RESOURCE ARTICLE

Species delimitation and geography

Bernhard Hausdorf[1] 🆔 | Christian Hennig[2]

[1]Zoological Museum, Center of Natural History, University of Hamburg, Hamburg, Germany

[2]Dipartimento di Scienze Statistiche "Paolo Fortunati", Università di Bologna, Bologna, Italy

Correspondence
Bernhard Hausdorf, Center of Natural History, Zoological Museum, University of Hamburg, Hamburg, Germany.
Email: hausdorf@zoologie.uni-hamburg.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: HA 2763/6-1

Abstract

Despite the importance of the geographical arrangement of populations for the inference of species boundaries, only a few approaches that integrate spatial information into species delimitation have thus far been developed. Persistent differentiation of sympatric groups of individuals is the best criterion for species status. Species delimitation becomes more prone to error if allopatric metapopulations are considered because it is often difficult to assess whether observed differences between allopatric metapopulations would be sufficient to prevent the fusion of these metapopulations upon contact. We propose a novel approach for testing the hypothesis that the multilocus genetic distances between individuals or populations belonging to two different candidate species are not larger than expected based on their geographical distances and the relationship of genetic and geographical distances within the candidate species. A rejection of this null hypothesis is an argument for classifying the two studied candidate species as distinct species. Case studies show that the proposed tests are suitable to distinguish between intra- and interspecific differentiation. The regression approach proposed here is more appropriate for testing species hypotheses with regard to isolation by distance than (partial) Mantel tests. Our tests assume a linear relationship between genetic and (transformed) geographical distances. This assumption can be compromised by a high genetic variability within populations as found in a case study with microsatellite markers.

KEYWORDS
genetic distance, geographical distance, isolation by distance, species delineation, taxonomy

1 | INTRODUCTION

Species delimitation remains a difficult task despite the increasing availability of genomic data and despite an increasing number of quantitative approaches for delimiting species (Carstens, Pelletier, Reid, & Satler, 2013; Hausdorf & Hennig, 2010; Sites & Marshall, 2004; Yang & Rannala, 2014). An integrative approach to taxonomy considering different kinds of data is often recommended because species delimitation approaches relying on a single kind of data may result in incorrect conclusions (Padial, Miralles, De la Riva, & Vences, 2010; Sauer & Hausdorf, 2012; Schlick-Steiner

et al., 2010). However, methods that can actually use different kinds of data as input to produce an automated species classification remain in their infancy (Edwards & Knowles, 2014; Guillot, Renaud, Ledevin, Michaux, & Claude, 2012; Solís-Lemus, Knowles, & Ané, 2015).

The geographical relationships between groups of individuals may be highly informative for the inference of species boundaries. If differentiated groups of individuals occur at the same locality, discontinuities in the distributions of their character states (other than polymorphisms or sexual dimorphism) demonstrate that these groups should be classified as distinct species. The criterion of

persistent differentiation of sympatric groups, at least with regard to specific characteristics, can be found in several species concepts such as the genotypic cluster definition of Mallet (1995), the genic species concept of Wu (2001) and the differential fitness species concept of Hausdorf (2011). Species delimitation becomes more prone to error if allopatric metapopulations are considered, because it is often difficult to assess whether observed differences between allopatric metapopulations would be sufficient to prevent the fusion of these metapopulations upon contact.

Despite the importance of geography for the inference of species boundaries and despite geographical data of the sampled individuals almost always being available, only a few approaches that integrate spatial information into species delimitation have been developed so far. These approaches can be classified into a priori methods that incorporate the geographical data into the protocol for delimiting candidate species, and a posteriori approaches that use geographical data to assess whether candidate species delimited with other approaches should be considered distinct species given the degree of differentiation of the candidate species and their geographical relationships.

Two a priori methods for considering geographical information directly in the species delimitation process have been proposed. Guillot et al. (2012) proposed a statistical model that can analyse genetic and phenotypic data and can incorporate geographical data in such a way that the clusters to be delimited tend to occupy only one or a few separate areas. Edwards and Knowles (2014) suggested a clustering approach based on a combination of nonmetrical multidimensional scalings of the different distance matrices that were derived from geographical, as well as genetic, morphological and ecological data. The implicit assumption of this approach is that populations that are further apart are more likely to evolve into separate species because of the decreasing gene flow and/or the more strongly differing environmental conditions with increasing geographical distance.

A posteriori approaches assess whether the observed relationships between geographical and genetic or morphological distances between candidate species determined with other approaches are compatible with the expectation based on the variation of genetic distances with increasing geographical distances within the candidate species. Such tests require a model that describes the relationships between geographical and genetic or morphological distances within species. The simplest model that describes this relationship is the "isolation by distance" (IBD) model introduced by Wright (1943).

Four studies have suggested different a posteriori approaches. Medrano, López-Perea, and Herrera (2014) used partial Mantel tests to assess whether a variable indicating the classification can explain a significant part of the variance in the genetic distances between populations in addition to the variance explained by geography. Gratton et al. (2016) formulated two operational criteria for recognizing "good" species, namely (a) a pattern of within-cluster IBD, and (b) a lack of dependence of the genetic differentiation between pairs of individuals belonging to different clusters on their

geographical distance. They compared the correlation between genetic and geographical distances within and between candidate species using Mantel tests. Spriggs et al. (2019) compared linear models with geographical distance and species identity as fixed effects to test whether the genetic divergence between candidate species was significant beyond what would be expected from geographical isolation alone.

It remains of debate whether Mantel and partial Mantel tests as used in previous approaches for assessing IBD are statistically appropriate (Frantz, Cellina, Krier, Schley, & Burke, 2009; Guillot & Rousset, 2013). Thus, here we develop a new regression-based protocol for testing whether the genetic distances between individuals or populations belonging to two different candidate species can be explained by their geographical distances given the variation of genetic distances with geography within the candidate species, or whether they indicate that the candidate species should be classified as distinct species. We discuss the underlying assumptions and methodological difficulties of this approach and compare it with previous a posteriori approaches for assessing the status of candidate species using geographical information.

2 | MATERIALS AND METHODS

We used multilocus data sets from four recent taxonomic studies to illustrate the performance of the IBD tests for evaluating the status of candidate species: AFLP data (Martínez-Ortega, Delgado, Albach, Elena-Rosselló, & Rico, 2004a) of speedwell Veronica (Pentasepalae) (Plantaginaceae) from the Iberian Peninsula and Morocco (Martínez-Ortega, Delgado, Albach, Elena-Rosselló, & Rico, 2004b), microsatellite data of Conradina (Lamiaceae) from Florida, Alabama and Tennessee (Edwards, Soltis, & Soltis, 2008), AFLP data of trumpet daffodils (Narcissus; Amaryllidaceae) from the southern Iberian Peninsula (Medrano et al., 2014), and haplotype data of RAD loci (Gratton et al., 2015) of brassy ringlets (Erebia; Lepidoptera: Nymphalidae) from the Alps (Gratton et al., 2016). These data sets are described in detail in File S1.

3 | RESULTS

3.1 | Outline of the IBD tests

We intend to test whether the genetic (or morphological) distances between units belonging to two candidate species delimited with other methods can be explained by IBD (i.e., by the increase in genetic distances with geographical distances observed within the candidate species). We derive the expected relationship of genetic distances and geographical distances from regressions of genetic distances of units belonging to the same candidate species against log-transformed geographical distances. The null hypothesis is that the genetic distances between units belonging to different candidate species are not larger than expected based on the within-group

regressions. A rejection of the null hypothesis is an argument for classifying two candidate species as distinct species.

As units either individuals or populations may be used. Given that the number of individuals is always equal to or larger than the number of populations, tests on the level of individuals have more power. However, inference based on individuals assumes that the units are independent samples. This assumption can be violated by relationships between individuals, which are especially close within populations. Thus, tests on the level of populations may be less affected by the violation of this assumption than tests on the level of individuals.

3.2 | Modelling setup

Assume that we have observations of n units $I_1,..., I_n$. These are characterized by a geographical distance measure $d_{ij} = d(I_i, I_j)$ and a genetic distance measure based on multilocus data $d_{ij}^* = d^*(I_i, I_j)$ for $i, j = 1,..., n$, both of which fulfil the standard axioms of a dissimilarity (non-negativity, symmetry, $d(I, I) = 0$ for all objects I; we do not require the triangle inequality to hold). Furthermore for $i = 1, ..., n$ we have group indicators $c_i \in \{1, 2\}$ indicating whether I_i belongs to candidate species 1 or 2. Let n_1 and n_2 be the number of units belonging to candidate species 1 and 2, respectively.

We use a linear regression approach, but we allow the distances to be transformed by known monotonic transformations f and f^*, i.e., $f_{ij} = f(d_{ij})$, $f_{ij}^* = f^*(d_{ij}^*)$, $i, j = 1, ..., n$. This allows for nonlinear relationships. Following Slatkin (1993) and Rousset (1997), we log-transform geographical distances. An issue is that geographical zero distances occur whereas log(0) is not defined. Thus, distances need to be transformed as $f_{ij} = \log(d_{ij} + c)$ with a constant $c > 0$. The choice of c will have an impact on the regression. c should depend on the value range of d_{ij}, because its impact is relative to that range. We choose c to be the 0.25-quantile of the geographical distances here. This makes the transformation invariant to the measurement units of the distances.

Within a candidate species, we assume that the following regression relationship holds:

$$f_{ij}^* = a_k + b_k f_{ij} + e_{ij}, \quad \text{where} \quad Ee_{ij} = 0. \tag{1}$$

Here i and j are from a set of indexes assumed to belong to the same candidate species. In the following we will assume that all units in candidate species k, $k = 1, 2$, belong to the same candidate species, characterized by regression parameters a_k (intercept) and b_k (slope). For statistical inference, we assume that the units are independent samples (see Outline of the IBD tests). However, we do not assume anything further regarding the distribution of e_{ij}, and particularly not that they are independent, which for different distances involving the same unit would not make sense. This means that the standard distribution theory of linear regression cannot be applied.

If all units belong to the same species and the regressions within the two candidate species are equal, we have $a = a_1 = a_2$ and

$b = b_1 = b_2$. A difference in the regressions might indicate that the two candidate species considered are different species with different dispersal abilities. However, the relationship between the genetic and the geographical distances does not depend only on species-specific characteristics such as dispersal ability, but also on other factors such as the terrain or history. For example, genetic distances may increase faster with geographical distance in a mountainous region than in a plain because the mountains inhibit dispersal. Furthermore, the genetic distances within a candidate species may be larger in long-standing populations in a refuge area than in populations in an area that has been colonized only recently. Thus, two candidate species may belong to the same species despite $a_1 \neq a_2$ and/or $b_1 \neq b_2$. There will not always be a simple regression (Equation 1) across the range of a species. Regional subgroups of a species may show different regression patterns between genetic and geographical distances. However, if the candidate species are in fact conspecific, we expect that the transformed between-group distances f_{ij}^*, $c_i \neq c_j$ are not larger than what is expected from at least one of the two regressions defined in Equation 1 for $k = 1, 2$. The possible causes for an inequality of regressions between two candidate species should be checked and considered in the interpretation of the results.

On the other hand, having $a_1 = a_2$ and $b_1 = b_2$ for the within-group distances in both candidate species does not necessarily imply that these candidate species belong to the same species, because this does not say that the between-group genetic distances are low enough to be explained by geographical distance alone. For this to be the case, the regression resulting from the within-group distances will need to fit the between-group distances as well.

3.3 | IBD tests

In the following, we describe three tests that investigate the equality of two different regressions of genetic versus geographical distances within and/or between two candidate species. The first test compares the regressions within the two candidate species. It does not test whether the overall pattern is compatible with IBD (see above), but indicates whether the second or the third test is appropriate given the structure of the data. These alternative tests were devised to test the hypothesis that the genetic distances between units belonging to two different candidate species can be explained by IBD.

The null hypothesis of the first test, H_{01}, is that the relationship between genetic and geographical distances within each candidate species can be modelled by a single regression for both candidate species: $a_* = a_1 = a_2$ and $b_* = b_1 = b_2$ (i.e., the regression coefficients are called a_* and b_* assuming that the regressions based on within-group distances are equal). This case is illustrated, for example, in Figure 1a: the within-group distances (red and black symbols) result in similar regressions (dotted black and red lines) that can (according to the results of the testing procedure explained below) be modelled by a single regression line (dotted green line).
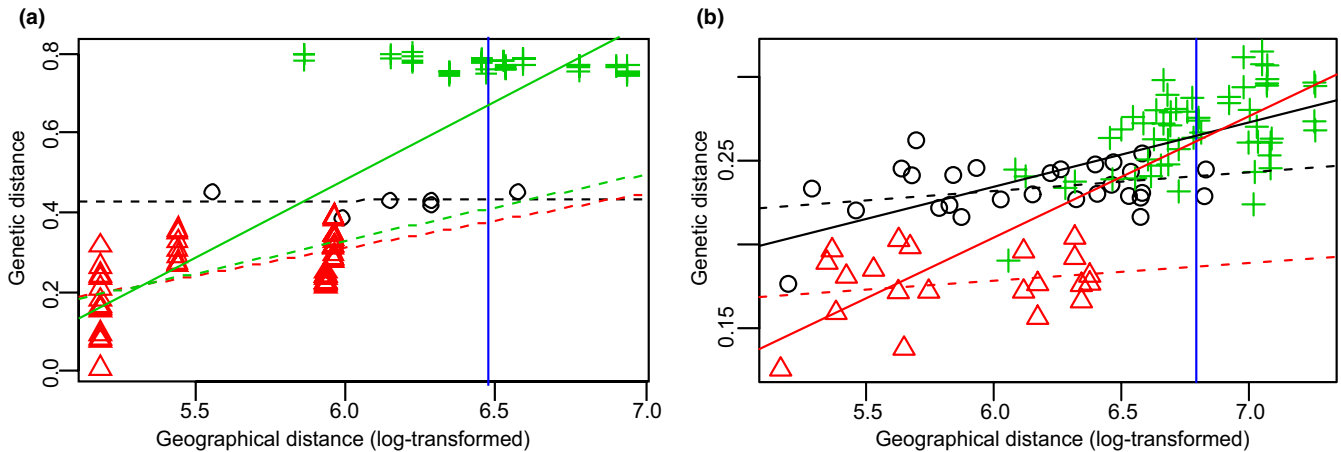
FIGURE 1 Relationships between genetic and log-transformed geographical distances in pairs of individuals or populations of two candidate species. Black circles and red triangles: distances between individuals or populations belonging to the first and second candidate species, respectively; green crosses: distances between individuals or populations belonging to different candidate species; black and red dotted lines: regression lines fitted within the first or second candidate species, respectively; blue lines: centres of the between-groups geographical distances. (a) The relationship between genetic and geographical distances within each candidate species can be modelled by a single regression for both candidate species (H_{01} not rejected). Green dotted line: regression line fitted on the within-group distances only (i.e., the black circles and red triangles taken together); green solid line: regression line fitted on all distances together. (b) The relationship between genetic and geographical distances within each candidate species cannot be modelled by a single regression for both candidate species (H_{01} rejected). Black and red solid lines: regression lines fitted to the distances within the first or second candidate species, respectively, together with the distances between the candidate species [Colour figure can be viewed at wileyonlinelibrary.com]

If H_{01} is not rejected, the hypothesis that the genetic distances between units belonging to two different candidate species can be explained by IBD can be investigated by checking whether the joint within-group regression also fits the distances between the two candidate species (green crosses in Figure 1), i.e., whether for all $i, j = 1, ..., n$:

$$H_{02} : f_{ij}^* = a + b f_{ij} + e_{ij}, \quad \text{where} \quad E e_{ij} = 0.$$

Here $a = a_1 = a_2$ and $b = b_1 = b_2$. This will be tested by comparing a regression fitted on all within-group distances (dotted green lines in Figure 1a) with another regression fitted on all distances (solid green lines in Figure 1a). If H_{02} is true, these should be equal (i.e., $a = a_*$ and $b = b_*$). If H_{02} is not rejected, the data provide no evidence for the specific distinctness of the candidate species.

If H_{01} is rejected (i.e., $a_1 \neq a_2$ and/or $b_1 \neq b_2$; for example the dotted black and red lines in Figure 1b), it would be invalid to fit a regression to all the within-group distances together. In this case, we compare the values f_{ij}^*, $c_i \neq c_j$, to what is predicted from each of the two regressions within the candidate species defined by the parameters (a_1, b_1), (a_2, b_2). The null hypothesis that the genetic distances between units belonging to two different candidate species can be explained by IBD is operationalized in this case as follows. We define another regression:

$$f_{ij}^* = a_k^* + b_k^* f_{ij} + e_{ij}, \quad \text{where} \quad E e_{ij} = 0$$

for i, j with $c_i = k$ whereas c_j may be either 1 or 2 (solid black and red lines, respectively, in Figure 1b). These regressions are based on the distances within candidate species k together with the distances between the candidate species, but without the distances within the respective other candidate species. Let $f_{between}$ be the centre of the

between-group transformed geographical distances (i.e., $f_{between} = \frac{1}{n_1 n_2} \sum_{c_i=1, c_j=2} f_{ij}$). With this, the null hypothesis of the third test, H_{03}, is $a_k^* + b_k^* f_{between} \leq a_k + b_k f_{between}$ for at least one $k \in \{1, 2\}$. If the genetic distances between the candidate species are too large to be compatible with the regression on the distances within candidate species k, putting the within-group and between-group distances together will result in a regression (solid black and red line, respectively, in Figure 1b) that fits a higher value at the centre of the between-group distances (blue lines in Figure 1) than the regression based on the within-group distances alone (dotted black and red line, respectively, in Figure 1b), that is $a_k^* + b_k^* f_{between} > a_k + b_k f_{between}$. If this is the case for both candidate species, H_{03} is rejected, indicating that the two candidate species probably represent distinct species. If this is the case for only one of the candidate species, the reasons have to be investigated.

H_{01} will be tested against the alternative that $a_1 - a_2 \neq 0$ or $b_1 - b_2 \neq 0$ (two-sided alternative). H_{02} will be tested against $a + b f_{between} > a_* + b_* f_{between}$ (one-sided alternative), because it should only be rejected if the genetic distances between candidate species are larger on average than what would be expected from the regression on within-group distances only. This holds for H_{03} as well, namely it is tested against the one-sided alternative $a_k^* + b_k^* f_{between} > a_k + b_k f_{between}$ for both $k = 1, 2$. We will use ordinary least squares regression to fit all the models and obtain parameter estimators $\hat{a}, \hat{a}_*, \hat{a}_k, \hat{a}_k^*, \hat{b}, \hat{b}_*, \hat{b}_k$ and $\hat{b}_k^*$, for $k = 1, 2$.

Although the three tests for assessing H_{01}, H_{02} and H_{03} are all based on comparing different regression lines, they differ from each other to some extent. When testing H_{01}, two regression lines computed from different sets of units (namely those in candidate

species 1 and 2, respectively) are compared. Here we test intercepts and slopes separately, and both need to be equal for the regression lines to be the same. The test statistics are $T_{1a} = \hat{a}_1 - \hat{a}_2$ and $T_{1b} = \hat{b}_1 - \hat{b}_2$. The variation of each of the two regression lines can be assessed independently, and the variation of T_{1a} and T_{1b} can be derived from those. This test can be generally applied to comparing two regressions between distances in two different independent groups.

When testing H_{02} and H_{03}, the regression lines that are compared are based on partly the same units, and we are interested in assessing differences at the centre of between-group distances $f_{between}$, rather than running separate tests for intercept and slope. The test statistic for H_{02} is $T_2 = \hat{a} + \hat{b}\, f_{between} - (\hat{a}_* + \hat{b}_*\, f_{between})$. There are two test statistics for H_{03} testing separately for the two candidate species, namely $T_{3k} = \hat{a}_k^* + \hat{b}_k^*\, f_{between} - (\hat{a}_k + \hat{b}_k\, f_{between})$, $k = 1, 2$. In terms of the test power, it is an advantage to have only a single test statistic, because if a test relies on a pair of test statistics, correction for multiple testing needs to be applied. Therefore, it makes sense to test H_{01} first to see whether there is an indication against testing H_{02}, and to test H_{02} if that is not the case, rather than testing H_{03} all the time, although this has lighter assumptions and could always be applied. Moreover, testing H_{02} uses all distances, and therefore it can be expected to be superior in terms of power even to every single test of H_{03}, although due to the lack of available distribution theory for distances, this currently cannot be assessed theoretically. Under these null hypotheses, all regressions involved in the corresponding tests are assumed to be equal, and therefore the expected values of all test statistics under the null hypotheses are zero.

For testing, we have to estimate the expected variation of the test statistics under H_0. We cannot use standard linear regression theory here because of the lack of distributional assumptions and particularly the lack of independence of e_{ij}. One possibility for assessing variability would be a nonparametric bootstrap (sampling with replacement n units from the empirically observed units). A nonparametric bootstrap will keep the sample size constant by sampling identical objects several times. This is problematic here, because it will lead to a number of pairs of (identical) units sampled with both geographical and genetic distance zero. This can have a strong effect on the regression estimation and is unrealistic unless the measurement of distances is imprecise and there are many such "both distances zero"-cases already in the data.

Because of these issues, we apply a different nonparametric statistical resampling principle, the jackknife (Quenouille, 1949). A simple nonparametric jackknife test has been proposed by Tukey (1958). The general idea is to define "pseudovalues" for the parameter estimators. If a parameter θ is estimated from n independent and identically distributed observations $X_1,..., X_n$ by an estimator $\hat{\theta}_n$, for $i = 1,..., n$ pseudovalues $\theta_{*i} = n\hat{\theta}_n - (n-1)\hat{\theta}_{n-1;i}$ are computed, where $\hat{\theta}_{n-1;i}$ is the estimator of θ computed with observation X_i omitted (in the context of distance data this means that all distances involving unit I_i are omitted). The pseudosample $\theta_{*i},..., \theta_{*n}$ can then be used to run a standard t test of the hypothesized value for θ (i.e., their mean is compared with the expected value under H_0, which here is zero).

For details about when this works see Miller (1974); the specific reasons given by Miller why such a procedure may not work, namely if involved estimators are not smooth enough in the observations, do not apply in our setup. See also Efron (1979) for more theoretical exploration.

The test statistic T_2 (taking the role of θ) allows a direct application of this principle. Some modification is required for T_{1a}, T_{1b}, T_{31}, T_{32}, because for these test statistics the role of observations differs between the two candidate species, and variances of θ_{*i} may differ between candidate species.

T_{1a} and T_{1b} are differences between parameter estimators from two different independent groups, and this is an analogous situation to Welch's (1947) two-sample t test allowing for different variances. The pseudovalues θ_{*i} can be used separately depending on whether c_i is 1 or 2, the two within-group variances can be combined and the test can be run in the same way as in Welch's t test.

In T_{31} and T_{32}, the two regression lines to be compared are not independent. The difference $\hat{a}_k^* + \hat{b}_k^*\, f_{between} - (\hat{a}_k + \hat{b}_k\, f_{between})$ needs to be evaluated omitting one unit I_i at a time to compute θ_{*i}. Again the variance of the θ_{*i} may differ depending on whether $c_i = 1$ or 2. This is because the units I_i with $c_i = k$ are used for both regressions, whereas the units from the other candidate species are used only for the regression that includes between-group distances. The variance of the mean $\frac{1}{n}\sum_{i=1}^{n}\theta_{*i}$ can be estimated as $\frac{n_1 V_1 + n_2 V_2}{n^2}$, where V_j, $j = 1, 2$ is the sample variance of θ_i for which $c_i = j$. This can be used in a t test, with degrees of freedom approximated by the Welch–Sattertwaithe equation (Welch, 1947), as in Welch's t test.

The p-values of the tests based on T_{1a} and T_{1b} can be aggregated using Bonferroni's rule (i.e., the smaller of the two p-values needs to be multiplied by 2) in order to have a test of H_{01}. This should be significant if a significant difference is found in at least one of the intercept and slope parameters. Aggregation is different for T_{31} and T_{32} because there the result should reject H_{03} significantly only if it can be rejected in both candidate species. Therefore, the maximum of the two p-values from T_{31} and T_{32} can be used as a p-value for H_{03}.

Before running the test of H_{01}, the geographical distances are centred by the mean within-cluster distance taken over both candidate species so that the regression intercept is located in a central place. This will not change the estimated regression lines, but enables a more precise estimation of the regression intercept than if it was located far away from the bulk of the data (i.e., its variance will be lower and the corresponding test will have a better power). This is not important for testing H_{02} and H_{03}, because the fitted value at $f_{between}$ does not depend on whether data are centred or not.

For testing H_{01} and H_{03} at least four units of a candidate species are necessary so that the variance can be estimated using the jackknife. For testing H_{02} at least two units of each candidate species are necessary. In addition, there has to be some variation in the geographical origin of the specimens.

The described IBD tests are implemented in the program package PRABCLUS (Hennig & Hausdorf, 2019), an add-on package for the free statistical software R (R Core Team, 2018). For all case studies, geographical distances were calculated as great-circle distances, the

shortest distances between two points on the surface of a sphere, measured along the surface of the sphere, from geographical coordinates using the function "coord2dist" of PRABCLUS. The genetic distances between individuals, Jaccard distances for AFLP data and shared allele distances (Bowcock et al., 1994) and $\hat{e}$ (Watts et al., 2007) for microsatellite data, haplotype data and single nucleotide polymorphisms (SNPs), can be calculated using the function "alleledist" of PRABCLUS. For testing IBD between populations, we implemented the chord distance (Cavalli-Sforza & Edwards, 1967), $F_{ST}/(1 - F_{ST})$ (Weir & Cockerham, 1984), Φ_{pt} (Peakall, Smouse, & Huff, 1995) and three variants of the shared allele distance (Bowcock et al., 1994) for populations (number of alleles shared by two populations summed over all loci divided by 2 × the number of loci compared; average linkage distance based on shared allele distance between individuals; and the variant described by Gutiérrez, Royo, Álvarez, and Goyache (2005), which divides 2 × the average proportion of shared alleles between individuals belonging to two different populations by the sum of the average proportions of shared alleles between individuals within each of the populations) in PRABCLUS.

3.4 | IBD tests of the case studies

These results of IBD tests of the case studies are described in detail in File S1.

4 | DISCUSSION

4.1 | Regression based IBD tests for assessing species status

Whereas the continued co-occurrence of differentiated groups without fusing can be considered as a proof of their species status, it is more difficult to assess the status of allopatric metapopulations. For example, approaches such as the multispecies coalescent model as implemented, for example, in BPP (Yang & Rannala, 2014) tend to overestimate the number of true species (Barley, Brown, & Thomson, 2018; Leaché, Zhu, Rannala, & Yang, 2018; Sukumaran & Knowles, 2017). Likewise, Bayesian clustering methods such as STRUCTURE (Pritchard, Stephens, & Donnelly, 2000), which have also been recommended for delimiting species (Hausdorf & Hennig, 2010; Shaffer & Thomson, 2007), may discern multiple clusters where there is only a single large metapopulation with IBD (Frantz et al., 2009). Sukumaran and Knowles (2017) concluded that candidate species delimited with such approaches should be considered hypotheses that require validation with multiple data types.

Here we have developed a novel approach that can provide evidence for the validation of candidate species based on geographical data. It assesses whether the differentiation between two candidate species can be explained by IBD. We derive expectations about the relationship of genetic distances based on multilocus markers and geographical distances from regressions of the genetic distances

between individuals or populations belonging to the same candidate species versus geographical distances. In principle, this approach might work also with morphological distances or distances based on premating signals (e.g., bird songs, sexual pheromones). Spriggs et al. (2019) have used linear models to test whether the genetic divergence between candidate species is significant beyond what would be expected from geographical isolation alone. However, their approach did not take the dependence between distances into account (see "IBD tests" above). Furthermore, their approach implicitly assumes that the regression slope between geographical and genetic distances is the same within both candidate species and between the candidate species; this assumption is violated in many cases (e.g., see Table S1 and Figure S1). We assess whether the genetic distances between individuals or populations belonging to different candidate species are not larger than expected based on their geographical distances by comparing the regression of the genetic distances between all individuals or populations or at least all distances within one of the candidate species and the distances between individuals or populations belonging to different candidate species versus geographical distances with the regressions within the candidate species. Significance is determined by jackknifing. A rejection of the null hypothesis that the distances between candidate species are not larger than expected based on the within-group regressions of genetic versus geographical distances provides evidence for a classification of the two studied candidate species as distinct species. Analyses of the *Veronica*, *Erebia* and *Narcissus* data (Table S1) showed that the IBD tests based on distances between individuals are suitable to distinguish between groups that were considered distinct species versus geographical subgroups within species based on other data (Algarra, Blanca, Cueto, & Fuentes, 2018; Gratton et al., 2016; Lattes, Mensi, Cassulo, & Balletto, 1994; Martínez-Ortega et al., 2004b; Medrano et al., 2014; see also the following discussion).

An important issue in IBD analyses is the assumption of linearity, or rather the assumption that we know appropriate transformations of the genetic and geographical distances which then should be linearly related. It is hard to argue why this should generally hold, although it may be considered appropriate to use a simple model if this is not rejected. Most genetic distances have maximum values of 1, so a linear model ultimately must be wrong at least if instances occur in which the distance reaches 1, as actually occurs for the distances between individuals of *Conradina* and *Narcissus*. Transformations that repair this and for which linearity is reasonable are hard to define. The power of the test regarding whether the between-group distances can be explained by IBD is the worse the faster high genetic distances are reached with increasing geographical distances. This is especially clear in the case of the distances between individuals based on microsatellite data of *Conradina*, which frequently show high distances even within one population (Figure S1d–f). Thus, the between-group distances can hardly be higher than predicted by the within-group regressions. The lack of significance for rejecting the null hypothesis that the between-group distances can be explained by IBD should not be interpreted as evidence for the conspecificity of

the examined groups if the within-group distances are already close to the maximum. Because of their high variability, microsatellites generally result in higher distances than AFLP or SNP data. Thus, and also because usually fewer loci are scored using microsatellites than with AFLP or SNPs, microsatellite data are less suitable for species delimitation than markers that result in distances that are less quickly "saturated" with increasing geographical distances and represent a larger portion of the genome.

In the case of the microsatellite data of *Conradina*, using populations instead of individuals as units strongly improved the linearity of the regression of the genetic versus the geographical distances (compare Figure S1d,e with Figure S2b,c). The IBD analyses based on distances between individuals (Figure S1d–f) did not permit conclusions about the status of the tested *Conradina* groups because of the lack of power of the tests due to the high distances within populations and between individuals of neighbouring populations. In contrast, the distances between populations increased linearly with log-transformed geographical distances (Figure S2b,c). The chord distances between *Conradina canescens* and *C. brevifolia* populations were not larger than expected considering the regression based on all within- and between-group distances taken together (H_{02} in Table S1). Thus, the data provide no evidence for their specific distinctness, and these two taxa might better be considered conspecific as suggested by Wunderlin (1998) and classified as subspecies.

With the other investigated data sets, IBD tests based on distances between populations proved to be problematic. It is clear that the sample size is smaller when populations instead of individuals are used as units. Several of the studied taxonomic problems could not be tested with IBD tests based on distances between populations because not enough populations were sampled to perform the tests. However, this is not only a problem of sampling. In some cases, locally endemic species comprise fewer populations than would be necessary for an IBD test at the population level. Even if enough populations for performing the tests were sampled, the results often remained inconclusive. A meta-analysis of intraspecific IBD analyses indicated that more than nine populations were needed to achieve more than 50% probability of significant IBD, more than 17 populations were needed to achieve 75% probability of significant IBD, and more than 24 populations were required to achieve 90% probability of significant IBD (Jenkins et al., 2010). Such high numbers of populations per species are rarely sampled across a group of species for systematic studies. One reason for the large numbers of populations that are necessary for demonstrating IBD and for the inconclusive results of our tests is a large scatter of the genetic distances depending on the geographical distances. In the data sets we re-analysed, this is probably at least partly caused by insufficient sampling within populations so that the distances between the populations cannot be accurately estimated, resulting in unreliable estimates of the regression coefficients. The standard sampling for taxonomic studies that often deal with rare and/or geographically restricted species is usually not adequate for IBD tests at the population level. In addition, IBD analyses based on populations have also more general problems. The distance measures between populations not only reflect the differentiation between populations but may also be affected by the variability within populations. The latter is not necessarily related to the geographical distances between populations. We used several statistics for quantifying the differentiation between populations (chord distance, Φ_{pt}, $F_{ST}/(1 - F_{ST})$, and three variants of the shared allele distance) in IBD tests, and they yielded mostly similar results. Chord and shared allele distances can also be calculated if a population is represented only by one individual, but several individuals of both populations are necessary for the calculation of $F_{ST}/(1 - F_{ST})$ and Φ_{pt}. Thus, more information is lost if the latter statistics are used.

Our tests indicate whether the differentiation between two candidate species can be explained by IBD. This is not necessarily a test for species status. As already mentioned, the population structure of a pair of species might also be compatible with IBD (e.g., if the two species originated from a widespread ancestral species that was structured by IBD across its range). An overlap of the ranges of the two species might nevertheless demonstrate their species status. The sympatry criterion (i.e., the continued co-occurrence of two differentiated groups without an erosion of their differentiation) is always the strongest proof of their species status. However, for allopatric candidate species additional criteria are necessary. In the case of peripatric taxa, the amount of admixture, the width of a hybrid zone and the abruptness of the changes across a hybrid zone may provide arguments for the classification. Apart from crossing experiments, IBD tests are the only tests that provide an argument for the classification of strictly allopatric candidate species without contact zones. Another criterion, which has not been implemented in a formal test so far, might be whether the differentiation of an allopatric pair of candidate species reaches the degree found in closely related sympatric species. However, differential adaptation to different environments may include different genetic changes that may or may not be associated with morphological changes. Thus, the "degree of differentiation" is difficult to measure and even more difficult to test, even between closely related taxa. Speciation is usually a gradual evolutionary process and, thus, the decision on at which point in this process two differentiating groups should be classified as species will remain arbitrary to some degree. The IBD tests are a tool to make this decision slightly more objective.

A geographical expansion of two candidate species (e.g., proceeding from refuges) leading to an approximation of their distribution areas may result in more large genetic distances between individuals or populations of the two candidate species at smaller geographical distances. This would increase the likelihood that the two candidate species are considered distinct species. However, as geographical distances are log-transformed, species must approach each other significantly before this affects the distribution of distances and IBD tests. If candidate species approach each other geographically, the probability increases that individuals or propagules will be exchanged at least from time to time. If the candidate species are not isolated, this will lead to gene flow and a decrease in genetic distances between candidate species. On the other hand, if we do not observe gene flow and a decrease in genetic distances between

candidate species, this will support their classification as distinct species.

Human-induced translocations, such as restocking of fish species, can disturb the natural pattern and decrease the informational value of the relationship between geographical and genetic distances. Thus, populations resulting from such translocations should not be used for IBD analyses.

4.2 | Comparison of regression-based IBD tests with approaches for assessing species status using Mantel tests

In contrast to the regression procedure proposed here, most previous approaches to assess whether the differentiation between candidate species can be explained by IBD (Gratton et al., 2016; Medrano et al., 2014) were based on permutation-based Mantel or partial Mantel tests (Mantel, 1967; Smouse, Long, & Sokal, 1986). Medrano et al. (2014) used permutation-based simple and partial Mantel tests "to determine the proportion of total variance of genetic distances between populations that could be attributed to long-term historical divergence or more recent and local isolation-by-distance processes." Decisive for the argumentation of Medrano et al. (2014) is whether a partial Mantel test indicates that a significant proportion of the genetic variation can be explained by the tested grouping after statistically accounting for the effect of the geographical distance matrix. Although Medrano et al. (2014) did not explicitly define a null hypothesis, what is tested by the partial Mantel test may be equivalent to our null hypothesis. However, we believe that it is more appropriate to frame this as a regression rather than a correlation problem because of the causal asymmetry between geography and genetics. Another possible permutation approach would be to fit the regression models presented here and to permute the group memberships of the individuals, which under H_0 should not change the regression parameters. Both of these approaches suffer from the same problem. In many cases most or all the within-group geographical distances are small and the between-groups geographical distances are large. Permuting the group labels (which implicitly also occurs in the partial Mantel test) means that some distances that were originally between-groups become within-group distances and vice versa. This will systematically change the distributions of geographical distances within groups, which in turn can have a strong effect on regression (and partial correlation) estimation, as regression estimation is less variable if there is more variation in the x (explanatory) variable whereas the variation in the y variable is unchanged. Therefore, such an approach is not appropriate to assess the expected variation for a real pattern in which within-group distances tend to be small. Similar problems regarding Mantel and partial Mantel tests have been reported by Frantz et al. (2009) and Guillot and Rousset (2013), who concluded that partial Mantel tests are not statistically valid.

Gratton et al. (2016) specified as operational criteria for classifying clusters as species "(1) a pattern of within-clusters IBD ..., and

(2) genetic differentiation between pairs of individuals belonging to different clusters shows no clear dependence on their geographical distance (i.e., individuals sampled in, or near to, contact zones do not tend to be genetically intermediate)." Criterion (1) is not suitable for testing species status because IBD is not a general property of species (Jenkins et al., 2010). After speciation, the interspecific distances may still be correlated with geographical distances if two species originated from a widespread ancestral species that was structured by IBD across its range (see above). Introgression might also contribute to the maintenance of this pattern. Thus, the condition described as criterion (2) is not mandatory for pairs of recently diverged species. Thus, neither a lack of a correlation of genetic and geographical distances within clusters, nor a significant correlation of genetic distances between groups with geographical distances, can be interpreted as an argument for lumping candidate species. The Mantel tests applied by Gratton et al. (2016) are not suitable to test specifically whether "individuals sampled in, or near to, contact zones do not tend to be genetically intermediate" (Gratton et al., 2016). The Mantel test assesses the correlation between genetic and geographical distances across the range occupied by the analysed individuals and not specifically the genetic distances of individuals from contact zones.

Gratton et al. (2016) did not apply their tests consequently. Their second criterion for distinct species, the lack of a correlation of the genetic differentiation between pairs of individuals belonging to different clusters with geographical distances, was not fulfilled for *Erebia tyndarus* and *E. nivalis*; that is, they found a significant correlation of the genetic distances between these species with geographical distances. Nevertheless, they classified them as good species. We agree with this decision because the ranges of the two species broadly overlap, they form clearly separated clusters in the principal components analysis (Gratton et al., 2016: fig. 2a) and a STRUCTURE analysis indicated only little admixture between co-occurring populations of the two species (Gratton et al., 2016: fig. 3). We consider the continued co-occurrence of two taxa without fusing as decisive evidence for their species status. Our test rejected the null hypothesis that the genetic distances between individuals belonging to two different candidate species are not larger than expected based on their geographical distance for *E. tyndarus* and *E. nivalis* as well as all other pairs of the four species of the *E. tyndarus* complex (Figure S1l–n; Table S1). The interspecific genetic distances are larger than expected based on the relationship between the genetic intraspecific distances and the geographical distances of the sampled specimens. The positive correlation of the genetic distances between *E. tyndarus* and *E. nivalis* specimens with their (least-cost path) geographical distances is not relevant for the test of the hypothesis that the magnitude of the interspecific distances can be explained by IBD as expected based on the intraspecific distances. Actually, the genetic distances between *E. tyndarus* and *E. nivalis* specimens are significantly larger than expected from the intraspecific IBD regression pattern (Figure S1l; Table S1). Thus, Mantel tests with the interspecific genetic distances as applied by Gratton et al. (2016) do not provide relevant evidence for or against the species status of the considered taxa.

Whereas whether the differentiation between two candidate species can be explained by IBD is always tested in our approach, Medrano et al. (2014) and Gratton et al. (2016) included three groups in one Mantel or partial Mantel test in some cases. The outcome of a test with more than two groups is difficult to interpret because the differentiation between two of the tested groups might be explained by IBD, whereas the third group could be more strongly differentiated. For example, Gratton et al. (2016) reported that a Mantel test showed a significant correlation between genetic and geographical distances within *E. cassioides* in the wide sense. They concluded that the three clusters that were identified by *k*-means clustering and STRUCTURE are conspecific. Our pair-wise regressions showed that the null hypothesis that the genetic distances between individuals belonging to two different candidate species are not larger than expected based on their geographical distance can actually not be rejected for the Western Alps + Pyrenees + Northern Apennines versus the Central + Southern Apennines cluster (Figure S1p; Table S1). Thus, the data provide no evidence for the specific distinctness of these two subgroups. The STRUCTURE analysis that showed that the genome of the single specimen collected in the Northern Apennines is composed to about equal portions of part of the Western Alps + Pyrenees cluster and the Central + Southern Apennines cluster (Gratton et al., 2016: fig. 3a) also indicated that the two subgroups should be considered conspecific. However, the null hypothesis was rejected for the comparisons of the population subgroup from the Orobian and Eastern Alps versus the subgroup from the Western Alps + Pyrenees + Northern Apennines and the subgroup from the Central + Southern Apennines as well as a combination of the two latter subgroups (Figure S1o,q,r; Table S1). Although the differentiation between these groups is smaller than between the other species of the *E. tyndarus* complex, these results suggest that the populations from the Orobian and Eastern Alps can be classified as a distinct species (albeit this should be corroborated with genetic data from additional samples). This conclusion is also supported by the result of the STRUCTURE analysis of the RAD data that showed little admixture between the cluster from the Orobian and Eastern Alps and the two other clusters (Gratton et al., 2016: fig. 3a). In particular, the single specimen from the Orobian Alps showed no admixture with the geographically close populations from the Western Alps. Lattes et al. (1994) had already recognized the distinction between western and eastern subgroups of *E. cassioides* based on allozyme data. Thus, the IBD tests and the STRUCTURE analysis of the RAD data together with allozyme data support the separation of the western populations as *E. arvernensis* (see Descimon & Mallet, 2009) from the eastern *E. cassioides*.

Concerning *Narcissus* from the Baetic Ranges, our analysis confirmed that the two major groups distinguished by Medrano et al. (2014), the blue group including *N. bujei* and the green group including *N. longispathus* and *N. nevadensis*, form two distinct species complexes. This was not surprising because these groups have overlapping ranges and are not sister groups in nuclear ITS and organellar phylogenies (Marques, Fuertes Aguilar, Martins-Louçao, Moharrek, & Nieto Feliner, 2017; Rønsted, Savolainen, Mølgaard, & Jäger, 2008). Using partial Mantel tests, Medrano et al. (2014) found that the classification into three subgroups remained a significant predictor of genetic distance after having statistically accounted for the effect of geographical distance for the green group, whereas the classification into three subgroups was able to explain only a small portion of the genetic variation after statistically accounting for the effect of geographical distance for the blue group. They concluded from the results of the partial Mantel tests that the subdivision within the green subgroup could be explained in terms of long-term historical processes rather than microevolutionary processes resulting from IBD, whereas IBD is the most parsimonious explanation to account for genetic differentiation within the blue group. Our regression analyses (Figures S1g–k and S2d–h) revealed that the relationships of genetic and geographical distances within and between the subgroups are more complicated. Our tests for H_{01} show that the regressions differ between all subgroups (Table S1). The general problems of Mantel tests (see above) are aggravated in these cases by combining subgroups (actually three in each test) with different relationships of genetic and geographical distances in a single correlation test. Such a test cannot provide evidence for the distinctness of the single subgroups and may even be misleading. Our tests showed that the pair-wise differentiation of subgroups within the green group as well as of blue_N and blue_C can be explained by IBD considering the relationship of genetic and geographical distances within one of the subgroups, but not from the perspective of the other subgroup (Figure S1h–k, Table S1). Thus, it might be preferable to classify the subgroups of the green group as subspecies of a single species, *N. nevadensis* (including *N. longispathus*), with three subspecies as recently proposed by Algarra et al. (2018) based on a morphometric analysis. In contrast to the results of the Mantel tests of Medrano et al. (2014), our analyses showed that there is no principal difference between the differentiation between the subgroups of the green group and of blue_N and blue_C, which might also be classified as subspecies of a single species, *N. bujei*.

The blue_S subgroup is represented in the data set only by a single population so that H_{01} could not be tested. However, the blue_S subgroup is exceptional in the composition of the genome, approximately three-quarters of which are from the blue group, but one-quarter originated from the green group according to the STRUCTURE analysis of Medrano et al. (2014). This composition is compatible with the hypothesis that the blue_S subgroup originated by a hybridization of the blue and the green group (which are not sister groups) and a backcross with the blue group. The very high frequency-downweighted marker value of this population (DW, see Medrano et al., 2014: table 1) indicates an accumulation of rare markers. This and the similar genome composition of the sampled individuals reveal that this population has probably been isolated for a long time. As a stabilized hybrid population with little gene exchange with the parental species (as indicated by the STRUCTURE analysis within the blue group and the low variability

of the genome composition), the blue_S subgroup may deserve species status.

ACKNOWLEDGMENTS

AUTHOR CONTRIBUTIONS

B.H. and C.H. designed the study; C.H. developed and programmed the tests; B.H. and C.H. analysed the data and wrote the manuscript.

DATA AVAILABILITY STATEMENT

Veronica data: TreeBASE, Study 1255. *Conradina* data: Supporting Information file 2. *Narcissus* data: Supporting Information File S3. *Erebia* data: Dryad http://dx.doi.org/10.5061/dryad.3n5c9.

ORCID

Bernhard Hausdorf https://orcid.org/0000-0002-1604-1689

REFERENCES

Algarra, J. A., Blanca, G., Cueto, M., & Fuentes, J. (2018). New data on daffodils of the *Narcissus nevadensis* complex (Amaryllidaceae) in SE Spain: *N. nevadensis* subsp. *herrerae* subsp. nov., and *N. nevadensis* subsp. *longispathus* comb. nov. *Phytotaxa*, *371*, 133–139. https://doi.org/10.11646/phytotaxa.371.2.8

Barley, A. J., Brown, J. M., & Thomson, R. C. (2018). Impact of model violations on the inference of species boundaries under the multispecies coalescent. *Systematic Biology*, *67*, 269–284. https://doi.org/10.1093/sysbio/syx073

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., & Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, *368*, 455–457. https://doi.org/10.1038/368455a0

Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species delimitation. *Molecular Ecology*, *22*, 4369–4383. https://doi.org/10.1111/mec.12413

Cavalli-Sforza, L. L., & Edwards, A. W. F. (1967). Phylogenetic analysis: Model and estimation procedures. *American Journal of Human Genetics*, *19*, 233–257.

Descimon, H., & Mallet, J. (2009). Bad species. In J. Settele, T. Shreeve, M. Konvička, & H. van Dyck (Eds.), *Ecology of butterflies in Europe* (pp. 219–249). Cambridge, UK: Cambridge University Press.

Edwards, C. E., Soltis, D. E., & Soltis, P. S. (2008). Using patterns of genetic structure based on microsatellite loci to test hypotheses of current hybridization, ancient hybridization and incomplete lineage sorting in *Conradina* (Lamiaceae). *Molecular Ecology*, *17*, 5157–5174.

Edwards, D. L., & Knowles, L. L. (2014). Species detection and individual assignment in species delimitation: Can integrative data increase efficacy? *Proceedings of the Royal Society B: Biological Sciences*, *281*, 20132765. https://doi.org/10.1098/rspb.2013.2765

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26. https://doi.org/10.1214/aos/1176344552

Frantz, A. C., Cellina, S., Krier, A., Schley, L., & Burke, T. (2009). Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: Clusters or isolation by distance? *Journal of Applied Ecology*, *46*, 493–505. https://doi.org/10.1111/j.1365-2664.2008.01606.x

Gratton, P., Trucchi, E., Trasatti, A., Riccarducci, G., Marta, S., Allegrucci, G., … Sbordoni, V. (2015). Data from: Testing classical species properties with contemporary data: how "bad species" in the brassy ringlets (*Erebia tyndarus* complex. Lepidoptera) turned good. Dryad. https://doi.org/10.5061/dryad.3n5c9

Gratton, P., Trucchi, E., Trasatti, A., Riccarducci, G., Marta, S., Allegrucci, G., … Sbordoni, V. (2016). Testing classical species properties with contemporary data: How "bad species" in the brassy ringlets (*Erebia tyndarus* complex, Lepidoptera) turned good. *Systematic Biology*, *65*, 292–303.

Guillot, G., Renaud, S., Ledevin, R., Michaux, J., & Claude, J. (2012). A unifying model for the analysis of phenotypic, genetic, and geographic data. *Systematic Biology*, *61*, 897–911. https://doi.org/10.1093/sysbio/sys038

Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, *4*, 336–344. https://doi.org/10.1111/2041-210x.12018

Gutiérrez, J. P., Royo, L. J., Álvarez, I., & Goyache, F. (2005). MOLKIN v2. 0: A computer program for genetic analysis of populations using molecular coancestry information. *Journal of Heredity*, *96*, 718–721. https://doi.org/10.1093/jhered/esi118

Hausdorf, B. (2011). Progress toward a general species concept. *Evolution*, *65*, 923–931. https://doi.org/10.1111/j.1558-5646.2011.01231.x

Hausdorf, B., & Hennig, C. (2010). Species delimitation using dominant and codominant multilocus markers. *Systematic Biology*, *59*, 491–503. https://doi.org/10.1093/sysbio/syq039

Hennig, C., & Hausdorf, B. (2019). Package 'prabclus', version 2.3-1. Functions for clustering of presence-absence, abundance and multilocus genetic data. Retrieved from http://cran.r-project.org/web/packages/prabclus/

Jenkins, D. G., Carey, M., Czerniewska, J., Fletcher, J., Hether, T., Jones, A., … Tursi, R. (2010). A meta-analysis of isolation by distance: Relic or reference standard for landscape genetics? *Ecography*, *33*, 315–320. https://doi.org/10.1111/j.1600-0587.2010.06285.x

Lattes, A., Mensi, P., Cassulo, L., & Balletto, E. (1994). Genotypic variability in western European members of the *Erebia tyndarus* group (Lepidoptera, Satyridae). *Nota Lepidopterologica, Suppl.*, *5*, 93–104.

Leaché, A. D., Zhu, T., Rannala, B., & Yang, Z. (2018). The spectre of too many species. *Systematic Biology*, *68*, 168–181. https://doi.org/10.1093/sysbio/syy051

Mallet, J. (1995). A species definition for the Modern Synthesis. *Trends in Ecology & Evolution*, *10*, 294–299. https://doi.org/10.1016/0169-5347(95)90031-4

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*, 209–220.

Marques, I., Fuertes Aguilar, J., Martins-Louçao, M. A., Moharrek, F., & Nieto Feliner, G. (2017). A three-genome five-gene comprehensive phylogeny of the bulbous genus *Narcissus* (Amaryllidaceae) challenges current classifications and reveals multiple hybridization events. *Taxon*, *66*, 832–854.

Martínez-Ortega, M. M., Delgado, L., Albach, D. C., Elena-Rosselló, J. A., & Rico, E. (2004a). Species boundaries and phylogeographic patterns in cryptic taxa inferred from AFLP markers: Veronica subgen. Pentasepalae (Scrophulariaceae) in the Western Mediterranean. TreeBASE; Study 1255.

Martínez-Ortega, M. M., Delgado, L., Albach, D. C., Elena-Rosselló, J. A., & Rico, E. (2004b). Species boundaries and phylogeographic patterns in cryptic taxa inferred from AFLP markers: *Veronica* subgen. *Pentasepalae* (Scrophulariaceae) in the Western Mediterranean. *Systematic Botany*, *29*, 965–986.

Medrano, M., López-Perea, E., & Herrera, C. M. (2014). Population genetics methods applied to a species delimitation problem: Endemic trumpet daffodils (*Narcissus* section *Pseudonarcissi*) from the Southern Iberian Peninsula. *International Journal of Plant Sciences*, *175*, 501–517.

Miller, R. G. (1974). The jackknife – a review. *Biometrika*, *61*, 1–15.

Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, *7*, 16. https://doi.org/10.1186/1742-9994-7-16

Peakall, R., Smouse, P. E., & Huff, D. R. (1995). Evolutionary implications of allozyme and RAPD variation in diploid populations of dioecious buffalograss *Buchloë dactyloides*. *Molecular Ecology*, *4*, 135–147. https://doi.org/10.1111/j.1365-294X.1995.tb00203.x

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multi-locus genotype data. *Genetics*, *155*, 945–959.

Quenouille, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics*, *20*, 355–375. https://doi.org/10.1214/aoms/1177729989

R Core Team (2018). *R: A language and environment for statistical computing. Version 3.5.1*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://cran.r-project.org/

Rønsted, N., Savolainen, V., Mølgaard, P., & Jäger, A. K. (2008). Phylogenetic selection of *Narcissus* species for drug discovery. *Biochemical Systematics and Ecology*, *36*, 417–422. https://doi.org/10.1016/j.bse.2007.12.010

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics*, *145*, 1219–1228.

Sauer, J., & Hausdorf, B. (2012). A comparison of DNA based methods for delimiting species in a Cretan land snail radiation reveals shortcomings of exclusively molecular taxonomy. *Cladistics*, *28*, 300–316. https://doi.org/10.1111/j.1096-0031.2011.00382.x

Schlick-Steiner, B. C., Steiner, F. M., Seifert, B., Stauffer, C., Christian, E., & Crozier, R. H. (2010). Integrative taxonomy: A multisource approach to exploring biodiversity. *Annual Review of Entomology*, *55*, 421–438. https://doi.org/10.1146/annurev-ento-112408-085432

Shaffer, H. B., & Thomson, R. C. (2007). Delimiting species in recent radiations. *Systematic Biology*, *56*, 896–906. https://doi.org/10.1080/10635150701772563

Sites, J. W. Jr, & Marshall, J. C. (2004). Operational criteria for delimiting species. *Annual Review of Ecology, Evolution, and Systematics*, *35*, 199–227. https://doi.org/10.1146/annurev.ecolsys.35.112202.130128

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, *47*, 264–279. https://doi.org/10.1111/j.1558-5646.1993.tb01215.x

Smouse, P. E., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Biology*, *35*, 627–632.

Solís-Lemus, C., Knowles, L. L., & Ané, C. (2015). Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution*, *69*, 492–507. https://doi.org/10.1111/evo.12582

Spriggs, E. L., Eaton, D. A. R., Sweeney, P. W., Schlutius, C., Edwards, E. J., & Donoghue, M. J. (2019). Restriction-site-associated DNA sequencing reveals a cryptic *Viburnum* species on the North American coastal plain. *Systematic Biology*, *68*, 187–203.

Sukumaran, J., & Knowles, L. L. (2017). Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences of USA*, *114*, 1607–1612. https://doi.org/10.1073/pnas.1607921114

Tukey, J. W. (1958). Bias and confidence in not quite large samples (abstract). *Annals of Mathematical Statistics*, *29*, 614.

Watts, P. C., Rousset, F., Saccheri, I. J., Leblois, R., Kemp, S. J., & Thompson, D. J. (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. *Molecular Ecology*, *16*, 737–751. https://doi.org/10.1111/j.1365-294X.2006.03184.x

Weir, B. S., & Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution*, *38*, 1358–1370.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, *34*, 28–35. https://doi.org/10.2307/2332510

Wright, S. (1943). Isolation by distance. *Genetics*, *28*, 114–138.

Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*, 851–865. https://doi.org/10.1046/j.1420-9101.2001.00335.x

Wunderlin, R. P. (1998). *Guide to the vascular plants of Florida*. Gainesville, FL: University Press of Florida.

Yang, Z., & Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple loci. *Molecular Biology and Evolution*, *31*, 3125–3135. https://doi.org/10.1093/molbev/msu279

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.