Short Title: Common Blue Diversity

Multiple colonisations and *Wolbachia* infections influence the fine-scale population structure of the widespread Common Blue butterfly (*Polyomattus icarus*) in the British Isles

Saad Arif[1,2*], Michael Gerth[2,3], William G. Hone-Millard[3,4], Maria D. S. Nunes[2], Leonardo Dapporto[5], Timothy G. Shreeve[3*]

[1]Centre for Functional Genomics, Oxford Brookes University, OX3 0BP Oxford, United Kingdom

[2]Department of Biological and Medical Sciences, Oxford Brookes University, OX3 0BP, United Kingdom

[3]Centre for Ecology and Environmental Conservation, Oxford Brookes University, OX3 0BP, United Kingdom

[4]Current address: John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK

[5]Department of Biology, University of Florence, Italy

*Authors for Correspondence

Keywords: *Polyommatus icarus*, Lepidoptera, CO1, RAD sequencing, *Wolbachia,* endosymbiont, mito-nuclear discordance, feminization

Correspondence:

Saad Arif

Centre for Functional Genomics, Oxford Brookes University, OX3 0BP Oxford, United Kingdom

Department of Biological and Medical Sciences, Oxford Brookes University, OX3 0BP, United Kingdom

Tel: +44 (0)1865 483437

Email: sarif@brookes.ac.uk

Timothy G. Shreeve

Centre for Ecology and Environmental Conservation, Oxford Brookes University, OX3 0BP, United Kingdom

Tel: +44 (0)1865 483957

Email: tgshreeve@brookes.ac.uk

Abstract

Deeply divergent mitochondrial genomes can reveal hidden diversity within species, however robust assessments of diversity require corroborative divergence in the nuclear genome. Previous phylogeographic analysis of the Common Blue butterfly (*Polyommatus icarus)* revealed a deeply divergent mitotype in Europe. Here, we reconstruct the phylogeography of this butterfly in the British Isles using both mitochondrial and nuclear DNA. We found strong geographical structuring of three distinct *CO*1 mitotypes with an ancient lineage restricted to the Outer Hebrides. The short time (*c*. 20,000 years) available for colonization of the British Isles and an expanded CO1 phylogeography from the species' entire range supports a hypothesis of at least 3 colonization events with sequential replacement of mtDNA. Analysis of ddRADSeq genomic markers from 184 individuals across the British Isles indicates that genetic clustering is similar but divergence weaker at the nuclear level, consistent with a biogeographic scenario of multiple colonisations. We were also able to use the ddRADSeq to infer the infection status of all individuals sequenced. Interestingly, we found evidence for high frequency (>70%) infections by two different *Wolbachia* strains, with strong mitotype associations, in the Northern populations of the British Isles. *Wolbachia*-induced cytoplasmic incompatibility may be sufficient to explain the mito-nuclear discordance observed across an admixture zone on the mainland and the complete dominance of a deeply divergent mitotype on the Outer Hebrides. Moreover, the strain (*w*Ica1), largely restricted to the Outer Hebrides potentially causes phenotypic feminisation of genetic males, a known but rare phenotypic effect of *Wolbachia* in Lepidoptera. Therefore, the impact of *Wolbachia* on the biology and genetic diversity of this butterfly merits further attention.

.

Introduction

Understanding within and between species diversity is fundamental to assessing and maintaining biodiversity and prioritising conservation effort. Advances in the application of molecular genetics for species identification and biodiversity monitoring (e.g. DNA barcoding projects) have revealed substantial genetic divisions within species that had previously been considered a single taxonomic species (historically based on phenotype, geographical isolation or some combination of both) giving rise to studies of cryptic biodiversity (Adams, Raadik, Burridge, & Georges, 2014; Bickford et al., 2007; Struck et al., 2018). Studies of intraspecific genetic diversity enhance estimates of species richness (Larsen, Miller, Rhodes, & Wiens, 2017), can help mitigate inadvertent loss of biodiversity (Bálint et al., 2011; Trontelj & Fišer, 2009) and potentially contribute to understanding of species' responses to climate change (Gotelli & Stanton-Geddes, 2015; Ikeda et al., 2017), especially when human activities are already impinging upon intraspecific diversity (Mimura et al., 2017). However, comprehensive and robust assessment of genetic diversity remains an exhausting and challenging task.

Many recent broad-scale taxonomic assessments of intraspecific genetic diversity have been based on mitochondrial DNA barcoding techniques, usually with a 650 base pair (bp) fragment of the mitochondrial *CO1* gene (e.g. Dincă et al., 2015; Hebert, Penton, Burns, Janzen, & Hallwachs, 2004; Lohman et al., 2010; Mendoza et al., 2016; Smith, Woodley, Janzen, Hallwachs, & Hebert, 2006). However, mtDNA markers can follow an evolutionary trajectory independent of the true demographic history. Due to the haploid nature and largely uniparental inheritance of the mitochondria (mtDNA), it has a fourfold lower effective population size compared to the nuclear genome. This lower effective population size means mtDNA loses genetic diversity via genetic drift at a faster rate than the nuclear genome (Charlesworth, 2009). Hence, mtDNA will differentiate faster than the nuclear genome during periods of isolation and will complete the process of lineage sorting more rapidly than the nuclear counterpart (Funk & Omland, 2003). During range expansions (e.g. during interglacial periods) differentiated mtDNA lineages can meet, and in the absence of reproductive isolation, nuclear genomes can recombine and homogenize (e.g. Hinojosa et al., 2019). Biogeographic inconsistencies between mtDNA and nuclear genomic variation can also result from adaptive introgression of mtDNA or sex-biased asymmetries, such as sex-biased dispersal (reviewed in Toews & Brelsford, 2012). Insects in particular are also prone to infections by reproduction manipulating endosymbionts (e.g.

Wolbachia) that can lead to cytoplasmic incompatibilities allowing mtDNA haplotypes to hitchhike to fixation without concomitant nuclear differentiation (Hurst & Jiggins, 2005). Whilst mtDNA divergence within taxa is informative, the additional use of neutral nuclear genetic markers can more precisely describe evolutionary events, facilitate phylogeographic interpretations and delineate independent evolutionary lineages (Galtier, Nabholz, Glémin, & Hurst, 2009). Obtaining a set of comprehensive and informative nuclear genetic markers has been made easier with the advent of next generation sequencing and reduced representation genome libraries (e,g RADseq, double digest RADseq) has greatly alleviated this obstacle (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016).

There has been rapid and extensive application of DNA barcoding to understand the diversity and biogeography of butterflies (Ashfaq, Akhtar, Khan, Adamowicz, & Hebert, 2013; Dincă et al., 2015; Hebert et al., 2004; Lavinia et al., 2017) which are popularly considered as environmental indicators for environmental change (van Swaay, Nowicki, Settele, & van Strien, 2008). Using *CO1* barcodes Dinca *et al.* (2015) showed that more than a quarter of the European butterfly species have deeply divergent mitochondrial haplotypes that could potentially represent cryptic lineages. Evaluating the hypothesis of cryptic species needs to be tested using nuclear genomic data on a case-by-case basis. For two species, with multiple deeply divergent mitochondrial lineages, the use of genomic data has largely revealed mito-nuclear discordance resulting from introgression and *Wolbachia* infections (Dincă, Lee, Vila, & Mutanen, 2019) or demographic disparities arsing from multiple lineage splitting/fusions through the Pleistocene glacial-interglacial cycles (Hinojosa et al., 2019). However, the status of potential cryptic diversity of many other species with divergent mitochondrial lineages still remains to be addressed, particularly in common species that are not conservation priorities.

The Palearctic butterfly *Polyommatus icarus* is common and widespread, occupying a range of open biotopes including grasslands, sand dune systems and waste sites over a range of elevational gradients but is a hostplant specialist, with larvae feeding on low growing Fabaceae (chiefly *Lotus corniculatus*). Phylogeogrpahic analysis of this species across continental Europe and Asia led to the discovery of strongly divergent *CO1* lineages in small pockets across Iberia and in Crete, presumably from multiple expansion/contraction cycles during the Pleistocene (Dincă, Dapporto, & Vila, 2011). The phylogeography of this species in the UK remains unknown but an allozyme analysis suggested UK populations may have undergone a bottleneck (de Keyser, Shreeve, Breuker, Hails, & Schmitt, 2012), probably during the Younger Dryas period following colonisation in the early Holocene. It has

121 also been suggested that the colonization of the British Isles by *P. icarus* could have involved more
122 than one period of establishment following the Last Glacial Maximum (Dennis, 1977). Physiological
123 differences (Howe, Bryant, & Shreeve, 2007) have been identified between populations in different
124 parts of the British Isles, with northern populations differing from those in the south and those on the
125 isles of the Outer Hebrides being the most different, flying with lower thoracic temperatures than
126 elsewhere. Additionally, modelling flight activity responses under climate change scenarios reveal
127 profound differences between Outer Hebridian and mainland populations (Howe et al., 2007).
128 Differences in life-history strategies also exist, with northern populations being (potentially obligate)
129 univoltine whilst southern ones are facultative polyvotine (de Keyser, 2012; Matechou, Dennis,
130 Freeman, & Brereton, 2014).
131 Here we construct a phylogeography of *P. icarus* using *CO1* data with a special emphasis on the
132 British Isles. We then use genome-wide ddRADseq markers to determine if geographic variation in
133 nuclear DNA is concordant with mitochondrial variation in the British Isles. We also investigated the
134 influence of *Wolbachia* infection on the concordance between mitochondrial and genomic divergence
135 across the British Isles. We use this information to propose a possible invasion sequence and to
136 determine if there are distinct evolutionary lineages within the British Isles.
137

138 Materials and Methods

139 *Phylogenetic reconstruction of CO1 lineages*

140 To determine the phylogenetic relationships of British *P. icarus* with those elsewhere in Europe (Dincă
141 et al., 2011) we used all available *P. icarus cytochrome c oxidase subunit 1* (*CO1*) DNA sequences
142 archived in the Barcode of Life Data Systems (http://www.boldsystems.org/) database (Ratnasingham
143 & Herbert, 2007). Sequences were aligned using AliView (Larsson, 2014) and truncated or potential
144 contaminant sequences were removed, resulting in a final alignment of 187 specimens (Table S1)
145 with a length of 654 bp composed of protein coding sequence. The best-fit substitution model for each
146 codon position was determined using modeltest-NG (Darriba et al., 2019) and the *CO1* phylogeny
147 was reconstructed using BEAST v2.5.2 (Bouckaert et al., 2019) using a constant coalescent
148 population size tree prior. To estimate divergence times between lineages we applied a strict clock
149 model with a log normal prior distribution centred on the average of two previously published

150 substitution rates for invertebrates: 1.5% and 2.3% uncorrected pairwise distance per million years
151 (Brower, 1994; Quek, Davies, Itino, & Pierce, 2004). We executed two independent runs with an
152 MCMC chain length of 200 million and convergence was assessed with a 10% burning using
153 TRACER v1.7 (Rambaut, Drummond, Xie, Baele, & Suchard, 2018). The final time-calibrated
154 consensus tree was visualized and annotated using the package ggtree v.1.17.4 (Yu, Smith, Zhu,
155 Guan, & Lam, 2017)in R v.3.6.2 (R Core Team, 2019). Additionally, we constructed mitochondrial
156 haplotype networks using the *CO1* alignments and TCS networks as implemented in PopART v1.7
157 (Leigh & Bryant, 2015). An additional 10 individuals were filtered from this analysis due to ambiguous
158 bases or missing data.

159

160 ***Sample Collection for ddRAD sequencing of Genomic DNA***

161 We sampled 190 butterflies from 14 sites spread across the British Isles (Figure 1, Table S2) together
162 with a single site in central-southern France (Table S2). Numbers of individuals collected per site
163 varied between 6-15, with an average of 13 per locality. We aimed to collect similar numbers of males
164 and females from each site, but our samples are male biased, due to cryptic female behaviour.
165 Butterflies were sexed based on wing colouring dimorphism and abdominal tip morphology. We
166 removed heads and legs of individuals anesthetized on ice, and stored these in 95% ethanol for DNA
167 extraction. Wings and bodies were dried and stored separately as specimen vouchers.

168

169 ***ddRADseq Library Construction, Sequencing and SNP Filtering***

170 DNA was extracted from head and legs using a salt extraction protocol (Miller, Dykes, & Polesky,
171 1988) and eluted in 60 µl of dH_20. DNA was quantified using a Qubit 2.0 flourometer (Life
172 Technologies) using a Qubit dsDNA high sensitivity assay kit (Life Technologies) and individual DNA
173 samples for ddRADseq libraries were normalized to10 ng/µl. Library preparation was performed at
174 Floragenex (Portland, Oregon) following a protocol similar to Han et al. (2018) using a digestion with
175 PstI/MseI along with a SBG 100-Kit v2.0 (Keygene N.V., Wageningen, the Netherlands). Barcoded
176 samples were sequenced over two lanes at the University of Oregon Genomics and Cell
177 Characterization Facility (Eugene, Oregon) on a HiSeq4000 with single-end 100bp chemistry.

178 We used Stacks 2.4 (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) to assemble RAD loci
179 and call SNP genotypes from the raw ddRADseq data. Raw reads were demultiplexed using the
180 process_radtags.pl script while discarding low quality reads (< 10 average Phred score) and removing
181 any restriction site tags. After demultiplexing and discarding low quality reads, three individuals were
182 removed from the final assembly due to low number of reads (<500,000; Table S3). We initially used
183 a subset of 24 individuals to determine optimal combinations of the major parameters (*m* :minimum
184 number of raw reads required to call a stack, *M*: number of mismatches allowed between stacks, *n*:
185 number of mismatches of allowed between loci of different individuals) involved in assembling RAD
186 loci with Stacks following guidelines in Paris, Stevens, and Catchen (2017) and Rochette and
187 Catchen (2017). We varied values of *m* from 2-12, while holding *M* and *n* constant at 2, and evaluated
188 how the number of RAD loci and number of polymorphic loci present in 80% of the samples (r80 rule;
189 Paris et al., 2017) stabilized as a function of m. After, obtaining a suitable value for *m,* we varied *M*
190 and *n* from 1-8, with the constraint that *M*=*n* (Rochette & Catchen, 2017), and used the r80 rule and
191 checked for stability of the proportion of loci with 1-5 SNPs to determine suitable values for *M* and *n*.
192 This procedure revealed that the number of total and polymorphic RAD loci begin to stabilize around a
193 value for 4 for all three major parameters (*m*, *M,* and *n;* Figure S1). The value of *M*=*n*=4 at *m*=4 was
194 sufficient to stabilize the distribution of loci with 1-5 SNPS (Figure S2). This combination of *m*:4-*M*:4-
195 *n*:4 was used to assemble RAD loci for the entire dataset of 187 individuals. The average number of
196 reads per individual was 2.95 million (standard deviation (sd): 1.09 million; Table S3). The average
197 coverage per locus after assembling stacks was 45.8x (sd: 14.57x). SNP markers with minimum allele
198 frequencies of < 0.05 and a maximum observed heterozygosity > 0.65 (to exclude potential
199 paralogues) were further excluded. To obtain a set of widely available loci for downstream population
200 genomic analysis, we only retained SNPs that were present in at least 50% (*r*-50) of the individuals
201 within the 15 sampled localities (*p*-15; Figure 1) which yielded 4852 loci, 1915 of which were
202 monomorphic.

203 Next, to remove any contamination of our ddRADseq markers with mitochondrial DNA we used
204 Centrifuge v.1.0.4 (Kim, Song, Breitwieser, & Salzberg, 2016) to search all RAD loci against NCBI's
205 database of mitochondrial RefSeq Genomes (https://www.ncbi.nlm.nih.gov/genome/organelle/), which
206 includes complete mitogenomes of several Lepidopteran species. We also excluded contamination
207 from *Wolbachia* using Centrifuge to search RAD loci against the Archaeal and Bacterial RefSeq

208 Genomes in the NCBI database

209 (ftp://ftp.ccb.jhu.edu/pub/infphilo/centrifuge/data/p_compressed_2018_4_15.tar.gz).

210 Finally, we used vcftools v0.1.17 (Danecek et al., 2011) to further exclude any loci with greater than 5
211 SNPs (potentially erroneous loci) and then generated two SNP data sets filtered on levels of missing
212 data per individual: one that excluded individuals with > 50% missing data (*p*15*r*50miss50) and
213 another excluding individuals with > 25% missing data (*p*15*r*50miss25). This resulted in two biallelic
214 SNP datasets with 2,824 loci and 5,592 SNPs. The dataset *p*15*r*50miss50 consisted of 176
215 individuals and the dataset *p*15*r*50miss25 had 148 individuals. We generated two additional marker
216 datasets, using the exact same procedure as above, but with a more stringent requirement to only
217 retain SNPS that were present in at least 60% (*p*15*r*60miss25) or 70 % (*p*15*r*70miss25) of the
218 individuals within each locality.

219

220 ***Population Structure based on ddRADseq data***

221 To examine genomic level population structure we conducted Principal Component Analysis (PCA) on
222 the Stacks derived and filtered SNP datasets using the package ade4 v.1.7-13 (Dray & Dufour, 2007)
223 in R v3.6.2 (R Core Team, 2019). To assess the impact of missing data in reconstructing population
224 structure we performed PCAs on the two datasets filtered for individuals with different thresholds of
225 missing data (*p*15*r*50miss50 or *p*15*r*50miss25). As linked SNPs can influence population clustering
226 techniques like PCA we performed a further PCA on the *p*15*r*50miss25 dataset but retaining only a
227 single SNP per RAD locus. Additionally, to assess the influence of varying number of RAD markers
228 on population structure, we also conducted PCA on SNP datasets generated by applying more
229 stringent inclusion criteria on the availability of loci (datasets *p*15*r*60miss25 or *p*15*r*70miss25). We
230 calculated pairwise Weir and Cockerham F_{st} between all 15 populations using the R package dartR
231 1.1.11 (Gruber, Unmack, Berry, & Georges, 2018) on the *p*15*r*50miss25 dataset. Statistical
232 significance between each pairwise F_{st} was determined using 10000 bootstrap replicates. We used
233 fineRadstructure and RADpainter (Malinsky, Trucchi, Lawson, & Falush, 2018) to assess fine-scale
234 population structure based on shared genetic co-ancestry also using the *p*15*r*50miss25 SNP dataset.

235

236

237 ***Predicting Wolbachia Infection in Individuals***

238 To predict *Wolbachia* infection in each individual, we searched each read from demultiplexed
239 individual fastq files (generated in the initial stages of the Stacks 2.4 pipeline using process_radtags)
240 against the index of NCBI's database of Archaeal and Bacterial RefSeq Genomes
241 (ftp://ftp.ccb.jhu.edu/pub/infphilo/centrifuge/data/p_compressed_2018_4_15.tar.gz) using Centrifuge
242 v.1.0.4 (Kim et al., 2016). We used Pavian (Breitwieser & Salzberg, 2019) to summarize results from
243 Centrifuge. To quantify *Wolbachia* infection level in a given individual we calculated the number of
244 reads mapping to a single (most common) *Wolbachia* sp. as a fraction of the total reads mapped to
245 any bacterial or archaeal genome. Previous predictions of *Wolbachia* infection status based on short
246 read data from whole genome shotgun libraries have been highly successful (Richardson et al., 2012;
247 98.8% concordant with PCR-based results). Additionally, Illumina read depth have been shown to be
248 a reliable proxy for *Wolbachia* titre, showing strong correlation with copy number estimated from
249 quantiative PCR (Early & Clark, 2013) . Differences in proportions of infected individuals between
250 localities were determined using a Fisher's Exact test with Bonferroni adjustment for multiple
251 comparisons in R v3.6.2.

252

253 ***Generating ddRADseq SNPs for Wolbachia***

254 In order to generate *Wolbachia* genotypes for the subset of infected individuals we used seqtk 1.3-
255 r106 (https://github.com/lh3/seqtk) to retain only reads tagged by Centrifuge as mapping to the most
256 common *Wolbachia* taxon. This set of filtered reads for the infected individuals were processed
257 through the same final Stacks 2.4 pipeline as above (except the initial run of process_tags.pl) with
258 parameters *m*:4-*M*:4-*n*:4 and excluding genotypes with allele frequencies of < 0.05 and a maximum
259 observed heterozygosity > 0.65. The assembled RAD loci were used as BLAST queries against all
260 available *Wolbachia* assemblies on NCBI to confirm their taxonomic identity and to determine
261 *Wolbachia* supergroup affiliation. Further, all loci were queried using BLAST against a database
262 containing all proteins from complete *Wolbachia* RefSeq assemblies.

263 To obtain a set of loci for assessing genetic variation in the *Wolbachia* genotypes, we retained SNPs
264 that were present in at least 50% (*r*-50) of the individuals in at least 3 of the populations (*p*-3). From
265 this we generated two SNP marker datasets, one in which loci with greater than 5 SNPs were

266 excluded and individuals with >50% missing data were removed and another included all individuals.
267 We performed PCA on the *Wolbachia* SNP dataset with the missing data individual exclusion criteria
268 to determine clustering of *Wolbachia* genotypes.

269

270 ***Testing association between Wolbachia strains and mitotypes***

271 To assess the congruence between *Wolbachia* genotypes and *P. icarus* mitochondrial haplotypes, we
272 further sequenced an additional 655 bp fragment of CO1 for a subset of 31 individuals. The fragment
273 was amplified by PCR using primers piLepF1 (5`-TCTACAAATCATAAAGATATTGGAAC-3`) and
274 LepR1 (5`-TAAACTTCTGGATGTCCAAAAAATCA-3`) (Hebert et al., 2004) using OneTaq
275 Mastermix with standard buffer (New England Biolabs) under standard cycling conditions. The
276 resulting sequences were trimmed for primers and aligned using AliView (Larsson, 2014). These
277 CO1sequences and concatenated SNPs from *Wolbachia* genotypes derived from the Stacks pipeline
278 (the dataset with >50% missing data individual exclusion criteria and without) were clustered
279 independently using bitwise distance (or Hamming's distance) with the unweighted pair group mean
280 averaging (UPGMA) algorithm and 1000 bootstrap replicates for support using the R package poppr
281 v.2.8.3 (Kamvar, Tabima, & Grünwald, 2014) and clustering dendograms were visualized and
282 annotated using the R package ggtree v.1.17.4

283

284 ***Identifying Sex-Linked loci to Investigate Wolbachia Induced Feminization***

285 In order to investigate potential feminization in *Wolbachia* infected females we sought to identify sex-
286 linked SNPs to establish genetic sex for each individual. Butterflies generally posses a chromosomal
287 ZW/ZZ sex determination mechanism where females are the heterogametic (ZW or ZO) sex (Traut,
288 Sahara, & Marec, 2007). Hence, female-specific sex marker (loci polymorphic in females but
289 homozygous in males), assuming partial homology between Z and W chromosomes, should help to
290 determine the genetic sex of an individual butterfly: discordance between genetic and morphological
291 sex of infected females would be consistent with physiological feminization. To identify female-specific
292 sex markers we first fitted a baseline generalized linear model (using a logit link function) with
293 morphological sex as the dependent variable and PC1 and PC2 from the PCA of SNPs as the
294 independent variables. We then fitted an additional model by adding a single SNP marker as an

295 additional independent variable, and iterated this for all SNPS in the dataset. A significant association
296 between a SNP marker and morphological sex was determined by performing a likelihood ratio test of
297 the baseline model and the baseline model with the added SNP term and using a strict Bonferroni
298 corrected 5% type I error rate. Significant markers were further excluded if they had more than 2
299 genotypic classes, were not homozygous for males or happened to be a single SNP from a RAD
300 locus harbouring multiple SNPS. Generalized linear models and likelihood ratio tests were performed
301 using the base *glm() and anova()* functions in R v3.6.2.

302 .

303 Results

304 *Mitochondrial phylogeography of P. icarus*

305 The Bayesian phylogenetic reconstruction of 187 CO1 sequences (Figure 2A, Figure S3) was
306 largely congruent with one published earlier for European continental and Eurasian CO1 haplotypes
307 by Dincă et al. (2011) identified five main monophyletic lineages. TCS haplotype networks
308 corroborated the genealogy for *CO1,* revealing three major splits (Figure S4) with the Iberian
309 lineages (Alicante-Provence and Sierra Nevada) consisting of the most divergent haplotypes.
310 Nineteen individuals from Spain, Italy and other Palaearctic regions could not be reliably assigned to
311 any clades (bottom left Figure 2A, Unclassified samples in Figure S4).

312 Estimates of node ages (Figure 2A) are also consistent with those from Dincă et al. (2011). The
313 estimated node age for the most recent common ancestor (MRCA) for the Cretean/Sierra
314 Nevada/Alicante Provence and Iberian-Italy/Palaearctic lineages was 1.81 million years ago (mya)
315 with a 95% Bayesian credible interval (BCI) of 0.93-2.99 mya. The Cretean lineage (here represented
316 by a single sequence) and Alicante-Provence/Sierra Nevada lineages emerged the earliest, around
317 1.11 mya (95%BCI: 0.53-1.90 mya), while the split between Alicante-Provence and Sierra Nevada
318 lineages is estimated at 0.7 mya (95% BCI: 0.29-1.23 mya). The Palaearctic and Iberia-Italy lineages
319 emerged around 0.73 mya (95% BCI: 0.35-1.25 mya).

320 Samples from the British Isles did not form a single lineage (Figure 2A), but exhibited strong
321 geographic clustering (Figure 2B). Those from the Outer Hebrides together with some from the
322 adjacent Scottish mainland were part of the Alicante-Provence lineage (also occurring in Kazakhstan,
323 Norway and Germany). Those from southwestern parts of the British mainland and Wales were part

324 of the main Palaearctic group, whilst those from central and northern parts of mainland Britain were
325 part of the Iberian-Italian lineage.

326

327 ***Population structure of British Isles P. icarus using Genome-wide ddRADseq SNPs***

328 PCA analyses of the ddRADseq SNP data filtered for individuals with either >25%
329 (*p*15*r*50miss25) or >50% (*p*15*r*50miss50) missing data and with the dataset filtered for 25% missing
330 data (*p*15*r*50miss25) but using unlinked SNPs produced virtually identical results (Figure 3 and Figure
331 S5A-B). Datasets with smaller number of markers based on more stringent inclusion criteria (datasets
332 (*p*15*r*60miss25 and (*p*15*r*70miss25) also yield similar results but the signal decays with decreasing
333 number of markers (Figure S5C-D).

334 The first component, accounting for 6.8% of the variation, separates the northern Scottish samples
335 (BER, TUL, MLG, DGC, and OBN) from all the southern and southwestern and Welsh samples
336 (Figure 3). The six individuals from southern France (FRN), RHD (northern England) and RVS
337 (southern Scotland) samples fell between these extremes along the first PC. The second component,
338 accounting for 1.7% of the variation, distinguishes the Outer Hebrides (BER, TUL) from the northern
339 Scottish mainland locations (MLG, DGC, OBN) (Figure 3). Pairwise F_{st} values between northern
340 Scottish and all southern populations suggest moderate levels of differentiation (0.083-.128; Figure
341 S6), while those between Outer Hebrides and the northern Scottish mainland locations suggest small
342 levels of differentiation (0.05-0.066, Figure S6). Analysis of shared genetic co-ancestry using
343 fineRADstructure presents qualitatively similar clustering of individuals (Figure S7). The resulting
344 clustered coancestry matrix (Figure S7) revealed three major populations consisting of the (i) northern
345 Scottish samples, (ii) southern and central Great British samples, and (iii) the 6 individuals from
346 southern France. There was evidence of further substructure in the northern Scottish population with
347 the Outer Hebrides (BER, TUL) samples forming their own distinct sub-cluster within the northern
348 Scottish cluster (Figure S7). Individuals from RVS and RHD clustered together with the French
349 population but did not display any distinct structure, sharing genetic ancestry with both the northern
350 mainland Scottish and southern British populations.

351

352

353 **Prediction of *Wolbachia* infection**

354 We used Centrifuge to match all demultiplexed reads, for all 190 individuals, to NCBI's
355 RefSeq genomes of archaeal and bacterial genomes. On average 5.07 % (sd: ±1.96%) of the total
356 reads per individual were classified as matching against an archaeal or bacterial genome (Table S4)
357 with most matches being to bacterial genomes (5.05 ± 1.96 %; Table S4). The most frequently
358 encountered *Wolbachia* genome was identified as endosymbiont of *Drosophila simulans, w*No (NCBI
359 txid: 77038). The total percentage of classified reads with matches to this genome varied by several
360 orders of magnitude across individuals (min= 0.00036%, max=53.56%, Table S4). There was no
361 relationship between the total number of raw reads and the percentage of classified reads mapping to
362 this taxon (*Spearman's rank correlation* = 0.131, *P*= 0.07544; Figure S8A). There was a natural
363 discontinuity in the percentage of classified reads mapping to *Wolbachia* across all individuals (Figure
364 S8B) and this metric clearly had a bimodal distribution. Using a threshold of $\log_2$ (percentage of
365 classified reads mapping to *Wolbachia*) > 0 to classify an individual as infected or not, only individuals
366 from BER, TUL, DGC, MLG, OBN and one individual from RHD were classified as infected with
367 infection percentages ranging from 87.5% (14/16 individuals), in TUL in the Outer Hebrides, to 8%
368 (1/13) in RHD near Durham (Figure 4A). The proportion of infected populations differed significantly
369 between RHD and all other localities; however, all other pairwise comparisons were insignificant
370 (Table S5). All females from the Outer Hebrides (BER, TUL) were infected with a high percentage of
371 classified reads mapping to *Wolbachia* (min=15.6%, max=53.56%; Figure 4B). However, no such
372 dimorphism in number infected or percentage mapped reads was apparent in the mainland
373 populations (DGC, MLG, OBN; Figure 4B).

374

375 **Clustering of *Wolbachia* ddRADseq genotypes**

376 For all individuals classified as infected, we selected only raw reads classified as matching
377 the most frequently encountered *Wolbachia* genome (NCBI txid: 77038). We then processed these
378 selected reads using the same Stacks pipeline as for *P. icarus* genotypes. Blast searches revealed
379 that 117 out of 124 loci were 100% identical to at least one *Wolbachia* genome assembly from NCBI,
380 and a further 5 were >=98.5% identical to *Wolbachia* sequences in NCBI Genbank. BLAST searches
381 against NCBI Genbank further showed that the remaining 2 loci best matched *Wolbachia* sequences,

382 however, identities were low (94.7% and 92.6%); however, these loci haboured no variants and hence
383 were not used in any downstream analsyes. The closest matching *Wolbachia* genomes for all loci
384 were from supergroup B strains and 115 out of 124 loci could be assigned to a *Wolbachia* protein with
385 sequence identities of >=95% (Table S6). Matching loci were mostly housekeeping genes, and
386 showed an enrichment for hypothetical proteins . These loci were further filtered to remove any loci
387 with >5 SNPs and any individuals with >50% missing data were removed. This final set of genotypes
388 included 37 loci with 68 SNPs across all 6 locations with infected individuals (Figure 4B). The first two
389 PCs of the *Wolbachia* SNP dataset explained 88.7% of the variation in the dataset (Figure S9), with
390 PC1 along accounting for 85% of the total variation. PC1 clearly distinguishes between *Wolbachia*
391 genotypes found in individuals from the Outer Hebrides (BER, TUL; hereafter referred to as *w*Ica1)
392 and the mainland locations with infected individuals (DGC, MLG, OBN, RHD; hereafter referred to as
393 *w*Ica2) apart from two individuals from the western coast of Scotland (MLGm002 &OBNm110; Figure
394 S9), which harboured *w*Ica1 (see below).

395

396 **Congruence between *CO1* haplotypes and *Wolbachia* strains**

397 In order to assess association of *Wolbachia* genotypes with CO1 haplotypes, we sequenced a
398 655 bp fragment of the CO1 gene from a subset of 31 infected individuals (Table S6). UPGMA
399 clustering of Hamming's distance between CO1 sequences recovered two clusters, with strong
400 support corresponding to the Alicante-Sierra Nevada and Iberia-Italy CO1 lineages (Figure 5). The
401 Alicante-Sierra Nevada cluster was composed entirely of individuals from Outer Hebrides except for
402 three individuals from the western coast of Scotland (MLGm002, MLGf010, & OBNm110; Figure 5).
403 UPGMA clustering of Hamming's distance between concatenated SNPs derived from the *Wolbachia*
404 genotypes dataset with and without the >50% missing data exclusion criteria were near identical, and
405 only the latter is shown (Figure 5). Clustering of *Wolbachia* SNPs also recovered two clusters, with
406 strong support, corresponding to strain *w*Ica1 (Outer Hebrides) and *w*Ica2 (mainland). The three
407 individuals from the mainland (MLGm002, MLGf010, & OBNm110; Figure 5) possessing the Alicante-
408 Sierra Nevada haplotype also carried the *w*Ica1 strains. Overall there is perfect association between
409 individuals bearing *Wolbachia* strain *w*Ica1 and the Alicante-Sierra Nevada CO1 haplotype and
410 *Wolbachia* strain *w*Ica2 and the Italy-Iberia CO1 haplotype (although not all individuals with this
411 haplotype are infected by *Wolbachia*) regardless of geographical locality (Figure 5).

412 ***Sex-specific markers and feminization in Outer Hebrides***

413 *Wolbachia* dosage can influence phenotypic outcomes in the host (Arai, Lin, Nakai, Kunimi, &
414 Inoue, 2020; Breeuwer & Werren, 1993), thus the high levels of *w*Ica1 levels observed in females of
415 the Outer Hebrides population could be indicative of *Wolbachia* induced feminization. In this
416 scenario, morphological males should all be homozygous and morphological females should be
417 heterozygous for female-specific markers (homozygous for Z but a novel allele for W chromosome).
418 Discordance between the latter would suggest potential feminization of males. We used a
419 combination of association analysis and filtering to identify such a set of 10 putatively female-specific
420 SNPs (Table S7) across 7 RAD loci. Quite strikingly 5 out of 10 SNPs (across 3 RAD loci) show
421 perfect concordance of morphological sex with genetic sex except for females carrying the *w*Ica1
422 strain (Figure 6), The additional 5 SNPs also show complete discordance for morphological and
423 genetic sex for *w*Ica1 infected females but also includes a small number of homozygote of females
424 from uninfected or *w*Ica2 carrying individuals (Table S7).

425

426 ## Discussion

427 Using both mitochondrial (*CO1*) (British Isles and European scales) and our putative nuclear
428 genetic markers (British Isles) (our ddRADseq data) we have identified regionalisation of genetic
429 lineages of the widespread *Polyommatus icarus*. Genetic structuring based on either sets of markers
430 are largely consistent in inferences of the number of genetic clusters but the magnitude of genomic
431 divergence is clearly not equivalent to mtDNA divergence with evidence for genomic introgression
432 and the existence of an admixture zone between two of the three lineages in the British Isles. This
433 pattern of admixture is consistent with the distribution of *CO1* lineages at the European scale. We
434 have also identified two strains of *Wolbachia* within the butterfly in the British Isles, with distinct
435 geographic occurrences and strict mitotype associations, one of which may be potentially feminizing
436 strain**.**

437 The butterfly is described as relatively mobile (Asher et al., 2001; Cowley et al., 2001) and early work
438 at the European scale using allozymes revealed little genetic differentiation over mainland Europe,
439 and thus no geographic regionalisation, although samples from the British Isles have lower allelic
440 diversity than in mainland Europe (de Keyser et al., 2012). Our data contrasts with this broad finding,

441 demonstrating a pattern of geographic variation over a smaller spatial scale: the British Isles. Studies
442 using mitochondrial *CO1* and the nuclear gene *ITS-1* In mainland Europe (Dinca et al. 2011) have
443 identified five lineages with geographic structuring within mainland Europe, Using a larger number
444 and geographical spread of samples, we recovered a similar topology and estimates of node ages to
445 those reported in Dinca et al. (2011) suggesting that splits within mtDNA lineages occurred within the
446 Pleistocene glacial-interglacial cycles. Within the British Isles, there is concordance between mtDNA
447 and ddRADseq data in the inference of genetic clusters. The Outer Hebredian genomic cluster is
448 entirely comprised of the Alicante-Provence mitotype; the Northern Scottish genomic cluster is largely
449 comprised of the Ibierian-Italian mitotype; and the southern English and Welsh populations are largely
450 comprosed of the Palearctic mitotype. However, the magnitude of divergence between mtDNA and
451 genomic markers is clearly not equivalent, and there is only weak divergence at the genomic level.
452 Hence, the observed concordance between mtDNA and genomic data does not represent the
453 existence of deeply divergence lineages. Instead we suggest that the degree of genomic divergence
454 between these mtDNA lineages reflects the order of colonization, founder effects and subsequent
455 time available for genomic homogenization.

456 The earliest time period that *P.icarus* could have occurred within the British Isles is *c.* 20,000 years
457 B.P, following rapid warming at the end of the Last Glacial Maximum. It is also likely that colonization
458 of the islands will only have occurred prior to the permanent opening of the Channel sea-straight*, c.*
459 9,800 years B.P. Whilst our data does not cover the whole of the British Isles, and sampling of *P.*
460 *icarus* for *CO1* in northern parts of mainland Europe is less intense than in the Mediterranean zone,
461 our *CO1* data point to colonization of the British Isles by three lineages which are geographically
462 structured with little evidence of widespread co-existence in this island region. The work of Dinca et
463 al. (2011) corroborated by our analysis of the occurrence of these lineages within mainland Europe
464 indicates that the main Palearctic lineage is spatially mixed with the other lineages in most of
465 mainland Europe, north of Mediterranean zone glacial refugia. Based on the occurrence of the
466 Alicante-Provence lineage on the Outer Hebrides, with a few individuals on the adjacent Scottish
467 mainland (where the Iberian-Italian lineage occurs), we suggest that the Alicante-Provence lineage
468 was an early colonizer that was replaced by the Iberian-Italian lineage on mainland Britain. There is
469 very weak genomic divergence between the Outer Hebrides and the adjacent mainland which could
470 reflect the longer time available for gene-flow between these lineages. That the Iberian-Italian

471 haplotype is not present on the Outer Hebrides also indicates that these islands are isolated. We also
472 suggest that the distribution of the Palearctic lineage in the southwest and southern parts of the
473 British mainland, might represent, more recent, sequential replacement of the Iberian-Italian lineage.
474 Where this has occurred it is the replaced *CO1* haplotype that is very infrequent in the areas where
475 we are suggesting replacement (Figure 2B: south-west England). The genomic divergence between
476 the southern and northern population is stronger (compared to that between the Outer Hebrides and
477 northern mainland), which may result from lesser time available for gene-flow. Genomic data also
478 provides evidence for the existence of an admixture zone between the northern Scottish populations
479 (largely Iberia-Italy haplotype) and southern English and Welsh populations (largely Palearctic)
480 consistent with more recent contact between lineages. In summary we see some concordance
481 between mtDNA and genomic data and this weak concordance might arise from the shorter duration
482 for genomic mixing. Longer mixing timings would allow for complete homogenization and starker mito-
483 nuclear discordance as has been observed in continental European Lepidoptera (Dincă et al., 2019;
484 Hinojosa et al., 2019). Importantly, both mtDNA and genomic data are consistent with a
485 phylogeographic scenario comprising multiple colonizations of the British Isles followed by historical
486 and on-going genomic mixing. This is consistent with a colonization scenario (based on *CO1*) for the
487 whole of Europe and its islands, from possible refugial areas occupied at the Last Glacial
488 Maximum,with the most ancient lineages being found on ecological and physical islands (Sierra
489 Nevada and Crete), with the Alicante-Provence lineage possibly spreading from Mediterranean
490 refugia occupied at the Last Glacial Maximum throughout Europe and now persisting on the Outer
491 Hebrides, northern Ireland and northern Scandinavia, and Belle Isle (France) and isolated (scarce)
492 elsewhere. This is then replaced by the Iberian-Italian lineage, now currently persisting in Iberia and
493 Italy and also in Ireland, northern/central England and mainland Scotland and later replaced over
494 most of Europe (including southern England) by the current main European haplotype, representing a
495 possible eastwards expansion from the Balkan region.

496 However, the persistence and geographic structuring of divergent mitotypes in the British Isles still
497 requires some further explanation. The replacement of the Alicante-Provence lineage by the Iberian-
498 Italian lineage in the north and of the Iberian-Italian lineage by the Palearctic mitotype in the south has
499 involved genomic homogenization, with the eventual loss of the *CO1* mitotype of the replaced lineage
500 This loss would involve either female biased dispersal of the invading lineage, greater fitness of the

501 mitotypes of the invading lineage or *Wolbachia*-mediated CI. Female biased dispersal is unlikely in
502 this species as females display cryptic behaviour and are rarely observed while sampling as opposed
503 to males which are frequently encountered on the wing (pers. obs.). There is some support for
504 directional selection on mitochondria in the case of northwestern Scotland. Individuals from the Outer
505 Hebrides fly with lower body temperatures than those from further south (Howe et al., 2007). This
506 increases the chance of flight in the generally cooler and cloudier and windier conditions in the Outer
507 Hebrides than elsewhere, Whether this confers a disadvantage in warmer conditions on mainland
508 Scotland is not known, but flight activity involves respiratory enzymes encoded by mitochondria,
509 There is the possibility that in mainland conditions the mitotype in the Outer Hebrides are at a
510 disadvantage to others. Associations of mitotypes with particular climates, against on-going gene
511 flow, has been identified in vertebrates (Morales et al., 2018; Pavlova et al., 2013). It is worth noting
512 that divergence between *CO1* barcodes of the Alicante-Provence lineage found on the Outer
513 Hebrides and the Iberian-Italy lineage found on adjacent mainland the includes 2 non-synonymous
514 substitutions.

515 However, a more general explanation for the persistence of divergent mitotypes, consistent with our
516 data, is a recent *Wolbachia* sweep (CI; Hurst & Jiggins, 2005). Our predictions of *Wolbachia* infection
517 status mirrored genomic differentiation based on ddRADseq SNPs. Populations in the north
518 harboured a substantial proportion (>50% in all cases) of infected individuals, whereas no predictions
519 of infection were made for individuals in the south (Figure 4A) suggesting that individuals in the south
520 are not infected or not infected to the same extent. We found evidence for two distinct *Wolbachia*
521 strains, *w*Ica1 and *w*Ica2, that show perfect association with the Alicante-Provence and Italy-Iberia
522 mitotypes, respectively. Such genetic structuring of *Wolbachia* could potentially explain the
523 persistence of older mitotypes of the host butterfly under male-biased dispersal and CI between
524 infected males and uninfected females (and between males and females infected with different
525 strains). This could account for the observation that the entire admixture zone between Iberian-Italian
526 and Palearctic lineages is dominated by the Iberia-Italian mitotype. Furthermore, bidirectional CI
527 between *w*Ica1 and *w*Ica2 individuals coupled with imperfect vertical transmission can promote stable
528 coexistence of both strains (Telschow, Yamamura, & Werren, 2005), which could account for the
529 persistence of the mitotype in the Outer Hebrides. We did not detect any double-infected individuals
530 (i.e. those carrying both strains) which would be consistent with bidirectional CI. However, it remains

plausible that are our sequencing data is not sensitive enough to detect double infections. A more systematic analysis, involving laboratory crosses between individuals with different infection status, is required to determine the extent of CI between infected and uninfected individuals (and between strains). Accurate dating of infection events and exploration on infections on continental Europe, could offer more conclusive insight into the impact of *Wolbachia* on genetic structuring in the Common Blue butterfly. To conclude, although, we cannot currently exclude the role of direct selection in explaining the persistence of mitotypes in the north, our data cannot exclude the current mitotype structure in the British Isles resulting from *Wolbachia*-mediated CI.

We currently have no evidence for the capability of either *w*Ica1 or *w*Ica2 to induce CI, two observations do indicate a potential phenotypic effect of the *w*Ica1 strain. Firstly, all females from Outer Hebrides were predicted as infected with *w*Ica but not all males, whereas infection status on the mainland exhibited no such dimorphism. Second, Outer Hebridean females carried a *Wolbachia* load, based on our proxy metric for copy number, an order of magnitude higher than in females. *Wolbachia* density can be indicative of its phenotypic effect on the host (Arai et al., 2020; Breeuwer & Werren, 1993). Our observations here are consistent with *Wolbachia*-induced feminization of genetic males (Stouthamer, Breeuwer, & Hurst, 1999). Such feminization would predict that female-specific markers in a species with a ZW sex determination system, with the caveat of partial homology between Z and W chromosomes would yield male (homozygous) genotypes in “feminized” males. As expected all morphological females carrying the *w*Ica1 had male genotypes.

Feminization is a well-known reproductive manipulation strategy deployed by *Wolbachia* (Werren, Baldo, & Clark, 2008), however, it has not been encountered frequently within Lepidoptera (Duplouy & Hornett, 2018). The best documented instance of feminization in butterflies has been initiated by the discovery of sex-biased female lines in two species of pierid *Eurema* butterflies in Japan (Kato, 2000). *w*Fem occurs at low frequencies in natural populations of *Eurema* and has not been detected in males. A causative role of *w*Fem in feminization has been suggested by antibiotic treatment of infected larva which results in intersex individuals (Narita, Kageyama, Nomura, & Fukatsu, 2007) and antibiotic treatment of adult females leads to all male progeny (Kern, Cook, Kageyama, & Riegler, 2015). The *w*Fem pattern of infection contrasts with that of *w*Ica1, where both females and males can be infected, but females carry higher bacterial dose. If *Wolbachia*-induced feminization indeed occurs in Outer Hebridean *P. icarus,* it will add to the small number of potential model system to study the

561 molecular basis underlying this poorly understood process in Lepidoptera. However, more direct
562 evidence for *w*Ica1- induced feminization is required (Hiroki et al., 2002; Kageyama et al., 2017).

563

564 ***Conclusions***

565 In conclusion, we have identified three distinct mitochondrial haplotypes in the British Isles.
566 We found partial concordance in nuclear genomic divergence using ddRADSeq makers. Both mtDNA
567 and genomic data are consistent with a biogeographical scenario involving at least three independent
568 colonizations of the British Isles and are consistent with a possible colonization history of most of
569 mainland Europe from refugia occupied at the Last Glacial Maximum. We observe mito-nuclear
570 discordance on the mainland with staggered clines of mtDNA and genomic introgression between two
571 mitochondrial lineages. *Wolbachia* infection status on the mainland suggests this discordance likely
572 results from *Wolbachia*-induced CI coupled with male-biased dispersal. In contrast, the Outer
573 Hebrides are divergent with respect to mtDNA and genomic markers, and the former may be
574 associated with climatic adaptation on the islands. However, bidirectional CI between the Outer
575 Hebrides and the mainland could also account for the persistence of parapatric mtDNA halplotypes.
576 Furthermore, the potential existence of feminizing *Wolbachia* indicates the distinctiveness of the
577 populations on the Outer Hebrides. Because of the current isolation we strongly suggest that the
578 populations on the Outer Hebrides represent a distinct genetic and ecological evolutionary unit, that
579 warrants consideration as being of conservation interest and monitoring in its own right. Finally, the
580 eco-evolutionary dynamics of this *Wolbachia*-host system and their impact on past and future biology
581 of these butterflies warrants further investigaion.

582

583 **Acknowledgements**

584 This work was funded by Oxford Brookes University, Oxford, UK. We thank Bruce J. Riddoch for help
585 collecting samples. We would like to thank the following for permissions to sample on their land: Mark
586 Dinning and the Durham Wildlife Trust, Paul Nunns at the UK Forestry Commission, Alastair Gardner
587 and Jenny Loring at Natural England, Simone Bullion and the Suffolk Wildlife Trust, and Simeon LD
588 Jones and the Carmarthenshire County Council.

Author Contributions

S.A., T.G.S and L.D. designed the study and conducted preliminary analysis. S.A., T.G.S, M.D.S.N and W.H.H-M performed the field collections. S.A. performed all laboratory work. S.A., L.D., W.G.H-M., M.G. and M.D.S.N. performed data analysis. S.A. and T.G.S. wrote the manuscript with input from all authors.

611 **References**

612 Adams, M., Raadik, T. A., Burridge, C. P., & Georges, A. (2014). Global Biodiversity
613 Assessment and Hyper-Cryptic Species Complexes: More Than One Species of Elephant in
614 the Room? *Systematic Biology*, *63*(4), 518–533. doi: 10.1093/sysbio/syu017

615 Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016).
616 Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews*
617 *Genetics*, *17*(2), 81–92. doi: 10.1038/nrg.2015.28

618 Arai, H., Lin, S. R., Nakai, M., Kunimi, Y., & Inoue, M. N. (2020). Closely Related Male-Killing
619 and Nonmale-Killing Wolbachia Strains in the Oriental Tea Tortrix Homona magnanima.
620 *Microbial Ecology*, *79*(4), 1011–1020. doi: 10.1007/s00248-019-01469-6

621 Asher, J., Warren, M., Fox, R., Harding, P., Jeffcoate, G., & Jeffcoate, S. (2001). *Millennium*
622 *Atlas of Butterflies in Britain and Ireland* (1st ed.). Oxford, UK: Oxford Univeristy Press.

623 Ashfaq, M., Akhtar, S., Khan, A. M., Adamowicz, S. J., & Hebert, P. D. N. (2013). DNA
624 barcode analysis of butterfly species from Pakistan points towards regional endemism.
625 *Molecular Ecology Resources*, *13*(5), 832–843. doi: 10.1111/1755-0998.12131

626 Bálint, M., Domisch, S., Engelhardt, C. H. M., Haase, P., Lehrian, S., Sauer, J., … Nowak,
627 C. (2011). Cryptic biodiversity loss linked to global climate change. *Nature Climate Change*,
628 *1*(6), 313–318. doi: 10.1038/nclimate1191

629 Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., … Das, I.
630 (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology &*
631 *Evolution*, *22*(3), 148–155. doi: 10.1016/j.tree.2006.11.004

632 Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M.,
633 Gavryushkina, A., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform
634 for Bayesian evolutionary analysis. *PLOS Computational Biology*, *15*(4), e1006650. doi:
635 10.1371/journal.pcbi.1006650

636 Breeuwer, J. A., & Werren, J. H. (1993). Cytoplasmic incompatibility and bacterial density in
637 Nasonia vitripennis. *Genetics*, *135*(2), 565–574.

638 Breitwieser, F. P., & Salzberg, S. L. (2019). Pavian: interactive analysis of metagenomics
639 data for microbiome studies and pathogen identification. *Bioinformatics*, *36*(4), 1303–1304.
640 doi: 10.1093/bioinformatics/btz715

641 Brower, A. V. (1994). Rapid morphological radiation and convergence among races of the
642 butterfly Heliconius erato inferred from patterns of mitochondrial DNA evolution. *Proceedings*
643 *of the National Academy of Sciences*, *91*(14), 6491–6495. doi: 10.1073/pnas.91.14.6491

644 Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an
645 analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. doi:
646 10.1111/mec.12354

647 Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and
648 variation. *Nature Reviews Genetics*, *10*(3), 195–205. doi: 10.1038/nrg2526

649 Cowley, M. J. R., Thomas, C. D., Roy, D. B., Wilson, R. J., León-Cortés, J. L., Gutiérrez, D.,
650 … Gaston, K. J. (2001). Density–distribution relationships in British butterflies. I. The effect
651 of mobility and spatial scale. *Journal of Animal Ecology*, *70*(3), 410–425. doi:
652 10.1046/j.1365-2656.2001.00508.x

653 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … 1000
654 Genomes Project Analysis Group. (2011). The variant call format and VCFtools.
655 *Bioinformatics*, *27*(15), 2156–2158. doi: 10.1093/bioinformatics/btr330

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2019). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, *37*(1), 291–294. doi: 10.1093/molbev/msz189

de Keyser, R. (2012). *Spatial structuring in trait variation in Polyommatus icarus in a functional context. PhD Thesis.* Oxford Brookes University: UK.

de Keyser, R., Shreeve, T. G., Breuker, C. J., Hails, R. S., & Schmitt, T. (2012). Polyommatus icarus butterflies in the British Isles: evidence for a bottleneck. *Biological Journal of the Linnean Society*, *107*(1), 123–136. doi: 10.1111/j.1095-8312.2012.01925.x

Dennis, R. L. H. (1977). *The British Butterflies: Their Origin and Establishment,*. Faringdon, Englnad: E. W. Classey.

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*(18), 3921–3935. doi: 10.1111/j.1365-294X.2011.05223.x

Dincă, V., Lee, K. M., Vila, R., & Mutanen, M. (2019). The conundrum of species delimitation: a genomic perspective on a mitogenetically super-variable butterfly. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1911), 20191311. doi: 10.1098/rspb.2019.1311

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*(1), 12395. doi: 10.1038/srep12395

Dray, S., & Dufour, A.-B. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, *22*(1), 1–20. doi: 10.18637/jss.v022.i04

Duplouy, A., & Hornett, E. A. (2018). Uncovering the hidden players in Lepidoptera biology: the heritable microbial endosymbionts. *PeerJ*, *6*, e4629. doi: 10.7717/peerj.4629

Early, A. M., & Clark, A. G. (2013). Monophyly of Wolbachia pipientis genomes within Drosophila melanogaster: geographic structuring, titre variation and host effects across five populations. *Molecular Ecology*, *22*(23), 5765–5778. doi: 10.1111/mec.12530

Funk, D. J., & Omland, K. E. (2003). Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 397–423. doi: 10.1146/annurev.ecolsys.34.011802.132421

Galtier, N., Nabholz, B., Glémin, S., & Hurst, G. D. D. (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology*, *18*(22), 4541–4550. doi: 10.1111/j.1365-294X.2009.04380.x

Gotelli, N. J., & Stanton-Geddes, J. (2015). Climate change, genetic markers and species distribution modelling. *Journal of Biogeography*, *42*(9), 1577–1585. doi: 10.1111/jbi.12562

Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2018). dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*, *18*(3), 691–699. doi: 10.1111/1755-0998.12745

Han, K., Lee, H.-Y., Ro, N.-Y., Hur, O.-S., Lee, J.-H., Kwon, J.-K., & Kang, B.-C. (2018). QTL mapping and GWAS reveal candidate genes controlling capsaicinoid content in Capsicum. *Plant Biotechnology Journal*, *16*(9), 1546–1558. doi: 10.1111/pbi.12894

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences*, *101*(41), 14812–14817. doi: 10.1073/pnas.0406166101

Hinojosa, J. C., Koubínová, D., Szenteczki, M. A., Pitteloud, C., Dincă, V., Alvarez, N., & Vila, R. (2019). A mirage of cryptic species: Genomics uncover striking mitonuclear discordance in the butterfly Thymelicus sylvestris. *Molecular Ecology*, *28*(17), 3857–3868. doi: 10.1111/mec.15153

Hiroki, M., Kato, Y., Kamito, T., & Miura, K. (2002). Feminization of genetic males by a symbiotic bacterium in a butterfly, Eurema hecabe (Lepidoptera: Pieridae). *Naturwissenschaften*, *89*(4), 167–170. doi: 10.1007/s00114-002-0303-5

Howe, P. D., Bryant, S. R., & Shreeve, T. G. (2007). Predicting body temperature and activity of adult Polyommatus icarus using neural network models under current and projected climate scenarios. *Oecologia*, *153*(4), 857–869. doi: 10.1007/s00442-007-0782-3

Hurst, G. D. D., & Jiggins, F. M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1572), 1525–1534. doi: 10.1098/rspb.2005.3056

Ikeda, D. H., Max, T. L., Allan, G. J., Lau, M. K., Shuster, S. M., & Whitham, T. G. (2017). Genetically informed ecological niche models improve climate change predictions. *Global Change Biology*, *23*(1), 164–176. doi: 10.1111/gcb.13470

Kageyama, D., Ohno, M., Sasaki, T., Yoshido, A., Konagaya, T., Jouraku, A., … Sahara, K. (2017). Feminizing Wolbachia endosymbiont disrupts maternal sex chromosome inheritance in a butterfly species. *Evolution Letters*, *1*(5), 232–244. doi: 10.1002/evl3.28

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2*, e281. doi: 10.7717/peerj.281

Kato, Y. (2000). Overlapping Distribution of Two Groups of the Butterfly Eurema hecabe Differing in the Expression of Seasonal Morphs on Okinawa-jima Island. *Zoological Science*, *17*(4), 539–547. doi: 10.2108/0289-0003(2000)17[539:ODOTGO]2.0.CO;2

Kern, P., Cook, J. M., Kageyama, D., & Riegler, M. (2015). Double trouble: combined action of meiotic drive and Wolbachia feminization in Eurema butterflies. *Biology Letters*, *11*(5), 20150095. doi: 10.1098/rsbl.2015.0095

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729. doi: 10.1101/gr.210641.116

Larsen, B. B., Miller, E. C., Rhodes, M. K., & Wiens, J. J. (2017). Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life. *The Quarterly Review of Biology*, *92*(3), 229–265. doi: 10.1086/693564

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276–3278. doi: 10.1093/bioinformatics/btu531

Lavinia, P. D., Bustos, E. O. N., Kopuchian, C., Lijtmaer, D. A., García, N. C., Hebert, P. D. N., & Tubaro, P. L. (2017). Barcoding the butterflies of southern South America: Species delimitation efficacy, cryptic diversity and geographic patterns of divergence. *PLOS ONE*, *12*(10), e0186845. doi: 10.1371/journal.pone.0186845

Leigh, J. W., & Bryant, D. (2015). popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution*, *6*(9), 1110–1116. doi: 10.1111/2041-210X.12410

Lohman, D. J., Ingram, K. K., Prawiradilaga, D. M., Winker, K., Sheldon, F. H., Moyle, R. G., … Meier, R. (2010). Cryptic genetic diversity in “widespread” Southeast Asian bird species suggests that Philippine avian endemism is gravely underestimated. *Biological Conservation*, *143*(8), 1885–1890. doi: 10.1016/j.biocon.2010.04.042

Malinsky, M., Trucchi, E., Lawson, D. J., & Falush, D. (2018). RADpainter and fineRADstructure: Population Inference from RADseq Data. *Molecular Biology and Evolution*, *35*(5), 1284–1290. doi: 10.1093/molbev/msy023

Matechou, E., Dennis, E. B., Freeman, S. N., & Brereton, T. (2014). Monitoring abundance and phenology in (multivoltine) butterfly species: a novel mixture model. *Journal of Applied Ecology*, *51*(3), 766–775. doi: 10.1111/1365-2664.12208

Mendoza, Á. M., Torres, M. F., Paz, A., Trujillo-Arias, N., López-Alvarez, D., Sierra, S., … Gonzalez, M. A. (2016). Cryptic diversity revealed by DNA barcoding in Colombian illegally traded bird species. *Molecular Ecology Resources*, *16*(4), 862–873. doi: 10.1111/1755-0998.12515

Miller, S. A., Dykes, D. D., & Polesky, H. F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research*, *16*(3), 1215–1215. doi: 10.1093/nar/16.3.1215

Mimura, M., Yahara, T., Faith, D. P., Vázquez-Domínguez, E., Colautti, R. I., Araki, H., … Hendry, A. P. (2017). Understanding and monitoring the consequences of human impacts on intraspecific variation. *Evolutionary Applications*, *10*(2), 121–139. doi: 10.1111/eva.12436

Morales, H. E., Pavlova, A., Amos, N., Major, R., Kilian, A., Greening, C., & Sunnucks, P. (2018). Concordant divergence of mitogenomes and a mitonuclear gene cluster in bird lineages inhabiting different climates. *Nature Ecology & Evolution*, *2*(8), 1258–1267. doi: 10.1038/s41559-018-0606-3

Narita, S., Kageyama, D., Hiroki, M., Sanpei, T., Hashimoto, S., Kamitoh, T., & Kato, Y. (2011). Wolbachia-induced feminisation newly found in Eurema hecabe, a sibling species of Eurema mandarina (Lepidoptera: Pieridae). *Ecological Entomology*, *36*(3), 309–317. doi: 10.1111/j.1365-2311.2011.01274.x

Narita, S., Kageyama, D., Nomura, M., & Fukatsu, T. (2007). Unexpected Mechanism of Symbiont-Induced Reversal of Insect Sex: Feminizing Wolbachia Continuously Acts on the Butterfly Eurema hecabe during Larval Development. *Applied and Environmental Microbiology*, *73*(13), 4332–4341. doi: 10.1128/AEM.00145-07

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution*, *8*(10), 1360–1373. doi: 10.1111/2041-210X.12775

Pavlova, A., Amos, J. N., Joseph, L., Loynes, K., Austin, J. J., Keogh, J. S., … 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*(12), 3412–3428. doi: 10.1111/evo.12107

Quek, S.-P., Davies, S. J., Itino, T., & Pierce, N. E. (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*(3), 554–570. doi: 10.1111/j.0014-3820.2004.tb01678.x

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. doi: 10.1093/sysbio/syy032

Ratnasingham, S., & Herbert, P. D. N. (2007). bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Richardson, M. F., Weinert, L. A., Welch, J. J., Linheiro, R. S., Magwire, M. M., Jiggins, F. M., & Bergman, C. M. (2012). Population Genomics of the Wolbachia Endosymbiont in Drosophila melanogaster. *PLOS Genetics*, *8*(12), e1003129. doi: 10.1371/journal.pgen.1003129

Rochette, N. C., & Catchen, J. M. (2017). Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols*, *12*(12), 2640–2659. doi: 10.1038/nprot.2017.123

Smith, M. A., Woodley, N. E., Janzen, D. H., Hallwachs, W., & Hebert, P. D. N. (2006). DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3657. doi: 10.1073/pnas.0511318103

Stouthamer, R., Breeuwer, J. A. J., & Hurst, G. D. D. (1999). Wolbachia Pipientis: Microbial Manipulator of Arthropod Reproduction. *Annual Review of Microbiology*, *53*(1), 71–102. doi: 10.1146/annurev.micro.53.1.71

Struck, T. H., Feder, J. L., Bendiksby, M., Birkeland, S., Cerca, J., Gusarov, V. I., … Dimitrov, D. (2018). Finding Evolutionary Processes Hidden in Cryptic Species. *Trends in Ecology & Evolution*, *33*(3), 153–163. doi: 10.1016/j.tree.2017.11.007

Telschow, A., Yamamura, N., & Werren, J. H. (2005). Bidirectional cytoplasmic incompatibility and the stable coexistence of two Wolbachia strains in parapatric host populations. *Journal of Theoretical Biology*, *235*(2), 265–274. doi: 10.1016/j.jtbi.2005.01.008

Toews, D. P. L., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*(16), 3907–3930. doi: 10.1111/j.1365-294X.2012.05664.x

Traut, W., Sahara, K., & Marec, F. (2007). Sex Chromosomes and Sex Determination in Lepidoptera. *Sexual Development*, *1*(6), 332–346. doi: 10.1159/000111765

Trontelj, P., & Fišer, C. (2009). Perspectives: Cryptic species diversity should not be trivialised. *Systematics and Biodiversity*, *7*(1), 1–3. doi: 10.1017/S1477200008002909

van Swaay, C. A. M., Nowicki, P., Settele, J., & van Strien, A. J. (2008). Butterfly monitoring in Europe: methods, applications and perspectives. *Biodiversity and Conservation*, *17*(14), 3455–3469. doi: 10.1007/s10531-008-9491-4

Werren, J. H., Baldo, L., & Clark, M. E. (2008). Wolbachia: master manipulators of invertebrate biology. *Nature Reviews Microbiology*, *6*(10), 741–751. doi: 10.1038/nrmicro1969

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. doi: 10.1111/2041-210X.12628

Data Availability Statement

CO1 sequences are deposited in NCBI Genbank with accession numbers listed in Table S6. Raw ddRADSeq reads can be found on the NCBI SRA under accessions SRRXXXXX and SRRXXXXX. Barcodes for de-multliplexing, processed data files and Bash and R scripts for the analyses can be found on github.

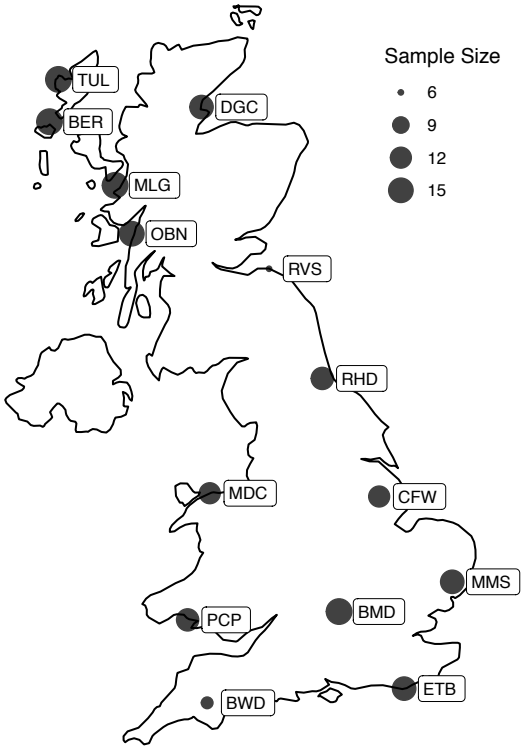


Figure 1. A map of the British Isles with geographical locations of *Polyommatus icarus* sampled for this study. An additional 6 individuals collected from southern France are not shown (see Table S2 for full details). Size of circles is proportional to the number of samples acquired at each locality.

Figure 2 (A) Bayesian ultrametric tree inferred from 187 CO1 *Polyommatus icarus* sequences. Numbers above nodes represent Bayesian posterior support (only where probabilities were > 0.75). The bars on each node represent the 95% Bayesian credible interval for estimates of node age. The scale is in millions of years. Clades with > 0.75 posterior support have been collapsed (see figure S3 for uncollapased tree) and coloured based on correspondence to the CO1 lineage classification from Dincă et al. (2011). **(B)** Bubble plot of the geographical distribution of the 51 Great British samples in the dataset. Colours are based on their correspondence to the lineage classification from Dincă et al. (2011). Size of the bubble represents the total number of CO1 lineages at a particular sitee

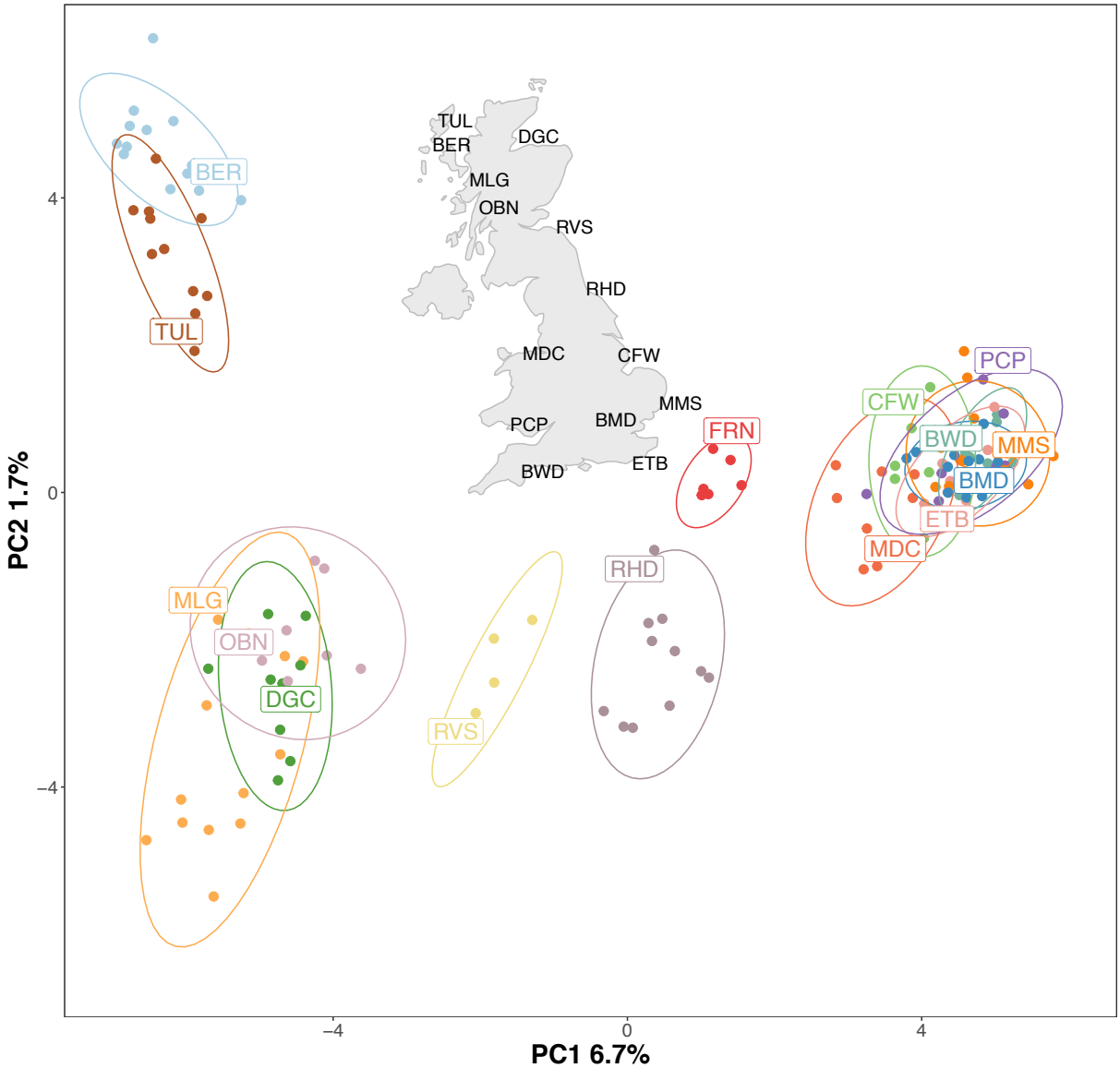


Figure 3 Population structure based on principal component analysis (PCA) of 5592 SNPs (2824 RAD loci) of 148 *Polyomattus icarus* individuals across the British Isles. Ninety-% Confidence Interval (CI) ellipses for PC1 and PC2 for each locality are also shown. Each locality is represented by a distinct colour and inset provides a labeled map of the sampled localities for quick geographical reference.

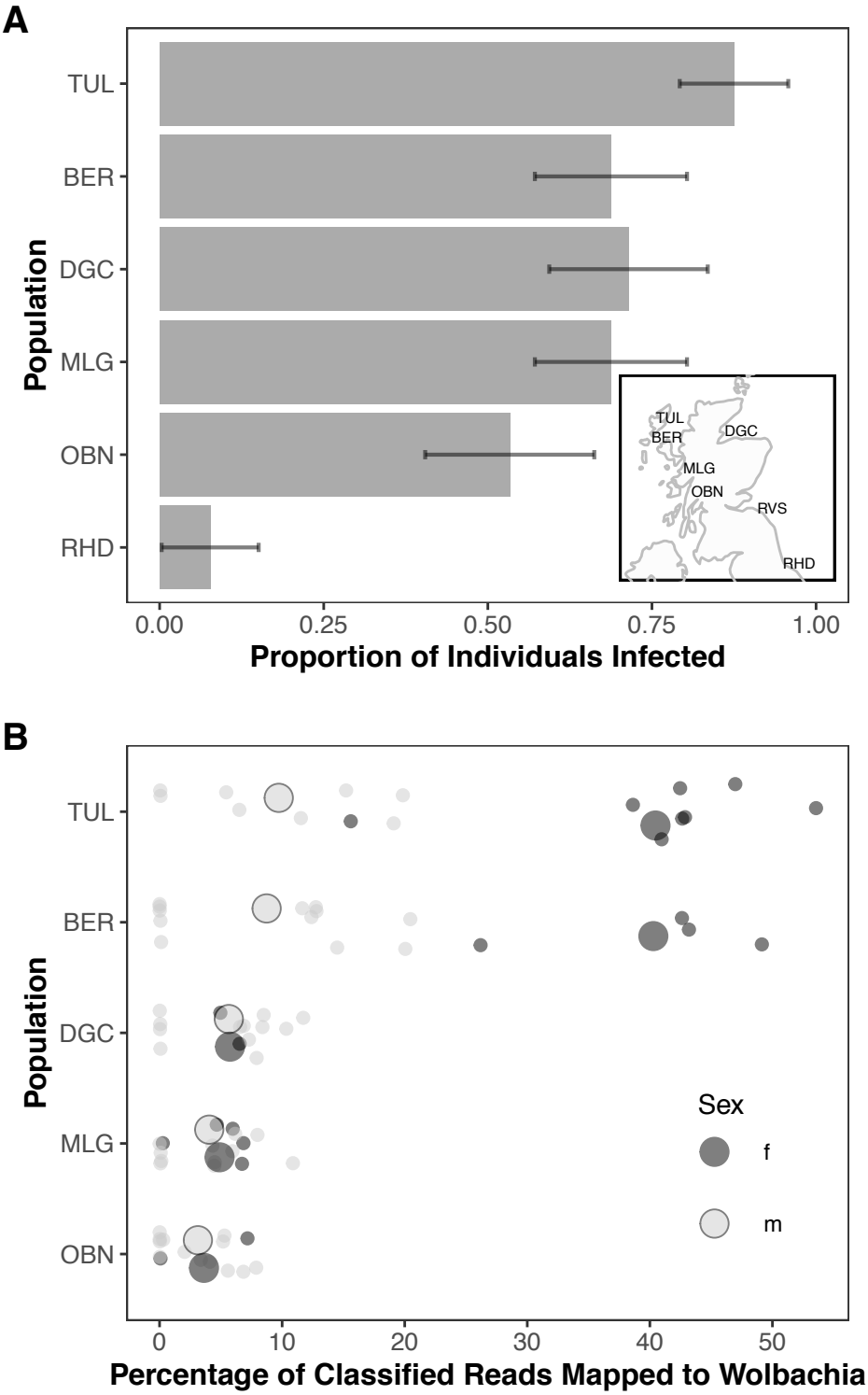
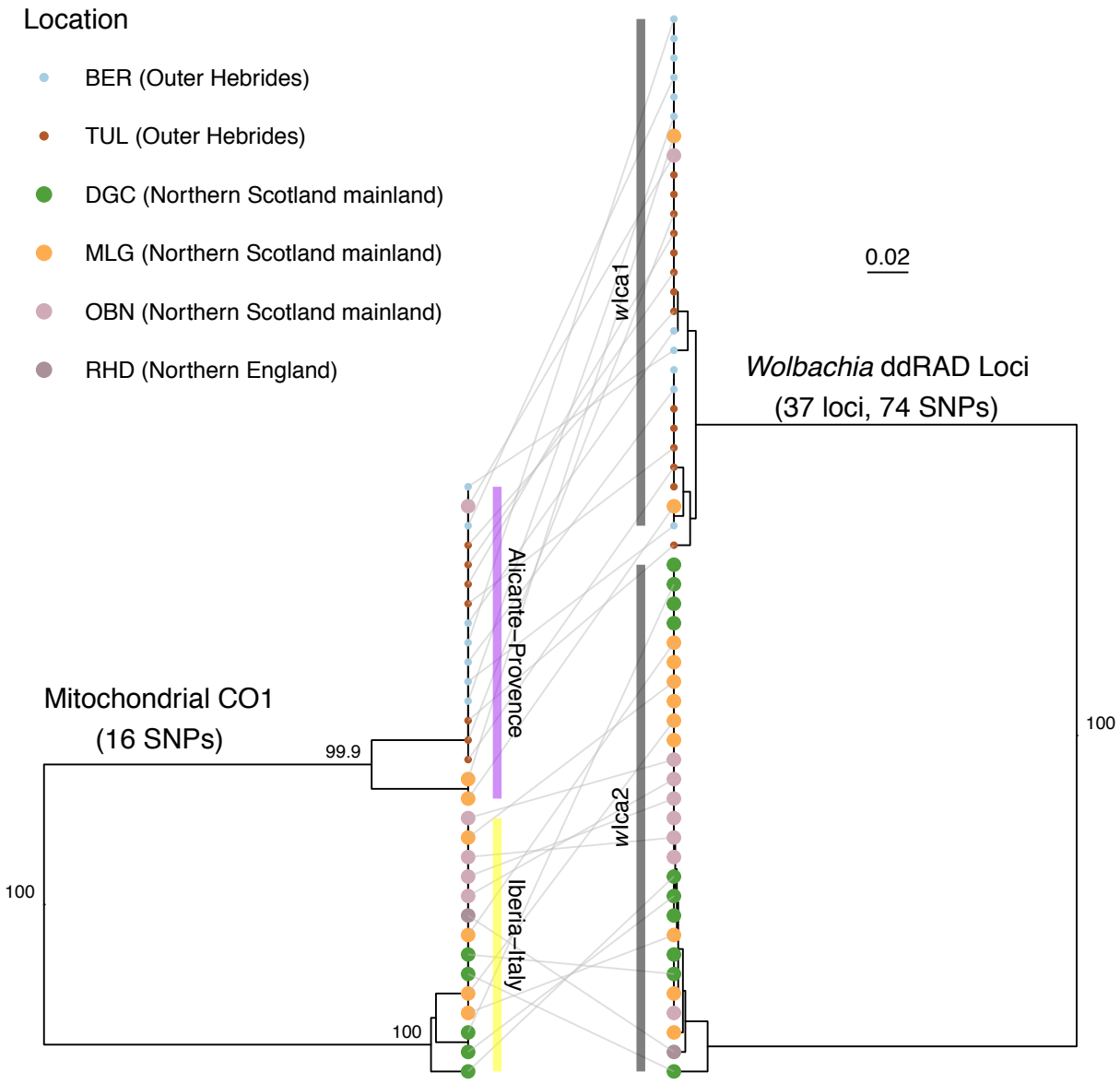


Figure 4 (A) Proportion of individuals classified as infected based on a threshold of $\log_2$(percentage of classified reads mapping to Wolbachia) > 0. Only populations with at least one infected individual are shown. Error bars are the standard errors of the estimated proportions. Inset map spans the geographical range of locations habouring infected individuals. **(B)** Distributions of the percentage of classified reads mapping to *Wolbachia* for locations with individuals classified as infected. RHD with a single infected male is not shown. Larger circles represent the average for each sex within locations.



873

874 **Figure 5** UPGMA clustering of mitochondrial CO1 fragments from 31 *Polyommatus icarus* individuals
875 based on bitwise distance (left). UPGMA clustering of 74 concatenated SNPs from ddRADseq
876 *Wolbachia* genotypes derived from reads mapping to *Wolbachia* from 55 infected individuals (right).
877 Numbers on nodes represent bootstrap branch support values based on 1000 bootstrap replicates,
878 values < 70% are not shown. Large circles represent individuals from mainland populations (DGC,
879 MLG, OBN, RHD) and smaller circles represent individuals from the Outer Hebrides (BER, TUL).
880 Lines between dendograms connect the *CO1* haplotype and *Wolbachia* strain for the same individual.
881 Scale bar reflects the proportion of loci that are different.

882

883

884

885

886

887

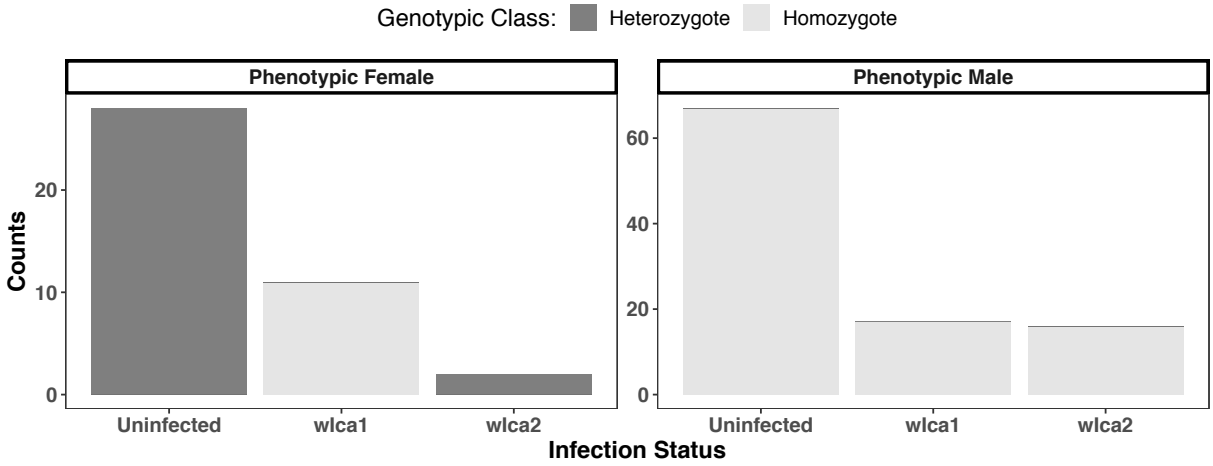


Figure 6 Feminization of *w*Ica1infected females as suggested by discordance between morphological sex, based on abdominal tip morphology and dimorphic wing patterning, and genetic sex based on female-specific markers. Discordance is exemplified by data on a single marker here (11011_27), see Table S7 for all sex-linked loci. All *w*Ica1 infected morphological females (all females in the Outer Hebrides and one from the mainland (MLGf010) are homozygous for female-specific markers (heterozygous ZW), consistent with feminization of males.