Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014 doi: 10.1111/2041-210X.12220

Comparing population patterns for genetic and morphological markers with uneven sample sizes. An example for the butterfly *Maniola jurtina*

Leonardo Dapporto[1]*,†, Raluca Vodă[2,3],†, Vlad Dincă[2,4,5] and Roger Vila[2]

[1]Department of Biological and Medical Sciences, Oxford Brookes University, Headington, Oxford OX3 0BP, UK; [2]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37, Barcelona 08003, Spain; [3]Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain; [4]Department of Zoology, Stockholm University, Stockholm, S-106 91, Sweden; and [5]Biodiversity Institute of Ontario, University of Guelph, Guelph, N1G 2W1 ON, Canada

Summary

1. Integrating genetic and/or phenotypic traits at population level is considered a fundamental approach in the study of evolutionary processes, systematics, biogeography and conservation. But combining the two types of data remain a complex task, mostly due to the high, and sometimes different, sample sizes required for reliable assessments of community traits. Data availability has been increasing in recent years, thanks to online resources, but it is uncommon that different types of markers are available for any given specimen.

2. We provide new R functions aimed at directly correlating traits at population level, even if data sets only overlap partially. The new functions are based on a modified Procrustes algorithm that minimizes differences between bidimensional ordinations of two different markers, based on a subsample of specimens for which both characters are known. To test the new functions, we used a molecular and morphological data set comprising Mediterranean specimens of the butterfly *Maniola jurtina*.

3. By using this method, we have been able to maximize similarities between genotypic and phenotypic configurations obtained after principal coordinate analysis for the model species and evaluated their degree of correlation at both individual and population level. The new recluster.procrustes function retained the information of the relative importance of different morphological variables in determining the observed ordinations and preserved it in the transformed configurations. This allowed calculating the best combination of morphological variables mirroring genetic relationships among specimens and populations. Finally, it was possible to analyse the modality and variance of the phenotypic characters correlated with the genetic structure among populations.

4. The genetic and phenotypic markers displayed high overall correlation in the study area except in the contact zone, where discrepancies for particular populations were detected. Interestingly, such discrepancies were spatially structured, with southern populations displaying typical western morphotype and eastern haplotypes, while the opposite occurred in the northern populations. The methodology here described can be applied to any number and type of traits for which bidimensional configurations can be obtained, and opens new possibilities for data mining and for meta-analyses combining existing data sets in biogeography, systematics and ecology.

Key-words: Cytochrome c oxidase subunit 1, geometric morphometrics, Lepidoptera, Mediterranean, *recluster* R package, recluster.procrustes function

Introduction

Integrating different types of data has become a usual procedure in the study of speciation, biogeography and conservation ecology (Pergams & Lacy 2008; Dincă, Dapporto & Vila 2011; Mila *et al.* 2011; Toews *et al.* 2014; Derryberry *et al.* in press). There is growing evidence that different types of markers can display contrasting spatial patterns (Dapporto *et al.* 2011; Toews & Brelsford 2012; Pavlova *et al.* 2013; Toews *et al.* 2014), and such discrepancies are hypothesized to encompass important evolutionary and population processes like local adaptation, character displacement and sex-biased asymmetries in hybridization processes (Toews & Brelsford 2012). Biogeographical, ecological and evolutionary studies typically combine different DNA sequences (usually nuclear and mitochondrial genes) or include both genetic and morphological traits to compare taxa or populations. While genetic patterns can reveal evolutionary relationships, hybridization processes and approximate divergence time, phenotypic traits represent

*Correspondence author. E-mail: leondap@gmail.com
†These authors equally contributed to this article.

the interface between genes and the environment and include key morphological traits for the interactions among individuals.

Nevertheless, comparing genetic and morphological patterns among populations in a spatial context is methodologically challenging because genetic and morphological analyses produce different raw data (e.g. DNA sequences vs. linear or shape measurements) and are usually examined with different techniques (Claude 2008; Paradis 2012). Moreover, while DNA sequencing follows highly standardized procedures producing directly comparable results, morphological assessments are affected by stochastic factors such as sample preparation, measurement errors and by a multitude of environmental events determining the observed individual's phenotype. As a consequence, the morphological assessment of populations usually requires a higher number of samples compared with genetic analyses. Yet, most of the existing methods for direct comparisons between different types of markers only allow correlations on specimens analysed for both characters. Importantly, the difficulty of directly comparing data sets that do not fully overlap (i.e. with different numbers of samples assessed for each marker) prevents a full use of data available in public repositories (e.g. GenBank, BOLD systems), where typically only one of the markers is available for any given specimen.

Even when an adequate sample size is gathered, the analyses for the spatial distribution of morphological and genetic markers among populations are often carried out independently (Pergams & Lacy 2008; Dincă, Dapporto & Vila 2011; Seraphim *et al.* 2013; Toews *et al.* 2014). Direct comparisons are complicated by the different approaches used to retrieve genetic and morphological patterns. Genetic patterns are usually represented by phylogenetic trees or haplotype networks, which are postulated to mirror the evolutionary relationships among taxa (Paradis 2012). However, although phylogenetic trees can be conveniently used to order taxa, they can be inefficient to sort hybrid individuals or to analyse the overall genetic structure of mixed populations (Kalinowski 2009). In these cases, less constrained representations avoiding bi- or polytomic ordinations should be preferred (Kalinowski 2009). Such continuous patterns for genetic data can be obtained by ordination methods, like multidimensional scaling, principal coordinates analysis, phylogenetic PCA (pPCA), etc., which reduce the dimensionality of any dissimilarity matrix obtained by different markers such as DNA sequences, microsatellites or allozymes (Pritchard, Stephens & Donnelly 2000; Paradis 2012). Interestingly, similar algorithms are the standard methods used to visualize patterns of variation in continuous morphological traits (Claude 2008). Such parallel approaches provide the basis for a direct comparison of genetic and morphological patterns. Direct correlations are usually tested through the overall correlation coefficients between original distance matrices or configurations obtained, for example, by means of Mantel tests or protest analyses (Renvoise *et al.* 2012). These analyses evaluate the degree of correlation between matrices and calculate the associated *P* value. However, they do not allow for: (i) the identification of particularly discrepant individuals or populations, (ii) the analysis of unimodal or bimodal distributions in phenotypic traits as associated with genetic ones and (iii) the representation of the variation of these characteristics in the geographical space.

Such analyses have been carried out only in a few studies that mainly compared single components (usually belonging to PCA analyses) of genetic and morphological variation between them and with environmental determinants (Gompert *et al.* 2010).

In this article, we developed new functions for the *recluster* R package that facilitate in-depth comparisons of bidimensional configurations (genotypic and phenotypic markers), even when only a limited series of specimens is analysed for the two traits. The main new function recluster.procrustes is a modified Procrustes algorithm which allows indicating the subset of specimens that overlap for the two data sets. The function minimizes the differences between two configurations by applying a classical Procrustes on individuals for which both traits have been assessed, and subsequently applies the same transformation to the specimens that were not analysed for both markers. Importantly, the new functions allow entering the coordinates of the variables determining the observed pattern (as it is usually obtained by PCA or MDS), and also applies the same transformation to them to maintain the information regarding their contribution.

To show the characteristics of the analysis and to provide a practical example, we tested the new method on two data sets for the butterfly *Maniola jurtina* (Nymphalidae) covering the west Mediterranean area and eastern Europe. We examined if the COI mitochondrial gene and the morphological data reveal with precision the location and nature of the contact zone between the two main lineages described for this species. Discrepancies between allozyme and morphological patterns have been described in *M. jurtina* (Dapporto *et al.* 2011), suggesting that this species is a suitable model for our analysis. We show how the proposed analysis can identify and model correspondences and discrepancies in a spatial framework. Due to the increasing availability of similar data for a large number of organisms, such a theoretical and practical assessment can represent a useful model and resource for future studies.

Methods

SPECIES, MARKERS AND BACKGROUND

Maniola jurtina

The meadow brown butterfly *Maniola jurtina* (Linnaeus, 1758) is a species that has often been targeted for phylogeography and speciation studies. It forms conspicuous populations over Europe and the Mediterranean basin including many islands. Two lineages, identified on the basis of male genitalia and allozyme analysis, occur in Europe: a western Atlantic-Mediterranean lineage (*M. j. jurtina*) in the Maghreb, Spain, western France, Sicily and Sardinia and an eastern-Mediterranean-Asian lineage (*M. j. janira*) widespread from Asia to eastern and central Europe, including the Italian Peninsula (Schmitt, Rober & Seitz 2005; Dapporto *et al.* 2009; Thomson 2011). On the basis of allozyme data, the two lineages seem to have diverged during the late Pleistocene

(Schmitt, Rober & Seitz 2005) and may have experienced a series of range contraction/expansion cycles during the following glacial–interglacial periods (Schmitt, Rober & Seitz 2005; Dapporto *et al.* 2011; Thomson 2011). Morphological and allozyme data also revealed the presence of a contact zone extending from a few western Mediterranean islands (Corsica, Elba, Giglio and Capri) to the western Alps and the Benelux region (Schmitt, Rober & Seitz 2005; Dapporto *et al.* 2011; Thomson 2011; but see also Habel, Dieker & Schmitt 2009). Modelling of continuous morphological variation and discrepancies between morphological and allozyme data have been used to infer the phylogeography of this species over the western Mediterranean area, and led to the formulation of a new post-glacial colonization paradigm over Europe (Dapporto *et al.* 2009, 2011; Habel, Dieker & Schmitt 2009; Dapporto & Bruschini 2012). Despite the considerable number of studies dealing with patterns in morphology and allozymes for this species, no comprehensive genetic assessments have been published and no direct evidence exists for a correspondence between morphology and DNA. We identified 39 geographic areas, comprising 20 insular and 19 mainland areas (Table 1). All areas included at least two specimens analysed for both markers (see Appendix S2 for details).

Genetic markers

We analysed the COI mitochondrial gene because it is a widely used marker in systematics and phylogeography (Avise 2009) and extensive libraries of publicly available sequences exist (e.g. GenBank and BOLD systems). A total of 218 *M. jurtina* COI sequences were used in the analysis, of which 45 were obtained from publicly available data in BOLD and 173 have been sequenced specifically for this study (GenBank accession numbers KM020807–KM020882, KJ994239–KJ994253 and KM033847–KM033941; see Appendix S2 for sampling localities). Well-assessed protocols described in Appendix S2 were used for DNA extraction and sequencing. A dissimilarity p-distance matrix among COI sequences was obtained with default settings in MEGA 5.05. The ordination of genetic markers has been obtained by projecting the dissimilarity matrix into a two-dimensional configuration through a principal coordinate analysis (PCoA). A neighbour-joining tree was constructed in MEGA 5.05 using p-distance and assessing node supports by 100 bootstrap pseudoreplicates. One sequence of *Hyponephele lupina* was used as outgroup.

Morphological markers

We selected the shape of male genitalia as morphological marker. We analysed 616 male *M. jurtina* specimens by means of geometric morphometrics (Bookstein 1991), a method that produces relative warps (PCs) representing continuous variables of shape variations. Variables from different structures of the genitalia can be combined in successive ordination analyses to reveal the overall patterns of variation among specimens and the shape variation associated with such patterns. Geometric morphometrics is a powerful method for analysing morphological species traits in a wide range of organisms (Viscosi & Cardini 2011; Madeira *et al.* 2012; Zelditch, Swiderski & Sheets 2012). It has been successfully used as a quantitative method to distinguish western and eastern morphotypes, as well as intermediate (presumably hybrid) individuals and populations of *M. jurtina* (Dapporto *et al.* 2009; Dapporto & Bruschini 2012). In this species, two structures (valva and brachium, Appendix S2) are mostly involved in differentiating morphotypes (Dapporto *et al.* 2009; Thomson 2011). Moreover, the continuous variation of these markers has been used to model the distribution of morphotypes (Dapporto *et al.* 2009; Dapporto & Bruschini 2012). We thus conducted geometric morphometrics of brachium and valva following the methodology employed by Dapporto *et al.* (2012), which is specifically described in Appendix S2.

Table 1. Studied areas and their abbreviations

Number	Area	Abbreviation
1	Eastern Europe	E_Eur
2	Dolomites	Dolom
3	Central-northern Italy	N_Ita
4	Argentario	Arg
5	Giglio Island	Giglio
6	Pianosa Island	Pian
7	Elba Island	Elba
8	Piombino	Piomb
9	Northern Corsica	N_Cor
10	Southern Corsica	S_Cor
11	Northern Sardinia	N_Sar
12	Central Sardinia	C_Sar
13	Southern Sardinia	S_Sar
14	Ischia Island	Ischia
15	Capri Island	Capri
16	Sorrento Peninsula	Sorren
17	Southern Italian Peninsula	S_Ita
18	Aspromonte	Aspr
19	Eastern Sicily	E_Sic
20	Vulcano Island	Vulc
21	Lipari Island	Lipari
22	Western Sicily	W_Sic
23	Gozo Island	Gozo
24	Tunisia	Tun
25	Algeria	Alg
26	Morocco	Mor
27	Southern Iberia	S_Ibe
28	Central Iberia	C_Ibe
29	Northern Iberia	N_Ibe
30	North-eastern Iberia	NE_Ibe
31	Ibiza Island	Ibiza
32	Mallorca Island	Mall
33	Menorca Island	Men
34	Southern France	S_Fra
35	Levant Island	Levant
36	North-western Alps	NW_Alps
37	South-western Alps	SW_Alps
38	Liguria	Liguria
39	Switzerland	Suisse

Similarly to what has been carried out for the genetic markers, we used relative warps obtained for the brachium and valva analysis to perform a PCoA. The contribution of the variables to the configuration has been obtained by computing the weighted average scores of variables by using the wascores *vegan* function. We retained the two most important variables for each of the two axes used.

THE NEW RECLUSTER.PROCRUSTES FUNCTION

The *recluster* package was originally created to solve a bias in the application of cluster analysis to turnover dissimilarity matrices (Dapporto *et al.* 2013). Subsequently, the package has been improved to allow the analysis of diversity patterns in a spatial context. This version of the *recluster* package enables coupling tree-based results with two-dimensional representations in RGB colour space, so that the nearest points in the bivariate configuration are represented with similar colours. This approach is becoming increasingly used in biogeography studies (Kreft & Jetz 2010; Holt *et al.* 2013).

The core of the proposed analysis is an algorithm maximizing similarities among bivariate configurations (in this case based on genotypic and phenotypic patterns). Such configurations can be obtained through any method for phenotypic traits (a crude combination of two measurements, the two main relative warps of a geometric morphometrics output, scatterplots of PCA, PCoA, non-metric multidimensional scaling (MDS), etc.). For genotypic data, bidimensional configurations can be obtained, for example, by multidimensional scaling or PCoA of any genetic distance matrix and by pPCA.

Some of these analyses (e.g. PCA, PCoA and MDS) not only order the specimens, but also provide the contribution of the variables to the obtained patterns (Legendre & Legendre 1998). This is much more common for morphological data where patterns are usually established by ordination of a multidimensional character matrix.

Two similar bidimensional configurations can appear highly different due to relative rotation, orientation and scaling (Appendix S1). The classic tools used to maximize similarity between configurations is the Procrustes analysis, which scales, flips and rotates a configuration to maximize its similarity to another one (Mardia, Kent & Bibby 1979). The *vegan* R package provides a Procrustes function to compute this analysis. However, it has two characteristics representing problematic limitations for our purposes:

1 The Procrustes *vegan* function only works when the specimens occurring in the two configurations are exactly the same.

2 The *vegan* Procrustes cannot maintain the correspondence between the rotated and rescaled configurations for specimens and the ordination of the variables.

The Procrustes analysis provides a rigid transformation of the second configuration. In theory, the second matrix can be thus transformed to maximally fit the first one on the basis of a subsample of corresponding specimens. Then, the same solid transformation can be applied to remnant (not shared) specimens, as well as to the coordinates of variables in the configuration to maintain the information about their contribution in the observed pattern. For this purpose, we created the recluster.procrustes function by modifying the Procrustes function of *vegan*. The recluster.procrustes function allows the user to indicate the number of common specimens, which must be listed first and in the same order in the two matrices. Moreover, it is possible to include other matrices containing variable contributions (coordinates) for each marker. An in-depth explanation of the recluster.procrustes function is given in Appendix S1.

Analyses

ORIENTING THE FIRST AXIS OF THE GENETIC CONFIGURATION ACCORDING TO A PHYLOGENETIC TREE

Using the recluster.group.col function, we computed the mean positions in the genetic PCoA configuration of specimens belonging to each of the two main clades in the phylogenetic tree (Fig. 1a). In the resulting matrix, all the specimens belonging to the same clade were collapsed to their barycentre in the PCoA configuration (Appendix S1). Subsequently, we created two functions: recluster.line and recluster.rotate. The first identifies the line connecting the most distant points in a configuration and computes its intercept and angular coefficient; the second rotates the points of a configuration to a new configuration where a line identified by its intercept and its angular coefficient is rotated to become horizontal. In practice, by using these functions in series, the specimens belonging to the two most diverging clades are polarized along the *x*-axis by maintaining the original configuration among individuals.

The first Procrustes analysis re-aligned the PCoA configuration obtained by genetic distances with the configuration of polarized clades. In such a way, the phylogenetic tree information is retained in the subsequent analyses as a horizontal ordination along a main genetic *x*-axis.

MINIMIZING DISSIMILARITIES BETWEEN CONFIGURATIONS AND COMPUTING POPULATION MEANS

We performed a second recluster.procrustes transformation of the morphological PCoA configuration based on the previously obtained

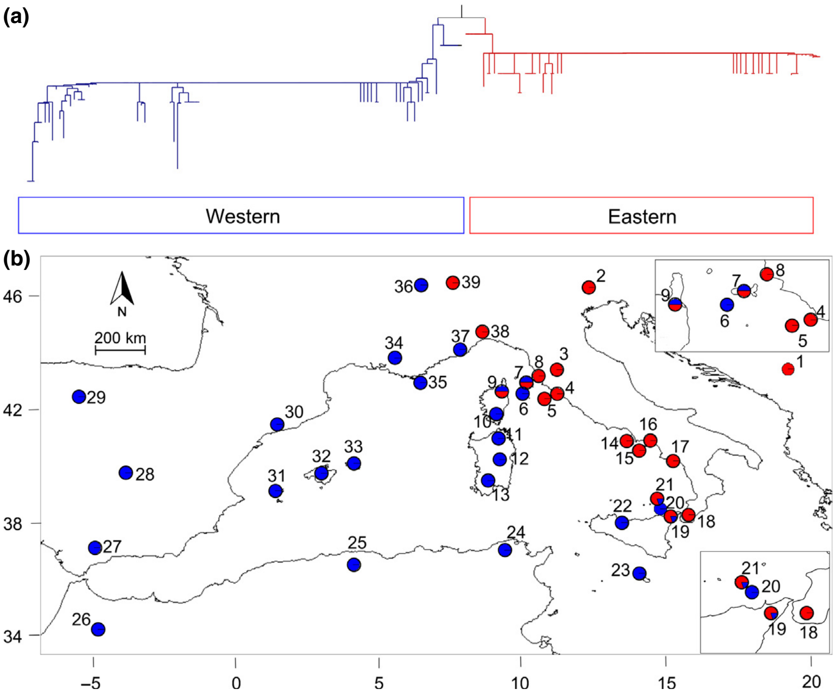

Fig. 1. (a) Structure of the neighbour-joining tree based on 218 COI sequences for *Maniola jurtina* confirming the existence of two main western and eastern clades (see Appendix S2 for a detailed representation). (b) Map of the study area indicating the proportion of individuals belonging to the western (blue) or eastern (red) COI lineages for sampled populations. Enlarged representations for the critical areas of Tuscan and Aeolian Islands are provided. Numbers correspond to areas listed in Table 1.

genetic one. Thus, the initial information fitting the phylogenetic variation along the *x*-axis was directly transferred to the resulting morphological configuration. The degree and significance of the correlation between the configuration of genetic and morphological data on the subset of shared specimens was evaluated by using the *vegan* protest function.

To study the genetic and morphological patterns at population level, we computed barycentre of specimens belonging to the same area in the two configurations. The protest function was applied to mean population coordinates to verify concordance between genetic and morphological configurations.

ASSESSING DISSIMILARITY LOSS AFTER COMPUTING POPULATION MEANS AND TESTING ITS SIGNIFICANCE

Because the two genetic PCoA axes were oriented to maximize differences among the most diverging phylogenetic clades along the first axis, we analysed results separately on the first and second axes. As a first step, we evaluated the amount of configuration variance lost along both axes by grouping specimens according to their population barycentre. If the distribution of specimens in the configuration is randomly scattered among populations, almost all population barycentres are expected to attain a rather central position with respect to the original individual points, which would result in a small mean distance between barycentres. The new recluster.test.dist function produces a ratio between the mean squared pairwise distance for all individuals and the mean squared pairwise distance for population barycentres (Bookstein 1991). This ratio is calculated for the overall configuration and for the two axes separately. Moreover, this function provides a test for the significance of the variation preserved by population means. To do this, 1000 matrices were obtained by randomly sampling the original vector defining population membership for each specimen. Then, we computed the frequency of mean squared distance ratios in random configurations higher than the observed ratio.

PROJECTING THE ROTATED CONFIGURATIONS IN RGB COLOUR SPACE AND COMPUTING POPULATION MEANS

We projected the two configurations together in the RGB space. For this purpose, the axis with the highest variance was standardized between 0 and 1 and the variance along the second axis rescaled according to the first one. Subsequently, the colours blue, green, yellow and red were assigned to the four corners. Finally, the contribution of each RGB colour to each site has been calculated on the basis of its position in the two-dimensional graph. This method is receiving an increasing interest in biogeography where it is employed to visually depict similarity patterns among elements (Kreft & Jetz 2010; Holt *et al.* 2013). We projected the colours for populations on a map, separately for genetic and morphological configurations.

ANALYSING MODALITY FOR AXIS VALUES IN POPULATIONS, REGRESSING THE GENETIC AND MORPHOLOGICAL POPULATION VALUES AND DEPICTING SPATIAL DISTRIBUTION OF RESIDUALS

We applied Hartigan's dip test of unimodality (Hartigan & Hartigan 1985) by using the R package *diptest* to verify whether the distribution along the axes significantly deviates from the unimodal distribution in all populations. Variance has been computed to verify whether populations hosting intermediate individuals show higher variance in shape. For this purpose, we made quadratic regressions for populations using mean values in morphological axes as predictors and their variance in the same axes as dependent variables. Finally, we regressed the mean values of morphological axes against the mean values of genetic axes and computed the residuals. In this case, we used major axis regression (MA) since it is specifically designed for minimizing errors for both variables. Thus, the existence of a dependent and independent variable is not presumed and the residuals reflect discrepancies in both genotypic and phenotypic traits (Claude 2008). Finally, we plotted these residuals on a map. To obtain a more conservative representation, we set to zero the residuals within the interval of standard deviation of absolute residual values.

Results

THE GENETIC MARKER

The 218 COI sequences obtained for *M. jurtina* represented 55 different haplotypes. The neighbour-joining phylogenetic tree supported the existence of two main clades largely matching the distribution of western and eastern populations previously recognized based on morphological and allozyme data (Dapporto *et al.* 2011; Thomson 2011) (Fig. 1a). In some areas located at contact zones, the two genetic lineages were found to coexist (Fig. 1b). A PCoA based on genetic distances confirmed a polarization between eastern and western lineages together with a higher differentiation among the haplotypes belonging to the western group (Fig. 2a, Appendix S1). The mean position in the PCoA configuration for the specimens belonging to each of the two clades was computed and the two points rotated to be aligned with the *x*-axis (Appendix S1). Subsequently, a second Procrustes aligned the original genetic configuration with this phylogenetic configuration of clades (see Appendix S1 for a step-by-step guide and for plots of intermediate configurations).

THE MORPHOLOGICAL MARKERS

We analysed 616 male specimens and obtained 12 and 42 relative warps from the brachium and valva analysis, respectively. A PCoA analysis based on these 53 variables revealed that most of the variance along both PCoA axes is explained by variation in the first PC of the valva and the brachium (as expected, since they explain most of the variance in their respective analyses, Appendix S1).

MINIMIZING DISSIMILARITIES BETWEEN CONFIGURATIONS

A series of 169 specimens was analysed for both genetic and morphological markers. A preliminary protest analysis on the rotated genetic and morphological PCoA configurations revealed that they are highly correlated (correlation coefficient 0·612, $P < 0{\cdot}001$). It should be noted that a higher correlation coefficient would be difficult to obtain due to the strict qualitative nature of COI. This resulted in high residuals for morphologically intermediate (potentially

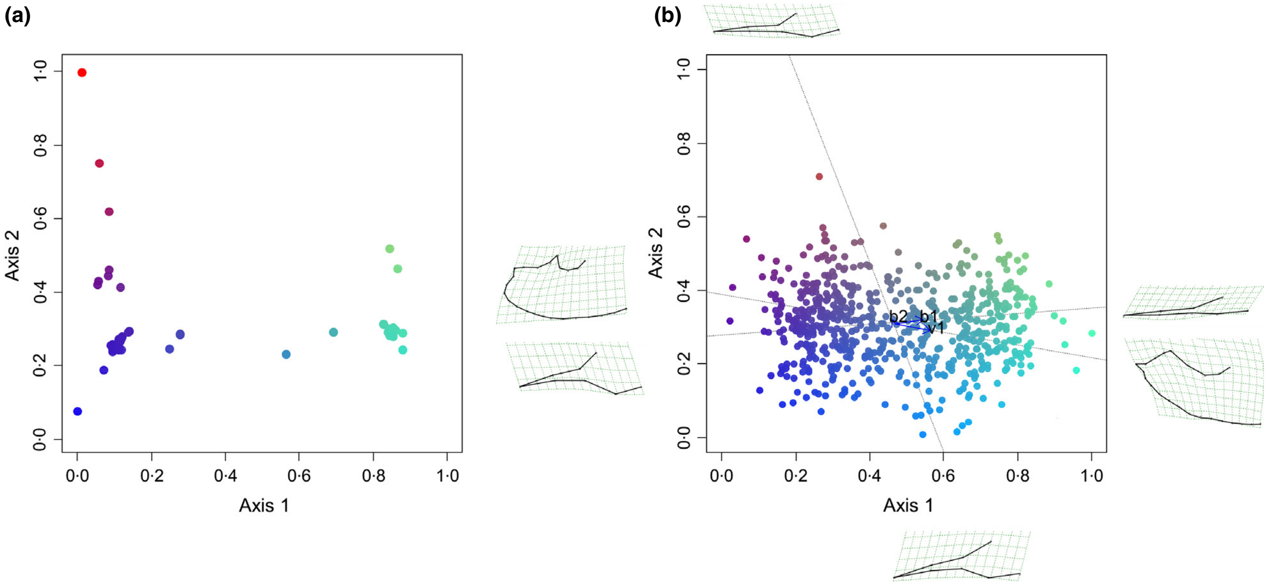

Fig. 2. (a) Principal coordinate analysis (PCoA) of genetic distance for sequenced specimens projected in the RGB colour space and rotated to maximize concordance with the existence of two phylogenetic clades over the *x*-axis (left, western lineage; right, eastern lineage). Specimens with identical haplotypes visually overlap and appear as a single dot. (b) PCoA of morphological data projected in the RGB colour space. Each point represents an individual with studied morphology. The arrows indicate the importance (length) and direction of the three main shape principal components (first PC for valva, v1, and first and second PCs for brachium, b1 and b2) in determining the configurations. The thin plate splines represent shape variation along the direction of main variables.

hybrid) specimens, which cannot have a corresponding intermediate genetic characteristic (see Appendix S1 for a graph). Subsequently, the morphological configuration with its variable coordinates was aligned with the phylogenetically rotated configuration of genetic distances by using the 169 shared specimens.

After computing the barycentre for populations, the correlation between genetic and morphological configurations resulted to be higher than for individuals. This was actually expected since, due to the presence of populations with both COI lineages, intermediate values for the genetic configuration are possible at population level (protest: correlation coefficient 0·797, P = 0·001).

ASSESSING DISSIMILARITY LOSS AFTER COMPUTING POPULATION MEANS AND TESTING ITS SIGNIFICANCE

After grouping individuals according to population means, most of the original variability of specimen configuration was maintained for genetic data (85·04% x-axis and 61·23% y-axis). Both axes maintained a significantly higher divergence than in random configurations (P < 0·001 and P = 0·007, as it can be also verified by visual comparisons of Figs 2a and 3a). Conversely, the grouped morphological configuration (Fig. 2b) only maintained a larger than random amount of variation in the first x-axis (61·79%, P < 0·001). In the y-axis, the maintained variance was only 14·15% with P = 0·079. The y-axis in Fig. 3b is much more flattened than in Fig 2b, suggesting that most of the variation in genitalia is linked to the x-axis, oriented according to the western–eastern clade membership.

PROJECTING THE ROTATED CONFIGURATIONS IN RGB COLOUR SPACE AND COMPUTING POPULATION MEANS

The transformed genetic and morphological configurations have been projected in the RGB colour space. For the morphological data, the points have been plotted together with the standardized contribution of shape variables. Maintaining the contribution of variables enabled the inspection of the main pattern of shape variation, as carried out by thin plate spline in geometric morphometrics (Bookstein 1991). The alignment with the genetic axis also allowed us to clearly recognize the direction of morphological variation with respect to genetic variation (Fig. 2a,b).

Mean population values have been projected in the RGB space (Fig. 3) and the colours obtained for each population have been plotted in the geographic space (Fig. 4). There is high concordance between genetic and morphological data over most of the study area, but some exceptions exist. They are mostly due to the existence of endemic lineages in Mallorca and the Aeolian islands and to strong discordances located from the Maritime Alps to the Messina strait, along the Tyrrhenian coast of the Italian Peninsula.

ANALYSING MODALITY FOR AXIS VALUES IN POPULATIONS, REGRESSING THE GENETIC AND MORPHOLOGICAL POPULATION VALUES AND DEPICTING THE SPATIAL DISTRIBUTION OF RESIDUALS

In the first morphological axis, no population showed a distribution that significantly differed from unimodal (Hartigan's dip test of unimodality P > 0·050 in each case, see Appendix

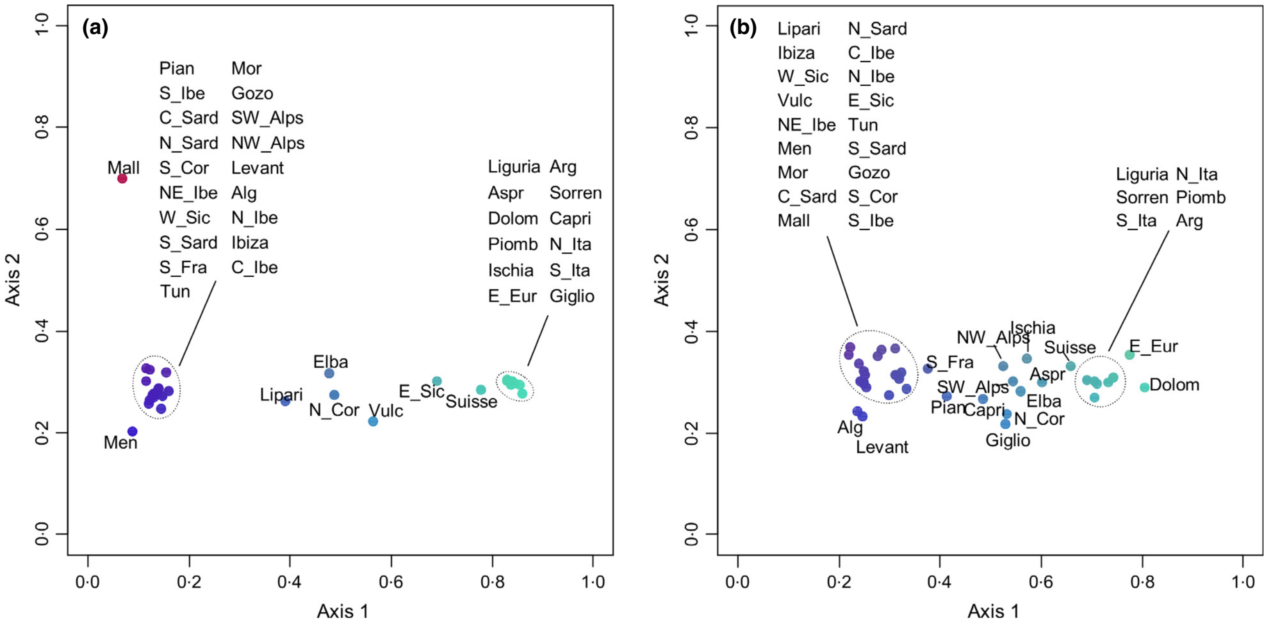

Fig. 3. Principal coordinate analysis (PCoA) projected in the RGB colour space representing mean population configurations for (a) genetic and (b) morphological data. See Table 1 for population abbreviations.

S1 for histograms). Variance in the *x*-axis of morphological configuration did not reveal higher values for intermediate populations, indicating that supposedly hybrid populations are not more morphologically variable (Appendix S1 for figure). Both quadratic coefficients revealed non-significant ($P > 0{\cdot}050$) effects on variance (Appendix S1).

A linear MA regression revealed a highly significant correlation between genetic and morphological *x*-axis values among populations (slope = 0·587, elevation = 0·194, $P < 0{\cdot}001$, $R^2 = 0{\cdot}732$, Fig. 5a). Nevertheless, the scatterplot revealed that several populations show large discrepancies (Fig. 5a). The values for residuals exceeding the standard deviation limits have been plotted on a map by using an appropriate colour scale for the dots (Fig. 5b). This allowed highlighting particularly discordant populations: populations showing more eastern genitalia with respect to their western genetic affiliation (north-western Alps, south-western Alps and, to a lesser extent, Pianosa) and vice versa (eastern Sicily, Vulcano, Capri and, to a lesser extent, Lipari, Giglio and Ischia). The sign of the residuals showed a highly coherent spatial pattern, with populations displaying more eastern morphology than expected based on genetic data being located in the northern Mediterranean, between the western Alps, Corsica and the Tuscan islands, while populations showing more western morphology than expected were located along the southern Tyrrhenian coast (Fig. 5b).

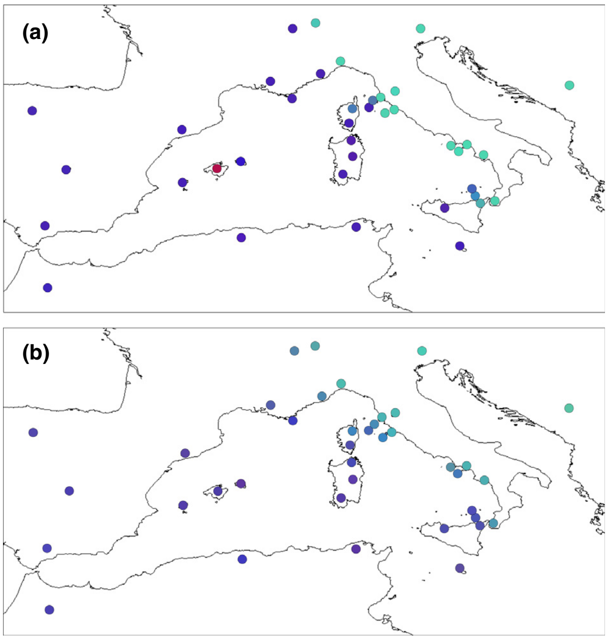

Fig. 4. Geographical location of populations with RGB colours as displayed in Fig. 3. (a) COI; (b) genitalia morphology.

Discussion

The method implemented by using the new *recluster* functions allowed a detailed assessment of the degree of co-variation between a genetic marker and a continuous morphological marker at population level. As a novelty, this method enabled the alignment and thus the simultaneous analysis of specimens for which data on either one or both traits was available. The possibility to use such heterogeneous data sets makes it possible, for example, to optimize the balance between the costs of DNA sequencing and the necessity to examine large numbers of specimens for morphological analyses. Undoubtedly, the new recluster.consensus function will also facilitate the use of public databanks and open the door to new data mining strategies. In particular, population studies such as the one performed in the model data set are becoming a powerful tool for biogeography and conservation biology and are facilitating the

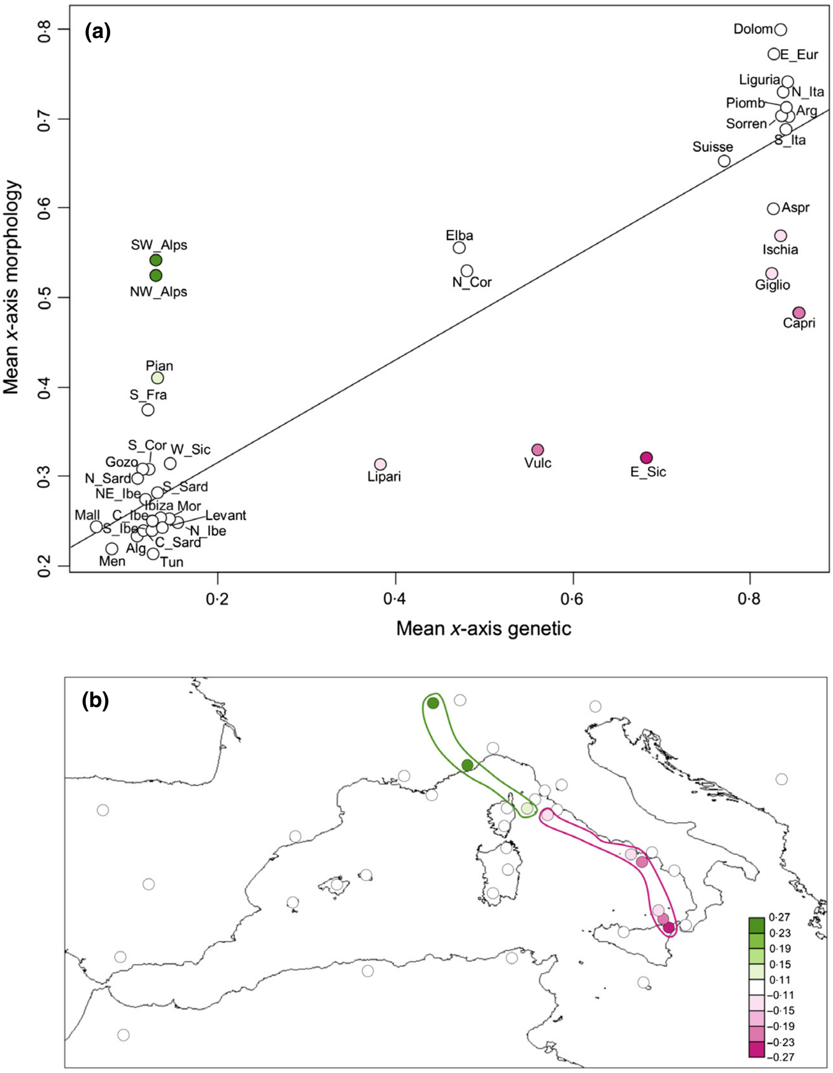

Fig. 5. (a) Major axis regression between genetic and morphological mean configurations for populations. Positive and negative residuals exceeding the standard deviation have been coloured with a green and purple scale, respectively. Values within the ± standard deviation interval are displayed in white. The positive residual for the completely eastern Dolomites population (Dolom) is represented in white. (b) Map showing the localization of large residuals over the study area.

recognition of speciation events, dynamics of taxa distribution and micro-evolutionary phenomena driven by various forces, including climate change (Schlick-Steiner *et al.* 2007; Dincă, Dapporto & Vila 2011; Dapporto *et al.* 2012; Renvoise *et al.* 2012; Toews & Brelsford 2012). Unlike previous assessments, the new procedure does not require the selection of morphological variables or components to be correlated with genetic patterns. Indeed, it has to be noted that the first morphological *x*-axis, to which most of our results belong, is profoundly different from a single morphological variable or PCA component since it is determined by a known contribution of several variables from two different genitalic structures. In most cases, several morphological components are correlated with the genetic signal and their contribution can vary. The first morphological axis we examined represents the best combination of such components to align the overall bidimensional morphological data to a genetic configuration obtained by combining phylogenetic relationships and overall genetic distance patterns.

The analysis allowed describing and evaluating the correlation between COI and morphological markers in *M. jurtina*. Most importantly, the possibility to align morphological and genetic data and to obtain mean values for populations allowed further examinations such as the modality and variance of the morphological traits as associated with genetic variation. Morphological traits revealed to be unimodal in all populations, thus excluding the possibility that the two lineages can cohabit while maintaining different shapes of genitalia, as found, for example, for the butterfly *Zerynthia cassandra* in the same geographic area (Zinetti *et al.* 2013). Populations at contact areas did not display higher variance than the rest, indicating that the mixing of the two genetic lineages produces morphologically homogeneous intermediate populations, a result that fully supports the subspecific status of the two taxa studied. However, the analysis revealed that some populations belong to a single genetic clade while showing intermediate morphology (like in the western Alps or in Capri). Some populations were also characterized by high residuals between morphological and genetic traits, like those in the western Alps and in eastern Sicily. The Vulcano population is especially interesting because, while it demonstrated a rather typical western morphology, it displayed an endemic genetic lineage (also detected in the neighbouring island of Lipari) that was sister to the rest of the western clade. Indeed, in the mean genetic PCoA configuration, the Vulcano population was placed at a rather intermediate position between the eastern and western clades, and this was not a consequence of hosting genetically mixed populations as was the case of northern Corsica, Elba, eastern Sicily and Lipari (Fig. 3a). This result suggests a rather old origin for the Vulcano–Lipari endemic genetic lineage, probably closely following the original split between the eastern and western clades. Mallorca hosts another example of an endemic genetic lineage, which was well placed within the western clade but showed signs of substantial drift, as indicated by the genetic PCoA y-axis (Fig. 3a). This lineage has been found at higher altitudes in the Serra de Tramuntana, while in lowland areas, the typical western lineage exists. In this apparently chaotic situation, a highly ordered spatial pattern of discordance was revealed, with discrepancies being located along the contact zone between the French Alps and the Messina strait. The sign of the residuals also clustered well in the two northern and southern halves of the contact zone. Several forces may determine such a pattern, for example, different selective pressures on the two latitudinal areas, differential introgression or dispersal between males and females (Descimon & Mallet 2009; Dasmahapatra *et al.* 2010; Gompert *et al.* 2010; Habel *et al.* 2011; Mallet, Wynne & Thomas 2011; Renvoise *et al.* 2012; Toews & Brelsford 2012; Mende & Hundsdoerfer 2013; Zinetti *et al.* 2013; Toews *et al.* 2014). What is important to point out is that a detailed study of the correlation between genetic and morphological markers within a geographic framework revealed that phylogeographic histories and ongoing dynamics are more multifaceted than hypothesized. The algorithms described in this article can be used with any kind of traits for which bidimensional configurations can be obtained, not necessarily a genetic and a morphological one. Moreover, more than two traits can be studied by using a series of Procrustes analyses and the possibility to handle missing data makes the new functions suitable for meta-analyses. The COI sequences and genitalia morphology of *M. jurtina* have provided an adequate model for this methodological study, and it is our hope that this procedure will be useful to many other organisms and characters.

Acknowledgements

We thank colleagues who provided samples used in this study. Financial support was provided by the Spanish Ministerio de Ciencia e Innovación project CGL2010-21226/BOS, by the project "Barcoding Italian Butterflies and by the project “Definizione dello status di conservazione delle falene e della malacofauna terrestre dell’Arcipelago Toscano”. Vlad Dincă was funded by a post-doctoral fellowship from the Wenner-Gren Foundation, Sweden and by a EU Marie Curie International Outgoing Fellowship for Career Development (project no. 625997). Raluca Vodă is funded by a FPU predoctoral fellowship (AP-2010-5409) from the Spanish Ministerio de Ciencia e Innovación. This study was performed in collaboration with the Parco Nazionale dell’Arcipelago Toscano.

Data accessibility

DNA sequences: GenBank accession numbers

KM020807–KM020882

KJ994239–KJ994253KM033847–KM033941

Morphological data: online supporting information

Geographic coordinates for specimens: online supporting information

R scripts: online supporting information and CRAN (http://cran.at.r-project.org/web/packages/recluster/index.html)

References

Avise, J.C. (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography*, **36**, 3–15.

Bookstein, F.L. (1991) *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, Cambridge UK; New York.

Claude, J. (2008) *Morphometrics with R*. Springer, New York, London.

Dapporto, L. & Bruschini, C. (2012) Invading a refugium: post glacial replacement of the ancestral lineage of a Nymphalid butterfly in the West Mediterranean. *Organisms Diversity & Evolution*, **12**, 39–49.

Dapporto, L., Bruschini, C., Baracchi, D., Cini, A., Gayubo, S.F., Gonzalez, J.A. & Dennis, R.L.H. (2009) Phylogeography and counter-intuitive inferences in island biogeography: evidence from morphometric markers in the mobile butterfly *Maniola jurtina* (Linnaeus) (Lepidoptera, Nymphalidae). *Biological Journal of the Linnean Society*, **98**, 677–692.

Dapporto, L., Habel, J.C., Dennis, R.L.H. & Schmitt, T. (2011) The biogeography of the western Mediterranean: elucidating contradictory distribution patterns of differentiation in *Maniola jurtina* (Lepidoptera: Nymphalidae). *Biological Journal of the Linnean Society*, **103**, 571–577.

Dapporto, L., Bruschini, C., Dinca, V., Vila, R. & Dennis, R.L.H. (2012) Identifying zones of phenetic compression in West Mediterranean butterflies (Satyrinae): refugia, invasion and hybridization. *Diversity and Distributions*, **18**, 1066–1076.

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.

Dasmahapatra, K.K., Lamas, G., Simpson, F. & Mallet, J. (2010) The anatomy of a ‘suture zone’ in Amazonian butterflies: a coalescent-based test for vicariant geographic divergence and speciation. *Molecular Ecology*, **19**, 4283–4301.

Derryberry, E.P., Derryberry, G.E., Maley, J.M. & Brumfield, R.T. (in press) hzar: hybrid zone analysis using an R software package. *Molecular Ecology Resources* **14**, 652–663.

Descimon, H. & Mallet, J. (2009) Bad species. *Ecology of Butterflies in Europe* (eds J. Settele, T.G. Shreeve, M. Konvicka & H. Van Dyck), pp. 219–249. Cambridge University Press, Cambridge.

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

Gompert, Z., Lucas, L.K., Fordyce, J.A., Forister, M.L. & Nice, C.C. (2010) Secondary contact between *Lycaeides idas* and L-melissa in the Rocky Mountains: extensive admixture and a patchy hybrid zone. *Molecular Ecology*, **19**, 3171–3192.

Habel, J.C., Dieker, P. & Schmitt, T. (2009) Biogeographical connections between the Maghreb and the Mediterranean peninsulas of southern Europe. *Biological Journal of the Linnean Society*, **98**, 693–703.

Habel, J.C., Lens, L., Rödder, D. & Schmitt, T. (2011) From Africa to Europe and back: refugia and range shifts cause high genetic differentiation in the Marbled White butterfly *Melanargia galathea*. *Bmc Evolutionary Biology*, **11**, 215.

Hartigan, J.A. & Hartigan, P.M. (1985) The dip test of unimodality. *Annals of Statistics*, **13**, 70–84.

Holt, B., Lessard, J.P., Borregaard, M.K., Fritz, S.A., Araujo, M.B., Dimitrov, D. *et al.* (2013) An Update of Wallace’s Zoogeographic Regions of the World. *Science*, **339**, 74–78.

Kalinowski, S.T. (2009) How well do evolutionary trees describe genetic relationships among populations? *Heredity*, **102**, 506–513.

Kreft, H. & Jetz, W. (2010) A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography*, **37**, 2029–2053.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*, 2nd English edn. Elsevier, Amsterdam; New York.

Madeira, C., Alves, M.J., Mesquita, N., Silva, S.E. & Paula, J. (2012) Tracing geographical patterns of population differentiation in a widespread mangrove gastropod: genetic and geometric morphometrics surveys along the eastern African coast. *Biological Journal of the Linnean Society*, **107**, 647–663.

Mallet, J., Wynne, I.R. & Thomas, C.D. (2011) Hybridisation and climate change: brown argus butterflies in Britain (Polyommatus subgenus Aricia). *Insect Conservation and Diversity*, **4**, 192–199.

Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979) *Multivariate Analysis*. Academic Press, London; New York.

Mende, M.B. & Hundsdoerfer, A.K. (2013) Mitochondrial lineage sorting in action – historical biogeography of the *Hyles euphorbiae* complex (Sphingidae, Lepidoptera) in Italy. *Bmc Evolutionary Biology*, **13**, 83.

Mila, B., Toews, D.P.L., Smith, T.B. & Wayne, R.K. (2011) A cryptic contact zone between divergent mitochondrial DNA lineages in southwestern North America supports past introgressive hybridization in the yellow-rumped warbler complex (Aves: Dendroica coronata). *Biological Journal of the Linnean Society*, **103**, 696–706.

Paradis, E. (2012) *Analysis of Phylogenetics and Evolution with R*, 2nd edn. Springer, New York.

Pavlova, A., Amos, J.N., Joseph, L., Loynes, K., Austin, J.J., Keogh, J.S., Stone, G.N., Nicholls, J.A. & Sunnucks, P. (2013) Perched at the mito-nuclear crossroads: divergent mitochondrial lineages correlate with environment in the face of ongoing nuclear gene flow in an Australian bird. *Evolution*, **67**, 3412–3428.

Pergams, O.R.W. & Lacy, R.C. (2008) Rapid morphological and genetic change in Chicago-area Peromyscus. *Molecular Ecology*, **17**, 450–463.

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

Renvoise, E., Montuire, S., Richard, Y., Quere, J.P., Gerber, S., Cucchi, T., Chateau-Smith, C. & Tougard, C. (2012) Microevolutionary relationships between phylogeographical history, climate change and morphological variability in the common vole (*Microtus arvalis*) across France. *Journal of Biogeography*, **39**, 698–712.

Schlick-Steiner, B.C., Seifert, B., Stauffer, C., Christian, E., Crozier, R.H. & Steiner, F.M. (2007) Without morphology, cryptic species stay in taxonomic crypsis following discovery. *Trends in Ecology & Evolution*, **22**, 391–392.

Schmitt, T., Rober, S. & Seitz, A. (2005) Is the last glaciation the only relevant event for the present genetic population structure of the meadow brown butterfly *Maniola jurtina* (Lepidoptera: Nymphalidae)? *Biological Journal of the Linnean Society*, **85**, 419–431.

Seraphim, N., Marín, M.A., Freitas, A.V.L. & Silva-Brandão, K.L. (2013) Morphological and molecular marker contributions to disentangling the cryptic *Hermeuptychia hermes species* complex (Nymphalidae: Satyrinae: Euptychiina). *Molecular Ecology Resources*, **14**, 39–49.

Thomson, G. (2011) *The Meadow Brown Butterflies*. Waterbeck, Scotland.

Toews, D.P.L. & Brelsford, A. (2012) The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, **21**, 3907–3930.

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

Viscosi, V. & Cardini, A. (2011) Leaf Morphology, Taxonomy and Geometric Morphometrics: a Simplified Protocol for Beginners. *PLoS ONE*, **6**, e25630.

Zelditch, M., Swiderski, D.L. & Sheets, H.D. (2012) *Geometric Morphometrics for Biologists: A Primer*, 2nd edn. Elsevier Academic Press.

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**, e65746.

Received 2 April 2014; accepted 10 June 2014
Handling Editor: Louise Johnson

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. R scripts (AppendixS1.doc) and data (geno.txt, morpho.txt, names.txt, new functions.R and gendist.csv) used to perform the analyses.

Appendix S2. Supplementary methods.