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).

43 Sequences were edited in CodonCode Aligner 3.0 or in GENEIOUS PRO 6.1.8 (Biomatters,
44 http://www.geneious.com/) and assembled using the latter.
45 The 47 COI sequences obtained in this study are available in GenBank (see Appendix S1 for
46 accession numbers), and all sequences are also publicly available in the dataset DS-MELARUSS
47 (dx.doi.org/10.5883/DS-MELARUSS) from the Barcode of Life Data Systems
48 (http://www.boldsystems.org/).
49
50 *Analyses of DNA sequences*
51 The 101 COI sequences of *M. russiae* used in this study were collapsed to 43 unique haplotypes
52 using TCS 1.21 (Clement et al. 2000). The same program was used to construct a maximum
53 parsimony haplotype network, with a 93% connection limit.
54 Phylogenetic relationships were inferred using Bayesian inference (BI) through the CIPRES
55 Science Gateway (Miller et al. 2010). For this analysis we used the 43 haplotypes of *M. russiae*
56 together with three outgroup sequences from *M. galathea*, *M. leda* and *M. ines* (see Nazari et al.
57 2010). Both BI analyses and the estimation of node ages were run in BEAST 1.8.0 (Drummond
58 & Rambaut 2007). The GTR + I + G substitution model was chosen according to the value of the
59 Akaike information criterion (AIC) obtained in JMODELTEST 2.1.3 (Darriba et al. 2012). Base
60 frequencies were estimated, six gamma rate categories were selected and a randomly generated
61 initial tree was used.
62 Rough estimates of node ages were obtained by applying two molecular clocks with: 1.5%
63 uncorrected pairwise distance per million years estimated for various invertebrates (Quek et al.
64 2004), and 2.3% estimated for the entire mitochondrial genome of several arthropods (Brower
65 1994). A strict clock and a normal prior distribution was used, centred on the mean between the
66 two substitution rates, and the standard deviation was tuned so that the 95% confidence interval
67 of the posterior density coincided with the 1.5% and 2.3% rates, respectively. Parameters were
68 estimated using two independent runs of 20 million generations each, and convergence was
69 checked using the program TRACER 1.6.
70
71 Wolbachia *infection analyses*
72 47 specimens of *M. russiae* were surveyed for the presence of the bacterial endosymbiont
73 *Wolbachia*.
74 DNA from half of the abdomen was extracted using Chelex 100 resin, 100–200 mesh, sodium
75 form (Biorad), under the following protocol: the abdomen piece was introduced into 100 μl of
76 Chelex 10%, and 5 μl of Proteinase K (20 mg/ml) were added. The samples were incubated
77 overnight at 55ºC and were subsequently incubated at 100ºC for 15 minutes. Samples were then
78 centrifuged for 10 s at 3000 rpm.
79 The samples where then tested for *Wolbachia* using polymerase chain reaction (PCR) primers
80 specific to *Wolbachia* genes wsp and coxA (Appendix S2). These genes are extensively used to
81 detect *Wolbachia* infection in a wide array of insects. Wsp was amplified by PCR in 25-μL
82 volume reactions containing: 16 μl autoclaved Milli-Q water, 5 μl 5x buffer, 1.4 μl 25 mM
83 $MgCl_2$, 0.5 μl 10 mM dNTPs, 0.5 μl of each primer (10 μM), 0.1 μl Taq DNA Polymerase
84 (Promega, 5U/ μl) and 1 μl of extracted DNA. The typical thermal cycling profile followed this
85 protocol: first denaturation at 94ºC for 120 s, followed by 35 cycles of 94ºC for 30 s, 59ºC for 30
86 s and 72ºC for 60 s and a final extension at 72ºC for 600 s.

87   CoxA was amplified by PCR in 25-μL volume reactions containing: 15.4 μl autoclaved Milli-Q
88   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),
89   0.1 μl Taq DNA Polymerase (Promega, 5U/ μl) and 1 μl of extracted DNA. The typical thermal
90   cycling profile followed this protocol: first denaturation at 94ºC for 120 s, followed by 35 cycles
91   of 94ºC for 30 s, 56ºC for 45 s and 72ºC for 90 s and a final extension at 72ºC for 600 s.
92   Samples with amplicons of the expected size were scored as positive for *Wolbachia*. To avoid
93   false negatives, samples that did not produce bands after electrophoresis of the PCR products
94   were additionally assayed using the primer pair LepF1/LepR1 that amplifies a 658-bp fragment
95   near the 5´ end of COI, to ascertain quality of the DNA extracts (PCR protocols were as
96   described under "DNA sequencing"). When this third PCR failed to produce a band, the sample
97   was removed from the assay. However if this PCR produced a band, the sample was declared
98   uninfected with regard to *Wolbachia*. Following this approach, of the 47 specimens screened, 37
99   could be reliably assessed for the presence/absence of *Wolbachia* (Appendix S1).
100   Subsequently, the *coxA* PCR products of 16 specimens, and the wsp PCR products of 14
101   specimens were purified and sequenced by Macrogen Inc. Sequences were then compared to
102   existing records using the *Wolbachia* MLST Database (pubmlst.org/wolbachia/) in order to
103   identify the sequence type for each gene locus. Sequences obtained during the screening are
104   available in GenBank (see Appendix S1 for accession numbers) and in the dataset DS-
105   MELARUSS (dx.doi.org/10.5883/DS-MELARUSS) from the Barcode of Life Data Systems
106   (http://www.boldsystems.org/).
107
108   *Mapping genetic diversity*
109   To examine spatial patterns of genetic differentiation, we calculated the genetic uncorrected p-
110   distances among all sequenced specimens of *M. russiae*. A Principal Coordinates Analysis
111   (PCoA) was applied to this matrix to reduce the dissimilarity matrix among specimens to two
112   dimensions. To visualize the pattern of genetic similarity over geographic space, we projected the
113   PCoA configuration in RGB space using the recluster.col function of the R package "recluster"
114   (Dapporto et al. 2014). Using this function, the color resemblance of the resulting dots was
115   directly proportional to the genetic distances among the specimens. The specimens with their
116   corresponding colors and geographic location were plotted on a map by using pie charts that
117   collapsed specimens belonging to the same square of 2x2 degrees of latitude and longitude.
118
119   *Climatic analyses*
120   We aimed to recognise the populations living in areas climatically similar to the Hungarian site
121   where the reintroduction is planned. We proceeded as follows: we downloaded the 19 climatic
122   layers from WorldClim (http://www.worldclim.org/, version 2.0 1970-2000) at a resolution of 5
123   minutes. Climatic variables tend to be highly correlated and we scaled them to obtain a mean
124   equal to zero and a standard deviation equal to one. Subsequently, we performed a Principal
125   Component Analysis (PCA) among the 19 layers by using the princomp R function. The PCA
126   produced a reduced number of layers composed by a combination of the 19 bioclim layers.
127   Among the layers generated by the PCA, we retained the four layers showing eigenvalues higher
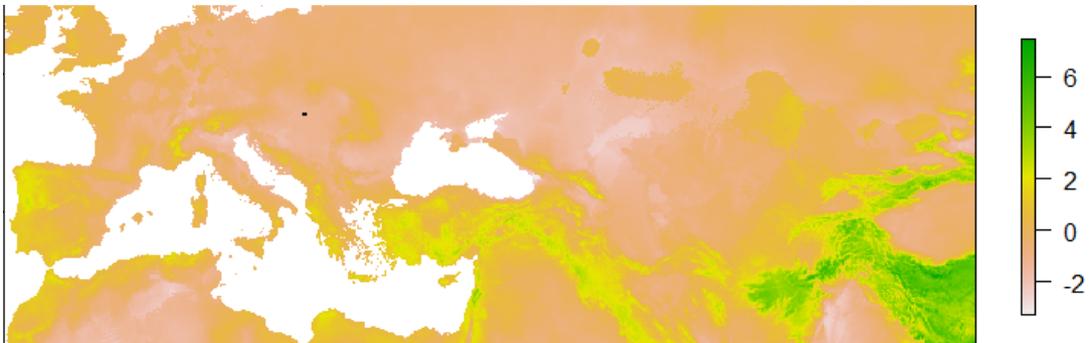128   than one.
129
130
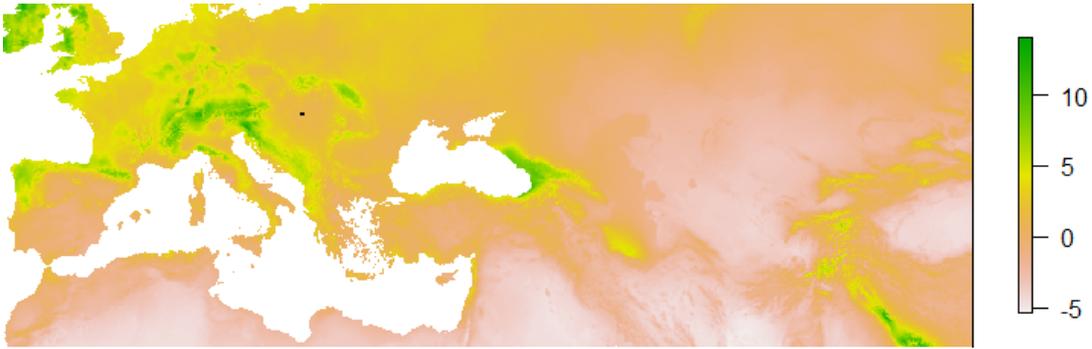
The first principal component layer mainly linked to summer precipitation (pink, dry; green, wet). The black square indicates the location of the Hungarian target area in this and in the following figures.
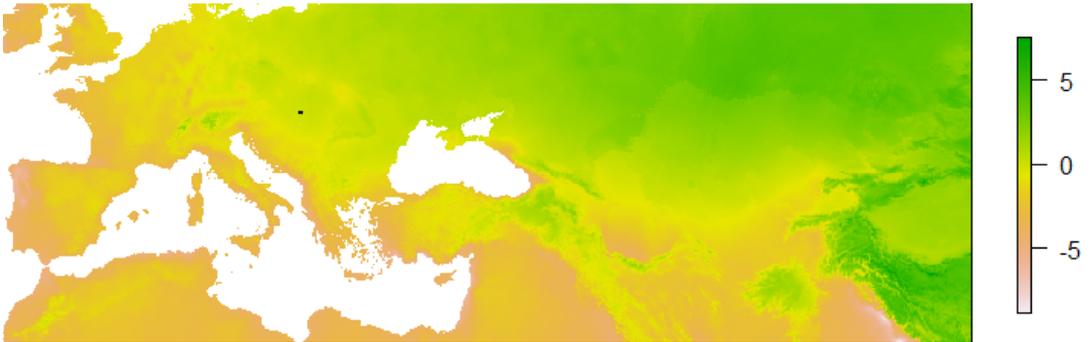
The second principal component layer mainly linked to overall temperature (pink, warm; green, cold).

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).

147
148 From the PCA layers, we extracted the values of the cells belonging to the area where the
149 reintroduction has been planned (target area) and averaged the values for each PCA layer. For
150 each cell in the study area, we calculated the Euclidean distances between it and the average
151 values of the PCA cells of the target area. The Euclidean distances were log-transformed to
152 provide a measure of climatic similarity of each cell with the target area. With these values, we
153 produced a new raster layer to understand which populations belong to areas climatically more
154 similar to the target one in Hungary.

155
156
157

158 **Literature cited**

159 Brower, AVZ. 1994. Rapid morphological radiation and convergence among races of the
160 butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution.
161 Proceedings of the National Academy of Sciences of the United States of America **91:** 6491-
162 6495.

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

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

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

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

173 Dincă V, Montagud S, Talavera G, Hernández-Roldán J, Munguira ML, García-Barros E, Hebert
174 PDN., Vila R. 2015. DNA barcode reference library for Iberian butterflies enables a
175 continental-scale preview of potential cryptic diversity. Scientific Reports **5:** 12395. DOI:
176 10.1038/srep12395

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

179 Lukhtanov VA, Sourakov A, Zakharov EV, Hebert PDN. 2009. DNA barcoding Central Asian
180 butterflies: increasing geographical dimension does not significantly reduce the success of
181 species identification. Molecular Ecology Resources **9(5):** 1302-1310. DOI: 10.1111/j.1755-
182 0998.2009.02577.x

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

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

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

192   of *Macaranga* (Euphorbiaceae). Evolution **58:** 554-570. DOI: 10.1111/j.0014-
193   3820.2004.tb01678.x