBD behaviour and sample size. 30 datasets per scenario, all with 40 loci.

IBD patterns: in datasets with IBD patterns, larger initial standard deviations emerge due to tiny genetic dissimilarities, which are rarely observed in quasi-panmictic species generated with these SLiM scenarios - see also Figure A.27.

In practice, the search for the value of ρ has to be constrained to some interval. Clearly, when $\rho \to 0$, the null dataset converges to the original one because all individuals $j \neq i$ have approximately zero weight in the resampling of the genetic information at location i. This is also why data from the original dataset was shown on the leftmost horizontal position in Figures 2.10, A.27 and A.28. Any difference still visible in this limit situation would be attributed to the multinomial law used to resample loci. Thus, while it is natural to consider positive values of ρ, the inclusion of values as low as 0.05 (default weight in TESS3) can be considered as last resort to retrieve the original slope observed in the dataset when larger values fail to do so. However, in Figure 2.10 the non-linear trend between slope estimates and ρ values suggests that the original slope can often be recovered with values of ρ above 0.3. This motivates a top-down search for ρ, and one can check experimentally how often this estimation procedure ends up selecting values close to the lower boundary.

Setting an upper boundary can also be hard, but heuristic rules can be adopted by observing how weights (Figure 2.9), slope estimates (Figure 2.10) and variability of $\mathbf{D}_y$ (Figure A.28) tend to vary with ρ. Since the scaled weights converge to $1/n$ as ρ grows, after some value of ρ null datasets will cease to evolve and will stabilize on

homogeneous versions of the original dataset with no IBD pattern. Hence, one can check for what value of ρ all scaled weights for all individuals are sufficiently close to their limit value $1/n$. Namely, the upper boundary for the search of ρ, denoted $\rho_{\max}$, can be set such that:

$$\left| \frac{\omega_{ij}}{2\boldsymbol{\omega}_i \mathbf{1}_n} - \frac{1}{n} \right| < \frac{2}{n} \quad \forall i, j\,.$$

This rule adapts to the total sample size n: for example, no scaled weight will be larger than $1/4$ with $n = 12$, or larger than $1/10$ with $n = 30$ or larger than $1/30$ with $n = 90$. For the SLiM datasets analyzed in this section, this criterion was always fulfilled with $\rho < 2$. Indeed, for similarly large values of ρ, null datasets tend to reach limit values of the regression slope between genetic and log-geographic distances and of the variability of genetic dissimilarities.

Once $\rho_{\max}$ is set, the following algorithm can be adopted to inform weighted null models with a data-driven estimate of parameter ρ. Given a dataset to be clustered:

i. obtain the OLS estimate of the slope of the regression between genetic and log-transformed geographic distances and set it as target slope to be replicated;

ii. simulate datasets based on the weighted null model with $\rho = 0.05$ and $\rho = \rho_{\max}$ plus, say, eight equally spaced values in between and interpolate the resulting estimates of the slope;

iii. starting from $\rho_{\max}$ and going down, use the interpolation to find the value of ρ that returns a null dataset with the slope closest to the target.

This allows to use a dataset-specific value of the ρ parameter instead having to rely upon a unique arbitrary value for all datasets to be analyzed.

Null datasets from simulators With different levels of sophistication, uniform and weighted null models generate datasets by resampling geographic and genetic information from the original dataset. An alternative approach to obtain null datasets can rely on the use of simulators like SLiM, GSpace or the algorithm described in section 1.5.1. This, however, poses the problem to inform all the parameter choices involved. As an example, pinning down a timespan for the forward simulation in SLiM that is compatible with the observed data can be thorny. Moreover, its default options do not allow to control the position of the sampled individuals. GSpace, on the other hand, allows to replicate the original positions of individuals, but several other parameters need to be estimated from the data anyway. In relative terms, using the simulator described in section 1.5.1 might be easier because fewer parameters are required. However, each locus would have to be associated with a linear gradient on the map, which can be

either incompatible with the data or technically cumbersome. As shown above for the weighted null model, even the estimation of a single parameter, ρ, can be challenging. The relationship between any null model parameter and observed quantities (e.g., the slope targeted above) can be complex and often arbitrary choices have to be made or heuristic criteria have to be adopted for the estimation. In this regard, null models that rely upon resampling strategies and involve one or no tuning parameter have the advantage of simplicity. In any case, the use of simulators for the calibration of test statistics can be addressed in future research.

In the next section, the performance of three different test statistics calibrated through uniform and weighted null models is assessed on a selection of SLiM datasets.

2.4.2 Test statistics

A wide range of statistics can be calibrated by means of the null models described in the previous section. Three options are:

- the β_3 coefficient in (2.7), whose OLS estimate for every K can be compared with that of a sufficiently large number of replicates obtained from null model datasets. Ideally, if null models successfully replicate all features of the original dataset that cannot be interpreted as clustering, the value of $\hat{\beta}_3$ on the original data will be larger than its null replicates only when individuals are truly clustered. Hence, if $\hat{\beta}_3$ on the original data is larger than, e.g., 95% percent of these replicates, the null hypothesis that $\beta_3 \leq 0$ can be rejected. Not only this statistic can be used to test the hypothesis that the dataset contains more than one species, it can also serve as method to select the best value of K;

- the partial Mantel correlation in (1.18) can be computed using a grouping distance based on TESS3 solution with $K = 2$. Analogously to $\hat{\beta}_3$, also $r(\mathbf{D}_y, \mathbf{D}_g | \mathbf{D}_x)$ can be compared with its null distribution to reject the hypothesis that its true value in the population is zero. As is, this test only provides an indication whether more than one species is present in the dataset and cannot be used to determine the best value of K;

- a different statistic can be constructed based on TESS3 RMSE values. Denote with $\text{RMSE}_{K,b}$ the minimum RMSE returned by TESS3 runs with K groups on

the b^{th} dataset generated under the null model, with $b = 1, \ldots, B$. Define:

$$\begin{aligned} \mathrm{RMSE}'_{K,b} &= \mathrm{RMSE}_{K,b} - \mathrm{RMSE}_{K-1,b}\,, \\ \mathrm{RMSE}''_{K,b} &= \mathrm{RMSE}'_{K+1,b} - \mathrm{RMSE}'_{K,b}\,, \\ \overline{\mathrm{RMSE}}''_K &= \frac{\sum_{b=1}^{B} \mathrm{RMSE}''_{K,b}}{B}\,, \\ s(\mathrm{RMSE}''_K) &= \sqrt{\frac{\sum_{b=1}^{B} (\mathrm{RMSE}''_{K,b} - \overline{\mathrm{RMSE}}''_K)^2}{(B-1)}}\,. \end{aligned} \tag{2.12}$$

Denote with RMSE''_K the analogue of $\mathrm{RMSE}''_{K,b}$ based on the minimum RMSE values across TESS3 runs with K groups on the original dataset. With B sufficiently large, RMSE''_K can be compared with the quantiles of its empirical distribution under the null model: the null hypothesis that the true value of RMSE''_K in the population is not larger than 0 can be rejected when its estimate is larger than, e.g., 95% of its null model replicates. The one-sided alternative is justified by the fact that positive second order differences are expected when an elbow occurs. With lower values of B, other options for significance assessment might be preferable. In general, for any real-valued random variable Y and for $t > 0$, the Chebyshev-Cantelli inequality prescribes that $\mathbb{P}(Y - \mathbb{E}(Y) \geq t) \leq \mathbb{V}(Y)/(\mathbb{V}(Y) + t^2)$ (see, e.g., Boucheron et al. 2013). If, given K, the expected value $\mathbb{E}$ and variance $\mathbb{V}$ of the second order difference under the null model are estimated with $\overline{\mathrm{RMSE}}''_K$ and $s(\mathrm{RMSE}''_K)^2$ respectively, this result can provide an indication for the upper boundary of the probability to observe deviations equal to $\mathrm{RMSE}''_K - \overline{\mathrm{RMSE}}''_K$ or larger. A rejection of the null hypothesis can follow if this upper boundary is lower than some threshold, e.g., 5%. Anyway, given that the true moments of the statistic remain unknown, the indication provided by this inequality is not guaranteed to be accurate. Alternatively, the value of the ratio $(\mathrm{RMSE}''_K - \overline{\mathrm{RMSE}}''_K)/s(\mathrm{RMSE}''_K)$ can be compared with the quantiles of a Student's t distribution with $B-1$ degrees of freedom. This is expected to yield more rejections than the aforementioned upper boundary.

Regardless of how the tail probability is computed in the test for every K, since $K_{\max} - 2$ many tests are carried out, potential multiple testing issues can arise. As a remedy, Bonferroni correction can be applied, adopting a rejection threshold that is a fraction of the significance level. If the latter is chosen to be 5%, the threshold can be computed as $0.05/(K_{\max} - 2)$. If no tail probability falls below it across the $K_{\max} - 2$ tests, there is not enough evidence to reject the hypothesis that $K = 1$, assuming that the best K is not larger than $K_{\max}$. Another strategy to aggregate these probabilities for an overall test of homogeneity is described in

Hennig and Lin (2015).
This test statistic can be used both for a homogeneity test and for determining the best value of K: for instance, the largest K such that a rejection of the null hypothesis occurs can be selected - see below.

In the following, three homogeneity tests based on calibrated versions of $\hat{\beta}_3$, $r(\mathbf{D}_y, \mathbf{D}_g|\mathbf{D}_x)$ and the quantities in (2.12) are compared in terms of type I error rate and power on SLiM datasets from a selection of the scenarios presented in section 2.2. One of these calibrated test statistics is used to estimate the best value for K.

Results on a selection of SLiM scenarios With the aim of investigating whether uniform and weighted null models can help calibrate statistics for testing the null hypothesis that the dataset contains only one species, SLiM datasets from three of the eight scenarios described in section 2.2 were analyzed. These are scenarios with one, two or three species inhabiting the same area, and were analyzed for three values of n, two IBD patterns and one value of P, i.e., when 40 loci are available for the computation of genetic dissimilarities. For each combination of these features, 30 datasets were considered, and each point in Figure 2.11 is the number of such datasets for which a rejections of the null hypothesis that $K = 1$ was recorded using a given combination of test and null model.

The four null models employed were: a uniform null model, as described at the beginning of section 2.4.1; two weighted null models with $\rho = 0.05$ and $\rho = 1$; a weighted null model where the value of ρ is estimated for each specific dataset with the algorithm described in the previous section. Whatever the null model, $B = 30$ null datasets were generated for the calibration of the test statistics. Whenever TESS3 was applied on the original or on a null dataset, ten runs for each K were executed and the minimum RMSE across them was recorded. Similarly to what whas done previously in this chapter, $K_{\max}$ was equal to 6 when $n = 12$, to 10 when $n = 30$ and to 20 when $n = 90$. All tests were carried out at a 5% significance level. Computational reasons informed most of these choices, as it is of course desirable to extend the analysis to additional SLiM scenarios and more datasets, and adopting a larger value for B.

In Figure 2.11, results from three tests are reported:

- the test labeled "dk vs dk+1 (Cantelli)" tested $\beta_3 \leq 0$ from model (2.7) specified with $K = 1$. The significance of $\hat{\beta}_3$, the OLS estimate of this coefficient on the original dataset, was assessed using the Chebyshev-Cantelli inequality. Let us denote with $\hat{\beta}_3^{(b)}$ the replicates of $\hat{\beta}_3$ based on null datasets - with $b = 1, \ldots, B$ - and let $\bar{\beta}_3 = n^{-1} \sum_b \hat{\beta}_3^{(b)}$ and $s(\hat{\beta}_3^{(b)}) = \sqrt{\sum_b (\hat{\beta}_3^{(b)} - \bar{\beta}_3)^2/(B-1)}$ be their empirical mean and standard deviation. The probability to observe values as large as $\hat{\beta}_3$

was quantified as:

$$\frac{s(\hat{\beta}_3^{(b)})^2}{s(\hat{\beta}_3^{(b)})^2 + (\hat{\beta}_3 - \bar{\beta}_3)^2},$$

and the null hypothesis that $\beta_3 \leq 0$ was rejected if this value was lower than 5%.

- The test labeled "SODs (Cantelli + Bonferroni)" was based on TESS3 RMSE second order differences. Each RMSE''_K, with $K = 2, \ldots, K_{\max} - 1$, obtained on the original data was compared with its null replicate using again the Chebyshev-Cantelli inequality. Based on quantities defined in (2.12), the tail probability

$$\frac{s(\text{RMSE}''_K)^2}{s(\text{RMSE}''_K)^2 + (\text{RMSE}''_K - \overline{\text{RMSE}}''_K)^2} \tag{2.13}$$

  was computed for $K = 2, \ldots K_{\max} - 1$ and compared with a Bonferroni-corrected threshold equal to $0.05/(K_{\max} - 2)$. If any of the $K_{\max} - 2$ tail probabilities was lower than this threshold, the null hypothesis that $K = 1$ was rejected. That is, calibrated SODs were too large to be compatible with the null hypothesis that only one species is represented in the dataset.

- For the test labeled "PMT (Cantelli)", TESS3 configurations with $K = 2$ were used to compute a grouping distance $\mathbf{D}_g$ and estimate the partial Mantel correlation $r(\mathbf{D}_y, \mathbf{D}_g | \mathbf{D}_x)$ both on the original dataset and on its null replicates. Also here, empirical mean and variance of these null replicates were computed and used to assess the significance of the observed partial correlation coefficient using the Chebyshev-Cantelli inequality, as done above for $\hat{\beta}_3$.

For the computation of $\hat{\beta}_3$ and $r(\mathbf{D}_y, \mathbf{D}_g | \mathbf{D}_x)$, log-transformed geographic distances were used in this analysis. In Figure 2.11, it is apparent how the tests based on these two statistics produced too many rejections in at least one of the considered combinations of scenario, sample size and IBD behaviour. This issue was exacerbated with larger n, meaning that, for these statistics, null model calibration did not succeed in offsetting the bias described in section 2.3.2, which is caused by testing data-driven null hypotheses. Low type I error rates could only be seen when these test statistics were calibrated via the weighted null model with $\rho = 0.05$ (pink lines), which however led to almost zero power. Interestingly, tests on partial Mantel correlations based on TESS3 configurations with $K = 2$ rejected more often in scenarios where the true K was 2 with respect to scenarios where K was 3. This fact remarks that homogeneity tests based on comparisons between a configuration with $K = 1$ and a configuration with $K = 2$ can show little power if more than two species are truly present in the dataset.

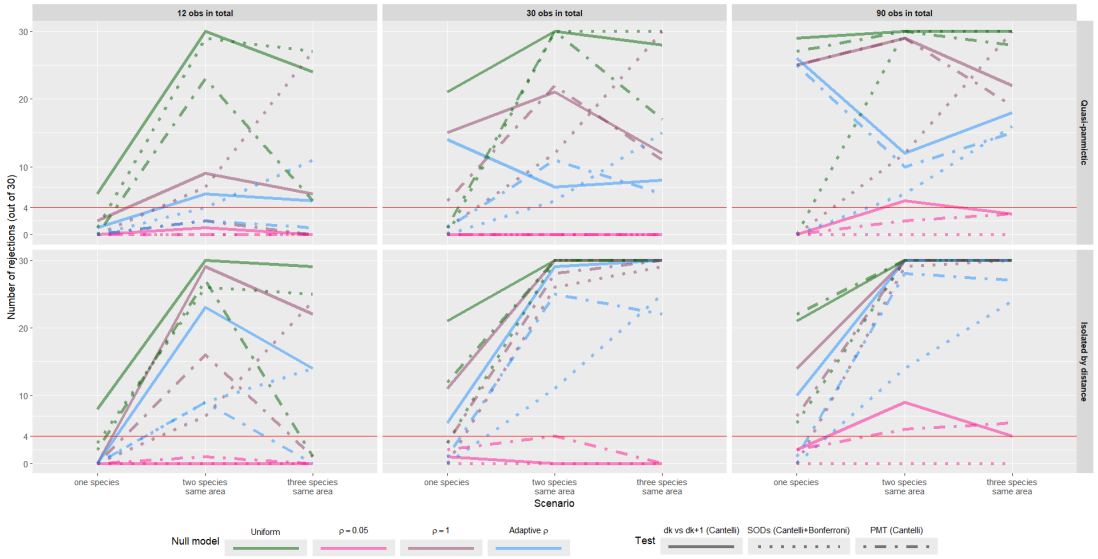

Fig. 2.11 Number of rejections (out of 30) of the null hypothesis that $K = 1$ tested on SLiM datasets with one, two or three overlapping species (first, second and fifth column in Figure 2.1, respectively). Colours by null model used: a uniform null model and three versions of the weighted null model, one where $\rho = 0.05$, one where $\rho = 1$ and one where ρ is estimated with the algorithm described in the previous section. Line type by test statistic: see text. Panels by total sample size n and IBD behaviour. All datasets with $P = 40$. A red horizontal line is superimposed at 4 rejections because $\mathbb{P}(Y \geq 4) < 0.05$, where Y is distributed as a Binomial RV with size 30 and success probability 0.05.

Only the test based on SODs consistently displayed sufficiently low type I error rates. As anticipated, a weighted null model with $\rho = 0.05$ led to tests that almost never rejected the null hypothesis that $K = 1$, regardless of the statistic employed. The null datasets generated with this value of ρ were too similar to the original ones, therefore the observed value of the test statistic could hardly deviate from what reproduced under the null model, yielding extremely conservative tests.

For the calibration of the SODs, the uniform null model (green dotted line) did surprisingly well in most scenarios, with excellent performance when quasi-panmictic species were involved. However, too many rejections occurred with $n = 90$ when IBD patterns were present in the data. This is because uniform null datasets show no association between genetic and geographic distances. Single species showing heterogeneity due to spatial patterns of differentiation thus constituted an anomaly with respect to what was generated under the null. However, the nice performance with quasi-panmictic datasets demonstrates how little information suffices to create null models that can successfully calibrate test statistics under certain conditions.

When calibrated via the weighted null model with the adaptive choice of ρ (blue dotted line in Figure 2.11), the test based on SODs made no type I error and showed fairly good power especially when true $K = 3$, with better results with large n. However,

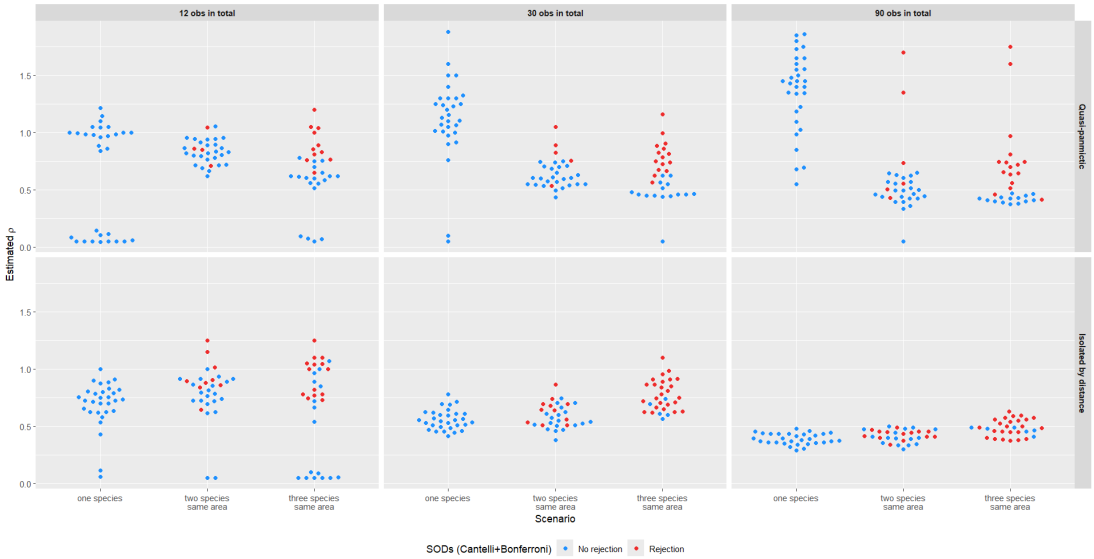

Fig. 2.12 Bee swarm plot of the estimated ρ values from the same scenarios in Figure 2.11. Given the scenario, each dot represents a dataset and its y coordinate equals the value of ρ that was estimated for that dataset through the procedure described in the previous section. Dots are coloured based on whether the null hypothesis that $K = 1$ was rejected by the test based on SODs for that dataset.

the best overall performance was achieved by this test when the weighted model with fixed $\rho = 1$ (brown dotted line) was used for calibration. Indeed, while keeping same type I error properties, with this setting the test was more powerful in all scenarios. In Figure 2.12, the distribution of the ρ value estimated with the algorithm described in the previous section is reported for the same SLiM datasets in Figure 2.11. Excluding the scenario with true $K = 1$, for the vast majority of the datasets values of $\rho < 1$ were selected. When IBD patterns were present, estimates were on average lower. This is not surprising given that in Figure 2.10 the largest slopes in weighted null datasets were typically observed when $\rho \sim 0.3$. Moreover, larger n was associated with lower estimates, too. In datasets with two or three quasi-panmictic species, despite the absence of spatial patterns of differentation within the species, the estimated ρ was often below 1. This was most likely caused by the range of between-groups dissimilarities, which was such that the overall trend between $\mathbf{D}_y$ and $\mathbf{D}_x$ appeared to be positive - see, e.g., the third plot in first column of Figure A.27. This issue would be even more evident if SLiM scenarios with species inhabiting separate areas were to be analyzed. To circumvent this, the estimation of ρ would have to be based merely on the trend observed in the within-group dissimilarities, but this information is not available when constructing null models. On top of this, problems would still remain if distinct species displayed different IBD behaviour. This issue partly explains the lack of power of this test in

quasi-panmictic scenarios and constitutes a limitation when achoring the estimation of ρ to the overall slope observed in the dissimilarities.

In Figure 2.12, data points are also coloured according to whether the test based on SODs rejected the null hypothesis of homogeneity for a given dataset. Especially when the true $K = 3$, larger estimated values of ρ corresponded to more rejections and when the species displayed IBD patterns, ρ values larger than 0.3 were often enough to lead to a rejection. In general, scenarios with two distinct species appeared to be more challenging, and the association between the value of ρ and the number of rejections was less evident. All in all, given that the adaptive null model seems to lead to too few rejections when the test based on SODs is employed, the lower boundary of the interval for the search of ρ might be increased. Future work may focus on how to improve the choice of this interval and whether a better target statistic exists than the overall slope in the regression between genetic and (log-transformed) geographic distances.

Table 2.2 Number of datasets (out of 30) for which the true value of K was recovered by looking at the *largest* K for which the tail probability in (2.13), calibrated with the weighted null model with $\rho = 1$, was lower than the Bonferroni-corrected threshold $0.05/(K_{\max} - 2)$.

Scenario	IBD pattern	Total sample size	Correct choices
Two species	Quasi-panmictic	90	10
		30	11
		12	5
	Isolated by distance	90	28
		30	25
		12	7
Three species	Quasi-panmictic	90	30
		30	30
		12	27
	Isolated by distance	90	30
		30	29
		12	24

Table 2.2 reports the number of times the test labeled "SODs (Cantelli + Bonferroni)" calibrated with the weighted null model with $\rho = 1$ recovered the correct number of species for each combination of scenario, IBD pattern and sample size. More precisely, to select a value of K, the largest K for which the tail probability in (2.13) was lower than the Bonferroni-corrected threshold $0.05/(K_{\max} - 2)$ was picked. It can be argued whether the smallest K such that this condition is met should be used to estimate the true number of species in the dataset instead. This resonates with the literature on the

search for "the uppermost hierarchical level of structure" in the dataset (Evanno et al. 2005), but as alternative criterion made little impact on the recovery rate - see Table A.1. In any case, since here the trend in the RMSE values is calibrated via null models, the largest K for which a low tail probability is recorded can be expected to correspond to the number of distinct species in the dataset. Indeed, lower levels of structure (i.e., spatial patterns of genetic differentiation within the species) should be captured in the null datasets. Thus, while they may be undistinguishable from species membership when K is lower than the true one, they should not produce significant deviations in the SODs - with respect to the null model - once the true value of K is overcome. In other words, looking at the maximum value of K for which a significant deviation in the SODs is observed seems a sensible criterion when using null model calibration.

In Table 2.2, low numbers of correct recoveries for K in scenarios with two species can be mainly imputed to the overestimation of the target slope (as explained above), and to low sample size. Apart from this, however, a rejection of the null hypothesis that only one species is present in the dataset also corresponded to the correct choice of K for most datasets with three species and for many datasets with two IBD species, provided that $n > 12$. A cursory comparison with information in Figure 2.7 for the relevant combinations of scenario and number of loci shows that the recovery rates reported in Table 2.2 are not better than those from the ΔK method. Anyway, given the promising performance when testing the null hypothesis of homogeneity ($K = 1$), a correctly calibrated test based on SODs can be used in tandem with the ΔK method. That is, if evidence from null models suggests that more than one species might be represented in the dataset, then the ΔK method might be employed to estimate the true K.

The following section briefly considers alternative strategies to null models for dealing with tests on data-driven null hypotheses.

2.4.3 Ideas from selective inference

Usually, in statistical inference the hypothesis to be tested is set before looking at the data. In Chapter 1, the putative grouping information employed in the conspecificity tests was fixed, i.e., it was not extracted from the data. Thus, no bias due to the double use of the data, null hypothesis selection and hypothesis testing, could be introduced in the tests. They could still differ in terms of conservativeness depending on how the distribution of their test statistics under the null was reconstructed - see, e.g., the discussion on partial Mantel tests in section 1.5.3. In section 2.3, instead, a clustering algorithm was first applied on the data in order to find groups and then null hypotheses were formulated in the light of the clustering configuration. Testing this kind of adaptive

null hypotheses led to biased tests, whose type I error rate was usually larger than the adopted significance level.

Methods from selective inference can deal with adaptive null hypotheses - see Lee et al. (2016) for an overview. In selective inference, the goal is to control the type I error rate when testing similar data-driven null hypotheses, namely to make sure that the probability to reject a null hypothesis *given that it was formulated* still falls below the adopted significance level (Fithian et al. 2014). This is done by conditioning the probability to reject under the null hypothesis on the fact that that specific null hypothesis was tested. Two contributions where this was achieved for tests on differences between groups identified by clustering algorithms are:

- Gao et al. (2024), where a test for the difference in mean between two clusters identified via hierarchical clustering is proposed;
- Y. T. Chen and Witten (2023), where the same is done for k-means clustering.

Both papers define a selective p-value, i.e., quantify the probability that differences between cluster means as large as the observed one emerge, given that the two clusters were identified. This amounts to browsing all data realizations that lead to the same clustering configuration the null hypothesis is based on, and assessing how unusual it is for two groups to be that far away. A key assumption is that the data follows a multivariate normal distribution. Then, all data realizations of interest are explored by recasting the selective p-value as the survival function of a scaled Chi-square random variable truncated to some set $\mathscr{S}$. This set contains all perturbations of the original data such that the same clustering configuration is recovered. This crucial move makes the computation of the selective p-value tractable because it reshapes it into characterizing the set $\mathscr{S}$. Here, distinctions between hierarchical and k-means clustering kick in - see the cited papers.

As far as the routine in section 2.3.1 is concerned, we are interested in quantifying the probability to find, e.g., a partial Mantel correlation between genetic and grouping distance, given geographic distance, as large as the observed one, where the grouping information stems from a clustering configuration output by sNMF. Similarly, for the model in (2.7), the goal would be to compute the probability that the estimate of the β_3 coefficient is as large as the observed one given that the regression was fitted based on two TESS3 configurations with K and $K+1$ groups.

Results from Gao et al. (2024) and Y. T. Chen and Witten (2023) cannot be directly appied in these setups because they are tailored to their specific combinations of data scenario and clustering algorithm. The largest obstacle in this sense lies in the nature of the data, since we are working with datasets made up of a pair of geographic coordinates

and P genetic loci. When conditioning on the event that a given clustering configuration was output by sNMF or TESS3, we need to explore all realizations of our dataset that would have led to the same configuration. The key theorems the aforementioned papers exploit to do so are modified versions of Theorem 3.1 in Loftus and Taylor (2015), where the data is assumed to be normal. This assumption is incompatible with the mixed nature of our data.

A simpler way of taking clustering configuration into account when testing the hypothesis that $\beta_3 \leq 0$ in section 2.3.2 could consist in an adaptation of jackknife testing. Instead of recomputing the value of $\hat{\beta}_3$ upon the removal of one individual at a time based on the same clustering solutions with K and $K+1$ groups, TESS3 optimization might be carried out on each dataset where one individual is dropped. This might make it possible for the jackknife replicates of $\hat{\beta}_3$ to better capture the variability associated with clustering configurations, offsetting the bias described in Figure 2.6.

This idea, together with the application of concepts from selective inference, is left for future work.

2.5 Closing remarks

This chapter considered the problem of determining how many species are represented in a spatially-explicit genetic dataset. The goal was to construct methods that can take into account the presence of spatial patterns of genetic differentiation (e.g., IBD) within the species, which may induce genetic heterogeneity in co-specific groups that inhabit separate geographic areas.

To this aim, two strategies were first conceived that integrate existing clustering algorithms with distance-based methods. In one of them, a clustering configuration based on sNMF (Frichot, Mathieu, et al. 2014) was used to initialize an iterative routine where pairs of putative groups are tested for merging using distance-based techniques from section 1.3. On SLiM simulations, this integration often improved the initial clustering that sNMF identified via its built-in cross-entropy criterion, which was found to be influenced by the total number of individuals and by the presence of IBD patterns. Five versions of the merging routine were compared, each based on a different test. Their performance was consistent with what observed in section 1.4, with jackknife-based and permutation-based partial Mantel tests delivering the best compromise between improvements and worsening of the sNMF initial configuration.

A second strategy revolved around TESS3 (Caye et al. 2018), whose clustering solutions are optimized both on genetic and geographic information, unlike sNMF. A regression model was proposed that embeds TESS3 output and can help determine

the number of species K in the dataset. Starting from the comparison of a TESS3 configuration with $K = 1$ vs one with $K = 2$, jackknife-based tests are carried out that check for the need of an additional group to explain genetic variability in the dataset, given geographic separation. On SLiM data, this strategy was found to more accurately estimate K than TESS3's built-in cross-entropy criterion, which overestimated K in all scenarios.

However, both the strategy based on sNMF and that based on TESS3 revealed a systematic shortcoming. In the former, genetic data was first used by sNMF to identify groups and then employed by distance-based methods to test whether the groups were too different to be merged. As regards TESS3, geographic and genetic information was involved both in its optimization and in the jackknife-based tests for the choice of K. The fact that the same data was used twice led to the overestimation of K, since distance-based methods tested data-driven null hypotheses. Putative groups were not fixed as in Chapter 1, but rather identified in a preliminary clustering step, and this made the hypotheses to be tested adaptive. Testing adaptive null hypotheses is known to invalidate inference (Fithian et al. 2014).

In the second part of the chapter, null models were explored as tools to calibrate tests for the presence of more than one species in the dataset and for the estimation of K. Three tests on $K = 1$ vs $K > 1$ were considered for calibration when using TESS3: the aforementioned jackknife-based test, a test on partial Mantel correlations and one based on the second order differences (SODs) of the RMSE from TESS3 solutions. SODs are connected to the ΔK statistic by Evanno et al. (2005), which is widely used to determine the best K in molecular ecology studies. Adapted to TESS3 output, SODs were found to retrieve the correct number of species in SLiM datasets more often than ΔK in most scenarios with more than two species. However, SODs cannot suggest that $K = 1$ in datasets containing only one species - unless null model calibration is used.

Therefore, a uniform and a weighted null model were introduced, that resample the geographic and genetic information in the dataset with different levels of sophistication. While the uniform null model returns dataset replicates with no IBD pattern, the effectiveness of the weighted null model in capturing observed spatial patterns of differentiation is related to the choice of parameter ρ, to be estimated from the data. An estimation procedure for ρ was thus proposed, whose goal is to produce dataset replicates where the relationship between genetic and log-transformed geographic distances between individuals is similar to that observed in the original dataset.

SLiM datasets from three scenarios were thus used to compare type I error and power properties of the three tests mentioned above when calibrated with four different null models: the uniform null model and three versions of the weighted one, i.e., one where ρ was estimated from the data and two where it was fixed. Results showed

that only the test based on SODs had type I error rate below the significance level in all scenarios, provided that calibration was done via the weighted null model. More precisely, even in scenarios with true $K > 1$, fixing $\rho = 0.05$ caused all calibrated tests to be unable to detect the presence of more than one species in most cases. The data-driven estimation of ρ made the calibrated test based on SODs more powerful, but the best results were observed when setting $\rho = 1$ for all datasets. Calibration with the uniform null model led to inflated type I error rates in scenarios with IBD species, as could be expected given that the null datasets it produces display no spatial patterns of genetic differentiation.

The calibrated test based on SODs was also used to estimate number of species in the dataset, but did not show recovery rates superior to those displayed by the ΔK method. However, given its promising performance at testing the hypothesis that the dataset contains only one species, it may constitute a valuable complementary tool to ΔK. That is, if calibrated SODs suggest that more than one species might be represented in the dataset, then the ΔK method might be employed to estimate the true K.

For future research, the effectiveness of null model calibration will have to be assessed on additional SLiM scenarios, but also on real data. Moreover, the estimation of the ρ parameter in weighted null models may be improved. Changes might relate the range on which the value of ρ is searched, but also the statistic targeted in the estimation, which here was the slope of the regression between genetic and log-geographic distances. Indeed, the choice of this target might have caused calibrated tests based on SODs to yield low power in quasi-panmictic scenarios.

In addition, different strategies might be pursued to control the type I error rate of the jackknife-based test introduced in section 2.3.2. These include approaches from selective inference and adaptations of the jackknife resampling scheme.

Furthermore, the use of software packages other than TESS3 may be investigated. In particular, conStruct (Bradburd, Coop, and Ralph 2018) can be expected to more accurately capture IBD patterns in the data, providing possibly better allocations of the individuals to the groups. Its output can replace that of TESS3 in virtually all of the tests proposed in this chapter, and their impact on the estimation of K might be systematically compared on simulated and real data.

Conclusions

This work explored methodologies to delimit species using genetic and geographic data in the presence of patterns of isolation by distance. Two main settings were considered, one per chapter. In Chapter 1, individuals were known to be divided in two putative groups and the goal was to test whether these belonged to the same species. Several methods were investigated that incorporate grouping information in models for the relationship between genetic and geographic distances. They were compared via an extensive simulation study based on two different software packages for the generation of spatially-explicit genetic data, SLiM (Haller and Messer 2023) and GSpace (Virgoulay et al. 2021b).

A key challenge was that of modeling the dependence between the dissimilarities: most methods that took it into account showed satisfactory type I error rates, while a multiple linear regression that treated the dissimilarities as independent was anti-conservative when the groups had unequal sample sizes. A newly proposed jackknife-based version of the partial Mantel test (Smouse et al. 1986) preserved the same marked conservativeness of the method by Hausdorf and Hennig (2020) while being more powerful. The permutation-based partial Mantel test as used by Medrano et al. (2014), however, was the most powerful method, but at the cost of borderline type I error rates in some setups.

Simulations from SLiM, GSpace and from a simulator developed specifically for this project showed that the relationship between the shared allele distance (the genetic distance employed in this study) and the geographic distance is seldom linear. This, however, did not seem to affect the distance-based delimitation methods. On the other hand, an identifiability issue with these methods was highlighted in setups where the two groups are geographically segregated and their within-group dissimilarities display a strong IBD pattern: in these cases it may become impossible to tell apart one species from two species. In addition, some probabilistic properties of the shared allele distance were described, such as the fact the it fulfills the triangle inequality if no genetic information is missing and its asymptotic normality under certain conditions.

In Chapter 2, no grouping information was available and the objective was to determine the number of species represented in the dataset, K. The built-in cross-

entropy criterion in existing model-based clustering algorithms such as sNMF (Frichot, Mathieu, et al. 2014) and TESS3 (Caye et al. 2018) was found to provide a poor indication for the correct number of species in SLiM datasets from a wide range of scenarios. By integrating these algorithms with distance-based methods from Chapter 1 more accurate estimates were obtained. However, these two-step strategies often overestimated the number of species in the dataset because they tested null hypotheses informed from the data.

To overcome this limitation, null models were conceived that can calibrate tests for the presence of more than one species in the dataset and for the estimation of K. A weighted null model was proposed and used to calibrate a test based on second order differences between the RMSE values output by TESS3. On a selection of SLiM scenarios, this combination proved to be an effective tool to check whether more than one species is present in the data. This constitutes a valid complement to methods like the ΔK (Evanno et al. 2005) which by construction cannot suggest that the correct number of species is 1.

For future research, several aspect of the simulations employed in this project can be extended and some (or all) of their parameters can be informed via estimates from real datasets. Versions of the distance-based methods from Chapter 1 can be explored where the assumption of linearity in the relationship between genetic and geographic distances is relaxed. Moreover, this work concerned analyses at individual level, with inference in most methods implicitly assuming even co-specific individuals to be independent. This would not be the case in population-level analyses, which thus constitute a direction for future work. The effectiveness of the weighted null model proposed in Chapter 2 has to be assessed on a wider range of SLiM scenarios and also on real data, where it might help solve the "$K = 2$ conundrum" (Janes et al. 2017). Furthermore, the software conStruct (Bradburd, Coop, and Ralph 2018) might be compared with TESS3 both in terms of their built-in cross-validation criteria for the choice of K and in combination with distance-based methods. Simulations may also show which one of the two leads to better estimates for K when null model calibration is used.

References

Adriaensen, F., J.P. Chardon, G. De Blust, E. Swinnen, S. Villalba, H. Gulinck, and E. Matthysen (2003). "The application of 'least-cost' modelling as a functional landscape model". In: *Landscape and Urban Planning* 64.4, pp. 233–247. ISSN: 0169-2046. DOI: https://doi.org/10.1016/S0169-2046(02)00242-6.

Adrion, J. R., C. B. Cole, N. Dukler, J. G. Galloway, A. L. Gladstein, G. Gower, C. C. Kyriazis, A. P. Ragsdale, G. Tsambos, F. Baumdicker, J. Carlson, R. A. Cartwright, A. Durvasula, I. Gronau, B. Y. Kim, P. McKenzie, P. W. Messer, E. Noskova, D. Ortega-Del Vecchyo, F. Racimo, T. J. Struck, S. Gravel, R. N. Gutenkunst, K. E. Lohmueller, P. L. Ralph, D. R. Schrider, A. Siepel, J. Kelleher, and A. D. Kern (2020). "A community-maintained standard library of population genetic models". In: *eLife* 9. Ed. by G. Coop, P. J. Wittkopp, J. Novembre, A. Sethuraman, and S. Mathieson, e54967. ISSN: 2050-084X. DOI: 10.7554/eLife.54967.

Alexander, D. H., J. Novembre, and K. Lange (2009). "Fast model-based estimation of ancestry in unrelated individuals". In: *Genome Research* 19.9, pp. 1655–1664. DOI: 10.1101/gr.094052.109.

Anderson, C. D., B. K. Epperson, M.-J. Fortin, R. Holderegger, P. M. A. James, M. S. Rosenberg, K. T. Scribner, and S. Spear (2010). "Considering spatial and temporal scale in landscape-genetic studies of gene flow". In: *Molecular Ecology* 19.17, pp. 3565–3575. ISSN: 0962-1083.

Balkenhol, N., S. A. Cushman, A. Storfer, and L. P. Waits (2015). "Introduction to landscape genetics – concepts, methods, applications". In: *Landscape Genetics*. Ed. by N. Balkenhol, S. A. Cushman, A. T. Storfer, and L. P. Waits. John Wiley & Sons, Ltd. Chap. 1, pp. 1–8. ISBN: 9781118525258. DOI: https://doi.org/10.1002/9781118525258.ch01.

Bamberger, S., J. Xu, and B. Hausdorf (2021). "Evaluating Species Delimitation Methods in Radiations: The Land Snail Albinaria cretensis Complex on Crete". In: *Systematic Biology* 71.2, pp. 439–460. ISSN: 1063-5157. DOI: 10.1093/sysbio/syab050.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Battey, C. J., P. L. Ralph, and A. D. Kern (2020). "Space is the place: effects of continuous spatial structure on analysis of population genetic data". In: *Genetics* 215.1. Epub 2020 Mar 24, pp. 193–214. DOI: 10.1534/genetics.120.303143.

Baumdicker, F., G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, B. Eldon, E. C. Ellerman, J. G. Galloway, A. L. Gladstein, G. Gorjanc, B. Guo, B. Jeffery, W. W. Kretzschumar, K. Lohse, M. Matschiner, D. Nelson, N. S. Pope, C. D. Quinto-Cortés, M. F. Rodrigues, K. Saunack, T. Sellinger, K. Thornton, H. van Kemenade, A. W. Wohns, Y. Wong, S. Gravel, A. D. Kern, J. Koskela, P. L. Ralph, and J. Kelleher (2021). "Efficient ancestry and mutation simulation with msprime 1.0". In: *Genetics* 220.3, iyab229. ISSN: 1943-2631. DOI: 10.1093/genetics/iyab229.

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

Billingsley, P. (1995). *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471007104.

Bohonak, A. J. (2002). "IBD (isolation by distance): a program for analyses of isolation by distance". In: *Journal of Heredity* 93.2, pp. 153–154. ISSN: 0022-1503. DOI: 10.1093/jhered/93.2.153.

Borcard, D., F. Gillet, and P. Legendre (2011). *Numerical ecology with R*. Vol. 2. Springer.

Borcard, D. and P. Legendre (2012). "Is the Mantel correlogram powerful enough to be useful in ecological analysis? A simulation study". In: *Ecology* 93.6, pp. 1473–1481.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: a nonasymptotic theory of independence*. OUP Oxford. ISBN: 9780199535255.

Bourgeois, Y. X. C. and B. H. Warren (2021). "An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes". In: *Molecular Ecology* 30.23, pp. 6036–6071.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, and L. L. Cavalli-Sforza (1994). "High resolution of human evolutionary trees with polymorphic microsatellites". In: *Nature* 368.6470, pp. 455–457. ISSN: 0028-0836.

Bradburd, G. S., G. M. Coop, and P. L. Ralph (2018). "Inferring continuous and discrete population genetic structure across space". In: *Genetics* 210.1, pp. 33–52. ISSN: 1943-2631. DOI: 10.1534/genetics.118.301333.

Bradburd, G. S., P. L. Ralph, and G. M. Coop (2013). "Disentangling the effects of geographic and ecological isolation on genetic differentiation". In: *Evolution* 67.11, pp. 3258–3273. DOI: https://doi.org/10.1111/evo.12193.

Burbrink, F. T. and S. Ruane (2021). "Contemporary philosophy and methods for studying speciation and delimiting species". In: *Ichthyology & Herpetology* 109.3, pp. 874–894. ISSN: 2766-1512. DOI: 10.1643/h2020073.

Carstens, B. C., T. A. Pelletier, N. M. Reid, and J. D. Satler (2013). "How to fail at species delimitation". In: *Molecular Ecology* 22.17, pp. 4369–4383.

Carvajal-Rodríguez, A. (2008). "Simulation of genomes: a review". In: *Current genomics* 9.3, pp. 155–159.

Caye, K., F. Jay, O. Michel, and O. François (2018). "Fast inference of individual admixture coefficients using geographic data". In: *The Annals of Applied Statistics* 12.1, pp. 586–608. ISSN: 19326157, 19417330.

Cayuela, H., Q. Rougemont, J. G. Prunier, J. Moore, J. Clobert, A. Besnard, and L. Bernatchez (2018). "Demographic and genetic approaches to study dispersal in wild animal populations: a methodological review". In: *Molecular Ecology* 27.20, pp. 3976–4010. ISSN: 0962-1083.

Chen, C., E. Durand, F. Forbes, and O. François (2007). "Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study". In: *Molecular Ecology Notes* 7.5, pp. 747–756. DOI: https://doi.org/10.1111/j.1471-8286.2007.01769.x.

Chen, Y. T. and D. M. Witten (2023). "Selective inference for k-means clustering". In: *Journal of Machine Learning Research* 24.152, pp. 1–41.

Clarke, R. T., P. Rothery, and A. F. Raybould (2002). "Confidence limits for regression relationships between distance matrices: estimating gene flow with distance". In: *Journal of Agricultural, Biological, and Environmental Statistics* 7.3, pp. 361–372.

Currat, M., M. Arenas, C. S. Quilodràn, L. Excoffier, and N. Ray (2019). "SPLATCHE3: simulation of serial genetic data under spatially explicit evolutionary scenarios including long-distance dispersal". In: *Bioinformatics* 35.21, pp. 4480–4483. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz311.

Day, C. C., E. L. Landguth, R. K. Simmons, and A. R. Bearlin (2023). "CDMetaPOP 2: a multispecies, eco-evolutionary simulation framework for landscape demogenetics and connectivity". In: *Ecography* 2023.8, e06566. DOI: https://doi.org/10.1111/ecog.06566.

De Queiroz, K. (1998). "The general lineage concept of species, species criteria, and the process of speciation". In: *Endless forms: species and speciation*.

— (2007). "Species concepts and species delimitation". In: *Systematic Biology* 56.6, pp. 879–886. ISSN: 1063-5157. DOI: 10.1080/10635150701701083.

Diniz-Filho, J. A. F., T. N. Soares, J. S. Lima, R. Dobrovolski, V. L. Landeiro, M. P. Telles, T. F. Rangel, and L. M. Bini (2013). "Mantel test in population genetics". In: *Genetics and molecular biology* 36, pp. 475–485.

Edwards, D. L. and L L. Knowles (2014). "Species detection and individual assignment in species delimitation: can integrative data increase efficacy?" In: *Proceedings of the Royal Society B: Biological Sciences* 281.1777, p. 20132765.

Efron, B. and R. Tibshirani (1993). *An introduction to the bootstrap*. New York London: Chapman & Hall. ISBN: 978-04-12-04231-7.

Epperson, B. K., B. H. McRae, K. Scribner, S. A. Cushman, M. S. Rosenberg, M.-J. Fortin, P. M. A. James, M. Murphy, S. Manel, Pierre Legendre, and M. R. T. Dale (2010). "Utility of computer simulations in landscape genetics". In: *Molecular Ecology* 19.17, pp. 3549–3564. DOI: https://doi.org/10.1111/j.1365-294X.2010.04678.x.

Ereshefsky, M. (2007). "Species, taxonomy and systematics". In: *Philosophy of biology*. Ed. by M. Matthen and C. Stephens. Handbook of the philosophy of science. Amsterdam: North-Holland, pp. 403–427. DOI: https://doi.org/10.1016/B978-044451543-8/50020-4.

Evanno, G., S. Regnaut, and J. Goudet (2005). "Detecting the number of clusters of individuals using the software structure: a simulation study". In: *Molecular ecology* 14.8, pp. 2611–2620. ISSN: 0962-1083.

Fithian, W., D. Sun, and J. Taylor (2014). "Optimal inference after model selection". In: *arXiv preprint arXiv:1410.2597*.

François, O. and L. P. Waits (2015). "Clustering and assignment methods in landscape genetics". In: *Landscape Genetics*. Ed. by N. Balkenhol, S. A. Cushman, A. T. Storfer, and L. P. Waits. John Wiley & Sons, Ltd. Chap. 7, pp. 114–128. ISBN: 9781118525258. DOI: https://doi.org/10.1002/9781118525258.ch07.

Frantz, A. C., S. Cellina, A. Krier, L. Schley, and T. Burke (2009). "Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance?" In: *Journal of Applied Ecology* 46.2, pp. 493–505. DOI: https://doi.org/10.1111/j.1365-2664.2008.01606.x.

Frichot, E. and O. François (2015). "LEA: An R package for landscape and ecological association studies". In: *Methods in Ecology and Evolution* 6.8, pp. 925–929. DOI: https://doi.org/10.1111/2041-210X.12382.

Frichot, E., F. Mathieu, T. Trouillon, G. Bouchard, and O. François (2014). "Fast and efficient estimation of individual ancestry coefficients". In: *Genetics* 196.4, pp. 973–983. ISSN: 1943-2631. DOI: 10.1534/genetics.113.160572.

Gao, L. L., J. Bien, and D. M. Witten (2024). "Selective inference for hierarchical clustering". In: *Journal of the American Statistical Association* 119.545. PMID: 38660582, pp. 332–342. DOI: 10.1080/01621459.2022.2116331.

Gaudeul, M., I. Till-Bottraud, F. Barjon, and S. Manel (2004). "Genetic diversity and differentiation in Eryngium alpinum L. (Apiaceae): comparison of AFLP and microsatellite markers". In: *Heredity* 92.6, pp. 508–518. ISSN: 1365-2540. DOI: 10.1038/sj.hdy.6800443.

Gordon, A. D. (1996). "Null models in cluster validation". In: *From data to knowledge*. Ed. by W. Gaul and D. Pfeifer. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 32–44. ISBN: 978-3-642-79999-0.

Goslee, S. C. and D. L Urban (2007). "The ecodist package for dissimilarity-based analysis of ecological data". In: *Journal of Statistical Software* 22, pp. 1–19.

Gratton, P., E. Trucchi, A. Trasatti, G. Riccarducci, S. Marta, G. Allegrucci, D. Cesaroni, and V. Sbordoni (2016). "Testing classical species properties with contemporary data: how "bad species" in the Brassy Ringlets (Erebia tyndarus complex, Lepidoptera) turned good". In: *Systematic biology* 65.2, pp. 292–303. ISSN: 1063-5157.

Guillot, G., F. Mortier, and A. Estoup (2005). "Geneland: a computer package for landscape genetics". In: *Molecular Ecology Notes* 5.3, pp. 712–715. DOI: https://doi.org/10.1111/j.1471-8286.2005.01031.x.

Guillot, G. and F. Rousset (2013). "Dismantling the Mantel tests". In: *Methods in Ecology and Evolution* 4.4, pp. 336–344. DOI: https://doi.org/10.1111/2041-210x.12018.

Haller, B. C. and P. W. Messer (2022). *SLiM Manual*. https://messerlab.org/slim/. Accessed: 2023-12-07.

— (2023). "SLiM 4: multispecies eco-evolutionary modeling". In: *The American Naturalist* 201.5, E127–E139.

Hardy, O. J. and X. Vekemans (2002). "SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels". In: *Molecular ecology notes* 2.4, pp. 618–620. ISSN: 1471-8278.

Hausdorf, B. (2011). "Progress toward a general species concept". In: *Evolution* 65.4, pp. 923–931. ISSN: 0014-3820. DOI: 10.1111/j.1558-5646.2011.01231.x.

Hausdorf, B. and C. Hennig (2019). *Package 'prabclus', version 2.3-2*. URL: https://CRAN.R-project.org/package=prabclus (visited on 07/25/2023).

— (2020). "Species delimitation and geography". In: *Molecular Ecology Resources* 20.4, pp. 950–960. DOI: https://doi.org/10.1111/1755-0998.13184.

Hedrick, P. W. (2009). *Genetics of populations*. Jones & Bartlett Publishers.

Hennig, C. and B. Hausdorf (2004). "Distance-based parametric bootstrap tests for clustering of species ranges". eng. In: *Computational statistics & data analysis* 45.4, pp. 875–895. ISSN: 0167-9473.

Hennig, C. and C. Lin (2015). "Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters". In: *Statistics and Computing* 25, pp. 821–833.

Huang, H., Y. Liu, D. N. Hayes, A. Nobel, J. S. Marron, and C. Hennig (2015). "Significance testing in clustering". In: *Handbook of cluster analysis*. Ed. by C. Hennig, M. Meila, F. Murtagh, and R. Rocci, pp. 315–336. ISBN: 9781466551886.

Hubert, L. and P. Arabie (1985). "Comparing partitions". In: *Journal of classification* 2, pp. 193–218.

Hutchison, D. W. and A. R. Templeton (1999). "Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of gene flow and

drift on the distribution of genetic variability". In: *Evolution* 53.6, pp. 1898–1914. ISSN: 0014-3820.
Ishida, Y. (2009). "Sewall Wright and Gustave Malécot on isolation by distance". In: *Philosophy of Science* 76.5, pp. 784–796. ISSN: 00318248, 1539767X.
Jain, A. K. and R. C. Dubes (1988). *Algorithms for clustering data*. USA: Prentice-Hall, Inc. ISBN: 013022278X.
Janes, J. K., J. M. Miller, J. R. Dupuis, R. M. Malenfant, J. C. Gorrell, C. I. Cullingham, and R. L. Andrew (2017). "The K = 2 conundrum". In: *Molecular Ecology* 26.14, pp. 3594–3602. DOI: https://doi.org/10.1111/mec.14187.
Jaquiéry, J., T. Broquet, A. H. Hirzel, J. Yearsley, and N. Perrin (2011). "Inferring landscape effects on dispersal from genetic distances: how far can we go?" In: *Molecular Ecology* 20.4, pp. 692–705. DOI: https://doi.org/10.1111/j.1365-294X.2010.04966.x.
Jukes, T. H. and C. R. Cantor (1969). "Evolution of protein molecules". In: *Mammalian Protein Metabolism*. Ed. by H. N. Munro. Academic Press, pp. 21–132. ISBN: 978-1-4832-3211-9. DOI: https://doi.org/10.1016/B978-1-4832-3211-9.50009-7.
Kaufman, L. and P. J. Rousseeuw (1990). *Finding groups in data: an introduction to cluster analysis*. New York...(etc.): J.Wiley & Sons. ISBN: 978-04-7187-876-6.
Kelleher, J., K. R. Thornton, J. Ashander, and P. L. Ralph (2018). "Efficient pedigree recording for fast population genetics simulation". In: *PLOS Computational Biology* 14.11, pp. 1–21. DOI: 10.1371/journal.pcbi.1006581.
Kimura, M. (1955). "Stochastic processes and distribution of gene frequencies under natural selection". In: *Cold Spring Harb. Symp. Quant. Biol.* 20.0, pp. 33–53.
Kimura, M. and G. H. Weiss (1964). "The stepping stone model of population structure and the decrease of genetic correlation with distance". In: *Genetics* 49.4, pp. 561–576. ISSN: 0016-6731.
Kingman, J.F.C. (1982). "The coalescent". In: *Stochastic processes and their applications* 13.3, pp. 235–248. ISSN: 0304-4149. DOI: https://doi.org/10.1016/0304-4149(82)90011-4.
Landguth, E., S. A. Cushman, and N. Balkenhol (2015). "Simulation modeling in landscape genetics". In: *Landscape Genetics*. Ed. by N. Balkenhol, S. A. Cushman, A. T. Storfer, and L. P. Waits. John Wiley & Sons, Ltd. Chap. 6, pp. 99–113. ISBN: 9781118525258. DOI: https://doi.org/10.1002/9781118525258.ch06.
Leaché, A. D., M. K. Fujita, V. N. Minin, and R. R. Bouckaert (2014). "Species delimitation using genome-wide SNP data". In: *Systematic biology* 63.4, pp. 534–542. ISSN: 1063-5157.
Leblois, R., A. Estoup, and F. Rousset (2009). "IBDSim: a computer program to simulate genotypic data under isolation by distance". In: *Molecular Ecology Resources* 9.1, pp. 107–109.
Lee, J. D., D. L. Sun, Sun Y, and J. E. Taylor (2016). "Exact post-selection inference, with application to the lasso". In: *The Annals of Statistics* 44.3, pp. 907–927. DOI: 10.1214/15-AOS1371.
Legendre, P. (2000). "Comparison of permutation methods for the partial correlation and partial Mantel tests". In: *Journal of statistical computation and simulation* 67.1, pp. 37–73.
Legendre, P. and M.-J. Fortin (2010). "Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data". In: *Molecular Ecology Resources* 10.5, pp. 831–844. DOI: https://doi.org/10.1111/j.1755-0998.2010.02866.x.

Legendre, P., M.-J. Fortin, D. Borcard, and P. Peres-Neto (2015). "Should the Mantel test be used in spatial analysis?" In: *Methods in ecology and evolution* 6.11, pp. 1239–1247. ISSN: 2041-210X.

Legendre, P. and L. Legendre (2012). *Numerical ecology*. 3rd English ed. Developments in environmental modelling ; 24. Amsterdam ; Boston: Elsevier. ISBN: 0-444-53868-2.

Liu, D. (2024). "We must train specialists in botany and zoology — or risk more devastating extinctions". In: *Nature* 633, p. 741. DOI: 10.1038/d41586-024-03072-3.

Liu, Q., C. Li, V. Wanga, and B. E. Shepherd (2018). "Covariate-adjusted Spearman's rank correlation with probability-scale residuals". In: *Biometrics* 74.2, pp. 595–605. DOI: https://doi.org/10.1111/biom.12812.

Loftus, J. R. and J. E. Taylor (2015). *Selective inference in regression models with groups of variables*.

Loiselle, B. A., V. L. Sork, J. Nason, and C. Graham (1995). "Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae)". In: *American journal of botany* 82.11, pp. 1420–1425.

MacDonald, Z. G., J. R. Dupuis, C. S. Davis, J. H. Acorn, S. E. Nielsen, and F. A. H. Sperling (2020). "Gene flow and climate-associated genetic variation in a vagile habitat specialist". In: *Molecular Ecology* 29.20, pp. 3889–3906. DOI: https://doi.org/10.1111/mec.15604.

Manel, S., O. E. Gaggiotti, and R. S. Waples (2005). "Assignment methods: matching biological questions with appropriate techniques". In: *Trends in Ecology & Evolution* 20.3, pp. 136–142. DOI: 10.1016/j.tree.2004.12.004.

Manel, S., M. K. Schwartz, G. Luikart, and P. Taberlet (2003). "Landscape genetics: combining landscape ecology and population genetics". In: *Trends in Ecology & Evolution* 18.4, pp. 189–197. DOI: 10.1016/S0169-5347(03)00008-9.

Mantel, N. (1967). "The detection of disease clustering and a generalized regression approach". In: *Cancer Research* 27.2 Part 1, pp. 209–220. ISSN: 0008-5472.

Mason, S. J. and G. M. Mimmack (1992). "The use of bootstrap confidence intervals for the correlation coefficient in climatology". In: *Theoretical and applied climatology* 45.4, pp. 229–233. ISSN: 0177-798X.

McRae, B. H. (2006). "Isolation by resistance". In: *Evolution* 60.8, pp. 1551–1561. ISSN: 00143820, 15585646.

McRae, B. H., B. G. Dickson, T. H. Keitt, and V. B. Shah (2008). "Using circuit theory to model connectivity in ecology, evolution and conservation". In: *Ecology (Durham)* 89.10, pp. 2712–2724. ISSN: 0012-9658.

Medrano, M., E. López-Perea, and C. M. Herrera (2014). "Population Genetics Methods Applied to a Species Delimitation Problem: Endemic Trumpet Daffodils (Narcissus Section Pseudonarcissi) from the Southern Iberian Peninsula". In: *International journal of plant sciences* 175.5, pp. 501–517. ISSN: 1058-5893.

Meirmans, P. G. (2012). "The trouble with isolation by distance". In: *Molecular Ecology* 21.12, pp. 2839–2846. DOI: https://doi.org/10.1111/j.1365-294X.2012.05578.x.

Meirmans, P. G. and P. W. Hedrick (2011). "Assessing population structure: FST and related measures". In: *Molecular Ecology Resources* 11.1, pp. 5–18. DOI: https://doi.org/10.1111/j.1755-0998.2010.02927.x.

Nei, M. (1972). "Genetic distance between populations". In: *The American naturalist* 106.949, pp. 283–292. ISSN: 0003-0147.

— (1973). "Analysis of gene diversity in subdivided populations". In: *Proceedings of the National Academy of Sciences - PNAS* 70.12, pp. 3321–3323. ISSN: 0027-8424.

Neuenschwander, S., F. Michaud, and J. Goudet (2018). "QuantiNemo 2: a Swiss knife to simulate complex demographic and genetic scenarios, forward and backward in time". In: *Bioinformatics* 35.5, pp. 886–888. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty737.

Oksanen, J., G. L. Simpson, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, P. Solymos, M. H. H. Stevens, E. Szoecs, H. Wagner, M. Barbour, M. Bedward, B. Bolker, D. Borcard, G. Carvalho, M. Chirico, M. De Caceres, S. Durand, H. B. A. Evangelista, R. FitzJohn, M. Friendly, B. Furneaux, G. Hannigan, M. O. Hill, L. Lahti, D. McGlinn, M.-H. Ouellette, E. Ribeiro Cunha, T. Smith, A. Stier, C. J. F. Ter Braak, and J. Weedon (2022). *vegan: Community Ecology Package*. R package version 2.6-4.

Pedraza-Marrón, C.d.R., R. Silva, J. Deeds, S. M. Van Belleghem, A. Mastretta-Yanes, O. Domínguez-Domínguez, R. A. Rivero-Vega, L. Lutackas, D. Murie, D. Parkyn, L. H. Bullock, K. Foss, H. Ortiz-Zuazaga, J. Narváez-Barandica, A. Acero, G. Gomes, and R. Betancur-R (2019). "Genomics overrules mitochondrial DNA, siding with morphology on a controversial case of species delimitation". In: *Proceedings of the Royal Society B: Biological Sciences* 286.1900, p. 20182924. DOI: 10.1098/rspb.2018.2924.

Peterman, W. E. (2018). "ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms". In: *Methods in Ecology and Evolution* 9.6, pp. 1638–1647. DOI: https://doi.org/10.1111/2041-210X.12984.

Peterman, W. E. and N. S. Pope (2021). "The use and misuse of regression models in landscape genetic analyses". In: *Molecular Ecology* 30.1, pp. 37–47. DOI: https://doi.org/10.1111/mec.15716.

Pope, L. C., L. Liggins, J. Keyse, S. B. Carvalho, and C. Riginos (2015). "Not the time or the place: the missing spatio-temporal link in publicly available genetic data". In: *Molecular Ecology* 24.15, pp. 3802–3809. DOI: https://doi.org/10.1111/mec.13254.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000). "Inference of population structure using multilocus genotype data". In: *Genetics (Austin)* 155.2, pp. 945–959. ISSN: 0016-6731.

Rannala, B. and Z. Yang (2020). "Species delimitation". In: *Phylogenetics in the genomic era*. Ed. by C. Scornavacca, F. Delsuc, and N. Galtier. Self published, 5.5:1–5.5:18.

Raxworthy, C. J., C. M. Ingram, N. Rabibisoa, and R. G. Pearson (2007). "Applications of ecological niche modeling for species delimitation: A review and empirical evaluation using day geckos (Phelsuma) from Madagascar". In: *Systematic Biology* 56.6, pp. 907–923. ISSN: 1063-5157. DOI: 10.1080/10635150701775111.

Rissler, L. J. and J. J. Apodaca (2007). "Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (Aneides flavipunctatus)". In: *Systematic Biology* 56.6, pp. 924–942. ISSN: 1063-5157. DOI: 10.1080/10635150701703063.

Rousset, F. (1997). "Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance". In: *Genetics* 145.4, pp. 1219–1228. ISSN: 1943-2631. DOI: 10.1093/genetics/145.4.1219.

— (2000). "Genetic differentiation between individuals". In: *Journal of evolutionary biology* 13.1, pp. 58–62. ISSN: 1010-061X.

— (2008). "genepop'007: a complete re-implementation of the genepop software for Windows and Linux". In: *Molecular Ecology Resources* 8.1, pp. 103–106. DOI: https://doi.org/10.1111/j.1471-8286.2007.01931.x.

Royston, P. (2007). "Profile likelihood for estimation and confidence intervals". In: *The Stata Journal* 7.3, pp. 376–387.

SAS Institute (2001). *SAS/STAT user's guide. Ver. 8*. Cary, NC.

Scapini, F., A. Aloia, M. F. Bouslama, L. Chelazzi, I. Colombini, M. ElGtari, M. Fallaci, and G. M. Marchetti (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". In: *Behavioral Ecology and Sociobiology* 51, pp. 403–414.

Sentinella, A. T., A. T. Moles, J. G. Bragg, M. Rossetto, and W. B. Sherwin (2022). "Detecting steps in spatial genetic data: which diversity measures are best?" In: *PloS one* 17.3, e0265110–e0265110. ISSN: 1932-6203.

Séré, M., S. Thévenon, A. M. G. Belem, and T. De Meeûs (2017). "Comparison of different genetic distances to test isolation by distance between populations". In: *Heredity* 119.2, pp. 55–63. ISSN: 0018-067X.

Shao, J. and D. Tu (2012). *The Jackknife and Bootstrap*. Springer Series in Statistics. Springer New York. ISBN: 9781461207955.

Shirk, A. J., E. L. Landguth, and S. A. Cushman (2018). "A comparison of regression methods for model selection in individual-based landscape genetic analysis". In: *Molecular Ecology Resources* 18.1, pp. 55–67. DOI: https://doi.org/10.1111/1755-0998.12709.

Slatkin, M. (1993). "Isolation by distance in equilibrium and non-equilibrium populations". In: *Evolution* 47.1, pp. 264–279. ISSN: 0014-3820.

Smouse, P. E., J. C. Long, and R. R. Sokal (1986). "Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence". In: *Systematic Zoology* 35.4, pp. 627–632. ISSN: 00397989.

Spriggs, E. L., D. A. R. Eaton, P. W. Sweeney, C. Schlutius, E. J. Edwards, and M. J. Donoghue (2018). "Restriction-site-associated DNA sequencing reveals a cryptic Viburnum species on the north American coastal plain". In: *Systematic Biology* 68.2, pp. 187–203. ISSN: 1063-5157. DOI: 10.1093/sysbio/syy084.

Storfer, A., M. A. Murphy, S. F. Spear, R. Holderegger, and L. P. Waits (2010). "Landscape genetics: where are we now?" In: *Molecular Ecology* 19.17, pp. 3496–3514.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). "Measuring and testing dependence by correlation of distances". In: *The Annals of Statistics* 35.6, pp. 2769–2794. DOI: 10.1214/009053607000000505.

Terasaki Hart, D. E., A. P. Bishop, and I. J. Wang (2021). "Geonomics: Forward-Time, Spatially Explicit, and Arbitrarily Complex Landscape Genomic Simulations". In: *Molecular Biology and Evolution* 38.10, pp. 4634–4646. ISSN: 1537-1719. DOI: 10.1093/molbev/msab175.

Thorndike, R. L. (1953). "Who belongs in the family?" In: *Psychometrika* 18.4, pp. 267–276. ISSN: 1860-0980. DOI: 10.1007/BF02289263.

Tibshirani, R., G. Walther, and T. Hastie (2001). "Estimating the number of clusters in a data set via the gap statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 411–423. DOI: https://doi.org/10.1111/1467-9868.00293.

Van Strien, M. J., R. Holderegger, and H. J. Van Heck (2015). "Isolation-by-distance in landscapes: considerations for landscape genetics". In: *Heredity* 114.1, pp. 27–37. ISSN: 0018-067X.

Van Strien, M. J., D. Keller, and R. Holderegger (2012). "A new analytical approach to landscape genetic modelling: least-cost transect analysis and linear mixed models". In: *Molecular Ecology* 21.16, pp. 4010–4023. ISSN: 0962-1083.

Vekemans, X. and O. J. Hardy (2004). "New insights from fine-scale spatial genetic structure analyses in plant populations". In: *Molecular Ecology* 13.4, pp. 921–935. DOI: https://doi.org/10.1046/j.1365-294X.2004.02076.x.

Venzon, D. J. and S. H. Moolgavkar (1988). "A method for computing profile-likelihood-based confidence intervals". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 37.1, pp. 87–94.

Vera, J. F. (2022). "Distance-based logistic model for cross-classified categorical data". In: *British Journal of Mathematical and Statistical Psychology* 75.3, pp. 466–492. ISSN: 2044-8317. DOI: 10.1111/BMSP.12264.

Verbeke, G. and G. Molenberghs (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.

Verity, R. and R. A. Nichols (2016). "Estimating the Number of Subpopulations (K) in Structured Populations". In: *Genetics* 203.4, pp. 1827–1839. ISSN: 1943-2631. DOI: 10.1534/genetics.115.180992.

Virgoulay, T., F. Rousset, and R. Leblois (2021a). *GSpace User Guide*. https://www1.montpellier.inrae.fr/CBGP/software/gspace/ressources/documentation/GSpace_User_Guide.pdf. Accessed: 2023-12-07.

— (2021b). "GSpace: an exact coalescence simulator of recombining genomes under isolation by distance". In: *Bioinformatics* 37.20, pp. 3673–3675.

Wagner, H. H. and M.-J. Fortin (2015). "Basics of spatial data analysis: linking landscape and genetic data for landscape genetic studies". In: *Landscape Genetics*. Ed. by N. Balkenhol, S. A. Cushman, A. T. Storfer, and L. P. Waits. John Wiley & Sons, Ltd. Chap. 5, pp. 77–98. ISBN: 9781118525258. DOI: https://doi.org/10.1002/9781118525258.ch05.

Waits, L. P. and A. Storfer (2015). "Basics of population genetics: quantifying neutral and adaptive genetic variation for landscape genetic studies". In: *Landscape Genetics*. Ed. by N. Balkenhol, S. A. Cushman, A. T. Storfer, and L. P. Waits. John Wiley & Sons, Ltd. Chap. 3, pp. 35–57. ISBN: 9781118525258. DOI: https://doi.org/10.1002/9781118525258.ch03.

Wang, I. J. and G. S. Bradburd (2014). "Isolation by environment". In: *Molecular Ecology* 23.23, pp. 5649–5662. ISSN: 0962-1083.

Watts, P. C., F. Rousset, I. J. Saccheri, R. Leblois, S. J. Kemp, and D. J. Thompson (2007). "Compatible genetic and ecological estimates of dispersal rates in insect (Coenagrion mercuriale: Odonata: Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator". In: *Molecular Ecology* 16.4, pp. 737–751. DOI: https://doi.org/10.1111/j.1365-294X.2006.03184.x.

Weir, B. S. and C. Clark Cockerham (1984). "Estimating F-statistics for the analysis of population structure". In: *Evolution* 38.6, pp. 1358–1370. ISSN: 0014-3820.

Welch, B. L. (1947). "The generalization of "Student's" problem when several different population variances are involved". In: *Biometrika* 34.1-2, pp. 28–35. ISSN: 0006-3444. DOI: 10.1093/biomet/34.1-2.28.

West, B. T., K. B. Welch, and A. T. Galecki (2022). *Linear mixed models: a practical guide using statistical software*. Crc Press.

Whitlock, M. C. (2011). "GST and D do not replace FST". In: *Molecular ecology* 20.6, pp. 1083–1091. ISSN: 0962-1083.

Wright, S. (1943). "Isolation by distance". In: *Genetics* 28.2, p. 114.

— (1949). "The genetical structure of populations". In: *Annals of Eugenics* 15.1, pp. 323–354. DOI: https://doi.org/10.1111/j.1469-1809.1949.tb02451.x.

Yang, R. (2004). “A Likelihood-based approach to estimating and testing for isolation by distance”. In: *Evolution* 58.8, pp. 1839–1845. DOI: https://doi.org/10.1111/j.0014-3820.2004.tb00466.x.

Yuan, X., D. J. Miller, J. Zhang, D. Herrington, and Y. Wang (2012). “An overview of population genetic data simulation”. In: *Journal of Computational Biology* 19.1, pp. 42–54.

Appendix A

Additional charts and tables

A.1 Approaches with known groups

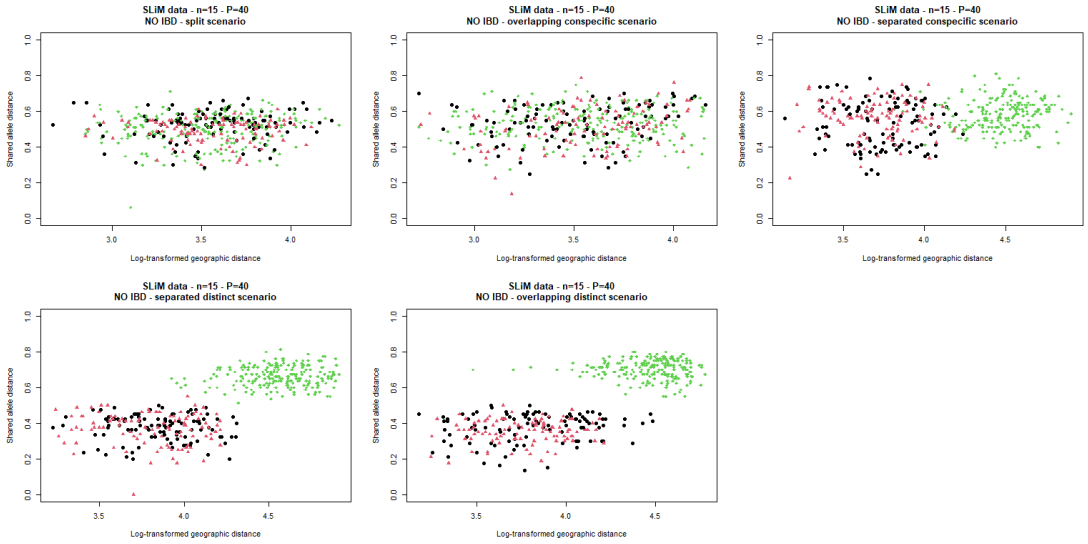

Fig. A.1 Distance-distance plots (shared allele distance against log-transformed geographic distance) based on exemplary datasets from the five scenarios simulated with **SLiM, without IBD behaviour,** 15 **individuals per group** and 40 available loci. Each scenario is specified in the plot title and discussed in section 1.4.1. In black the dissimilarities among individuals belonging to the first group, in red those for the second group and in green the dissimilarities among individuals belonging to different groups.

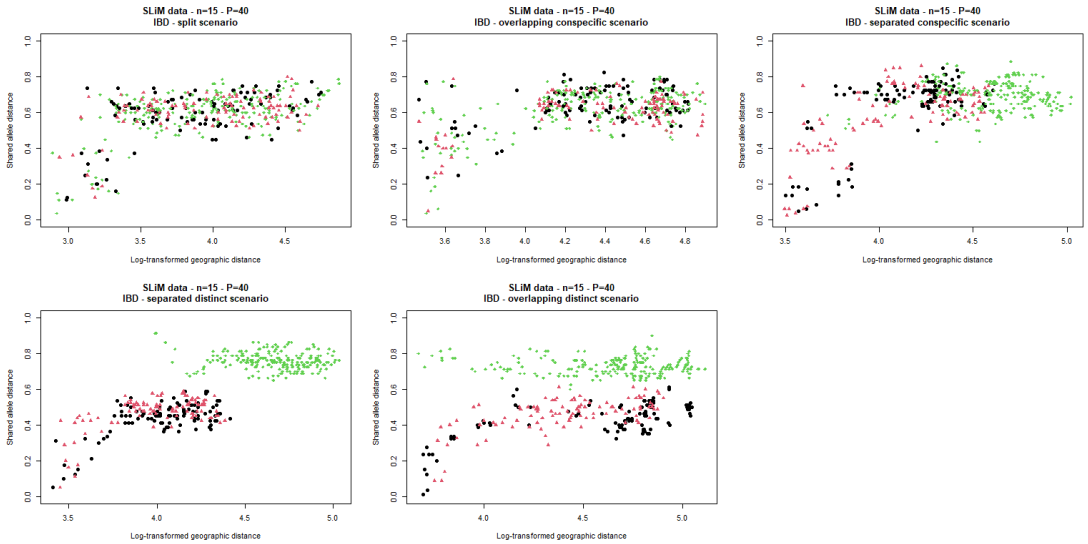

Fig. A.2 Distance-distance plots based on exemplary datasets from the five scenarios simulated with **SLiM, with IBD behaviour,** 15 **individuals per group** and 40 available loci. See caption to Figure A.1 for further description.

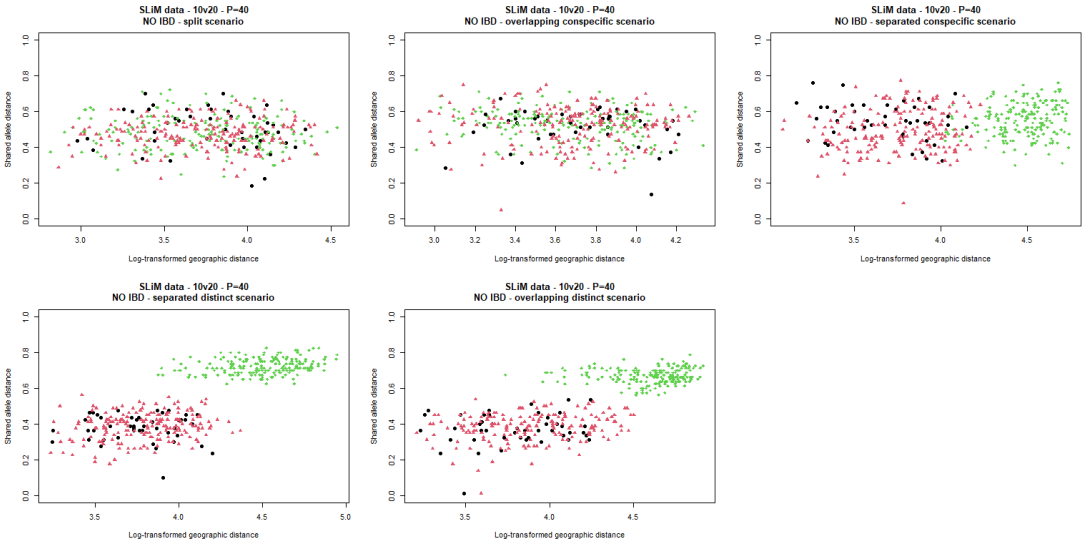

Fig. A.3 Distance-distance plots based on exemplary datasets from the five scenarios simulated with **SLiM, without IBD behaviour,** 10 **versus** 20 **individuals** and 40 available loci. See caption to Figure A.1 for further description.

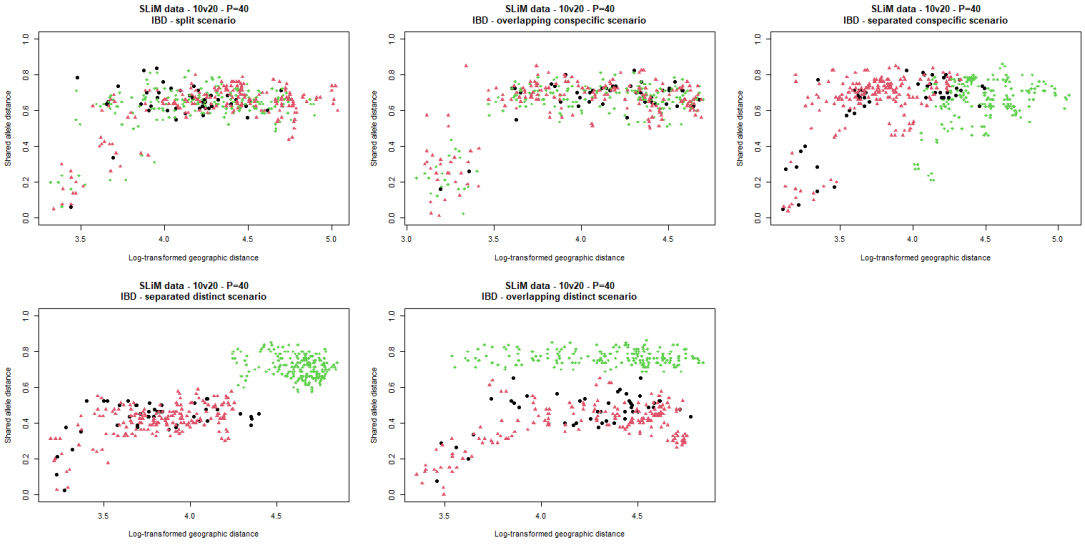

Fig. A.4 Distance-distance plots based on exemplary datasets from the five scenarios simulated with **SLiM, with IBD behaviour,** 10 **versus** 20 **individuals** and 40 available loci. See caption to Figure A.1 for further description.

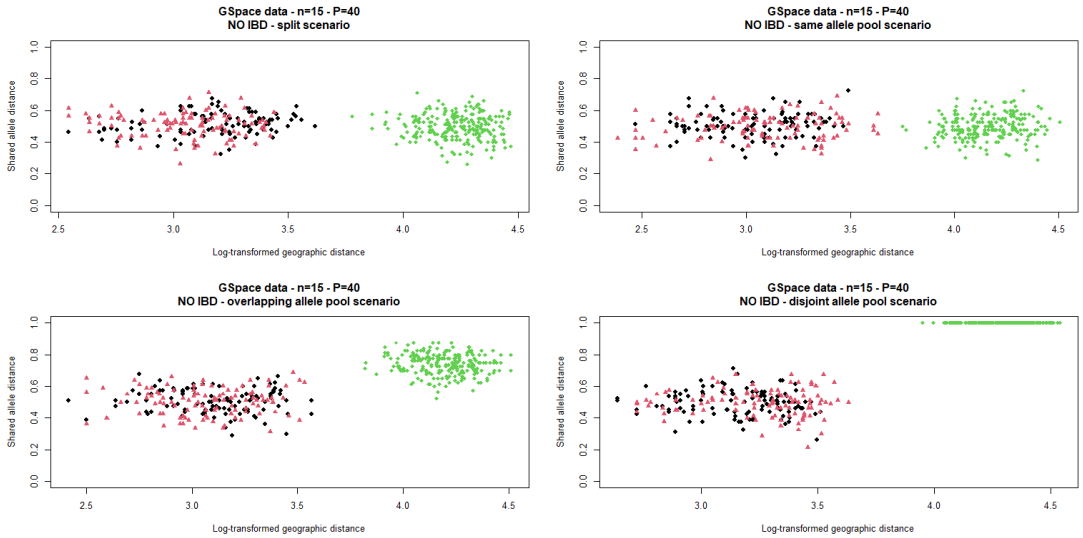

Fig. A.5 Distance-distance plots based on exemplary datasets from the four scenarios simulated with **GSpace, without IBD behaviour, with** 15 **individuals per group** and 40 available loci. Each scenario is specified in the plot title and discussed in section 1.4.2. In black the dissimilarities among individuals belonging to the first group, in red those for the second group and in green the dissimilarities among individuals belonging to different groups.

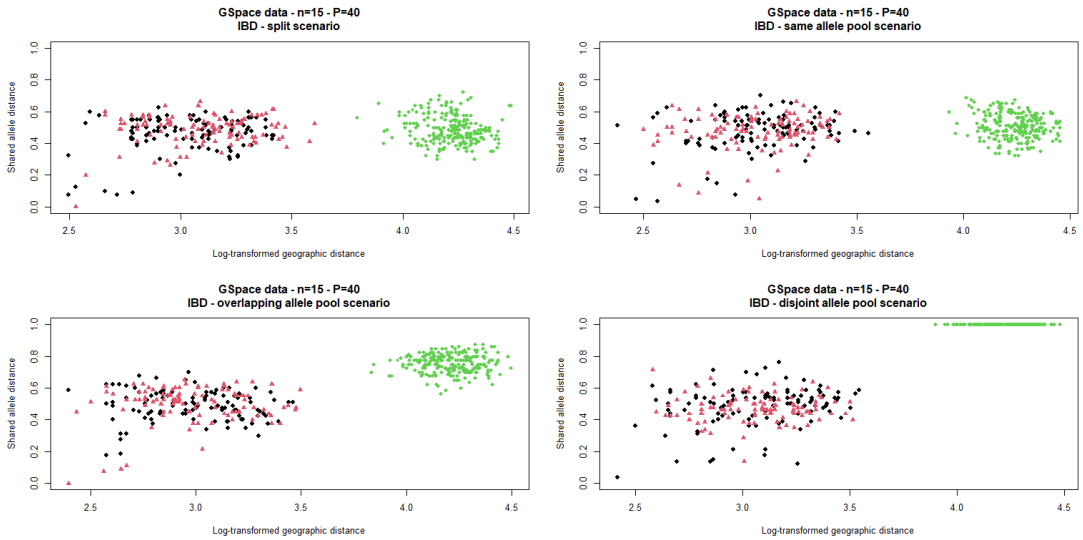

Fig. A.6 Distance-distance plots based on exemplary datasets from the four scenarios simulated with **GSpace, with IBD behaviour,** 15 **individuals per group** and 40 available loci. See caption to Figure A.5 for further description.

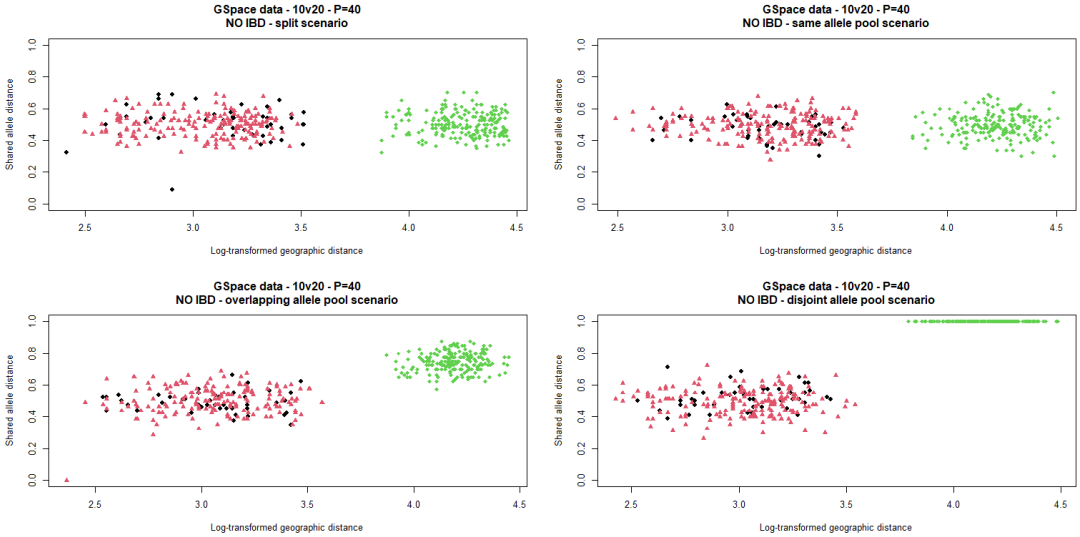

Fig. A.7 Distance-distance plots based on exemplary datasets from the four scenarios simulated with **GSpace, without IBD behaviour, with** 10 **versus** 20 **individuals** and 40 available loci. See caption to Figure A.5 for further description.

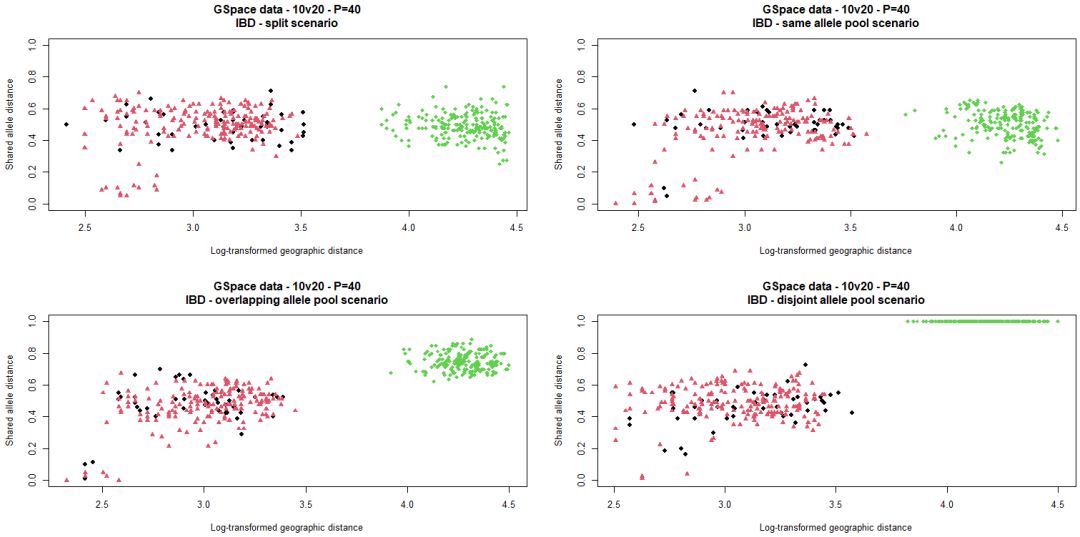

Fig. A.8 Distance-distance plots based on exemplary datasets from the four scenarios simulated with **GSpace, with IBD behaviour,** 10 **versus** 20 **individuals** and 40 available loci. See caption to Figure A.5 for further description.

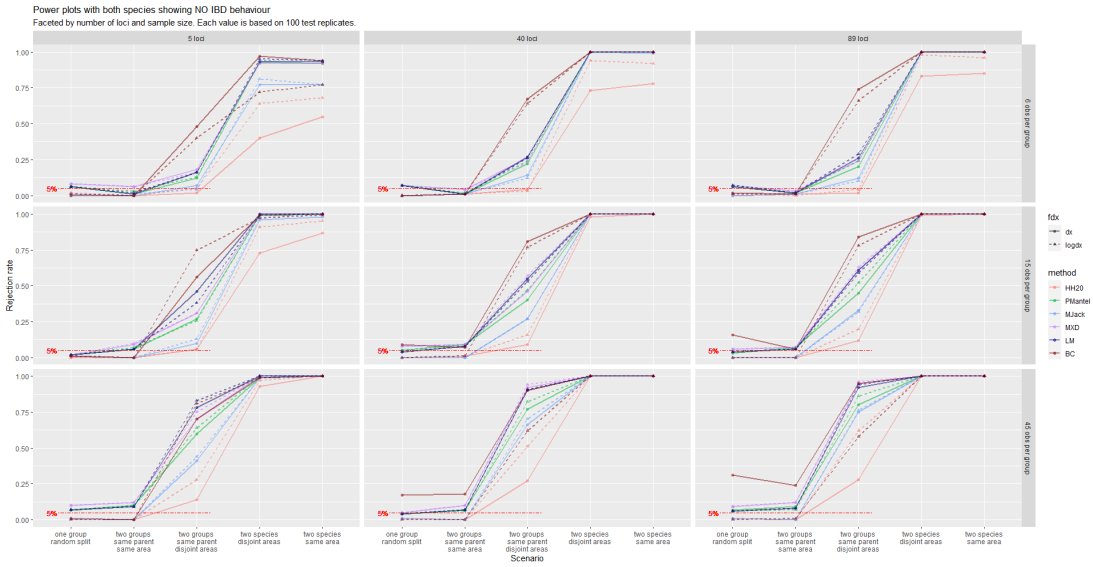

Fig. A.9 Power plot based on **SLiM** data simulated **with equal-sized quasi-panmictic groups**. This extends the left-hand half of Figure 1.3. Facets by number of individuals per group (rows) and number of available loci (columns). Each rejection rate is based on 100 simulations with same parameter settings. Circles and solid lines refer to analyses with untransformed geographic distances, whereas triangles and dashed lines refer to analyses with log-transformed ones. Acronyms are explained in section 1.4.

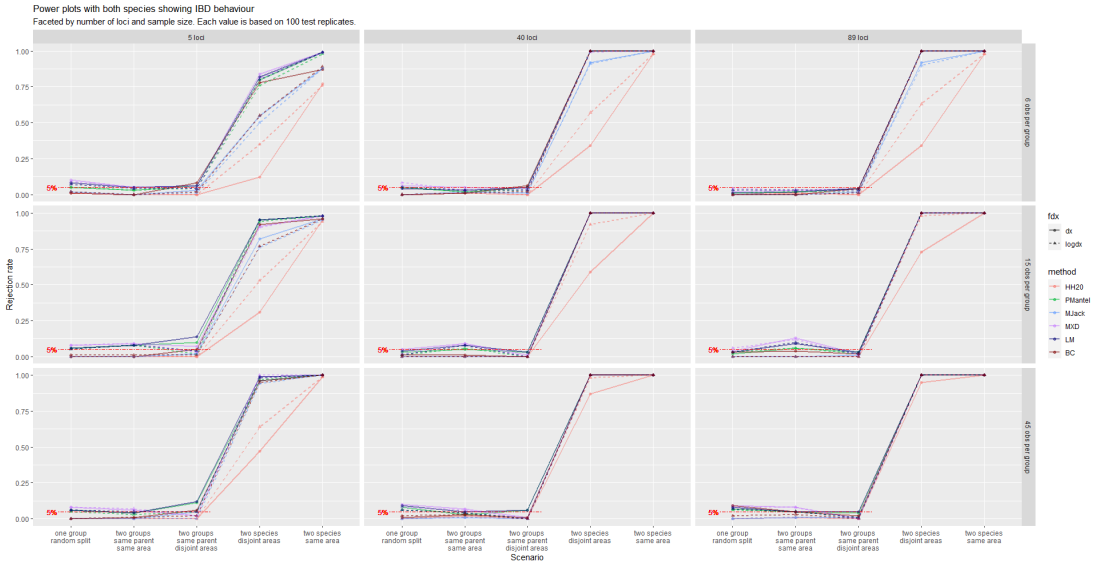

Fig. A.10 Power plot based on **SLiM** data simulated **with equal-sized isolated by distance groups**. This extends the right-hand half of Figure 1.3. See the caption to Figure A.9 for further description.

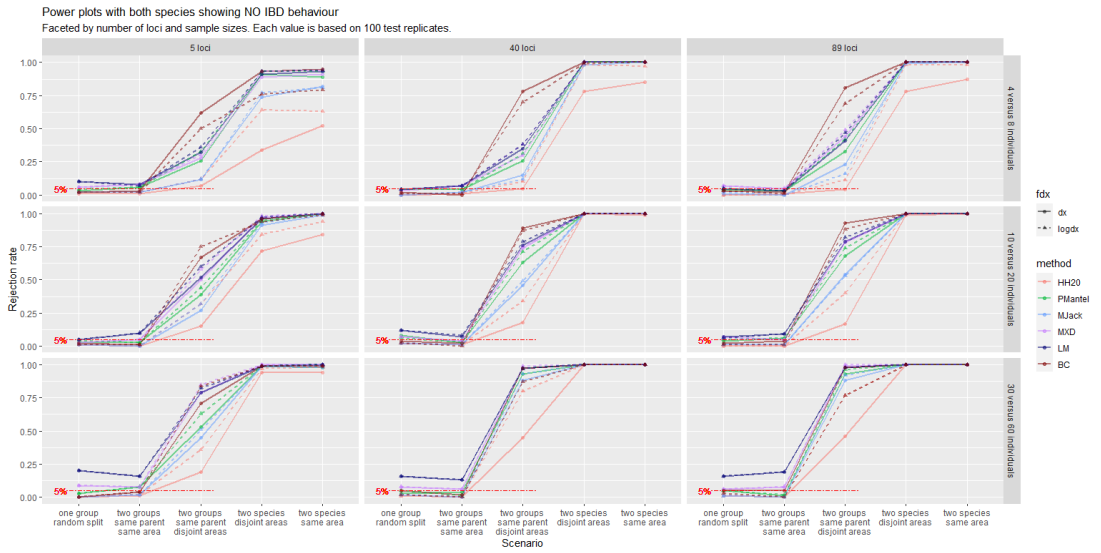

Fig. A.11 Power plot based on **SLiM** data simulated **with unequal-sized quasi-panmictic groups**. This extends the left-hand half of Figure 1.4. See the caption to Figure A.9 for further description.

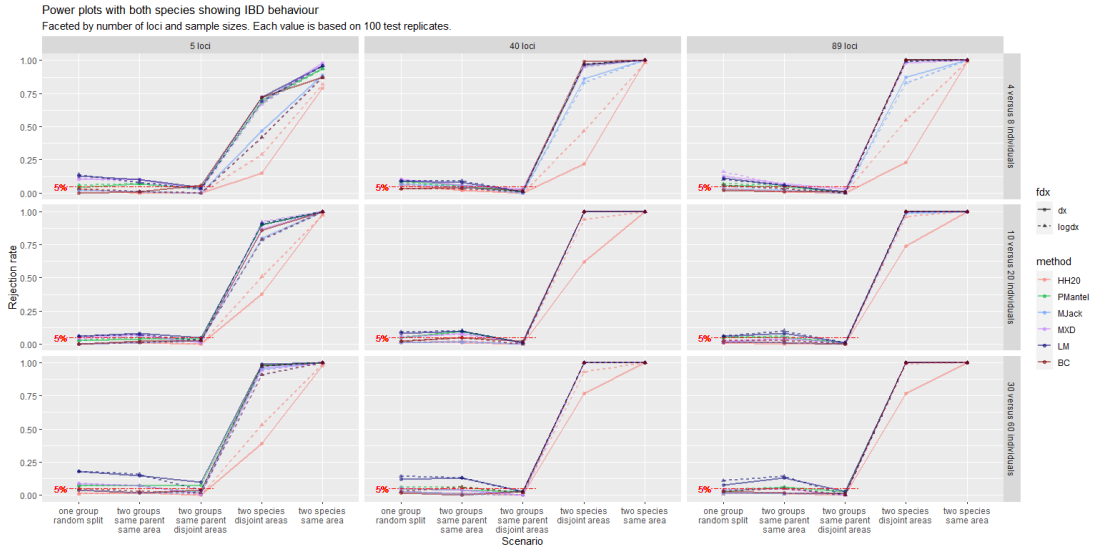

Fig. A.12 Power plot based on **SLiM** data simulated **with unequal-sized isolated by distance groups**. This extends the right-hand half of Figure 1.4. See the caption to Figure A.9 for further description.

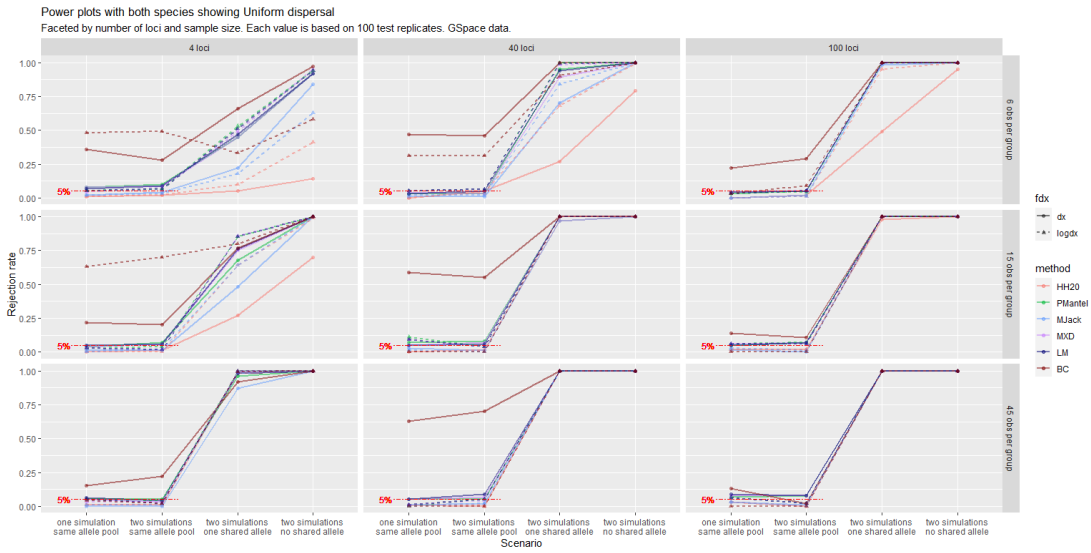

Fig. A.13 Power plot based on **GSpace** data simulated **with equal-sized quasi-panmictic groups**. This extends the left-hand half of Figure 1.5. See the caption to Figure A.9 for further description.

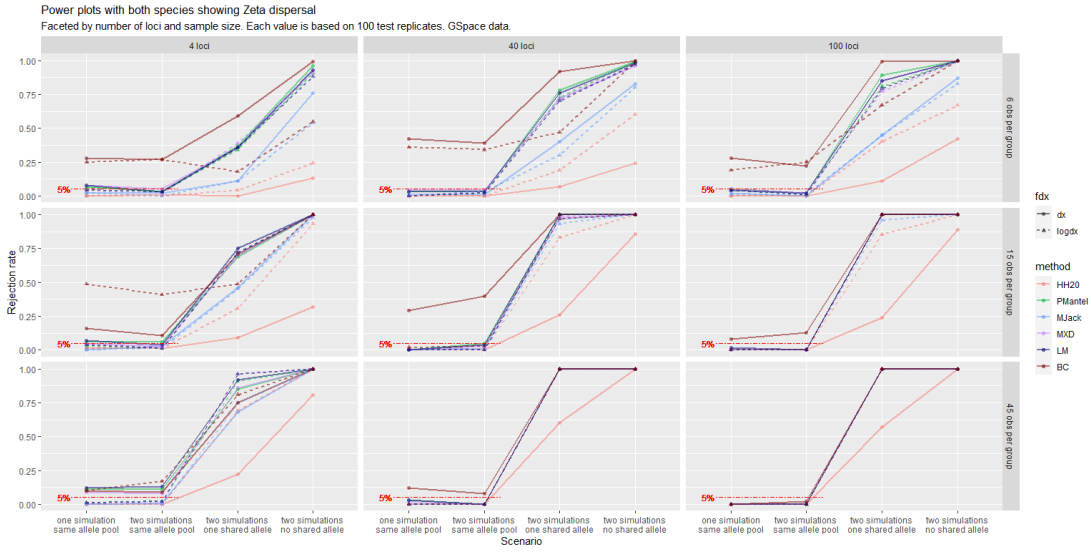

Fig. A.14 Power plot based on **GSpace** data simulated **with equal-sized isolated by distance groups**. This extends the right-hand half of Figure 1.5. See the caption to Figure A.9 for further description.

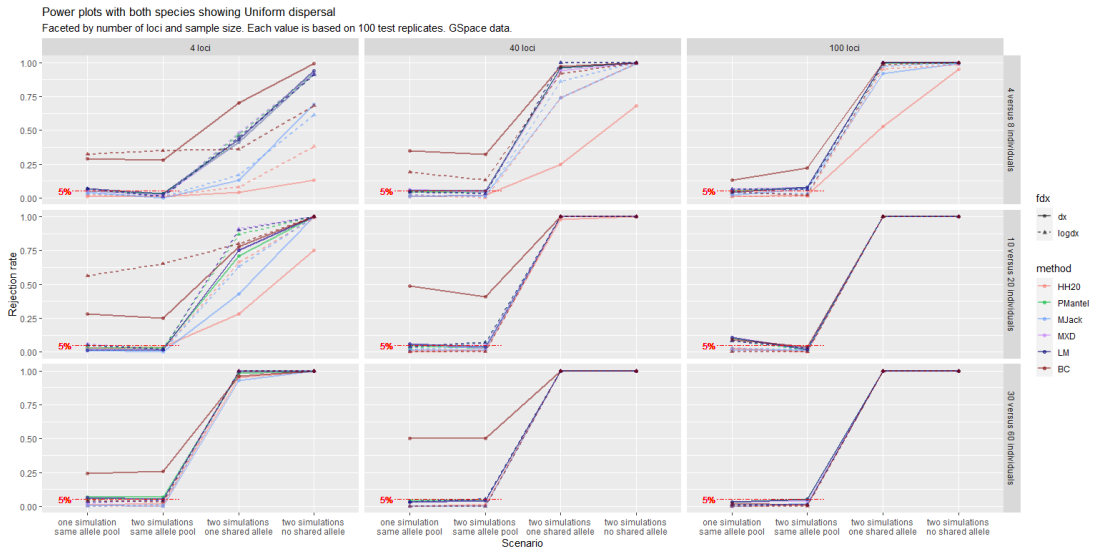

Fig. A.15 Power plot based on **GSpace** data simulated **with unequal-sized quasi-panmictic groups**. This extends the left-hand half of Figure 1.6. See the caption to Figure A.9 for further description.

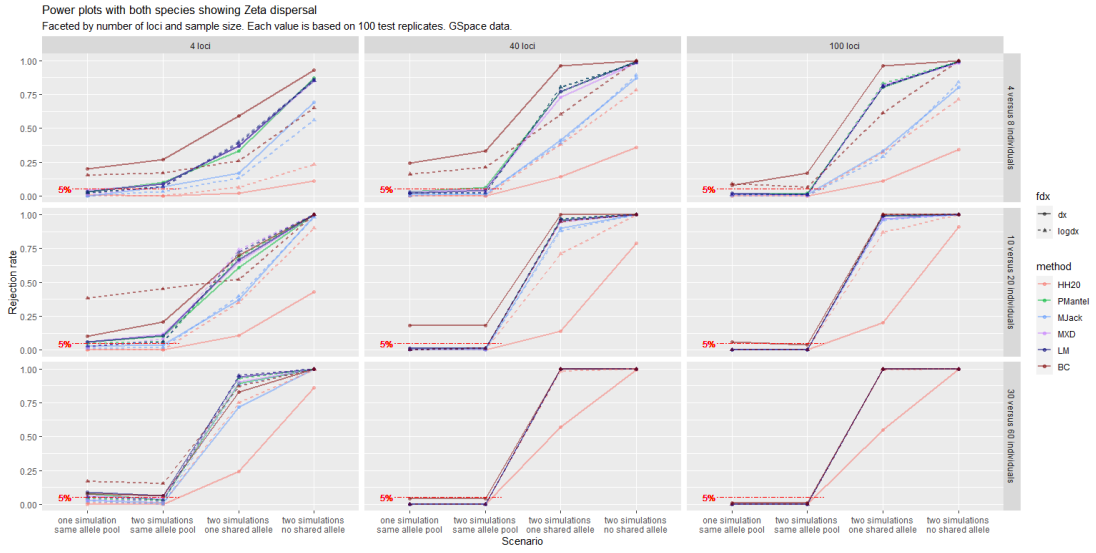

Fig. A.16 Power plot based on **GSpace** data simulated **with unequal-sized isolated by distance groups**. This extends the right-hand half of Figure 1.6. See the caption to Figure A.9 for further description.

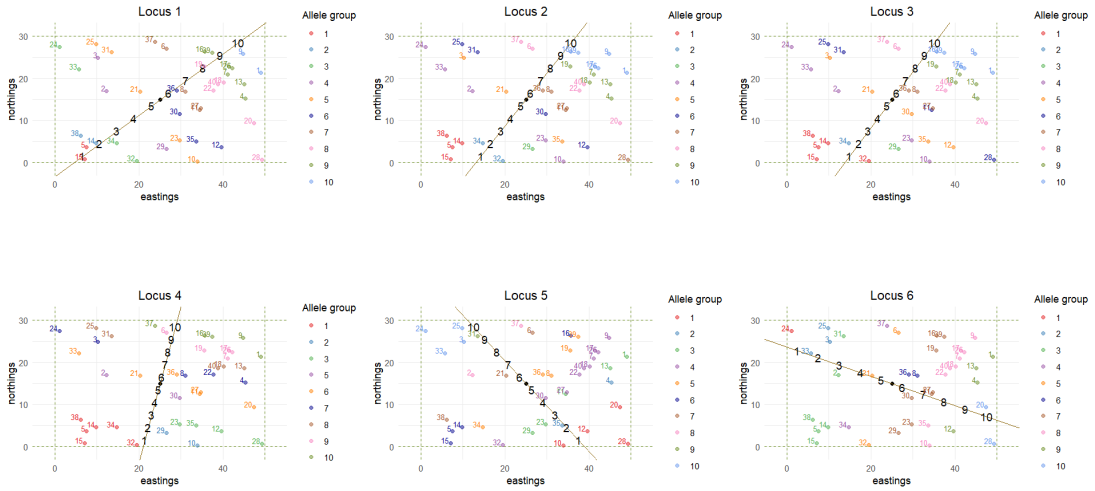

Fig. A.17 Geographic locations of 40 individuals simulated with the algorithm described in section 1.5.1 on a rectangular map. For each locus p, with $p = 1,\ldots,6$, its associated linear gradient is plotted as a black line. Based on the length of the segment of this line that falls within the map boundaries, $M = 10$ equidistant points are identified, one for each allelic status. Each individual is assigned to the allele group whose point on the gradient is closest. This will inform the multinomial draw for the allelic stata in the p^{th} locus for that individual (i.e., allelic status assignment is not deterministic). Back to formula (1.23).

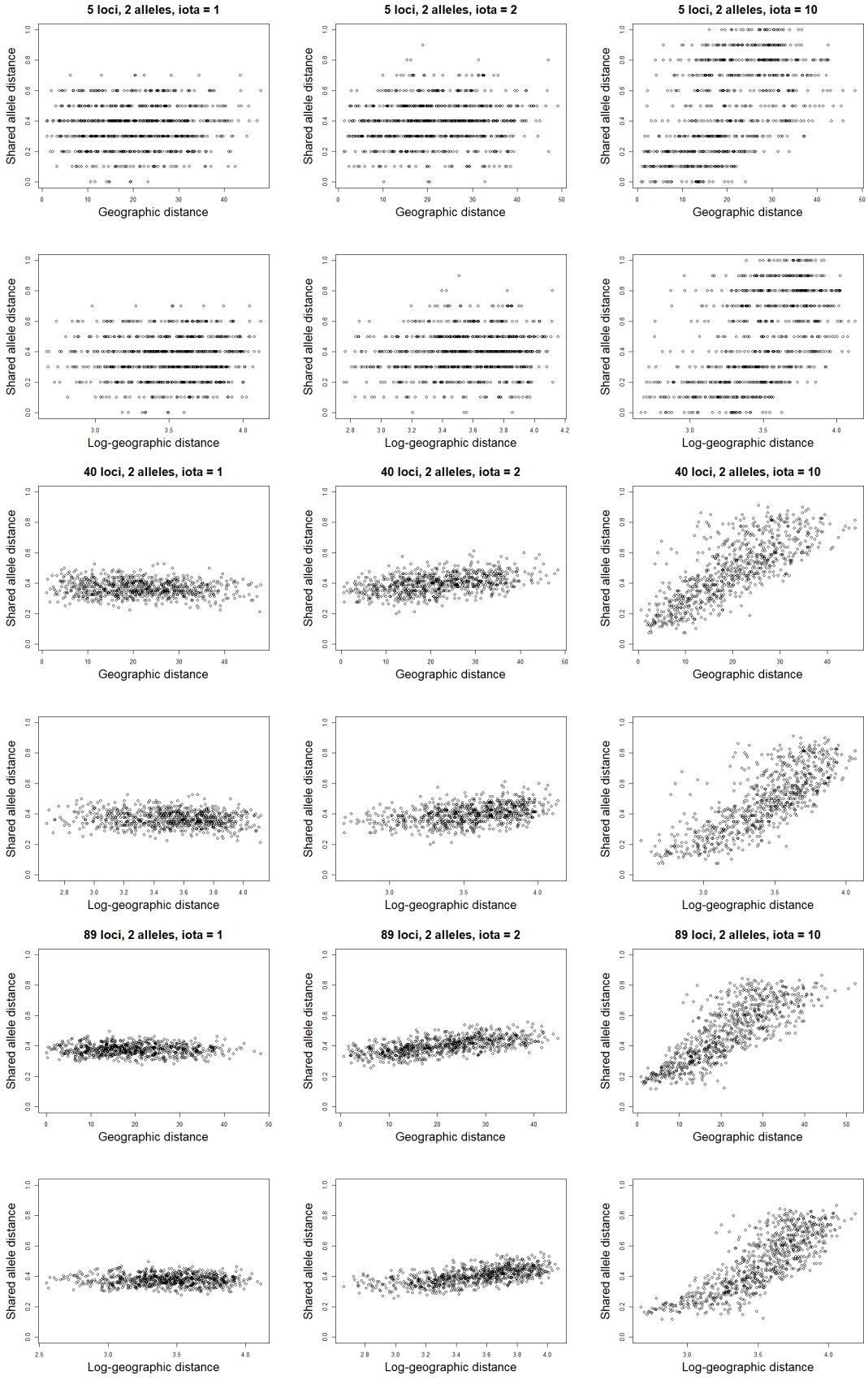

Fig. A.18 Distance-distance plot both with untransformed and log-transformed geographic distance based on data simulated with the algorithm described in section 1.5.1 using 40 individuals on a square map, **2 allelic stata** and $\iota \in \{1, 2, 10\}$.

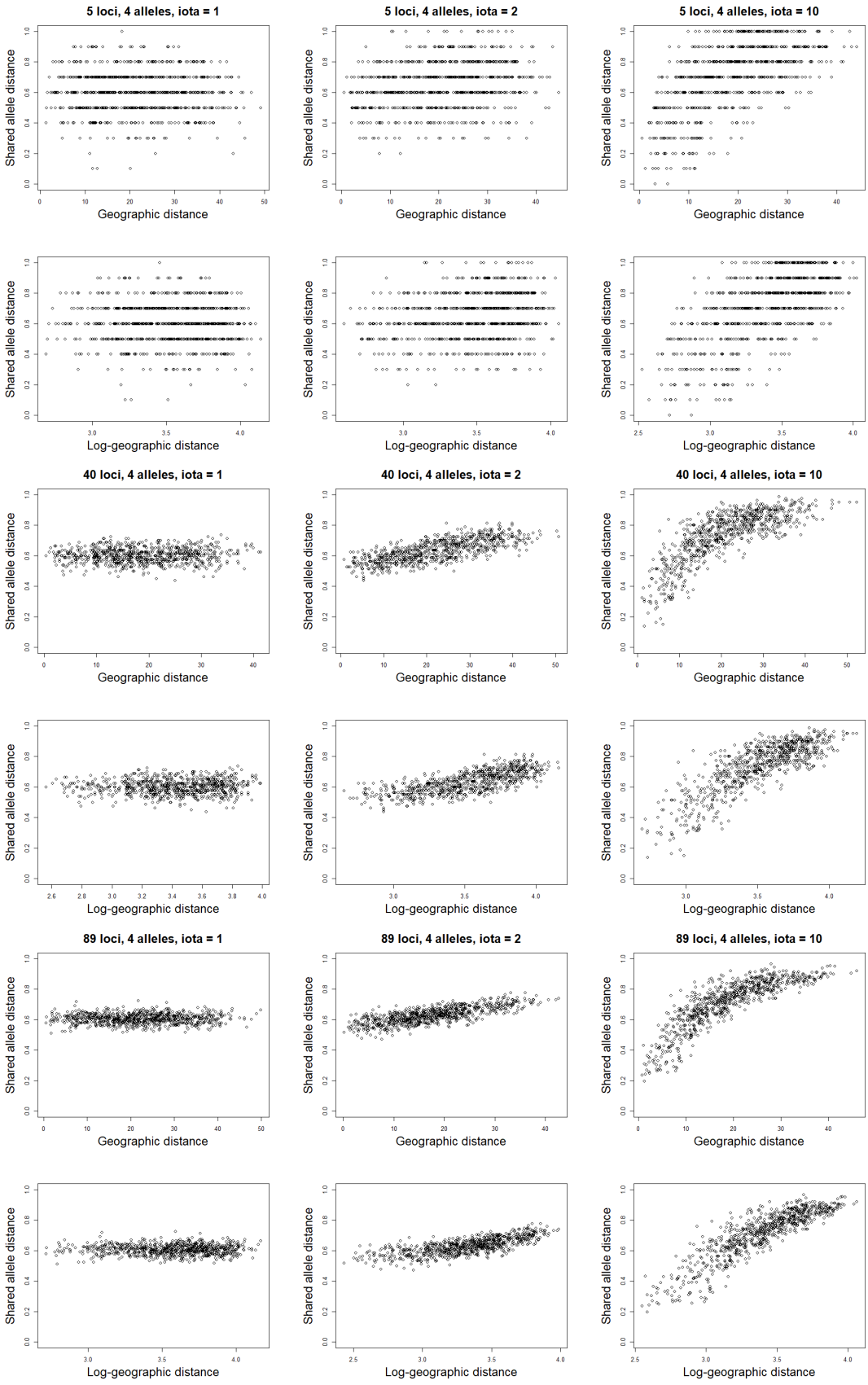

Fig. A.19 Distance-distance plot both with untransformed and log-transformed geographic distance based on data simulated with the algorithm described in section 1.5.1 using 40 individuals on a square map, **4 allelic stata** and $\iota \in \{1, 2, 10\}$.

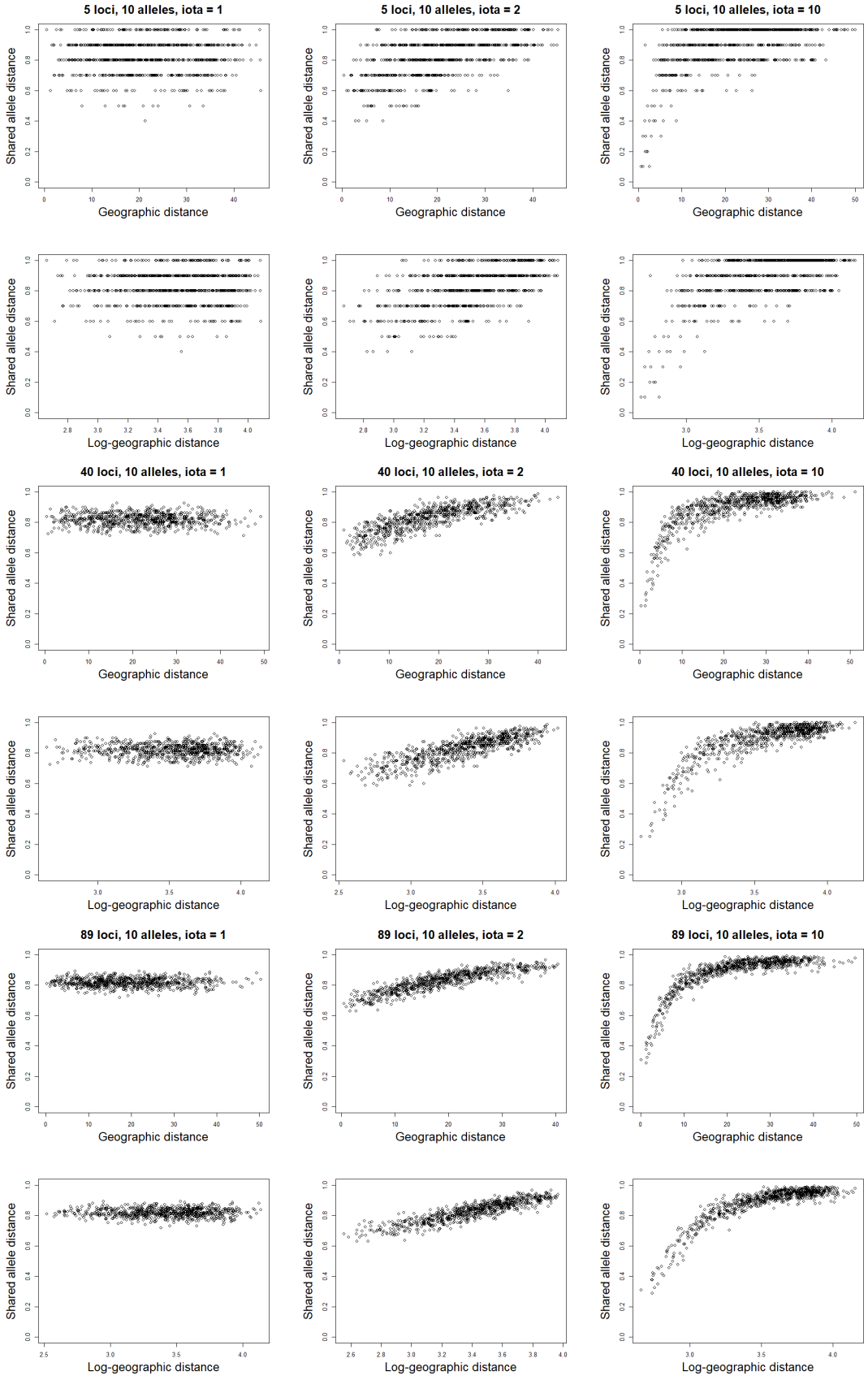

Fig. A.20 Distance-distance plot both with untransformed and log-transformed geographic distance based on data simulated with the algorithm described in section 1.5.1 using 40 individuals on a square map, **10 allelic stata** and $\iota \in \{1, 2, 10\}$.

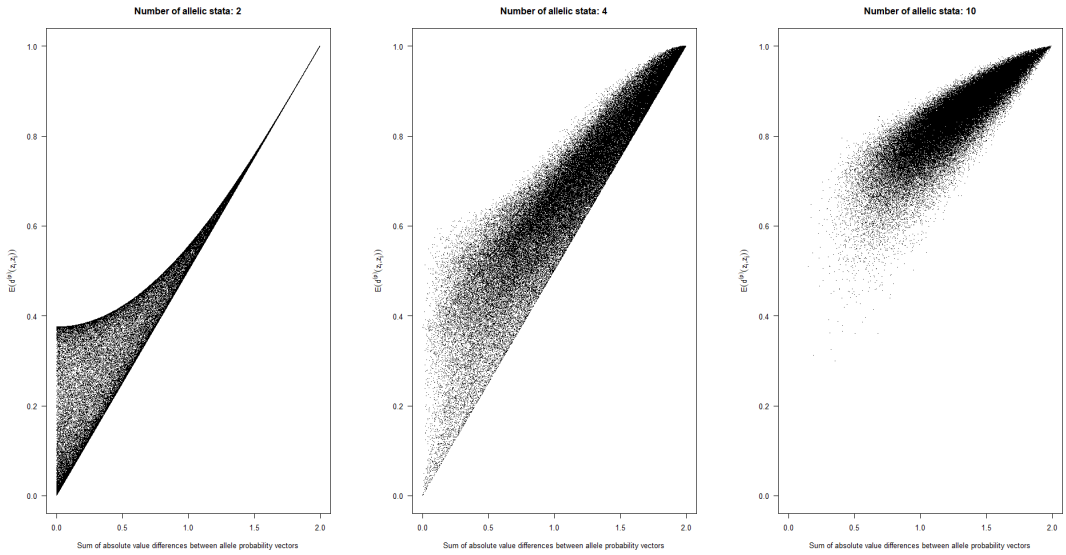

Fig. A.21 Expectation of the shared allele distance between two loci as a function of the Manhattan distance between the probability vectors on which their multinomial draw was based. Panels by number of allelic stata M. Back to formula (1.24).

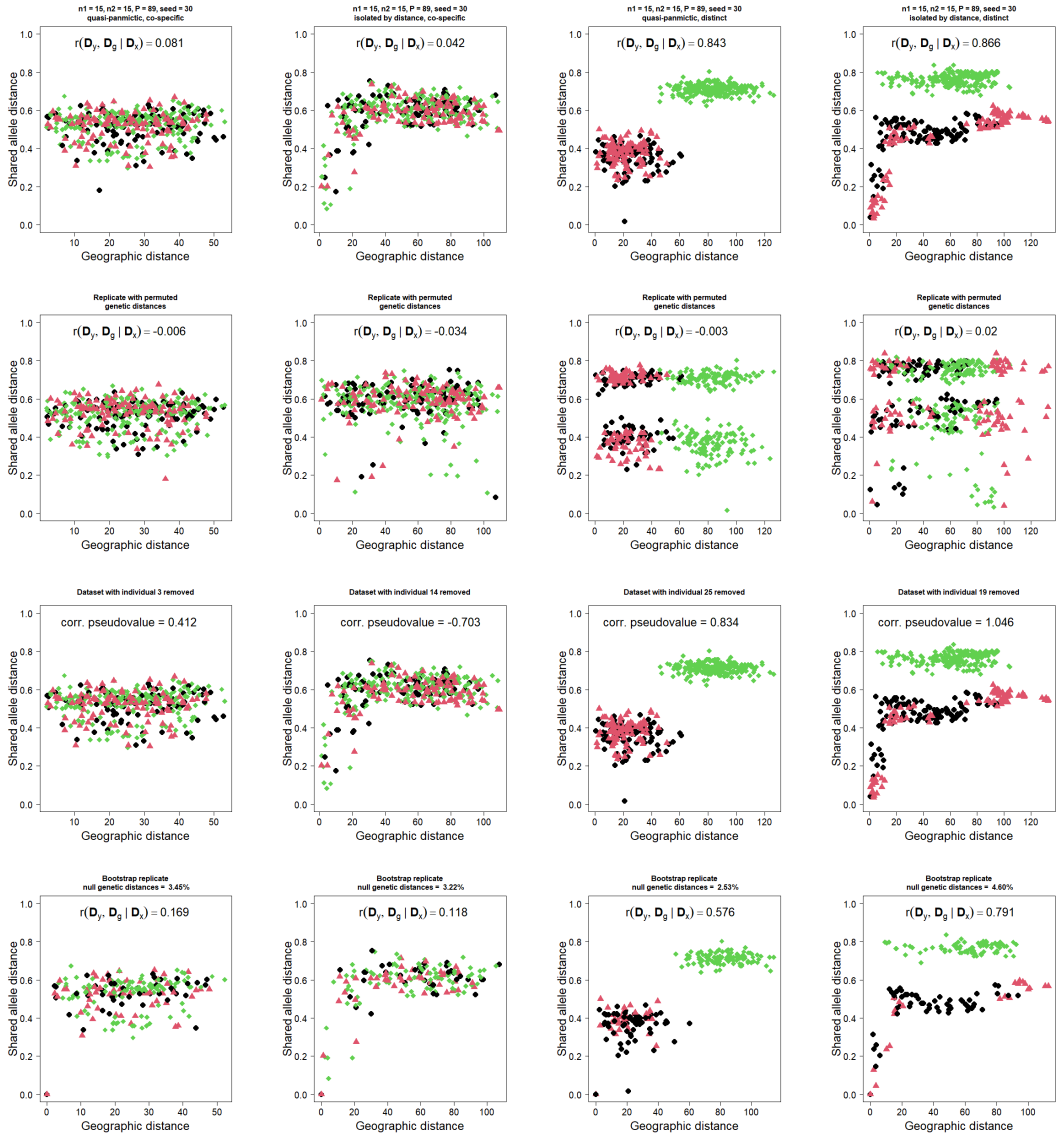

Fig. A.22 Distance-distance plots based on four **SLiM datasets with equally sized groups**, 30 individuals in total and 89 loci (second and last plot in the middle row of Figure 1.3). First two columns refer to datasets from the *split* scenario, the last two to datasets from the *overlapping distinct* scenario. In both cases, a dataset with quasi-panmictic and one with IBD species is shown. For each of the four datasets (columns), the first row reports original dissimilarities, the second row reports dissimilarities from a dataset with permuted $\mathbf{D}_y$, the third row those from a jackknife replicate of the dataset, the last row those from a bootstrap replicate of the dataset. For each plot, the estimated partial correlation coefficient (or its jackknife pseudovalue) is superimposed. Color coding as in Figure 1.1. Back to section 1.5.3.

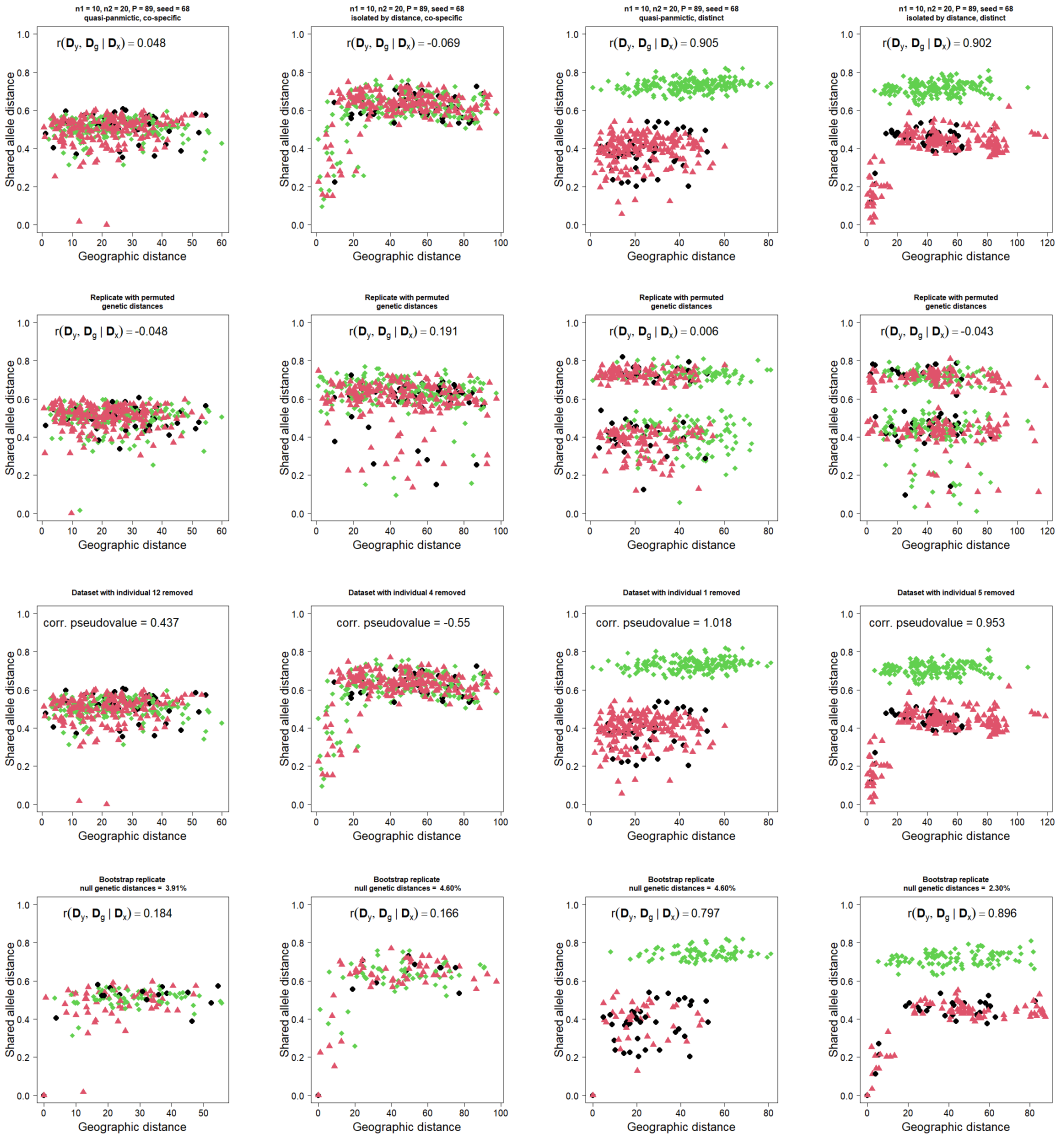

Fig. A.23 Distance-distance plots based on four **SLiM datasets with unequally sized groups**, 30 individuals in total and 89 loci (second and last plot in the middle row of Figure 1.4). For further description, see Figure A.22. Back to section 1.5.3.

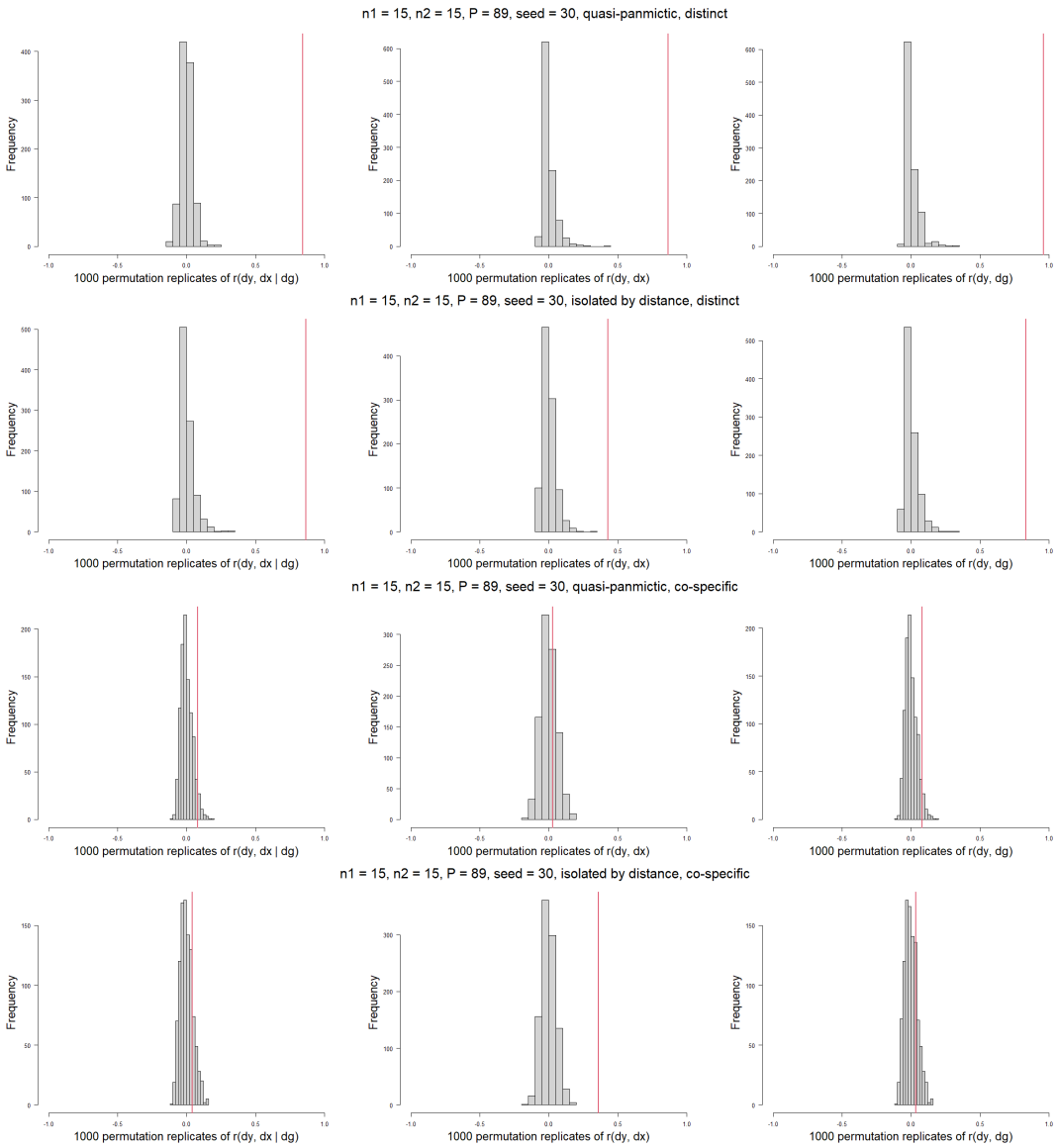

Fig. A.24 Empirical distribution of permutation replicates of the partial correlation between $\mathbf{D}_y$ and $\mathbf{D}_g$ given $\mathbf{D}_x$, simple correlation between $\mathbf{D}_y$ and $\mathbf{D}_x$ and simple correlation between $\mathbf{D}_y$ and $\mathbf{D}_g$ for the same SLiM datasets in Figure A.22. In each plot, the vertical red line indicates the original value of the replicated correlation. Back to section 1.5.3.

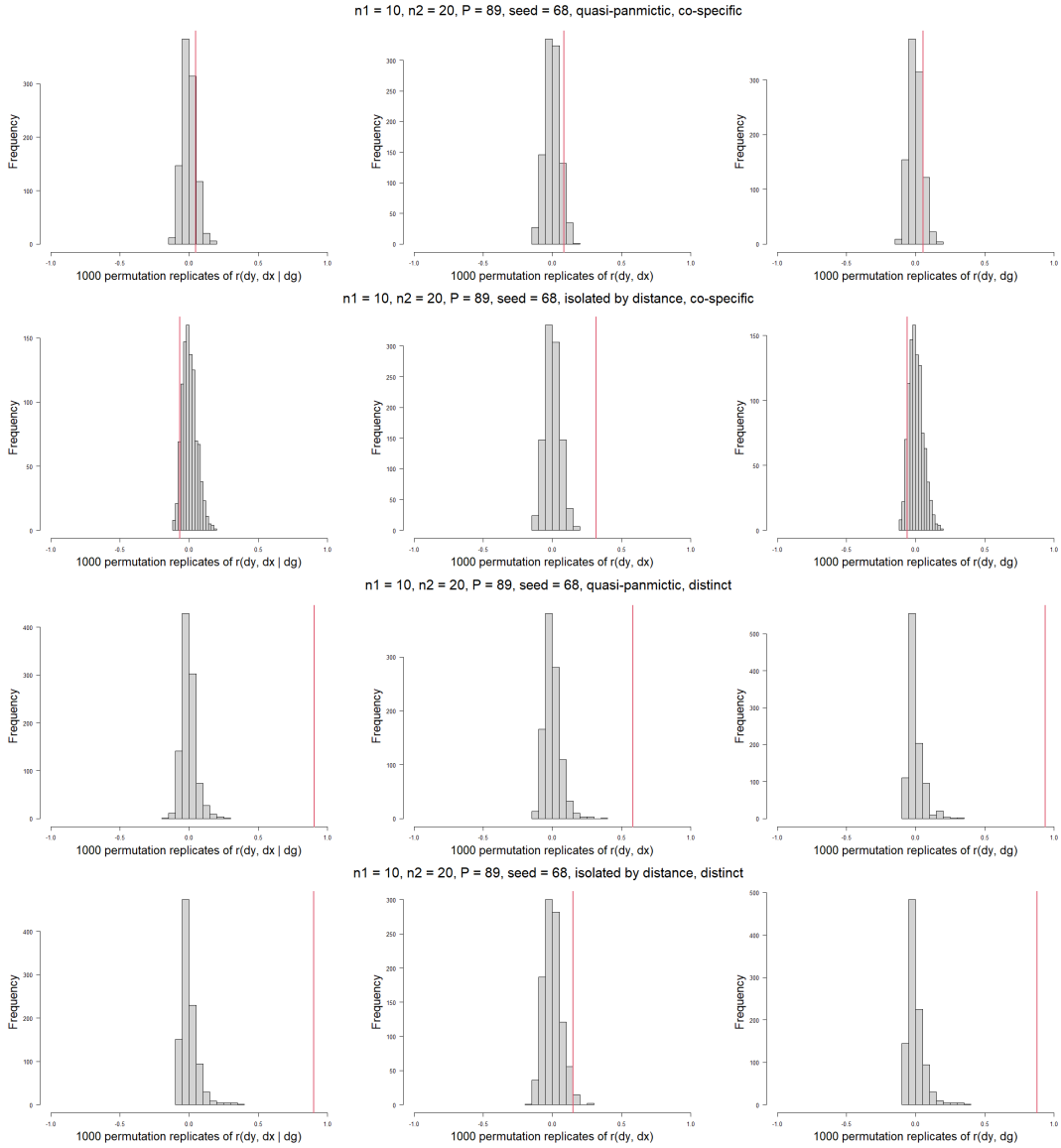

Fig. A.25 Empirical distribution of permutation replicates of the partial correlation between $\mathbf{D}_y$ and $\mathbf{D}_g$ given $\mathbf{D}_x$, simple correlation between $\mathbf{D}_y$ and $\mathbf{D}_x$ and simple correlation between $\mathbf{D}_y$ and $\mathbf{D}_g$ for the same SLiM datasets in Figure A.23. In each plot, the vertical red line indicates the original value of the replicated correlation. Back to section 1.5.3.

A.2 Approaches with unknown groups

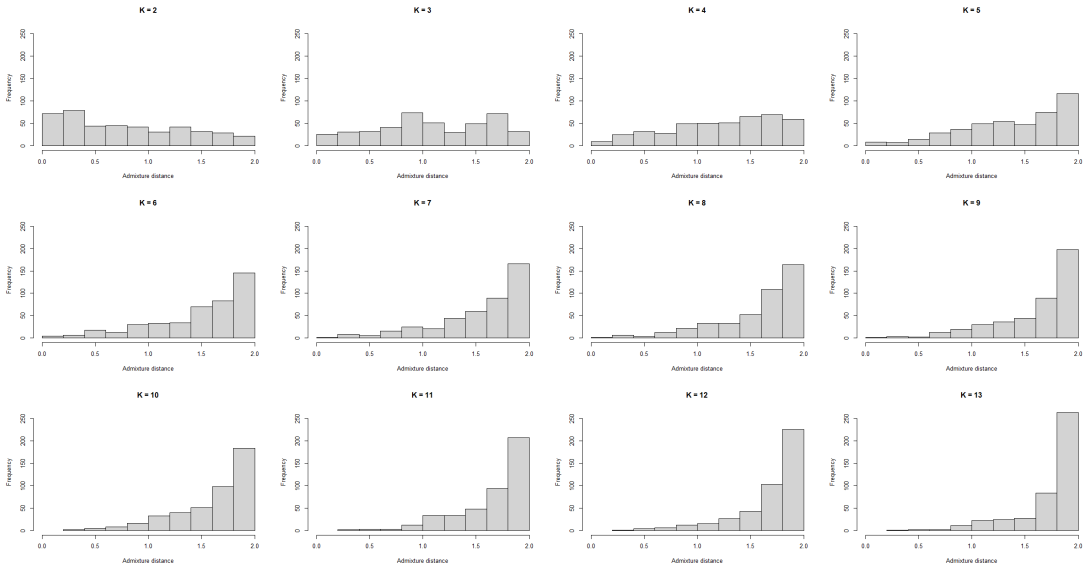

Fig. A.26 Empirical distribution of the admixture dissimilarity defined in (2.6) for varying K from a SLiM scenario with 30 individuals from a quasi-panmictic species.

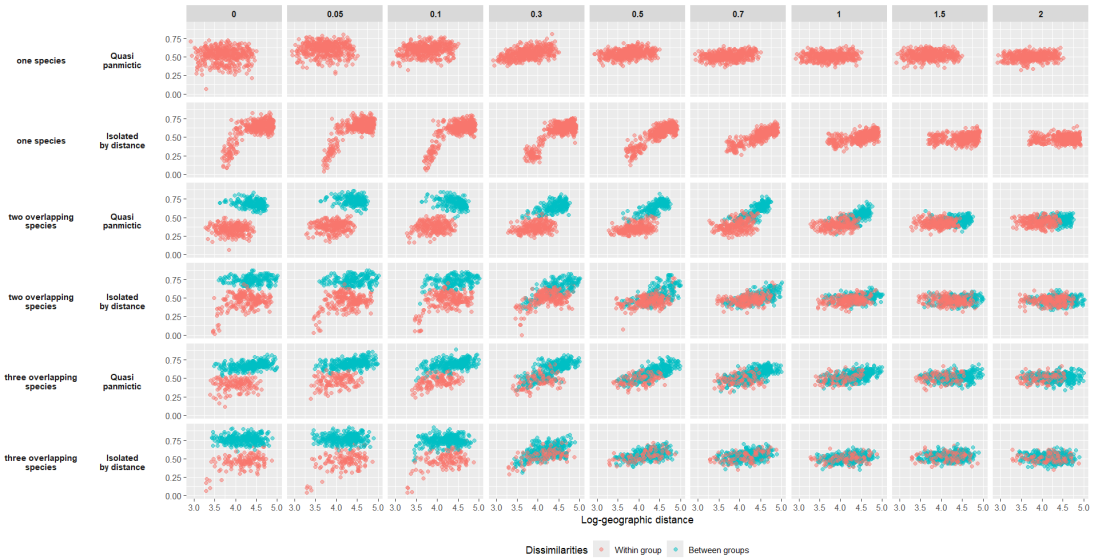

Fig. A.27 Genetic versus log-geographic distance from datasets based on scenarios with one, two or three species inhabiting the same area. Horizontal panels by scenario and IBD behaviour. First vertical panel (0) for original dataset, then for value of ρ. 30 datasets per scenario, all with average sample sizes and 40 loci. Back to Figure 2.10.

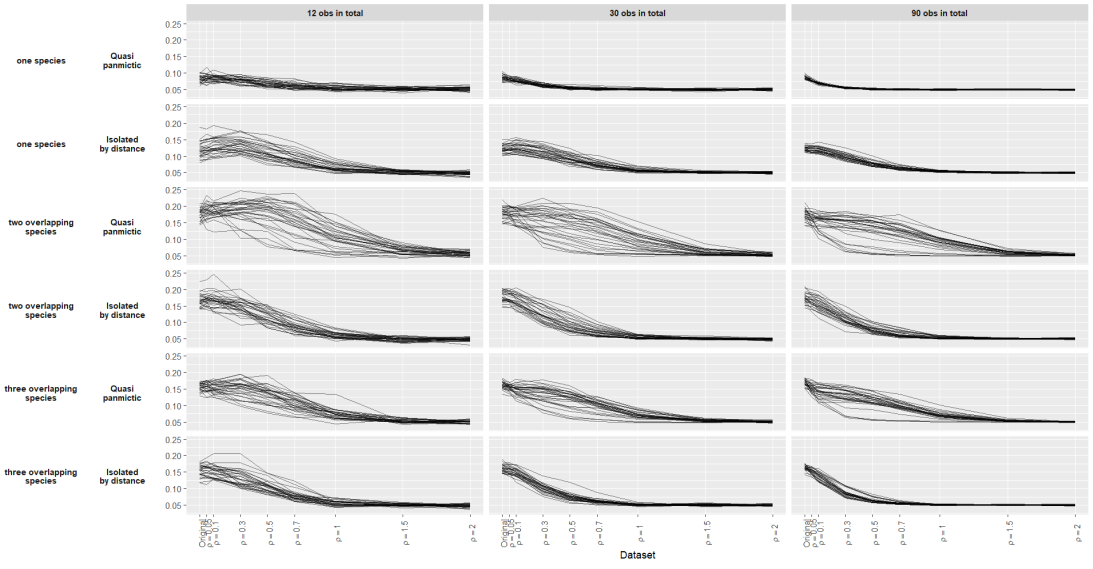

Fig. A.28 Standard deviation of genetic dissimilarities from the same data in Figure 2.10. Connected estimates pertain to the same SLiM dataset. Panels by scenario, IBD behaviour and sample size. 30 datasets per scenario, all with 40 loci.

Table A.1 Number of datasets (out of 30) for which the true value of K was recovered by looking at the *lowest* K for which the tail probability in (2.13), calibrated with the weighted null model with $\rho = 1$, was lower than the Bonferroni-corrected threshold $0.05/(K_{\max} - 2)$. Back to Table 2.2.

Scenario	**IBD pattern**	**Total sample size**	**Correct choices**
Two species	Quasi-panmictic	90	12
		30	11
		12	5
	Isolated by distance	90	28
		30	26
		12	7
Three species	Quasi-panmictic	90	28
		30	30
		12	27
	Isolated by distance	90	29
		30	29
		12	24

Appendix B

Proof that shared allele distance fulfills triangle inequality

For this proof, it is assumed that loci are diploid and non-missing, but any number of different allelic stata is allowed.

Triangle inequality is first proved for the single-locus shared allele distance, defined in (1.26). Consider three individuals $\mathbf{z}_i$, $\mathbf{z}_j$ and $\mathbf{z}_k$. It is required that:

$$
\begin{aligned}
d_y^{(p)}(\mathbf{z}_i,\mathbf{z}_j) &\leq d_y^{(p)}(\mathbf{z}_i,\mathbf{z}_k) + d_y^{(p)}(\mathbf{z}_j,\mathbf{z}_k) \\
1-\frac{1}{2}|Z_i^p \cap Z_j^p|\left[1+\mathbb{1}\left(|Z_i^p|+|Z_j^p|=2\right)\right] &\leq 1-\frac{1}{2}|Z_i^p \cap Z_k^p|\left[1+\mathbb{1}\left(|Z_i^p|+|Z_k^p|=2\right)\right] \\
&\quad +1-\frac{1}{2}|Z_j^p \cap Z_k^p|\left[1+\mathbb{1}\left(|Z_j^p|+|Z_k^p|=2\right)\right] \\
|Z_i^p \cap Z_j^p|\left[1+\mathbb{1}\left(|Z_i^p|+|Z_j^p|=2\right)\right] &\geq |Z_i^p \cap Z_k^p|\left[1+\mathbb{1}\left(|Z_i^p|+|Z_k^p|=2\right)\right] \\
&\quad +|Z_j^p \cap Z_k^p|\left[1+\mathbb{1}\left(|Z_j^p|+|Z_k^p|=2\right)\right]-2
\end{aligned}
$$

The maximum value the right-hand side (RHS) of the last inequality above can take is 2. Now:

- if the left-hand side (LHS) of the last inequality above is equal to 2, i.e., if $Z_i^p = Z_j^p$, the inequality always holds.
- If the LHS is 1, this means that Z_i^p and Z_j^p share one allele, such as when $Z_i^p = \{A\}$ and $Z_j^p = \{A,B\}$ or when $Z_i^p = \{A,B\}$ and $Z_j^p = \{B,C\}$. The inequality holds unless the RHS is 2, but this can only happen if Z_k^p is equal to both Z_i^p and Z_j^p, which contradicts the assumption that the LHS is 1. Hence the inequality stands.

- If the LHS is 0, Z_i^p and Z_j^p share no allele, as when $Z_i^p = \{A,B\}$ and $Z_j^p = \{C,D\}$ or $Z_i^p = \{A,B\}$ and $Z_j^p = \{C\}$ or $Z_i^p = \{A\}$ and $Z_j^p = \{C,D\}$ or $Z_i^p = \{A\}$ and $Z_j^p = \{B\}$. In order for the RHS to be as large as 1 and falsify the inequality, Z_k^p would need to be identical to Z_i^p and share one allele with Z_j^p, or viceversa. However, these two conditions cannot be met at the same time, since that would imply that Z_i^p and Z_j^p share one allele, too, violating the assumption that the LHS is 0. Thus, also in this case, the inequality holds.

Therefore, the triangle inequality for the single-locus shared allele distance is proved to hold for all the possible values the LHS can take.

Because this applies to all diploid loci, assuming no missing locus is present, it is possible to write a system of inequalities:

$$\begin{cases} d_y^{(1)}(\mathbf{z}_i,\mathbf{z}_j) \leq d_y^{(1)}(\mathbf{z}_i,\mathbf{z}_k) + d_y^{(1)}(\mathbf{z}_j,\mathbf{z}_k) \\ \dots \\ d_y^{(P)}(\mathbf{z}_i,\mathbf{z}_j) \leq d_y^{(P)}(\mathbf{z}_i,\mathbf{z}_k) + d_y^{(P)}(\mathbf{z}_j,\mathbf{z}_k) \end{cases}.$$

All terms on the two sides of the inequalities can be summed, leading to:

$$\begin{aligned} \sum_{p=1}^{P} d_y^{(p)}(\mathbf{z}_i,\mathbf{z}_j) &\leq \sum_{p=1}^{P} d_y^{(p)}(\mathbf{z}_i,\mathbf{z}_k) + \sum_{p=1}^{P} d_y^{(p)}(\mathbf{z}_j,\mathbf{z}_k) \\ \frac{1}{P}\sum_{p=1}^{P} d_y^{(p)}(\mathbf{z}_i,\mathbf{z}_j) &\leq \frac{1}{P}\sum_{p=1}^{P} d_y^{(p)}(\mathbf{z}_i,\mathbf{z}_k) + \frac{1}{P}\sum_{p=1}^{P} d_y^{(p)}(\mathbf{z}_j,\mathbf{z}_k) \\ d_y(\mathbf{z}_i,\mathbf{z}_j) &\leq d_y(\mathbf{z}_i,\mathbf{z}_k) + d_y(\mathbf{z}_j,\mathbf{z}_k)\,. \end{aligned}$$

This proves that the shared allele distance computed for diploid non-missing loci fulfills the triangle inequality.

Appendix C

Software code

C.1 SLiM

Script editing, software executions and data analysis were carried out in `R`. The variables defined with the `defineConstant` command were set at each simulation iteration before feeding the script to the SLiM executable: for instance, `i1sd` would take value 1 or 5 according to the scenario at hand.

Listing C.1 SLiM script for the *split* scenario

```
initialize() {
initializeSLiMOptions(keepPedigrees = T, dimensionality="xy",
    nucleotideBased=T);
  defineConstant("L", 1e3);
  initializeAncestralNucleotides(randomNucleotides(L));
  initializeMutationTypeNuc("m1", 0.5, "f", 0.0);
  initializeGenomicElementType("g1", m1, 1.0, mmJukesCantor(2.5e
   -3));
  initializeGenomicElement(g1, 0, L-1);
  initializeRecombinationRate(rates = 1e-8);

  defineConstant(symbol="nsim", value=200);
  defineConstant(symbol="nsam", value=45);
  defineConstant(symbol="i1sd", value=1);
  defineConstant(symbol="i2maxd", value=3);
  defineConstant(symbol="nneigh", value=3);
  defineConstant(symbol="childsd", value=9);

  // spatial competition
  initializeInteractionType(1, "xy", reciprocal=T, maxDistance=3*
   i1sd);
    i1.setInteractionFunction("n", i1sd/2, i1sd);
```

```
20
21
22    // spatial mate choice
23    initializeInteractionType(2, "xy", reciprocal=T, maxDistance=
       i2maxd);
24  }
25
26  1 late() {
27    sim.addSubpop("p0", nsim*2);
28    p0.setSpatialBounds(c(50.00, 50.00, 150.00, 150.00));
29    p0.individuals.setSpatialPosition(p0.pointUniform(nsim*2));
30  }
31
32  1:100 late() {
33    i1.evaluate(p0);
34    inds = p0.individuals;
35    competition = i1.totalOfNeighborStrengths(inds) / size(inds);
36    competition = pmin(competition, 0.99);
37    inds.fitnessScaling = 1 - competition;
38  }
39
40  2:100 first() {
41    i2.evaluate(p0);
42  }
43
44  mateChoice(p0) {
45    // nearest-neighbor mate choice
46    neighbors = i2.nearestNeighbors(individual, count = nneigh);
47    return (size(neighbors) ? sample(neighbors, 1) else float(0));
48  }
49  modifyChild(p0) {
50    do pos = parent1.spatialPosition + rnorm(2, 0, childsd);
51    while (!p0.pointInBounds(pos));
52    child.setSpatialPosition(pos);
53
54    return T;
55  }
56
57  100 late() { // last generation
58    sampledIndividuals = p0.sampleIndividuals(nsam*2);
59
60    out = paste("subpopulation", "pedigreeID", "x", "y", "genome1",
        "genome2");
61    for (i in sampledIndividuals){
62      info = paste(i.subpopulation, i.pedigreeID, i.spatialPosition
       , i.genome1.nucleotides(), i.genome2.nucleotides());
```

```
63      out = c(out, info);
64    }
65    writeFile("coordgenomes.txt", out);
66  }
```

Listing C.2 SLiM script for the *conspecificity* scenario

```
1   initialize() {
2       initializeSLiMOptions(keepPedigrees = T, dimensionality="xy",
        nucleotideBased=T);
3       defineConstant("L", 1e3);
4       initializeAncestralNucleotides(randomNucleotides(L));
5       initializeMutationTypeNuc("m1", 0.5, "f", 0.0);
6       initializeGenomicElementType("g1", m1, 1.0, mmJukesCantor(2.5
      e-3));
7       initializeGenomicElement(g1, 0, L-1);
8       initializeRecombinationRate(rates = 1e-8);
9
10      defineConstant(symbol="nsim", value=200);
11      defineConstant(symbol="nsam", value=45);
12      defineConstant(symbol="i1sd", value=1);
13      defineConstant(symbol="i2maxd", value=3);
14      defineConstant(symbol="nneigh", value=3);
15      defineConstant(symbol="childsd", value=9);
16      defineConstant(symbol="separation", value=F);
17
18      // spatial competition
19      initializeInteractionType(1, "xy", reciprocal=T, maxDistance
      =3*i1sd);
20      i1.setInteractionFunction("n", i1sd/2, i1sd);
21
22      // spatial mate choice
23      initializeInteractionType(2, "xy", reciprocal=T, maxDistance=
      i2maxd);
24  }
25  1 late() {
26      sim.addSubpop("p0", nsim*2);
27      p0.setSpatialBounds(c(50.00, 50.00, 150.00, 150.00));
28      p0.individuals.setSpatialPosition(p0.pointUniform(nsim*2));
29  }
30
31  1:100 late() {
32      i1.evaluate(p0);
33      inds = p0.individuals;
34      competition = i1.totalOfNeighborStrengths(inds) / size(inds);
35      competition = pmin(competition, 0.99);
```

```
    inds.fitnessScaling = 1 - competition;
}
2:100 first() { i2.evaluate(p0); }

90 early() {
    sim.addSubpopSplit("p1", nsim, p0);
    sim.addSubpopSplit("p2", nsim, p0);
    p1.setSpatialBounds(c(50.00, 50.00, 150.00, 150.00));
    p2.setSpatialBounds(c(50.00, 50.00, 150.00, 150.00));
    p1.individuals.setSpatialPosition(p1.pointUniform(nsim));
    p2.individuals.setSpatialPosition(p2.pointUniform(nsim));
}

91: late() {
    i1.evaluate(p1);
    inds = p1.individuals;
    competition = i1.totalOfNeighborStrengths(inds) / size(inds);
    competition = pmin(competition, 0.99);
    inds.fitnessScaling = 1 - competition;

    i1.evaluate(p2);
    inds = p2.individuals;
    competition = i1.totalOfNeighborStrengths(inds) / size(inds);
    competition = pmin(competition, 0.99);
    inds.fitnessScaling = 1 - competition;
}

92: first() {
    i2.evaluate(p1);
    i2.evaluate(p2);
}

101 early() { p0.setSubpopulationSize(0); }


101:150 early() {
    if(separation){
p1.setSpatialBounds(p1.spatialBounds + c(0, 0, -1, -1));
p2.setSpatialBounds(p2.spatialBounds + c(1, 1, 0, 0));
    }
}

92: early() {
    migrationProgress = runif(1, min=0.8, max=1);
    p1.setMigrationRates(p2, 0.5 * migrationProgress);
    p2.setMigrationRates(p1, 0.5 * migrationProgress);
```

```
82  }
83
84
85  // NEAREST NEIGHBORS MATE CHOICE
86  2:91 mateChoice(p0) {
87      // nearest-neighbor mate choice
88      neighbors = i2.nearestNeighbors(individual, count = nneigh);
89      return (size(neighbors) ? sample(neighbors, 1) else float(0))
        ;
90  }
91  2:91 modifyChild(p0) {
92      do pos = parent1.spatialPosition + rnorm(2, 0, childsd);
93      while (!p0.pointInBounds(pos));
94      child.setSpatialPosition(pos);
95      return T;
96  }
97
98  92: mateChoice(p1) {
99      // nearest-neighbor mate choice
100     neighbors = i2.nearestNeighbors(individual, count = nneigh);
101     return (size(neighbors) ? sample(neighbors, 1) else float(0))
        ;
102 }
103 92: mateChoice(p2) {
104     // nearest-neighbor mate choice
105     neighbors = i2.nearestNeighbors(individual, count = nneigh);
106     return (size(neighbors) ? sample(neighbors, 1) else float(0))
        ;
107 }
108
109 92: modifyChild(p1) {
110     counter = 1;
111     do{
112         pos = parent1.spatialPosition + rnorm(2, ifelse(
        separation, -0.5, 0.0), childsd);
113         counter = counter + 1;
114     }
115     while (!p1.pointInBounds(pos) & counter < 100);
116     child.setSpatialPosition(pos);
117     return T;
118 }
119 92: modifyChild(p2) {
120     counter = 1;
121     do{
122         pos = parent1.spatialPosition + rnorm(2, ifelse(
        separation, 0.5, 0.0), childsd);
```

```
123          counter = counter + 1;
124      }
125      while (!p2.pointInBounds(pos) & counter < 100);
126      child.setSpatialPosition(pos);
127      return T;
128  }
129  150 late() { // last generation
130          allind1 = p1.individuals[p1.pointInBounds(p1.individuals.
     spatialPosition)];
131          allind2 = p2.individuals[p2.pointInBounds(p2.individuals.
     spatialPosition)];
132      sampledIndividuals = c(sample(allind1, nsam), sample(allind2,
      nsam));
133      sampledIndividuals.genomes.outputVCF(filePath="tmp.VCF",
     outputMultiallelics = F, simplifyNucleotides=T);
134
135      out = paste("subpopulation", "pedigreeID", "x", "y", "genome1
     ", "genome2");
136      for (i in sampledIndividuals){
137          info = paste(i.subpopulation, i.pedigreeID, i.
     spatialPosition, i.genome1.nucleotides(), i.genome2.
     nucleotides());
138    out = c(out, info);
139      }
140      writeFile("coordgenomes.txt", out);
141  }
```

Listing C.3 SLiM script for the *distinctness* scenario (more than one species)

```
1  species all initialize() {
2      defineConstant("L", 1e3);
3      defineConstant(symbol="nsim", value=200);
4      defineConstant(symbol="nsam", value=45);
5      defineConstant(symbol="i1sd", value=1);
6      defineConstant(symbol="i2maxd", value=3);
7      defineConstant(symbol="nneigh", value=3);
8      defineConstant(symbol="childsd", value=9);
9
10     // spatial competition
11     initializeInteractionType(1, "xy", reciprocal=T, maxDistance
     =3*i1sd);
12     i1.setInteractionFunction("n", i1sd/2, i1sd);
13
14     // spatial mate choice
15     initializeInteractionType(2, "xy", reciprocal=T, maxDistance=
     i2maxd);
```

```
16 }
17
18 species sunflower initialize() {
19     initializeSpecies(tickModulo=1, tickPhase=1, avatar="S");
20     initializeSLiMOptions(keepPedigrees = T, dimensionality="xy",
       nucleotideBased=T);
21     initializeAncestralNucleotides(randomNucleotides(L, c(1, 1,
       1, 1)));
22     initializeMutationTypeNuc("m1", 0.5, "f", 0.0);
23     initializeGenomicElementType("g1", m1, 1.0, mmJukesCantor(2.5
       e-3));
24     initializeGenomicElement(g1, 0, L-1);
25     initializeRecombinationRate(rates = 1e-8);
26 }
27
28 species tulip initialize() {
29     initializeSpecies(tickModulo=1, tickPhase=1, avatar="T");
30     initializeSLiMOptions(keepPedigrees = T, dimensionality="xy",
       nucleotideBased=T);
31     initializeAncestralNucleotides(randomNucleotides(L, c(1, 1,
       1, 1)));
32     initializeMutationTypeNuc("m2", 0.5, "f", 0.0);
33     initializeGenomicElementType("g2", m2, 1.0, mmJukesCantor(2.5
       e-3));
34     initializeGenomicElement(g2, 0, L-1);
35     initializeRecombinationRate(rates = 1e-8);
36 }
37
38 ticks all 1 early() {
39     sunflower.addSubpop("p1", nsim);
40     tulip.addSubpop("p2", nsim);
41     p1.setSpatialBounds(c(50.00, 50.00, 100.00, 100.00));
42     p2.setSpatialBounds(c(100.00, 100.00, 150.00, 150.00));
43     p1.individuals.setSpatialPosition(p1.pointUniform(nsim));
44     p2.individuals.setSpatialPosition(p2.pointUniform(nsim));
45 }
46
47 ticks all 1: late() {
48     i1.evaluate(p1);
49     inds = p1.individuals;
50     competition = i1.totalOfNeighborStrengths(inds) / size(inds);
51     competition = pmin(competition, 0.99);
52     inds.fitnessScaling = 1 - competition;
53
54     i1.evaluate(p2);
55     inds = p2.individuals;
```

```
56      competition = i1.totalOfNeighborStrengths(inds) / size(inds);
57      competition = pmin(competition, 0.99);
58      inds.fitnessScaling = 1 - competition;
59  }
60
61  ticks all 2: first() {
62    i2.evaluate(p1);
63    i2.evaluate(p2);
64  }
65
66  // NEAREST NEIGHBORS MATE CHOICE
67  species sunflower 2: mateChoice(p1) {
68      // nearest-neighbor mate choice
69      neighbors = i2.nearestNeighbors(individual, count = nneigh);
70      return (size(neighbors) ? sample(neighbors, 1) else float(0))
        ;
71  }
72  species tulip 2: mateChoice(p2) {
73      // nearest-neighbor mate choice
74      neighbors = i2.nearestNeighbors(individual, count = nneigh);
75      return (size(neighbors) ? sample(neighbors, 1) else float(0))
        ;
76  }
77  species sunflower modifyChild(p1) {
78      do pos = parent1.spatialPosition + rnorm(2, 0, childsd);
79      while (!p1.pointInBounds(pos));
80      child.setSpatialPosition(pos);
81
82      return T;
83  }
84  species tulip modifyChild(p2) {
85      do pos = parent1.spatialPosition + rnorm(2, 0, childsd);
86      while (!p2.pointInBounds(pos));
87      child.setSpatialPosition(pos);
88
89      return T;
90  }
91
92  ticks all 50 late() { // last generation
93      allind1 = p1.individuals;
94      allind2 = p2.individuals;
95      sampled1 = sample(allind1, nsam);
96      sampled2 = sample(allind2, nsam);
97      sampledIndividuals = c(sampled1, sampled2);
98
```

```
99      out = paste("subpopulation", "pedigreeID", "x", "y", "genome1
        ", "genome2");
100     for (i in sampledIndividuals){
101         info = paste(i.subpopulation, i.pedigreeID, i.
        spatialPosition, i.genome1.nucleotides(), i.genome2.
        nucleotides());
102         out = c(out, info);
103   }
104     writeFile("coordgenomes.txt", out);
105 }
```

C.2 GSpace

Script editing, software executions and data analysis were carried out in `R`. Variables like `Sequence_Size` or `Dispersal_Distribution` were modified according to the scenario at hand.

When only one software execution was involved (trivial conspecificity scenario), all individuals from the two groups were simulated at once, using a longer sequence of random coordinates next to the `Sample_Coordinate` variables. In all other situations, the code was run twice with a reduced number of coordinate pairs (e.g., six or four for the first group with equal-sized and unequal-sized groups, respectively), each time fulfilling the geographic separation explained in the main text: e.g., individuals from the first group would have coordinates bound within point $(70, 70)$ and point $(90, 90)$.

Listing C.4 Common structure of the GSpace script

```
%%%%%%%% SIMULATION SETTINGS %%%%%%%%%%%%%%%
Setting_Filename = GSpaceSettings.txt
Random_seeds = 11000
Run_Number = 1

%%%%%%%% OUTPUT FILE FORMAT SETTINGS %%%%%%%
Output_Dir = .
Data_File_Name = trial
Data_File_Extension = .txt

Genepop = True
Genepop_ind_file = F
Genepop_Group_All_Samples = T

Approximate_time = F

%%%%%%%% MARKERS SETTINGS %%%%%%%%%%%%%%%%%%
Ploidy = Diploid
Chromosome_number = 1
Mutation_Rate = 0.005
Mutation_Model = KAM
Allelic_Lower_Bound = 240
Allelic_Upper_Bound = 241
Sequence_Size = 100

%%%%%%%% RECOMBINATION SETTINGS %%%%%%%%%%%%
Recombination_Rate = 0.005

%%%%%%%% DEMOGRAPHIC SETTINGS %%%%%%%%%%%%%%
%% LATTICE
```

```
31 Min_Sample_Coord_X = 50
32 Min_Sample_Coord_Y = 50
33 Lattice_Size_X = 200
34 Lattice_Size_Y = 200
35 Ind_Per_Node = 1
36
37 %% DISPERSAL
38 Dispersal_Distribution = p
39 Pareto_Shape = 5
40
41 Edge_Effects = circular
42 Total_Emigration_Rate = 0.5
43 Disp_Dist_Max = 200, 200
44
45 %%%%%%%% SAMPLE SETTINGS %%%%%%%%%%%%%%%%%%%
46 Sample_Coordinate_X = 82,84,70,82,75,71,115,127,130,122,128,124
47 Sample_Coordinate_Y = 89,76,89,72,73,70,111,116,128,120,117,129
48
49 % STATS
50 Dist_Class_Nbr = 1
51 Ind_Per_Node_Sampled = 1
52 Pause = Never
```