Title: Assigning occurrence data to cryptic taxa improves climatic niche assessments: Biodecrypt, a new tool tested on European butterflies

Leonardo Platania[1]†, Mattia Menchetti[2]†, Vlad Dincă[3], Cecília Corbella[1], Isaac Kay-lavelle[1] Roger Vila[1], Martin Wiemers[4,5], Oliver Schweiger[5], Leonardo Dapporto[2]*

[1] Institut de Biologia Evolutiva (CSIC - Universitat Pompeu Fabra), Barcelona, Spain

[2] ZEN Lab, Dipartimento di Biologia dell'Università di Firenze, via Madonna del Piano 6 50019 Sesto Fiorentino, Italy

[3] Ecology and Genetics Research Unit, PO Box 3000, University of Oulu, 90014, Finland

[4] Senckenberg Deutsches Entomologisches Institut, Eberswalder Str. 90, 15374 Müncheberg, Germany

[5] Helmholtz Centre for Environmental Research - UFZ, Department of Community Ecology, 06120 Halle, Germany

Corresponding author: leonardo.dapporto@unifi.it

Running title: *biodecrypt*, assigning occurrence data

Abstract

Aim

Occurrence data are fundamental to macroecology, but accuracy is often compromised when multiple units are lumped together (e.g. in recently separated cryptic species or citizen science records). Using amalgamated data leads to inaccuracy in species mapping, to biased beta-diversity assessments and to potentially erroneously predicted responses to climate change. We provide a set of R functions (biodecrypt) to objectively attribute undetermined occurrences to the most probable taxon based on a subset of identified records.

Innovation

Biodecrypt assumes that unknown occurrences can only be attributed at certain distances from areas of sympatry. The function draws concave hulls based on the subset of identified records; subsequently, based on hull geometry, it attributes (or not) unknown records to a given taxon. Concavity can be imposed with an alpha value and sea or land areas can be excluded. A cross-validation function tests attribution reliability and another function optimizes the parameters (alpha, buffer, distance ratio between hulls). We applied the procedure to 16 European butterfly complexes recently separated into 33 cryptic species for which most records were amalgamated. We compared niche similarity and divergence between cryptic taxa, and we re-calculated and contributed updated CLIMBER variables for climatic preferences.

Main conclusions

Biodecrypt showed a cross-validated correct attribution of known records always ≥98% and attributed more than 80% of unknown records to the most likely taxon in parapatric species. The functions determined where records can be assigned even for largely sympatric species, and highlighted areas where further sampling is required. All the cryptic taxa showed significantly diverging climatic niches, reflected in different values of mean temperature and precipitation compared to the values originally provided in the CLIMBER database. The substantial fraction of cryptic taxa existing across different taxonomic groups and their divergence in climatic niches highlights the importance of using reliably assigned occurrence data in macroecology.

Key words. Biodecrypt, climatic niches, CLIMBER variables, occurrence data, cryptic taxa, European butterflies

Introduction

A solid record of species occurrence data is key to understand the multiple factors defining their large-scale geographic distributions and, by means of ecological niche modelling, to assess and project their responses to changing environmental conditions in terms of range expansion or contraction (Franklin, 2010; Schweiger et al., 2012; Hortal et al., 2015; Thuiller et al., 2016). In addition, resulting species-specific niche characteristics have provided conservation biogeography with a powerful set of species features, such as measures of mean and variation in multiple climatic characteristics (e.g. CLIMBER variables; Schweiger, Harpke, Wiemers, & Settele 2014), useful for assessing community-wide responses to global change (Devictor et al., 2012; Zografou et al., 2014; Herrando et al., 2019).

Generating reliable species occurrence data at continental scale requires an enormous effort and such databases have been assembled over decades of field research, often based on, or improved by, citizen science projects (Dennis, Morgan, Brereton, Roy, & Fox, 2017; Titeux et al., 2017). In addition, proper definition and discrimination of species are necessary for reliable niche modelling, as well as for the identification of environmental preferences of species and the derived specific indices. However, the existence of a considerable fraction of cryptic species in almost all groups of living organisms (Bickford et al., 2007) poses a serious challenge to our understanding of diversity in general, and to this line of research in particular. When a taxon, believed to represent a single species, is recognized as two or multiple cryptic species, the occurrence data accumulated for decades suddenly become obsolete. Researchers often amalgamate the occurrence data for cryptic taxa, but this approach ignores a substantial fraction of diversity in terms of species identity, distribution, evolution, and its potential dynamics in changing environments.

Most complexes of cryptic taxa are parapatric with minimal areas of sympatry, frequently because they evolved in allopatry and the achievement of secondary sympatry is delayed by

(i) limited dispersal and competition due to a still incomplete separation of ecological niches or by (ii) reproductive interference due to the lack of a pre-mating barrier (Pigot & Tobias, 2013, 2015; Vodă, Dapporto, Dincă, & Vila, 2015). Only a minor fraction of cryptic taxa are largely sympatric and these typically show strong reproductive barriers (Dincă, Lukhtanov, Talavera, & Vila, 2011; Dincă et al., 2013). Because cryptic species tend to show parapatric distributions (Waters, 2011; Vodă et al., 2015; Dapporto et al., 2017; Scalercio et al., 2020), they encompass a conspicuous fraction of beta-diversity (Vodă et al., 2015) and, since they inhabit different areas, are expected to be adapted to different climates (Toews, Mandic, Richards, & Irwin, 2014). For this reason, cryptic species are supposed to react differently to climatic changes and any modelling based on amalgamated records is likely inaccurate (Lecocq, Harpke, Rasmont, & Schweiger, 2019).

We provide a methodology to objectively attribute occurrence data previously referring to a single taxon to the most likely taxon among two or more newly recognized entities, based on a subset of ascertained records and on justifiable geographic rules. We applied this procedure to the cryptic butterfly species of Europe recently separated into different taxa (Wiemers et al., 2018) that were amalgamated in the Distribution Atlas of Butterflies in Europe (Kudrna et al., 2011; Kudrna, Pennerstorfer, & Lux, 2015). This atlas, of which several editions have been published, represents the most comprehensive source of occurrence data for European butterflies, and the data from the 2011 edition were used to calculate the widely used CLIMBER variables describing species ranges and their climatic preferences over Europe (Schweiger et al., 2014).

We provide six new R functions, added to the recluster R package (Dapporto et al., 2013), for parameter optimisation, record attribution to potential cryptic taxa and for testing the reliability of the procedure. Finally, we provide new climatic variables for the species included in this study and show that cryptic taxa differ substantially in their climatic niches.

Methods

The algorithm

The objective of the algorithm is to reliably attribute species membership to a set of ambiguous records belonging to two (or more) cryptic species based on the distribution of a subset of accurately determined records. The main idea is that records from an area where only one taxon occurs can be attributed with confidence, while records from the areas of sympatry or too far from any ascertained record cannot be reliably attributed. For this purpose, we developed a series of R (R Core Team, 2019) functions added to the recluster R package (biodecrypt, biodecrypt.view, biodecrypt.cross, biodecrypt.wrap, optimize.biodecrypt, plot.biodecrypt). The main inputs for the functions are a matrix with longitude and latitude for all the occurrence data and a vector (in the same order) providing their identification. The records identified to species-level (identified records) must be indicated in the vector with a sequential numeric value (1, 2... n), which represents the verified membership to the nth species. The occurrence data with unknown identification (unidentified records) are marked with a 0 (Fig. 1). Based on this vector and on the geographical coordinates of identified records, biodecrypt builds hulls of distribution for each species. In a highly simplified hypothesis, the distribution of a species can be approximated by a convex hull among the geographic coordinates of identified records. Nevertheless, areas of distribution can be largely concave, mostly in geomorphologically highly heterogeneous regions. This is the case for Southern Europe, characterized by the presence of three major peninsulas and several insular systems with contrasting species assemblages (Vodă et al., 2015), as well as by a complex quaternary biogeography (Schmitt, 2007; Dapporto et al., 2019). For this reason, biodecrypt and biodecrypt.view use the function getDynamicAlphaHull of the rangeBuilder R package (https://github.com/ptitle/rangeBuilder, accessed 2020/02/05) to construct concave alpha-hulls.

An alpha-hull is a piecewise series of linear simple curves in the Euclidean plane associated with the shape of a finite set of points (Edelsbrunner, Kirkpatrick, & Seidel, 1983). Alpha-hulls generalize the concept of the convex hull since every convex hull is an alpha-hull, whereas not every alpha-hull is a convex hull. Alpha-hulls are not necessarily convex and two points inside an alpha-hull can be connected by a segment not completely lying inside the hull itself. The boundary of the alpha-hull is formed by arcs with radius α (see Fig. 2 for a polygon with α = 3). For α=0 the hull is reduced to the set of points. For increasing values, the area encompassed by the alpha-hull increases in the form of separate concave polygons connecting an increasing number of points, which in some cases remain isolated. For a very high value of α, an alpha-hull corresponds to the convex hull connecting the set of points. For these reasons, alpha-hulls are particularly suitable to model disjunct organism distributions including dot-like populations, which – using convex hulls – appear as continuous areas.

The getDynamicAlphaHull function supports an initial alpha value that determines a starting custom degree of concavity, which is increased until a given fraction of identified records are included in the resulting hulls (default 95%) and the number of separate polygons is lower than a custom number (default 10). This function can also remove sea or ground areas from the hulls when terrestrial or aquatic organisms, respectively, are under study, thus improving the precision of the hull geometry. After the construction of the alpha-hulls, biodecrypt attempts the attribution of unidentified records to the most likely species (Fig. 1). For this aim, biodecrypt also requires a buffer and a ratio value (explained below). Based on hull geometry and their relative position, each unidentified record could be either: i) inside more than one hull, ii) inside a single hull, or iii) outside all hulls. The three cases are treated separately.

Cases inside more than one hull

In this case, the function cannot attribute the unidentified records to a species (case 1 in Fig. 2) and only the a priori identified records belonging to intersection areas are passed to the final vector as identified.

Cases inside a single hull

The unidentified records falling inside a single hull are attributed to that species if their distance to any other hull is higher than the buffer value (in km) provided by the user (case 2 in Fig. 2). Unidentified records inside the buffer of another hull are not attributed (case 3 in Fig. 2).

Cases outside all hulls

The unidentified records which do not fall inside any hull are attributed to the closest hull if: i) the distance from the second nearest hull is higher than the buffer and if ii) the ratio between the minimum distance to the second closest hull and to the closest hull is more than the ratio value indicated by the user. For example, in Fig. 2 point 4 is not attributed while point 5 is attributed to *Polyommatus celina*.

Check for distances from the nearest identified record

As described above, the attribution of unknown records is strictly determined by the distance from the hulls. The biodecrypt function also contains an option ("checkdist") to check if cases attributed to a given species based on relative distance from hulls are closer to an identified record of another species, which may occasionally occur. If this option is selected (default) these cases are not attributed to any species (Fig. 2, case 6).

Different alpha values can better fit the distribution of a given cryptic species and the optimal alpha value can be evaluated by series of cross-validation analyses using biodecrypt.wrap (see below), or according to the researcher's perception. For this reason, we implemented the biodecrypt.view function providing a visual representation of the alpha-hulls for the different cryptic taxa and given alpha values. The alpha values can be modified until an optimal representation is obtained (Fig. 1). The biodecrypt and biodecrypt.view functions also provide three measures: i) the area occupied by each hull (in km^2), ii) the area of all the possible intersections between pairs of hulls (in km^2) and iii) the fraction of area of intersection between pairs of hulls.

Cross-validation

A third function (biodecrypt.cross) wraps the biodecrypt function to carry out cross-validation of identified cases and to verify the robustness of the attribution of unknown cases (Fig. 2). This function requires the same input of biodecrypt (coordinates and vector with attribution and distance ratio, buffer and alpha values) and a "runs" value defining the number of different runs, thus the fraction of test records in each run. The analysis is repeated as many times as defined in "runs" (a "runs" value of 10 will perform a ten-fold cross-validation). In each run, a randomly selected fraction of 1/"runs" identified records are regarded as unidentified (0 value) and the biodecrypt function is carried out to attribute them. The blind attribution of identified records is compared with their membership and two values are provided: the percentages of cases attributed to a wrong species (misidentified records, MIR) and the percentage of cases not attributed to any species (non-attributed identified records, NIR). MIR and NIR represent measures for the power of the function to correctly attribute unknown records to a given species (NIR) and to avoid mis-identification (MIR).

The function also has an option to calculate the percentage of non-attributed unidentified records (NUR) representing the fraction of unknown records that could not be attributed to a species after the biodecrypt function was completed using the parameters provided by the user and the complete set of identified and non-identified records.

We also provide a function (biodecrypt.wrap) that replicates the cross-validation analysis by using all possible combinations of a series of distance ratio, alpha and buffer values to compare their resulting MIR, NIR and NUR. To optimise the three parameters for each species, we introduced a combination of MIR^2+NIR+NUR as a penalty value for the different combinations of the parameters. Since the method showing the lowest penalty in cross-validation might not necessarily be the optimal value for the final analysis, all the combinations showing a penalty value not higher than a certain threshold compared with the analysis showing the lowest penalty should be considered as similarly good. We provided a value of 10 as a default, representing a variation of about 3 for each addendum of the penalty. The optimal parameters can then be calculated as mean values of distance ratio, alpha and buffer among those used in these cross-validation analyses, weighted by 1/penalty to provide an increasing contribution to the solutions with low penalty values. This is done by optimise.biodecrypt, calculating the optimal values of alpha, buffer and distance ratio based on biodecrypt.wrap results.

A plot.biodecrypt function can be applied to the results of biodecrypt to inspect the solution of the analysis and to locate the NUR records. The same function can be applied to the cross-validation results to locate NIR and MIR records.

Occurrence data used in this study

As a main source for occurrence data of amalgamated data we used the Distribution Atlas of Butterflies in Europe (Kudrna et al., 2011) which also served as the basis for the calculation of the CLIMBER variables describing climatic preferences of European butterflies (Schweiger et al., 2014). An earlier edition of this atlas (Kudrna, 2002) was also used to generate the Climatic Risk Atlas of European Butterflies (Settele et al., 2008). As a supplementary source of occurrence data for both amalgamated and split species, we used specimens belonging to Roger Vila's collection (Institut de Biologia Evolutiva, Barcelona). The main source for occurrence of cryptic species with known attribution was published data, in most cases represented by genitalia assessment and/or by mitochondrial DNA sequences (Appendix S1 for details). For ten species, a series of 52 specimens from the contact zones have also been specifically sequenced (DNA barcoded) for this study (sequencing methods in Appendix S1) that are included in the ”DS-WEUP” BOLD project.

Case study species

Following the latest taxonomic assessment for European butterflies (Wiemers et al., 2018), 20 cases amalgamated in the Distribution Atlas should be divided into 41 distinct taxa with parapatric or sympatric ranges. Of these, we compiled sufficient identified records for 16 cases amalgamated in the above-mentioned atlas, which represent 33 cryptic species. These species were used in this study : 1) *Carcharodus alceae* and *C. tripolinus*, which co-exist in southern Iberia (Dincă et al., 2015); 2) *Spialia sertorius* and *S. rosae*, largely sympatric in Iberia (Hernández-Roldán et al., 2016), 3) *Pyrgus malvae* and *P. malvoides*, parapatric with a contact along central France and the Alps (Koren, Beretta, Črne, & Verovnik, 2013; Litman et al., 2018), 4) *Iphiclides podalirius* and *I. feisthamelii*, parapatric with a contact zone in the Pyrenees and southern France (Wiemers & Gottsberger, 2010; Gaunet et al., 2019); 5) *Zerynthia polyxena* and *Z. cassandra*, parapatric with contact zone in northern Italy (Zinetti et al., 2013); 6) *Pontia daplidice* and *P. edusa,* parapatric with a contact zone in northwestern Italy (Porter, Wenger, Geiger, Scholl, & Shapiro, 1997); 7) *Leptidea sinapis/reali/juvernica*, with *L. reali* and *L. juvernica* being allopatric, but each sympatric with respect to *L. sinapis* (Dincă, Lukhtanov et al., 2011, Dinca et al., 2013); 8) *Lycaena tityrus* and *L. bleusei* parapatric in Iberia (Dincă et al., 2015); 9) *Polyommatus icarus* and *P. celina*, parapatric with contact in southern Iberia (Dincă, Dapporto, & Vila, 2011); 10) *Lysandra coridon* and *L. caelestissima*, parapatric in central Iberia (Talavera, Lukhtanov, Rieppel, Pierce, & Vila, 2013); 11) *Melitaea athalia* and *M. celadussa* parapatric with a contact zone from southern France through the Alps. We also applied the assignment procedure to a series of species showing almost complete allopatric distribution that were split in Wiemers et al. (2018) but not yet considered in the CLIMBER dataset: 12) *Aglais urticae* and *A. ichnusa*, 13) *Iolana iolas* and *I. debilitata,* 14) *Pseudochazara anthelea* and *P. amalthea.* For two largely allopatric species: 15) *Erebia hispania* and *E. rondoui* and 16) *Zizeeria knysna* and *Z. karsandra*, applying the procedure was pointless because of their clear allopatry (Appendix S1), but we separated their records and re-calculated the CLIMBER variables (see below). In-depth descriptions of the markers used to generate the subset of identified records of each species are provided in the Appendix S1. We excluded four species groups from the study because knowledge regarding their distribution was still incomplete. This is caused by uncertainty in the identification of records due to the absence of unequivocal morphological markers, sharing of DNA barcodes, and/or to their insufficiently assessed distribution (*Hipparchia semele/neapolitana/blachieri*, *Pieris napi/balcana*, *Lycaena hippothoe/candens, Melitaea phoebe/ornata*).

To identify the best combination of alpha, buffer and distance ratio parameters, we ran the biodecrypt.wrap function for each species in 80 possible combinations of four alpha values (1, 5, 10, 15), four distance ratio values (2, 3, 4, 5) and five buffer values (0, 40, 80, 120, 160).

Dependency of attribution on range overlap and on parameters

The most problematic cases are represented by records in or close to areas of sympatry between species, which cannot be attributed with confidence. To verify the effect of the degree of sympatry, we correlated the percentage of MIR, NIR and NUR and of optimised distance ratio, alpha and buffer values with the percentage of the overlapping area between species. The significance of the correlations was tested with Spearman tests. We also evaluated the relationship of MIR, NIR and NUR from the three different parameters by using Generalized Additive Mixed Models. We collated the output of the wrap analyses of the 12 species to which biodecrypt.wrap was applied (960 biodecrypt.cross analyses) and modelled MIR, NUR and NIR in three separate analyses against smoothed (k=2) alpha, ratio and buffer using species as a random factor.

Evaluation of niche overlap among cryptic taxa and calculation of CLIMBER variables

To evaluate the potential impact of separation of cryptic taxa on macroecological studies, we evaluated climatic niche overlap among the taxa separated in this study. We used an approach based on a PCA of the climate space in Europe and a density smoothing of the occurrence points for each target species within this space (Broennimann et al., 2012). This is followed by the calculation of niche overlap based on Schoener's D (Schoener, 1968) and the modified Hellinger metric I (Appendix S1 for details). Both indices range from 0 to 1, indicating no niche overlap (0) to full overlap (1). We verified whether the observed overlap is significantly different among separated taxa as done by Warren, Glor, & Turelli (2008). We performed two one-sided tests based on a randomised null model approach, one for niche conservatism and one for niche divergence by testing (i) niche equivalency, i.e. without considering overall available niche space, and (ii) niche similarity, i.e. accounting for the available niche space in Europe (Warren et al., 2008). Analyses were performed with the R package ecospat (Broennimann, Di Cola, & Guisan, 2016) and ade4 (Dray & Dufour, 2007). We also recalculated the CLIMBER variables for the improved distribution data of 33 cryptic taxa following the same procedure applied by Schweiger et al. (2014). For each cryptic group, the former values of mean temperature and precipitation in the distribution area for the amalgamated species complex and for the separated cryptic species were plotted in bivariate plots, together with the values of all the other European species, to illustrate divergence in climatic preferences among cryptic taxa. In order to assess if cryptic pairs have smaller or larger differences in mean temperature and precipitation compared to their congeneric species, we proceeded as follows: We scaled and centered the complete dataset of CLIMBER for mean temperature and precipitation (mean=0, sd=1). Within each genus of the 16 complexes examined we calculated the Euclidean distances in scaled mean temperature and precipitation between all pairs of congeneric species included in CLIMBER. Then, we compared the distances between pairs of cryptic taxa separated in this study and between all the other congeneric taxa by using a generalized linear mixed model with a gamma family distribution (glmer function of the lme4 R package) (Bates et al., 2015), including genus as a random factor. Script and data are uploaded in Dryad (https://doi.org/10.5061/dryad.hmgqnk9dh).

Results

Cross-validation analyses obtained by biodecrypt.wrap identified the models under a penalty threshold of 10 compared to the model with the lowest penalty, and allowed setting of appropriate alpha, buffer and distance ratio for each species (Fig. 1, Table 1). Compared to the relatively large range of the three parameters tested, the optimization by penalty produced similar values for all the species, with optimal ratio ranging from 2.1 to 3.2, alpha from 5.6 to 11.0 and buffer from 31.7 to 96.1 (Table 1). When using the optimised alpha values, the parapatric taxa revealed very limited overlap, with intersections among the hulls usually lower than 5% (Table 1). All species showed very low values of misidentified records (MIR), which were at most 2.0% in the case of *Z. polyxena/cassandra* and *I. podalirius/feisthamelii*. The percentage of non-attributed identified records (NIR) was much higher and ranged between 1.2 and 84.7%, with very high values in sympatric species (*Spialia* and *Leptidea* groups, Fig. 3; Table 1) because, based on the assignment algorithm, all the test records belonging to the overlap areas cannot be attributed in a cross-validation analysis. The percentage of non-attributed unknown records (NUR) also varied considerably but, except for the sympatric *Leptidea* (71.1% NUR) it did not exceed 20.0% of the unknown records (Fig. 3; Table 1). The percentage NIR and NUR correlated with the percentage of area of overlap (Spearman test: NIR, Rho=0.932, P<0.001; NUR, Rho=0.587, P=0.027; Supplementary Figure S1). Conversely, the percentage MIR, and the optimised parameters revealed no correlations with the area of overlap (Spearman test: MIR, Rho=0.432, P=0.160; alpha, Rho=0.030, P=0.919; buffer, Rho=0.048, P=0.870; distance ratio, Rho=-0.073, P=0.804; Supplementary Figure S1).

When the different solutions obtained by biodecrypt.wrap for all species were compared in GAMM analyses it emerged that ratio had no significant effect on MIR, while it strongly increased NIR and NUR with an almost linear trend (Fig. 4 and effective degrees of freedom close to 1 in Table 2). Increasing the buffer had a strong effect in reducing MIR, but it also increased NUR and NIR (with a strongly curvilinear effect for MIR flattening around 100 km), while high values of alpha reduced the number of MIR and NUR and slightly increased NIR (Fig. 4; Table 2).

The analysis of niche overlap (Table 3) showed that climatic niches are more similar than expected by chance for all species pairs, as indicated by non-significant divergence for the niche similarity test (SDD and SDI in Table 3), but only *S. sertorius/rosa*e and *L.*

sinapis/reali also showed a significant conservatism in terms of niche similarity (SCD and SCI in Table 3). However, significant divergence for the niche equivalency tests indicated that the climatic niches differ considerably for all species pairs (EDD and EDI in Table 3). Taken together, these results show that the climatic niches of all species pairs are significantly different, but still more similar than expected by a random distribution across Europe. These results were consistent for both measures of niche overlap, D and I.

CLIMBER variables were calculated for the 33 cryptic species and are available in Appendix S2. When the mean temperature and precipitation for the formerly amalgamated species were plotted together with the data of the newly separated cryptic taxa, it became obvious that the values for the separated species diverge considerably, in some cases spreading all across the main sector of the space occupied by most European species (Fig. 5a-f). A Generalised Linear Mixed Model revealed that differences in mean temperature and precipitation among the cryptic taxa separated here are not smaller than the differences among all congeneric species included in CLIMBER (cryptic taxa, mean difference=1.29=±0.63; congeneric taxa mean difference=1.68±1.21, Estimate=-0.020, Standard Error=0.121, df=1825, t=-0.161, P=0.872).

Discussion

Recently diverged taxa (Pigot & Tobias, 2013), as well as cryptic species of butterflies (Vodă et al., 2015; Dapporto et al., 2017; Scalercio et al., 2020), tend to have parapatric distributions with narrow contact zones. This phenomenon facilitates the reliable attribution of ambiguous records to newly discovered species following a hull-based procedure based on the distribution of a subset of identified records. After the parameters used to build the hulls and to attribute records to species had been optimised in a series of cross-validation analyses, the procedure showed a very high attribution of identified records to their correct species

(MIR always lower or equal to 2%). Moreover, following this procedure, the fraction of unknown records which remained non-attributed (NUR) was at most 20% for parapatric taxa. The number of incorrectly attributed specimens does not considerably increase with the degree of sympatry. However, the values of optimised parameters did not depend on the degree of overlap, and they are more likely imposed by the geometry of the distribution areas. The comparison of a large set of cross validation analyses (GAMM approach) revealed how the parameters can be varied to impose stricter or wider inclusion possibilities for a given taxon. Theoretically, higher distance ratio and buffer, and lower alpha values are expected to decrease the number of incorrectly attributed records, but to increase the number of non-attributed records. Accordingly, increasing buffer strongly reduced MIR but increased NIR and NUR (Fig. 4), while the ratio and the alpha values had much lighter effects on MIR. This is probably due to the relatively simple (not interdigitated) shape of the distribution boundaries among cryptic species of butterflies for which a high alpha value (producing slightly concave hulls) and an optimized distance from other hulls (buffer) can avoid most MIR. These observations can help in refining the solutions. In the genus *Iphiclides*, unavailability of identified specimens from northern areas did not allow to attribute northernmost unidentified records (most probably belonging to *I. podalirius*) and this could have slightly affected the assessment of climatic niches. When including a few misclassified specimens does cause more problems than generating several non-attributed unrecognized record (NUR), a computational strategy could be excluding the squared exponent of MIR from the penalty calculation in biodecrypt.optimise (e.g. MIR+NIR+NUR). This can reduce the maximum ratio for attribution, thus lowering the number of NUR with a very small increase in MIR (Fig. 4). Finally, the values of the procedure parameters can also be adjusted according to researcher’s perception.

The framework presented here provides a standardised method that allows researchers to take advantage of previous efforts of data collection even when modern investigation methods identify new layers of biodiversity. This is not necessarily restricted to cryptic species but can also be relevant for analyses of evolutionary significant units such as those highlighted by species delimitation methods (Lecocq et al., 2019). Moreover, biodecrypt can also be used in cases where very similar species are treated as a single unit in citizen science projects.

Clearly, in the presence of largely sympatric species the approach described here is less efficient because all the occurrence data for the overlapping zones must be directly verified. In general, this can happen when climatic distance governs limits between species distributions more than geographic distance, as revealed by the existence of several species with small ranges being restricted to climatically rare areas (Ohlemüller et al., 2008). In our dataset this could be the case of two pairs, *Leptidea sinapis*/*L. reali* and *Spialia sertorius*/*S. rosae*. In fact, *L. reali* and *S. rosae* are specialized to mountain areas, while *L. sinapis* and *S. sertorius* are climate generalists. Although the distance-based biodecrypt algorithm is less efficient in attributing unknown records in areas where the pairs of species are mostly separated by climatic distances, it remains conservative in avoiding mis-identifications. Moreover, even predominantly sympatric taxa often do not co-occur in large areas of their distributions (Fig. 3) and in these cases, biodecrypt can highlight areas where the attribution to a given taxon is reliable based on distances as opposed to areas where climatic distances might exert an important complementary effect. The detection of poorly studied areas is another notable result stemming from our analysis. In fact, the distribution of non-attributed unknown records can highlight where further research efforts are needed to confirm the distribution of recently recognized species by means of dedicated field research or by morphological examination and DNA sequencing from existing natural history collections (e.g. Kharouba, Lewthwaite, Guralnick, Kerr, & Vellend, 2019). An example of such regions is north-western France for *Pontia daplidice*/*edusa*, *Melitaea athalia*/*celadussa*, and *Pyrgus malvae*/*malvoides* (Fig. 3).

It has already been demonstrated that considering cryptic taxa as amalgamated entities overlooks a large fraction of beta-diversity (Vodă et al., 2015). Our results also indicate that amalgamating cryptic taxa may affect macroecological studies investigating the response of species and communities to climatic changes (Settele et al., 2008; Devictor et al., 2012).We found significant niche divergence for all the pairs of cryptic species and a particularly low level of niche conservatism (two species pairs out of sixteen). This is reflected in the strong differences in the values of CLIMBER variables, which are widely used in macroecology to assess butterfly responses to current and future climatic conditions. The differences in mean temperature and precipitation experienced by cryptic species pairs in Europe are considerable and in the same order of magnitude as the differences among other congeneric species. In some cases, the variables describing climatic preferences of cryptic taxa were so divergent that they spread all over the space defined by the mean temperature and precipitation of most European species.

Failing to consider the divergent climatic niches of cryptic taxa can have considerable impacts on climatic risk assessments based on species distribution modelling. It has been shown, in the case of locally adapted subspecies, that individual responses of taxa to climate change do not necessarily resemble the modelled response of the amalgamated group, which might lead to severe over- or underestimation of the risks (Lecocq et al., 2019). Also, assessments of the response of communities, such as based on the community-weighted mean temperature index, might be improved by considering cryptic species. The replacement of a cool-adapted species by a warm-adapted species of an initially amalgamated group will clearly contribute to an increase in the community-weighted mean temperature index, while ignoring climatic niche divergence within the amalgamated group will be interpreted as no change. Such an effect might, partly, contribute to an observed time lag in the response of butterfly communities to climate change (Devictor et al., 2012). Our dataset comprised 33 species, representing 6.7% of the 496 species recorded in Europe (Wiemers et al., 2018). Notably, some of these complexes represent widely distributed and common butterflies in Europe (*Polyommatus icarus/celina*, *Aglais urticae/ichnusa, Carcharodus alceae/tripolinus, Melitaea athalia/celadussa, Pontia daplidice/edusa* and *Leptidea sinapis/reali/juvernica*). Given the range, density and the large differences in climatic preferences displayed by cryptic taxa, their separation can be an important improvement for monitoring and analysing community-level responses to climate change. This information can also help in connecting large-scale and long-term evolutionary processes with ecological processes through the analysis of species interactions with their abiotic and biotic environments, or shed light on the phylogenetic conservation of niche characteristics relevant to climate-change (e.g. Devictor et al., 2012; Kharouba et al., 2019; Ohlemüller et al., 2008).

In conclusion, it is our hope that the framework here presented, which allows the objective attribution of undetermined occurrence data to recently assessed taxonomic units, will benefit biodiversity mapping, highlight gaps in current knowledge and improve macroecological analyses.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R.H.B., Singmann, H., … Green, P. (2015) Package 'lme4.' *Convergence*, **12**, 2.

Bickford, D., Lohman, D.J., Sodhi, N.S., Ng, P.K.L., Meier, R., Winker, K., … Das, I. (2007) Cryptic species as a window on diversity and conservation. *Trends in ecology & evolution*, **22**, 148–155.

Broennimann, O., Di Cola, V. & Guisan, A. (2016) ecospat: Spatial ecology miscellaneous methods. R package version 2.1. 1.

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., … Zimmermann, N.E. (2012) Measuring ecological niche overlap from occurrence and spatial environmental data. *Global ecology and biogeography*, **21**, 481–497.

Dapporto, L., Cini, A., Menchetti, M., Vodă, R., Bonelli, S., Casacci, L.P., … Biermann, H. (2017) Rise and fall of island butterfly diversity: Understanding genetic differentiation and extinction in a highly diverse archipelago. *Diversity and Distributions*, **23**, 1169–1181.

Dapporto, L., Cini, A., Vodă, R., Dincă, V., Wiemers, M., Menchetti, M., … Vila, R. (2019) Integrating three comprehensive data sets shows that mitochondrial DNA variation is linked to species traits and paleogeographic events in European butterflies. *Molecular Ecology Resources*, **19**, 1623–1636.

Dapporto, L., Ramazzotti, M., Fattorini, S., Talavera, G., Vila, R. & Dennis, R.L.H. (2013) recluster: an unbiased clustering procedure for beta-diversity turnover. *Ecography*, **36**, 1070–1075.

Dennis, E.B., Morgan, B.J.T., Brereton, T.M., Roy, D.B. & Fox, R. (2017) Using citizen science butterfly counts to predict species population trends. *Conservation biology*, **31**, 1350–1361.

Devictor, V., Van Swaay, C., Brereton, T., Brotons, L., Chamberlain, D., Heliölä, J., …

Lindström, Å. (2012) Differences in the climatic debts of birds and butterflies at a continental scale. *Nature climate change*, **2**, 121–124.

Dincă, V., Dapporto, L. & Vila, R. (2011a) A combined genetic-morphometric analysis unravels the complex biogeographical history of *Polyommatus icarus* and *Polyommatus celina* Common Blue butterflies. *Molecular Ecology*, **20**, 3921–3935.

Dincă, V., Lukhtanov, V.A., Talavera, G. & Vila, R. (2011b) Unexpected layers of cryptic diversity in wood white *Leptidea* butterflies. *Nature communications*, **2**, 1–8.

Dincă, V., Montagud, S., Talavera, G., Hernández-Roldán, J., Munguira, M.L., García-Barros, E., … Vila, R. (2015) DNA barcode reference library for Iberian butterflies enables a continental-scale preview of potential cryptic diversity. *Scientific Reports*, **5**, 12395.

Dincă, V., Wiklund, C., Lukhtanov, V.A., Kodandaramaiah, U., Norén, K., Dapporto, L., … Friberg, M. (2013) Reproductive isolation and patterns of genetic differentiation in a cryptic butterfly species complex. *Journal of Evolutionary Biology*, **26**, 2095–2106.

Dray, S. & Dufour, A.-B. (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, **22**, 1–20.

Edelsbrunner, H., Kirkpatrick, D. & Seidel, R. (1983) On the shape of a set of points in the plane. *IEEE Transactions on information theory*, **29**, 551–559.

Franklin, J. (2010) *Mapping species distributions: spatial inference and prediction*, Cambridge University Press.

Gaunet, A., Dincă, V., Dapporto, L., Montagud, S., Vodă, R., Schär, S., … Vila, R. (2019) Two consecutive *Wolbachia*-mediated mitochondrial introgressions obscure taxonomy in Palearctic swallowtail butterflies (Lepidoptera, Papilionidae). *Zoologica Scripta*, **48**, 507–519.

Hernández-Roldán, J.L., Dapporto, L., Dincă, V., Vicente, J.C., Hornett, E.A., Šíchová, J., … Vila, R. (2016) Integrative analyses unveil speciation linked to host plant shift in *Spialia* butterflies. *Molecular Ecology*, **25**, 4267–4284.

Herrando, S., Titeux, N., Brotons, L., Anton, M., Ubach, A., Villero, D., … Stefanescu, C.

(2019) Contrasting impacts of precipitation on Mediterranean birds and butterflies. *Scientific reports*, **9**, 1–7.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, **46**, 523–549.

Kharouba, H.M., Lewthwaite, J.M.M., Guralnick, R., Kerr, J.T. & Vellend, M. (2019) Using insect natural history collections to study global change impacts: challenges and opportunities. *Philosophical Transactions of the Royal Society B*, **374**, 20170405.

Koren, T., Beretta, S., Črne, M. & Verovnik, R. (2013) On the distribution of *Pyrgus malvoides* (Elwes & Edwards, 1897) (Lepidoptera: Hesperiidae) at the eastern part of its range. *Entomologist's Gazette*, **64**, 225–234.

Kudrna, O. (2002) The distribution atlas of European butterflies. *Oedippus*, **20**, 1–343.

Kudrna, O., Harpke, A., Lux, K., Pennerstorfer, J., Schweiger, O., Settele, J. & Wiemers, M. (2011) *Distribution atlas of butterflies in Europe*, Gesellschaft für Schmetterlingsschutz eV Halle.

Kudrna, O., Pennerstorfer, J. & Lux, K. (2015) *Distribution atlas of European butterflies and skippers*, Peks.

Lecocq, T., Harpke, A., Rasmont, P. & Schweiger, O. (2019) Integrating intraspecific differentiation in species distribution models: Consequences on projections of current and future climatically suitable areas of species. *Diversity and Distributions*, **25**, 1088–1100.

Litman, J., Chittaro, Y., Birrer, S., Praz, C., Wermeille, E., Fluri, M., … Gonseth, Y. (2018) A DNA barcode reference library for Swiss butterflies and forester moths as a tool for species identification, systematics and conservation. *PLoS ONE*, **13**.

Ohlemüller, R., Anderson, B.J., Araújo, M.B., Butchart, S.H.M., Kudrna, O., Ridgely, R.S. & Thomas, C.D. (2008) The coincidence of climatic and species rarity: high risk to small-range species from climate change. *Biology letters*, **4**, 568–572.

Pigot, A.L. & Tobias, J.A. (2015) Dispersal and the transition to sympatry in vertebrates.

Proceedings of the Royal Society B: Biological Sciences, **282**, 20141929.

Pigot, A.L. & Tobias, J.A. (2013) Species interactions constrain geographic range expansion over evolutionary time. *Ecology letters*, **16**, 330–338.

Porter, A.H., Wenger, R., Geiger, H., Scholl, A. & Shapiro, A.M. (1997) The *Pontia daplidice-edusa* hybrid zone in northwestern Italy. *Evolution*, **51**, 1561–1573.

R Core Team (2019) R: a language and environment for statistical computing, version 3.0. 2. Vienna, Austria: R Foundation for Statistical Computing; 2013.

Scalercio, S., Cini, A., Menchetti, M., Vodă, R., Bonelli, S., Bordoni, A., … Vila, R. (2020) How long is 3 km for a butterfly? Ecological constraints and functional traits explain high mitochondrial genetic diversity between Sicily and the Italian Peninsula. *Journal of Animal Ecology*.

Schmitt, T. (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in zoology*, **4**, 11.

Schoener, T.W. (1968) The *Anolis* lizards of Bimini: resource partitioning in a complex fauna. *Ecology*, **49**, 704–726.

Schweiger, O., Harpke, A., Wiemers, M. & Settele, J. (2014) CLIMBER: Climatic niche characteristics of the butterflies in Europe. *ZooKeys*, 65.

Schweiger, O., Heikkinen, R.K., Harpke, A., Hickler, T., Klotz, S., Kudrna, O., … Settele, J. (2012) Increasing range mismatching of interacting species under global change is related to their ecological characteristics. *Global Ecology and Biogeography*, **21**, 88–99.

Settele, J., Kudrna, O., Harpke, A., Kühn, I., Van Swaay, C., Verovnik, R., … Hickler, T. (2008) *Climatic risk atlas of European butterflies*, Pensoft Sofia.

Talavera, G., Lukhtanov, V.A., Rieppel, L., Pierce, N.E. & Vila, R. (2013) In the shadow of phylogenetic uncertainty: the recent diversification of *Lysandra* butterflies through chromosomal change. *Molecular Phylogenetics and Evolution*, **69**, 469–478.

Thuiller, W., Georges, D., Engler, R., Breiner, F., Georges, M.D. & Thuiller, C.W. (2016) Package ‘biomod2’. Species distribution modeling within an ensemble forecasting framework. *Ecography*, **32**, 369–373.

Titeux, N., Maes, D., Van Daele, T., Onkelinx, T., Heikkinen, R.K., Romo, H., … van Swaay, C.A.M. (2017) The need for large-scale distribution data to estimate regional changes in species richness under future climate change. *Diversity and Distributions*, **23**, 1393–1407.

Toews, D.P.L., Mandic, M., Richards, J.G. & Irwin, D.E. (2014) Migration, mitochondria, and the yellow-rumped warbler. *Evolution*, **68**, 241–255.

Vodă, R., Dapporto, L., Dincă, V. & Vila, R. (2015) Cryptic matters: overlooked species generate most butterfly beta-diversity. *Ecography*, **38**, 405–409.

Warren, D.L., Glor, R.E. & Turelli, M. (2008) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution: International Journal of Organic Evolution*, **62**, 2868–2883.

Waters, J.M. (2011) Competitive exclusion: phylogeography's 'elephant in the room'? *Molecular Ecology*, **20**, 4388–4394.

Wiemers, M., Balletto, E., Dincă, V., Fric, Z.F., Lamas, G., Lukhtanov, V., … Verovnik, R. (2018) An updated checklist of the European Butterflies (Lepidoptera, Papilionoidea). *ZooKeys*, **811**, 9–45.

Wiemers, M. & Gottsberger, B. (2010) Discordant patterns of mitochondrial and nuclear differentiation in the Scarce Swallowtail *Iphiclides podalirius feisthamelii* (Duponchel, 1832)(Lepidoptera: Papilionidae). *Entomologische Zeitschrift*, **120**, 111–115.

Zinetti, F., Dapporto, L., Vovlas, A., Chelazzi, G., Bonelli, S., Balletto, E. & Ciofi, C. (2013) When the rule becomes the exception. No evidence of gene flow between two *Zerynthia* cryptic butterflies suggests the emergence of a new model group. *PLoS ONE*, **8**.

Zografou, K., Kati, V., Grill, A., Wilson, R.J., Tzirkalli, E., Pamperis, L.N. & Halley, J.M. (2014) Signals of climate change in butterfly communities in a Mediterranean protected area. *PLoS ONE*, **9**.

Data Accessibility Statement:

Occurrence data and R scripts to replicate the analysis are available in Dryad https://doi.org/10.5061/dryad.hmgqnk9dh.

Newly generated COI data are available in the “DS-WEUP” BOLD project.

The biodecrypt functions are available in the recluster R package and in https://cran.r-project.org/web/packages/recluster/index.html and in Dryad https://doi.org/10.5061/dryad.hmgqnk9dh.

Table 1. Results of cross-validation, parameters estimated by optimization and range overlap.

Species	MIR	NIR	NUR	Alpha	Buffer	Ratio	Overlap%
S. sertorius/rosae	1.9	41.2	18.2	11.0	96.1	2.3	15.6
P. malvae/malvoides	0.0	33.6	7.6	9.7	68.7	2.4	2.3
C. alceae/tripolinus	0.5	35.2	5.0	7.3	92.4	2.4	4.4
Z. polyxena/cassandra	2.0	12.0	7.6	5.6	35.3	2.3	0.1
I. podalirius/feisthamelii	2.0	9.1	6.2	7.7	39.0	2.1	0.2
P. daplidice/edusa	0.6	4.3	3.9	9.0	50.5	2.6	0.0
L. sinapis/reali/juvernica	1.9	84.7	71.1	8.5	52.5	3.2	Lr-Lj 0.0 Lr-Ls 4.1 Ls-Lj 50.1
L. tityrus/bleusei	0.0	2.7	2.0	9.2	74.0	2.9	0.0
I. iolas/debilitata	0.0	4.5	1.5	7.6	80.5	2.4	0.0
P. icarus/celina	0.7	11.4	3.1	8.5	52.5	3.2	1.9
L. coridon/caelestissima	0.0	1.2	0.3	8.9	76.4	2.7	0.0
M. athalia/celadussa	1.9	10.2	7.2	10.4	31.7	2.7	1.0
A. urticae/ichnusa	NA	NA	0.2	8.0	60.0	2.5	0.0
P. anthelea/amalthea	NA	NA	20.0	8.0	60.0	2.5	0.0

Table 1. Misidentified identified records (MIR), non-attributed identified records (NIR) and non-attributed non-identified records (NUR) obtained for each species after optimization of alpha, buffer and distance ratio parameters. The parameters have been estimated with series of cross-validation analysis. The alpha, buffer and ratio of the solutions showing a penalty (MIR^2+NIR+NUR) not higher than 10 compared to the model with lowest penalty have been averaged weighted by their inverse penalty by biodecrypt.optimise. The percentage of range overlap among hulls in the solution obtained by using these parameters is also reported. For *Leptidea*, the pairwise overlap values between *L. sinapis* (Ls), *L. reali* (Lr) and *L. juvernica* (Lj) are provided.

Table 2. Generalized Additive Mixed Model (GAMM) results for the influence of the three parameters on record attribution.

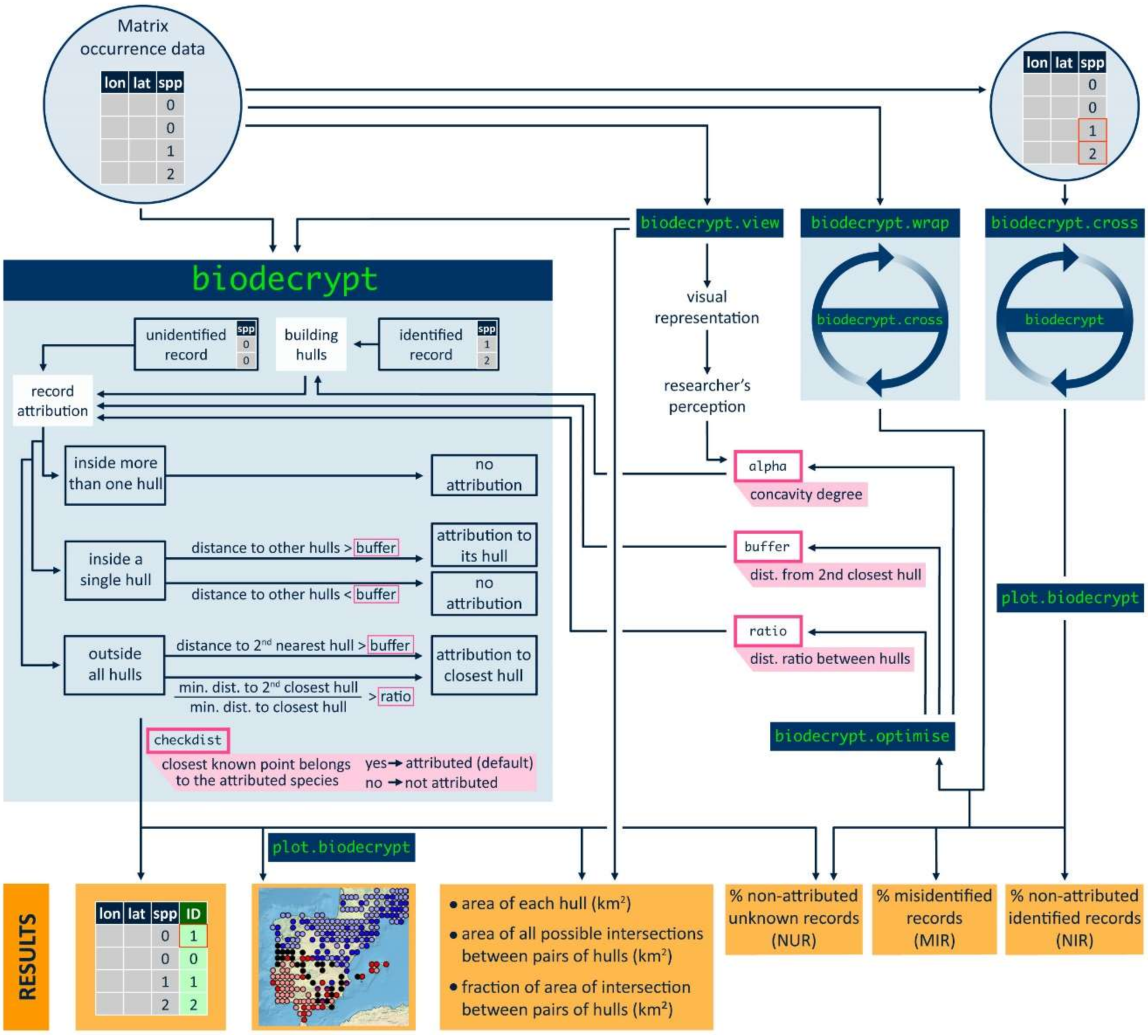
	MIR			NIR			NUR		
	edf	F	P	edf	F	P	edf	F	P
Ratio	1	0.486	0.486	1	219.2	<0.001	1.168	933.36	<0.001
Buffer	1.992	430.06	<0.001	1.546	832.8	<0.001	1.825	186.48	<0.001
Alpha	1.611	16.972	<0.001	1.54	11.7	<0.001	1.967	95.09	<0.001

Table 2. The effect of ratio, buffer and alpha of misidentified records (MIR), non-attributed identified records (NIR) and non-attributed unidentified records (NUR) as verified by GAMMs (edf, effective degrees of freedom, F and P values are provided).

Table 3. Tests for niche equivalency, divergence and conservatism.

Group	D	I	ECD P	ECI P	EDD P	EDI P	SCD P	SCI P	SDD P	SDI P
C. alceae/tripolinus	0.179	0.367	1.000	1.000	**0.010**	**0.010**	0.069	0.089	0.881	0.842
C. sertorius/rosae	0.155	0.376	1.000	1.000	**0.010**	**0.010**	**0.030**	**0.030**	0.950	0.960
P. malvae/malvoides	0.252	0.429	1.000	1.000	**0.010**	**0.010**	0.178	0.208	0.772	0.762
I. podalirius/feisthamelii	0.351	0.485	1.000	1.000	**0.010**	**0.010**	0.248	0.267	0.752	0.743
Z. polyxena/cassandra	0.341	0.538	1.000	1.000	**0.010**	**0.010**	0.059	0.069	0.931	0.931
P. daplidice/edusa	0.481	0.608	1.000	1.000	**0.010**	**0.010**	0.079	0.119	0.960	0.931
L. sinapis/reali	0.274	0.507	1.000	1.000	**0.020**	**0.030**	0.069	**0.050**	0.970	0.970
L. sinapis/juvernica	0.266	0.491	1.000	1.000	**0.010**	**0.010**	0.129	0.079	0.782	0.861
L. reali/juvernica	0.179	0.303	1.000	1.000	**0.010**	**0.010**	0.208	0.178	0.832	0.861
L. tityrus/bleusei	0.131	0.326	0.990	1.000	**0.010**	**0.010**	0.168	0.158	0.812	0.832
I. iolas/debilitata	0.255	0.429	1.000	1.000	**0.010**	**0.010**	0.149	0.168	0.901	0.871
P. icarus/celina	0.115	0.294	1.000	1.000	**0.010**	**0.010**	0.347	0.347	0.634	0.653
L. coridon/caelestissima	0.012	0.104	1.000	1.000	**0.010**	**0.010**	0.238	0.248	0.703	0.693
M. athalia/celadussa	0.271	0.429	1.000	1.000	**0.010**	**0.010**	0.257	0.317	0.723	0.703
A. urticae/ichnusa	0.054	0.226	1.000	1.000	**0.010**	**0.010**	0.267	0.238	0.723	0.812
P. anthelea/amalthea	0.060	0.120	1.000	1.000	**0.010**	**0.010**	0.059	0.277	0.941	0.752

Table 3. Niche overlap metrics Shoener's D (D) and the modified Hellinger metric I (I) calculated in multivariate climatic space for pairs of the studied taxa. Four tests have been performed per metric and respective P values are provided for conservatism based on niche equivalency for D and I (ECD P; ECI P), divergence based on niche equivalency (EDD P; EDI P), conservatism based on niche similarity (SCD P, SCI P) and divergence based on niche similarity (SDD P, SDI P). Significant values are indicated in bold.


Matrix
occurrence data
biodecrypt.view
biodecrypt.wrap
biodecrypt.cross
biodecrypt
unidentified record
building hulls
identified record
record attribution
inside more than one hull
no attribution
inside a single hull
distance to other hulls > buffer
attribution to its hull
distance to other hulls < buffer
no attribution
outside all hulls
distance to 2nd nearest hull > buffer
attribution to closest hull
min. dist. to 2nd closest hull
min. dist. to closest hull
> ratio
checkdist
closest known point belongs to the attributed species
yes → attributed (default)
no → not attributed
visual representation
researcher's perception
alpha
concavity degree
buffer
dist. from 2nd closest hull
ratio
dist. ratio between hulls
biodecrypt.cross
biodecrypt
plot.biodecrypt
biodecrypt.optimise
plot.biodecrypt
RESULTS
• area of each hull (km²)
• area of all possible intersections between pairs of hulls (km²)
• fraction of area of intersection between pairs of hulls (km²)
% non-attributed unknown records (NUR)
% misidentified records (MIR)
% non-attributed identified records (NIR)

lon	lat	spp
		0
		0
		1
		2

lon	lat	spp	ID
		0	1
		0	0
		1	1
		2	2

Figure 1. A diagram showing the workflow of the six functions of the biodecrypt family. Function names are indicated as blue boxes. Circles with arrows indicate wrapping of a function inside another one. Options/Parameters are indicated as pink boxes and a short description is given. Results are shown in orange boxes.

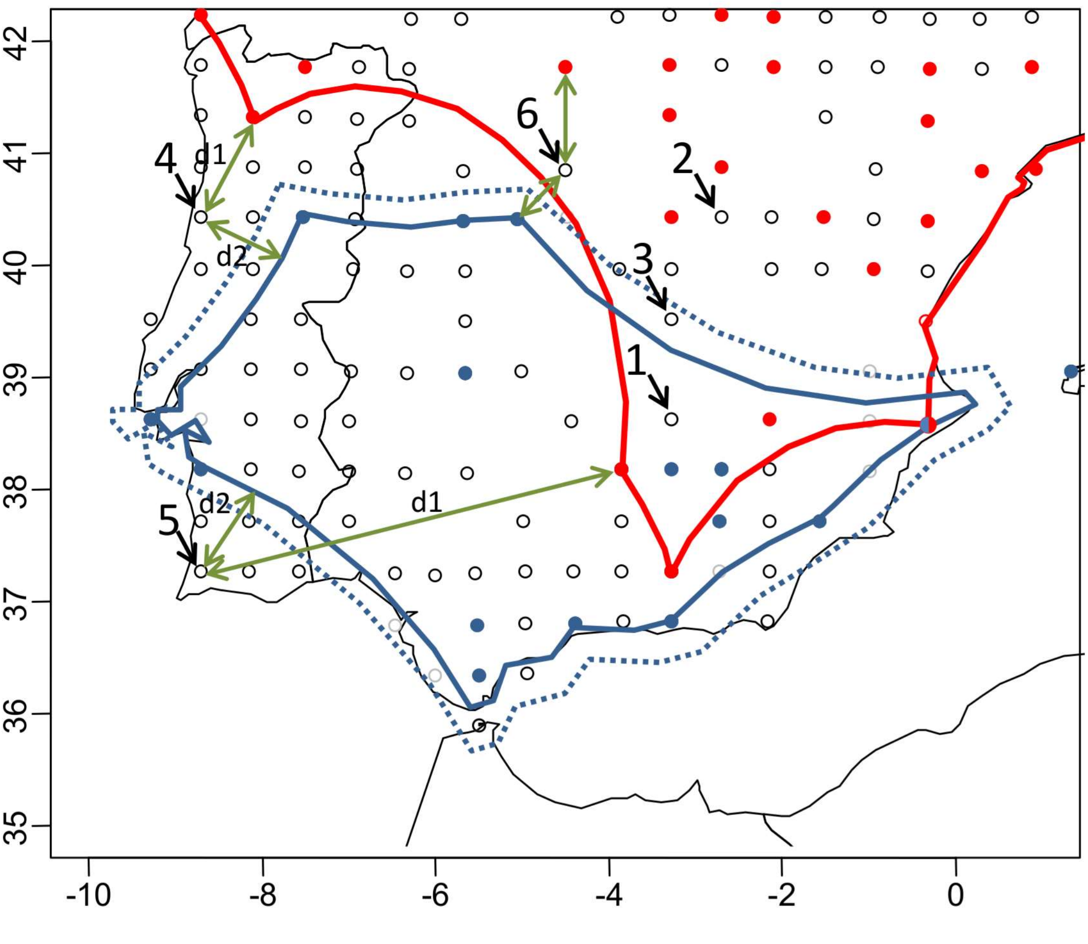


Figure 2. An example of the assignment procedure based on the overlapping distributions of *Polyommatus icarus* (red) and *P. celina* (blue) in Iberia. Red and blue dots represent the sites from where specimens of the respective species have been sequenced, empty circles represent sites with unidentified records. The continuous red and blue lines represent the hulls obtained for the two species based on occurrence of sequenced specimens (alpha = 3 in this example compared to the optimized value of 8.5 to show the effect of a low alpha value on polygon shape), the dotted blue line represents the buffer of the *P. celina* hull (for clarity, the buffer of the *P. icarus* hull is not represented). Unidentified record 1 is not attributed since it falls inside both hulls; record 2 is attributed to *P. icarus*, record 3 is not attributed since it falls in the buffer of the *P. celina* hull; record 4 (external to both hulls) is not attributed because the distance ratio (d1/d2; where d1 represents the larger distance) is smaller than the default ratio

of 2.5; while record 5 is attributed to *P. celina* because d1/d2 is larger than 2.5 (default in biodecrypt). If checkdist is selected, point 6 is not attributed to *P. icarus* because the closest known record is of *P. celina*.

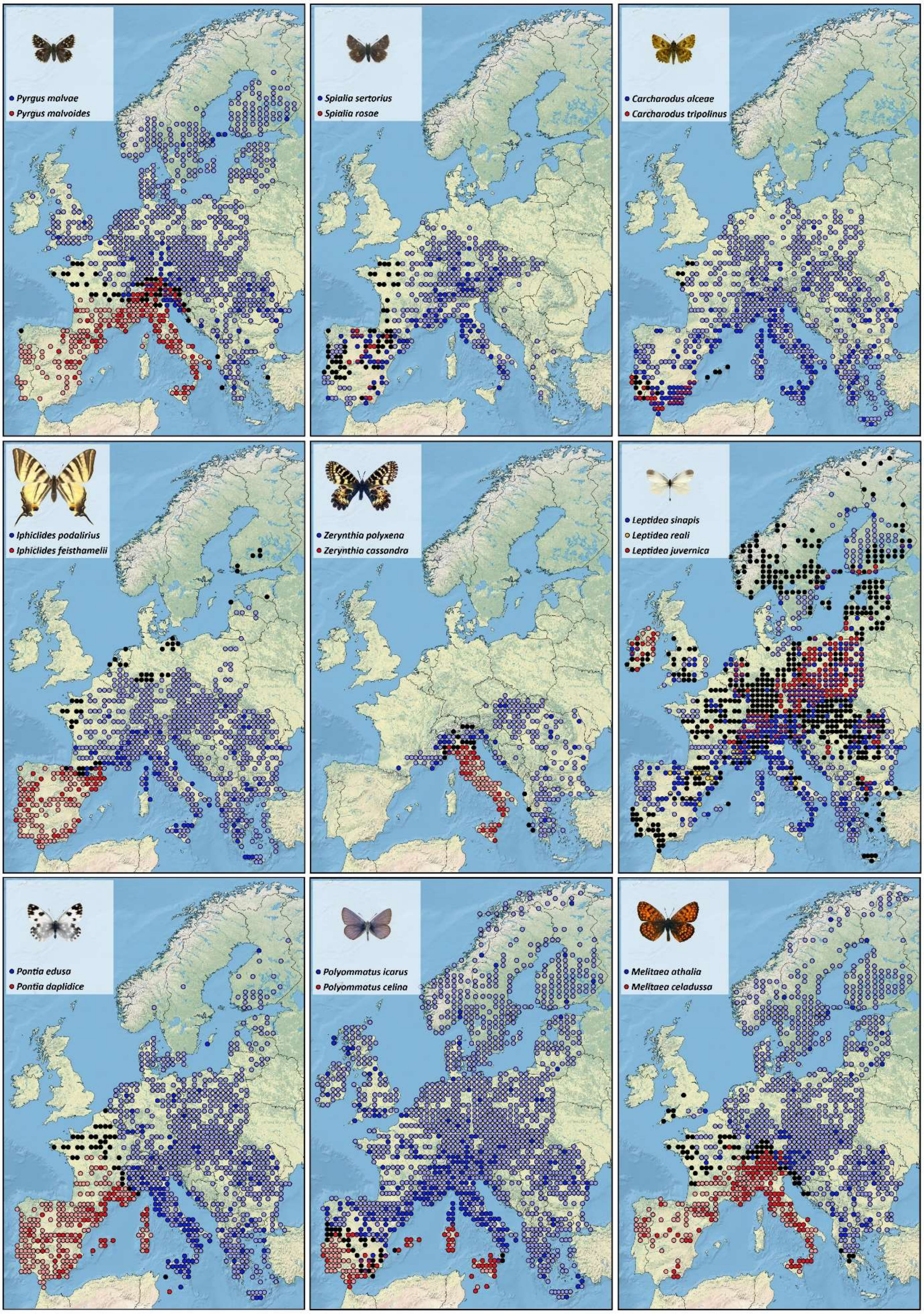


Figure 3. Distribution maps obtained for nine groups of cryptic species by using the optimised parameters reported in Table 1. Dots with darker colour represent records

identified based on DNA sequences (mostly mitochondrial), genitalic morphology or other markers. Circles with paler colours represent records attributed by the algorithm. Black circles are non-attributed records. The physical map used is freely available from Natural Earth (www.naturalearthdata.com).

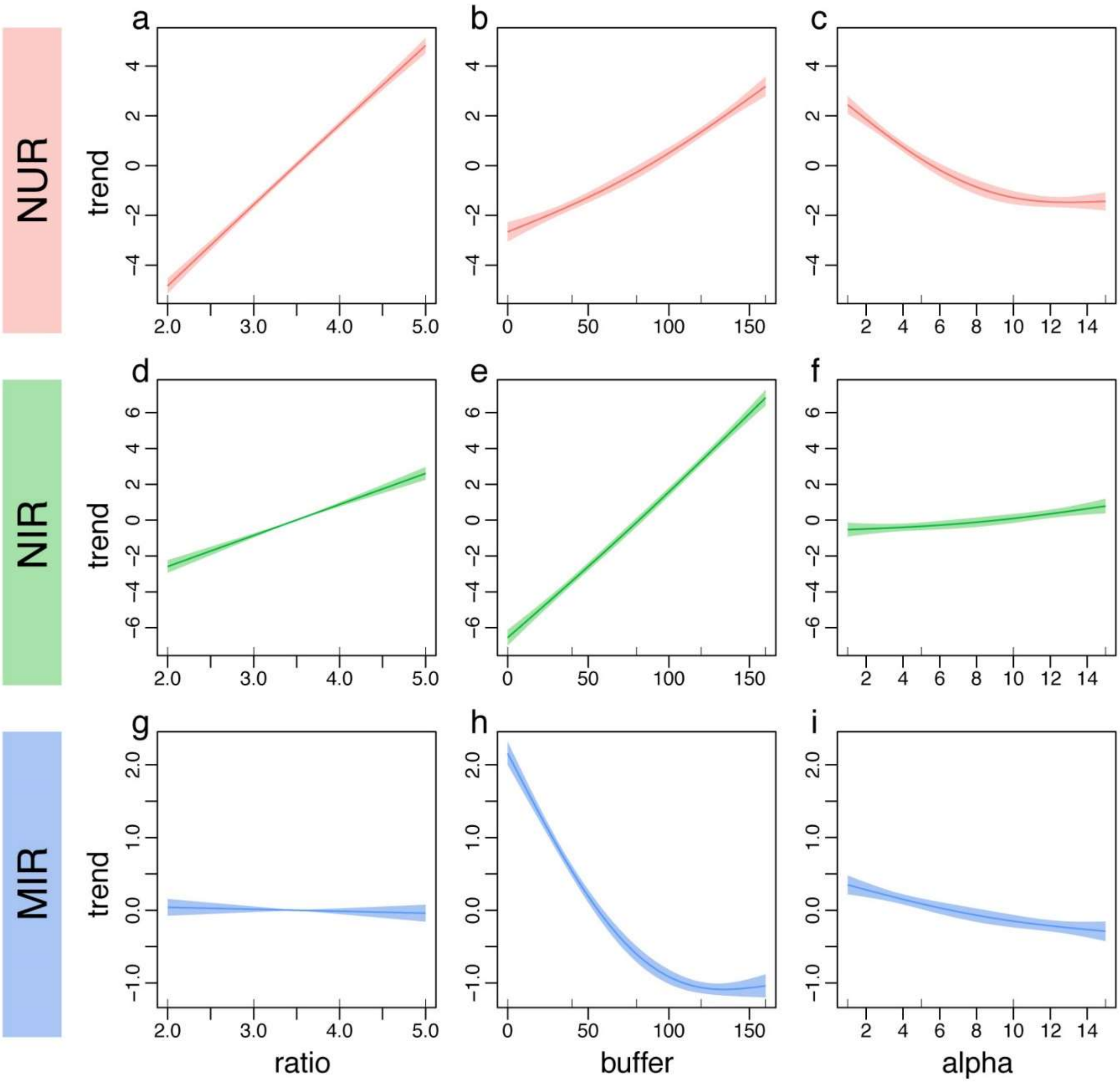


Figure 4. The standardized effect (trends) of ratio, buffer and alpha obtained after Generalized Additive Mixed Model on non-attributed unknown records (NUR), non-attributed known records (NIR) and misidentified known records (MIR). Data belongs to series of biodecrypt.wrap analyses.

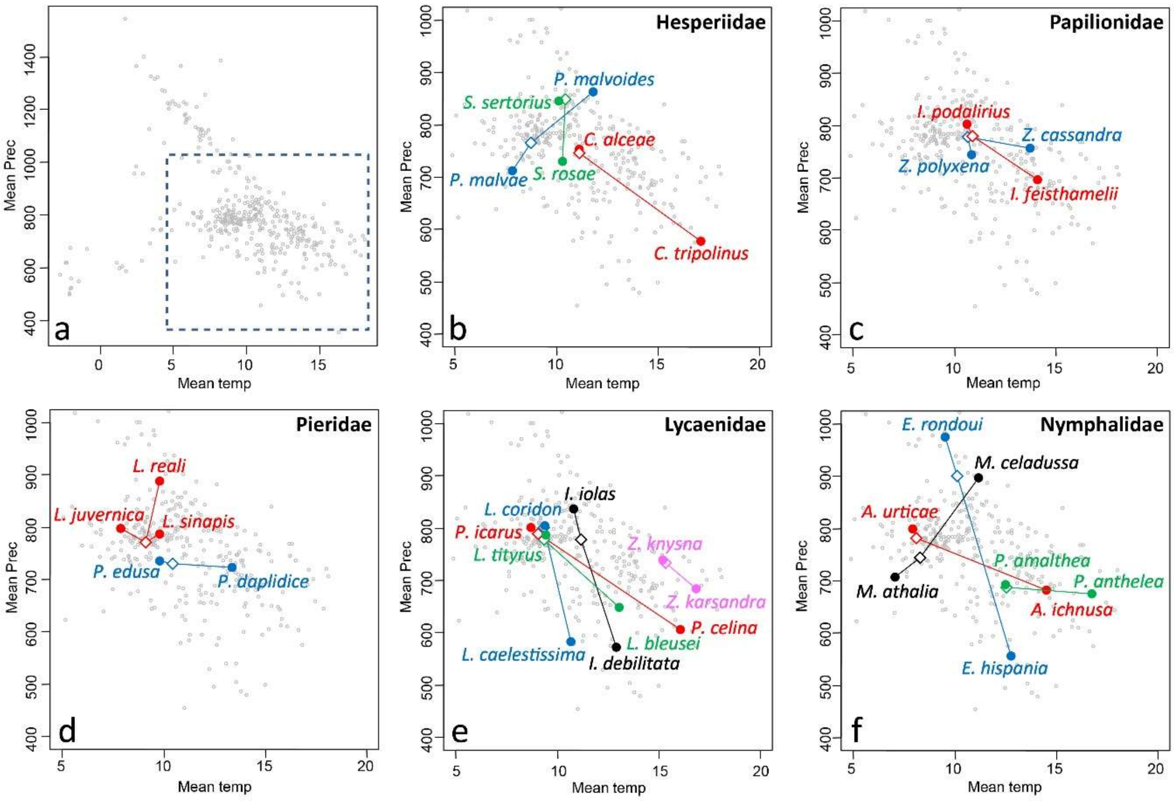


Figure 5. Climatic space defined by mean annual temperature and annual precipitation for all European species included in the CLIMBER dataset (a). The blue rectangle represents the climate space where the species under study are located and which is used for figures b-f. For the five families, the resulting values for the cryptic taxa separated in this study are represented by filled dots connected with a line to the former values of the amalgamated taxon (diamonds). Each cryptic complex within a family is represented by a different colour (b-f) and separated cryptic species names are reported.