Supporting Information

Use of genetic, climatic, and microbiological data to inform reintroduction of a regionally extinct butterfly

Supplementary methods

Dataset used for molecular analyses

The dataset included 101 cytochrome *c* oxidase subunit I (COI) sequences of *M. russiae*, representative for the range of this species (Appendix S1), and obtained as follows: 54 sequences were mined from GenBank (Nazari et al. 2010, 35 sequences; Dincă et al. 2015, 17 sequences; Lukhtanov et al. 2009, 2 sequences), and 47 sequences obtained for this study. We did not use GenBank sequences shorter than 600 base pairs (bp), or those that lacked locality data of reasonable precision (e.g. less than 10 km error).

The above-mentioned dataset included three COI sequences of *M. russiae* from the extinct Hungarian population (all collected in 1912). These sequences were the only usable ones (see below) obtained from ten specimens collected between 1898 and 1912, that are stored in the Hungarian Natural History Museum (HNHM) Budapest.

DNA sequencing

Twelve of the 47 COI sequences generated for this study were obtained at the Butterfly Diversity and Evolution Lab of the Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Spain. In this case, total genomic DNA was extracted using Chelex 100 resin, 100–200 mesh, sodium form (Biorad), under the following protocol: one leg was removed and introduced into 100 µl of Chelex 10% and 5 µl of Proteinase K (20 mg/ml) were added. The samples were incubated overnight at 55ºC and were subsequently incubated at 100ºC for 15 minutes. Samples were then centrifuged for 10 s at 3000 rpm. A 658-bp fragment near the 5´ end of COI was amplified by polymerase chain reaction using the primers LepF1 and LepR1 (Appendix S2). Double-stranded DNA was amplified in 25-µL volume reactions containing: 14.4 µl autoclaved Milli-Q water, 5 µl 5x buffer, 2 µl 25 mM $MgCl_2$, 0.5 µl 10 mM dNTPs, 0.5 µl of each primer (10 µM), 0.1 µl Taq DNA Polymerase (Promega, 5U/ µl) and 2 µl of extracted DNA. The typical thermal cycling profile followed this protocol: first denaturation at 92ºC for 60 s, followed by five cycles of 92ºC for 15 s, 48ºC for 45 s and 62ºC for 150 s, and then by 35 cycles of 92ºC for 15 s, 52ºC for 45 s and 62ºC for 150 s and a final extension at 62ºC for 420 s. PCR products were purified and sequenced by Macrogen Inc.

The remaining 35 sequences were generated at the Biodiversity Institute of Ontario, Canada following standard protocols for DNA barcoding (deWaard et al. 2008), and DNA sequencing was performed on an ABI 3730XL capillary sequencer (Applied Biosystems). In the case of the old specimens of *M. russiae* from Hungary, a full DNA barcode (658-bp) for one specimen was obtained combining amplicons obtained using the primers [LepF1 + MLepR2] + [MLepF1 + LepR1]. The other two specimens had reliable sequences only for the 407-bp amplicons obtained using the primers MLepF1 + LepR1 (Appendix S2).

Sequences were edited in CodonCode Aligner 3.0 or in GENEIOUS PRO 6.1.8 (Biomatters, http://www.geneious.com/) and assembled using the latter.

The 47 COI sequences obtained in this study are available in GenBank (see Appendix S1 for accession numbers), and all sequences are also publicly available in the dataset DS-MELARUSS (dx.doi.org/10.5883/DS-MELARUSS) from the Barcode of Life Data Systems (http://www.boldsystems.org/).

Analyses of DNA sequences

The 101 COI sequences of *M. russiae* used in this study were collapsed to 43 unique haplotypes using TCS 1.21 (Clement et al. 2000). The same program was used to construct a maximum parsimony haplotype network, with a 93% connection limit.

Phylogenetic relationships were inferred using Bayesian inference (BI) through the CIPRES Science Gateway (Miller et al. 2010). For this analysis we used the 43 haplotypes of *M. russiae* together with three outgroup sequences from *M. galathea*, *M. leda* and *M. ines* (see Nazari et al. 2010). Both BI analyses and the estimation of node ages were run in BEAST 1.8.0 (Drummond & Rambaut 2007). The GTR + I + G substitution model was chosen according to the value of the Akaike information criterion (AIC) obtained in JMODELTEST 2.1.3 (Darriba et al. 2012). Base frequencies were estimated, six gamma rate categories were selected and a randomly generated initial tree was used.

Rough estimates of node ages were obtained by applying two molecular clocks with: 1.5% uncorrected pairwise distance per million years estimated for various invertebrates (Quek et al. 2004), and 2.3% estimated for the entire mitochondrial genome of several arthropods (Brower 1994). A strict clock and a normal prior distribution was used, centred on the mean between the two substitution rates, and the standard deviation was tuned so that the 95% confidence interval of the posterior density coincided with the 1.5% and 2.3% rates, respectively. Parameters were estimated using two independent runs of 20 million generations each, and convergence was checked using the program TRACER 1.6.

Wolbachia *infection analyses*

47 specimens of *M. russiae* were surveyed for the presence of the bacterial endosymbiont *Wolbachia*.

DNA from half of the abdomen was extracted using Chelex 100 resin, 100–200 mesh, sodium form (Biorad), under the following protocol: the abdomen piece was introduced into 100 µl of Chelex 10%, and 5 µl of Proteinase K (20 mg/ml) were added. The samples were incubated overnight at 55ºC and were subsequently incubated at 100ºC for 15 minutes. Samples were then centrifuged for 10 s at 3000 rpm.

The samples where then tested for *Wolbachia* using polymerase chain reaction (PCR) primers specific to *Wolbachia* genes wsp and coxA (Appendix S2). These genes are extensively used to detect *Wolbachia* infection in a wide array of insects. Wsp was amplified by PCR in 25-µL volume reactions containing: 16 µl autoclaved Milli-Q water, 5 µl 5x buffer, 1.4 µl 25 mM $MgCl_2$, 0.5 µl 10 mM dNTPs, 0.5 µl of each primer (10 µM), 0.1 µl Taq DNA Polymerase (Promega, 5U/ µl) and 1 µl of extracted DNA. The typical thermal cycling profile followed this protocol: first denaturation at 94ºC for 120 s, followed by 35 cycles of 94ºC for 30 s, 59ºC for 30 s and 72ºC for 60 s and a final extension at 72ºC for 600 s.

CoxA was amplified by PCR in 25-µL volume reactions containing: 15.4 µl autoclaved Milli-Q water, 5 µl 5x buffer, 2 µl 25 mM $MgCl_2$, 0.5 µl 10 mM dNTPs, 0.5 µl of each primer (10 µM), 0.1 µl Taq DNA Polymerase (Promega, 5U/ µl) and 1 µl of extracted DNA. The typical thermal cycling profile followed this protocol: first denaturation at 94ºC for 120 s, followed by 35 cycles of 94ºC for 30 s, 56ºC for 45 s and 72ºC for 90 s and a final extension at 72ºC for 600 s.

Samples with amplicons of the expected size were scored as positive for *Wolbachia*. To avoid false negatives, samples that did not produce bands after electrophoresis of the PCR products were additionally assayed using the primer pair LepF1/LepR1 that amplifies a 658-bp fragment near the 5´ end of COI, to ascertain quality of the DNA extracts (PCR protocols were as described under "DNA sequencing"). When this third PCR failed to produce a band, the sample was removed from the assay. However if this PCR produced a band, the sample was declared uninfected with regard to *Wolbachia*. Following this approach, of the 47 specimens screened, 37 could be reliably assessed for the presence/absence of *Wolbachia* (Appendix S1).

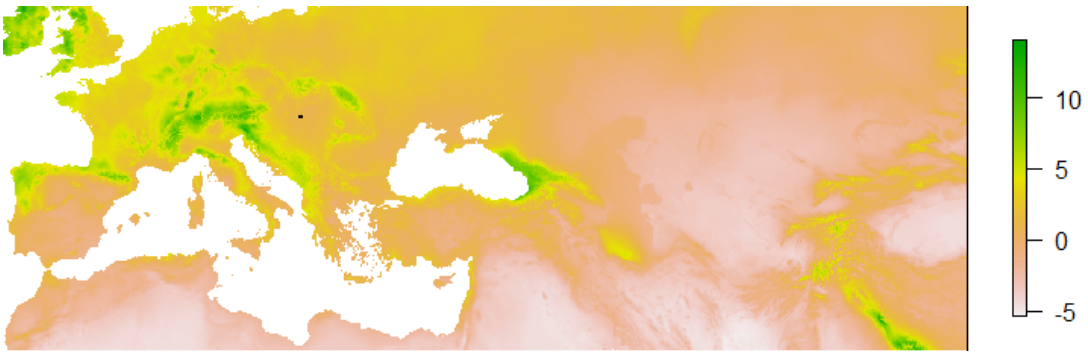
Subsequently, the *coxA* PCR products of 16 specimens, and the wsp PCR products of 14 specimens were purified and sequenced by Macrogen Inc. Sequences were then compared to existing records using the *Wolbachia* MLST Database (pubmlst.org/wolbachia/) in order to identify the sequence type for each gene locus. Sequences obtained during the screening are available in GenBank (see Appendix S1 for accession numbers) and in the dataset DS-MELARUSS (dx.doi.org/10.5883/DS-MELARUSS) from the Barcode of Life Data Systems (http://www.boldsystems.org/).

Mapping genetic diversity

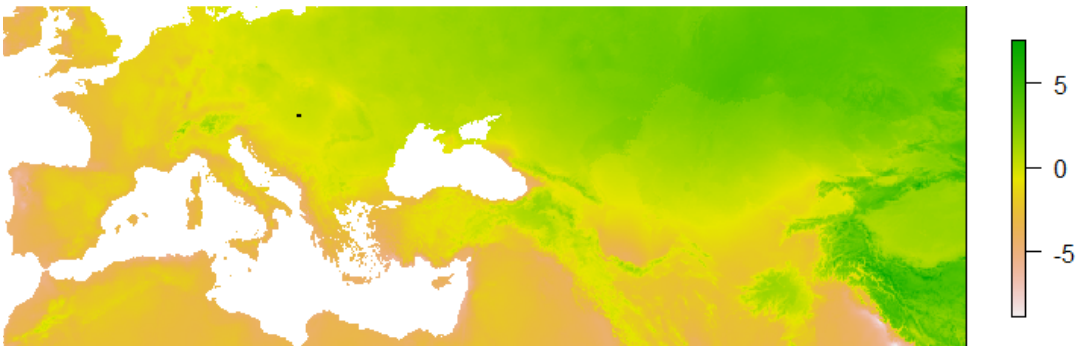
To examine spatial patterns of genetic differentiation, we calculated the genetic uncorrected p-distances among all sequenced specimens of *M. russiae*. A Principal Coordinates Analysis (PCoA) was applied to this matrix to reduce the dissimilarity matrix among specimens to two dimensions. To visualize the pattern of genetic similarity over geographic space, we projected the PCoA configuration in RGB space using the recluster.col function of the R package "recluster" (Dapporto et al. 2014). Using this function, the color resemblance of the resulting dots was directly proportional to the genetic distances among the specimens. The specimens with their corresponding colors and geographic location were plotted on a map by using pie charts that collapsed specimens belonging to the same square of 2x2 degrees of latitude and longitude.

Climatic analyses

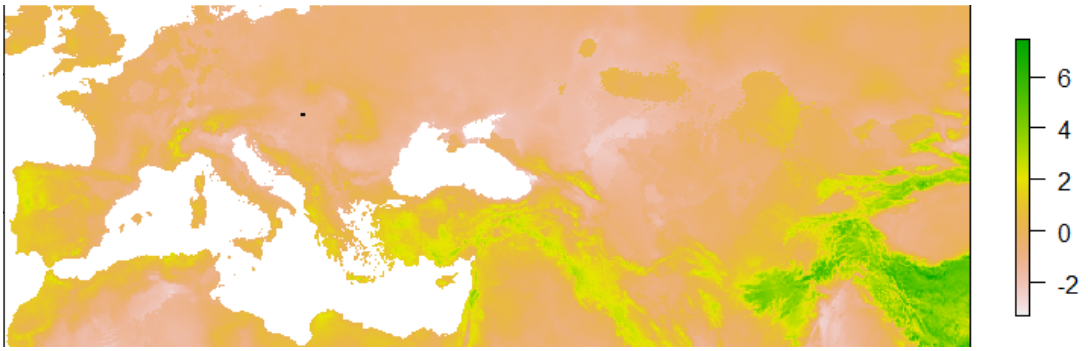
We aimed to recognise the populations living in areas climatically similar to the Hungarian site where the reintroduction is planned. We proceeded as follows: we downloaded the 19 climatic layers from WorldClim (http://www.worldclim.org/, version 2.0 1970-2000) at a resolution of 5 minutes. Climatic variables tend to be highly correlated and we scaled them to obtain a mean equal to zero and a standard deviation equal to one. Subsequently, we performed a Principal Component Analysis (PCA) among the 19 layers by using the princomp R function. The PCA produced a reduced number of layers composed by a combination of the 19 bioclim layers. Among the layers generated by the PCA, we retained the four layers showing eigenvalues higher than one.



The first principal component layer mainly linked to summer precipitation (pink, dry; green, wet). The black square indicates the location of the Hungarian target area in this and in the following figures.



The second principal component layer mainly linked to overall temperature (pink, warm; green, cold).



The third principal component layer, mainly linked to overall precipitation (pink, dry; green, wet).

The fourth principal component layer, mainly linked to annual variation in precipitation (pink, low variation; green, high variation).

From the PCA layers, we extracted the values of the cells belonging to the area where the reintroduction has been planned (target area) and averaged the values for each PCA layer. For each cell in the study area, we calculated the Euclidean distances between it and the average values of the PCA cells of the target area. The Euclidean distances were log-transformed to provide a measure of climatic similarity of each cell with the target area. With these values, we produced a new raster layer to understand which populations belong to areas climatically more similar to the target one in Hungary.

Literature cited

Brower, AVZ. 1994. Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution. Proceedings of the National Academy of Sciences of the United States of America **91:** 6491-6495.

Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. Molecular Ecology **9:** 1657–1660. DOI: 10.1046/j.1365-294x.2000.01020.x

Dapporto L, Vodă R, Dincă V, Vila R. 2014. Comparing population patterns for genetic and morphological markers with uneven sample sizes. An example for the butterfly *Maniola jurtina*. Methods in Ecology and Evolution **5:** 834-843. doi: 10.1111/2041-210X.12220

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods **9:** 772. doi:10.1038/nmeth.2109

deWaard JR, Ivanova NV, Hajibabaei M, Hebert PDN. 2008. Assembling DNA Barcodes: Analytical Protocols. p. 275-293. In: Cristofre M. (Hrsg.), *Methods in Molecular Biology: Environmental Genetics*. Humana Press Inc., Totowa, USA, 364 p.

Dincă V, Montagud S, Talavera G, Hernández-Roldán J, Munguira ML, García-Barros E, Hebert PDN., Vila R. 2015. DNA barcode reference library for Iberian butterflies enables a continental-scale preview of potential cryptic diversity. Scientific Reports **5:** 12395. DOI: 10.1038/srep12395

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7**:** 214. https://doi.org/10.1186/1471-2148-7-214

Lukhtanov VA, Sourakov A, Zakharov EV, Hebert PDN. 2009. DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification. Molecular Ecology Resources **9(5):** 1302-1310. DOI: 10.1111/j.1755-0998.2009.02577.x

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans, LA, Nov. 14, 2010*. Institute of Electrical and Electronics Engineers, p. 1-8.

Nazari V, Ten Hagen W, Bozano GC. 2010. Molecular systematics and phylogeny of the 'Marbled Whites' (Lepidoptera: Nymphalidae, Satyrinae, *Melanargia* Meigen). Systematic Entomology **35(1):** 132-147. DOI: 10.1111/j.1365-3113.2009.00493.x

Quek SP, Davies SJ, Itino T, Pierce NE. 2004. Codiversification in an ant-plant mutualism: stem texture and the evolution of host use in *Crematogaster* (Formicidae: Myrmicinae) inhabitants

of *Macaranga* (Euphorbiaceae). Evolution **58:** 554-570. DOI: 10.1111/j.0014-3820.2004.tb01678.x