Whole-chromosome hitchhiking driven by a male-killing endosymbiont

Simon H. Martin*[1,2], Kumar Saurabh Singh[3], Ian J. Gordon[4], Kennedy Saitoti Omufwoko[5,6], Steve Collins[7], Ian A. Warren[2], Hannah Munby[2], Oskar Brattström[2], Walther Traut[8], Dino J. Martins[5,6], David A. S. Smith[9], Chris D. Jiggins[2], Chris Bass[3], Richard H. ffrench-Constant[3]

[1]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom

[2]Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, United Kingdom

[3]Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Penryn, TR10 9EZ, United Kingdom

[4]BirdLife International Kigali Office, Box 2527, Kigali Post Office, Kigali, Rwanda

[5]Mpala Research Centre, P O Box 555, Nanyuki, 10400 Kenya.

[6]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA

[7]African Butterfly Research Institute, Box 14308-0800, Nairobi, Kenya

[8]Institut für Biologie, Zentrum für medizinische Struktur- und Zellbiologie, Universität Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany

[9]Natural History Museum, Eton College, Windsor SL4 6DW, UK

*Correspondence to: simon.martin@ed.ac.uk

Abstract

Neo-sex chromosomes are found in many taxa, but the forces driving their emergence and spread are poorly understood. The female-specific neo-W chromosome of the African monarch (or queen) butterfly *Danaus chrysippus* presents an intriguing case study because it is restricted to a single 'contact zone' population, involves a putative colour patterning supergene, and co-occurs with infection by the the male-killing endosymbiont *Spiroplasma*. We investigated the origin and evolution of this system using whole genome sequencing. We first identify the 'BC supergene', a broad region of suppressed recombination across nearly half a chromosome, which links two colour patterning loci. Association analysis suggests that the genes *yellow* and *arrow* in this region control the forewing colour pattern differences between *D. chrysippus* subspecies. We then show that the same chromosome has recently formed a neo-W that has spread through the contact zone within ~2200 years. We also assembled the genome of the male-killing *Spiroplasma*, and find that it shows perfect genealogical congruence with the neo-W, suggesting that the neo-W has hitchhiked to high frequency as the male killer has spread through the population. The complete absence of female crossing-over in the Lepidoptera causes whole-chromosome hitchhiking of a single neo-W haplotype, carrying a single allele of the BC supergene, and dragging multiple non-synonymous mutations to high frequency. This has created a population of infected females that all carry the same recessive colour patterning allele, making the phenotypes of each successive generation highly dependent on uninfected male immigrants. Our findings show how hitchhiking can occur between the unlinked genomes of host and endosymbiont, with dramatic consequences.

Introduction

Structural changes to the genome play an important role in evolution by altering the extent of recombination among loci. This is best studied in the context of chromosomal inversions that cause localised recombination suppression, and can be favoured by selection if they help to maintain clusters of co-adapted alleles (or 'supergenes') in the face of genetic mixing [1–4]. A greater extent of recombination suppression occurs in the formation of heteromorphic sex chromosomes, which can link sex-specific alleles similarly to supergenes [5]. However, suppressed recombination can also have costs. In particular, male-specific Y and female-specific W chromosomes can be entirely devoid of recombination, making them vulnerable to genetic hitchhiking and the accumulation of deleterious mutations through 'Muller's ratchet', which may explain their deterioration over time [6–8]. These contrasting benefits and costs of recombination suppression are of particular interest in the evolution of neo-sex chromosomes, which can form through fusion of autosomes to existing sex chromosomes. There is accumulating evidence that neo-sex chromosomes are common in animals [9–15], but the processes underlying their emergence, spread and subsequent evolution have not been widely studied. In particular, there are few studied examples of recently-formed neo-sex chromosomes that are not yet fixed in a species.

The African monarch (or queen) butterfly *Danaus chrysippus*, provides a unique test case for the causes and consequences of changes in genome architecture and recombination suppression. Like its American cousin (*D. plexippus*), it feeds on milkweeds and has bright colour patterns that warn predators of its distastefulness. However, within Africa *D. chrysippus* is divided into four subspecies with distinct colour patterns and largely distinct ranges (Fig. 1A). Predator learning should favour the maintenance of a single monomorphic warning in any single area. For this reason, researchers have long been puzzled by the large polymorphic contact zone in East and Central Africa where all four *D. chrysippus* subspecies meet and interbreed [16–18] (Fig. 1A). Crosses have shown that colour pattern differences between the subspecies are controlled by Mendelian autosomal loci, including the tightly linked 'B' and 'C' loci (putatively a 'BC supergene' [19]) that define three common forewing patterns [20,21] (Fig. 1A). However, crosses with females from the contact zone revealed that the BC chromosome has become sex linked, forming a neo-W that is unique to this population [19,22]. Since female meiosis is achiasmatic (it lacks crossing-over) in the Lepidoptera, the formation of a neo-W would

instantaneously cause perfect linkage, not just of the B and C loci, but of an entire non-recombining chromosome, along with other maternally-inherited DNA.

What is particularly striking is that the presence of the neo-W coincides with infection by a maternally-inherited 'male killer' endosymbiont related to *Spiroplasma ixodetis,* which kills male offspring and leads to highly female-biased sex-ratios where infection is common [22–24]. The combination of neo-W and male killing is expected to dramatically alter the inheritance and evolution of the BC chromosome [22,25]: Infected females typically give rise to all-female broods who should always inherit the same colour patterning allele on their neo-W, along with the male-killer, while the other maternal allele is systematically eliminated in the dead sons (Fig. 1B), forming a genetic sink for all colour pattern alleles not on the neo-W. It has been suggested that the restriction of male killing to females with the neo-W, and only in the region in which hybridisation occurs between subspecies, may not be a coincidence [19,22,25–27]. However, the genomic underpinnings of this system – the genetic controllers of colour pattern, the source and spread of the neo-W, and its relationship with the male killer – have until now remained a mystery. We generated a reference genome for *D. chrysippus* and used whole genome sequencing of population samples to uncover the interconnected evolution of the BC supergene, neo-W and *Spiroplasma.* Our findings reveal a recent whole-chromosome selective sweep caused by hitchhiking between the host and endosymbiont genomes.

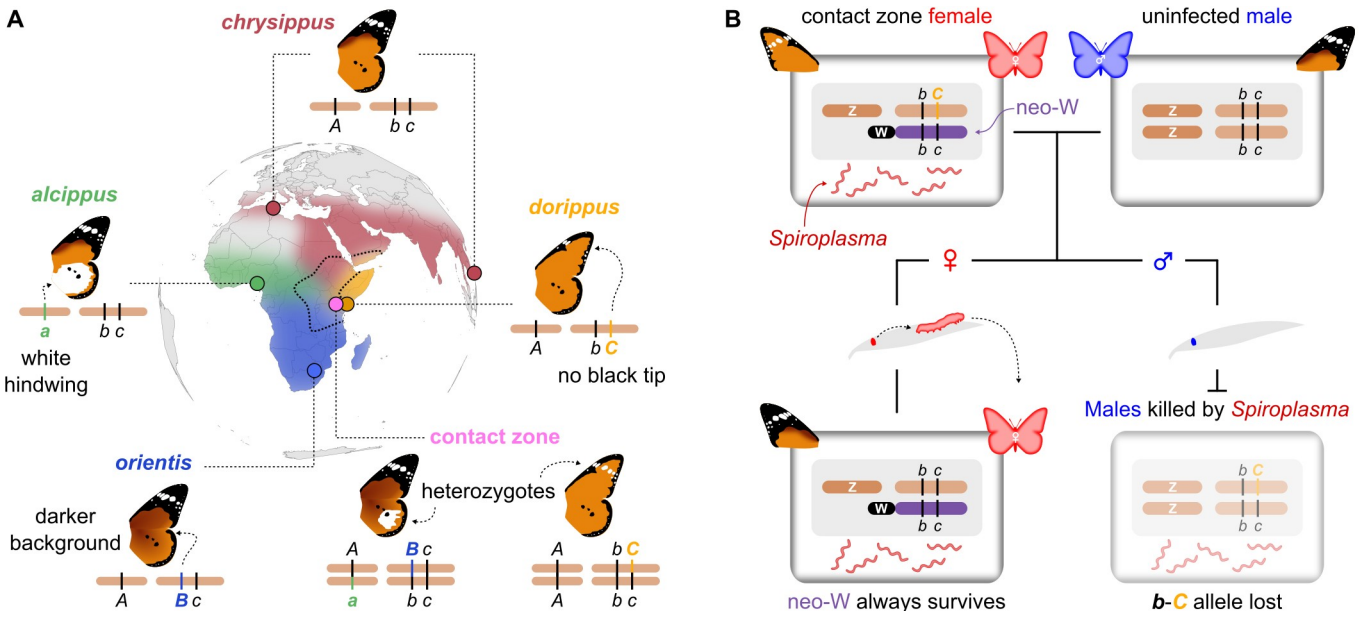


Fig 1. Geography and genetics of colour pattern. (**A**) Approximate ranges of the four subspecies of *D. chrysippus*, with the contact zone outlined. Sampling locations for each of the subspecies and the contact zone are indicated. Cartoon chromosomes show the genotypes of each subspecies at the A (white hindwing patch), B (brown background colour) and C (forewing tip) colour patterning loci, based on previous crosses [20]. Note the linkage of B and C, putatively forming a 'BC supergene' [19]. Two examples of heterozygotes that can be found in the contact zone are shown. Note that *Cc* heterozygotes can exhibit the *transiens* phenotype with white markings on the forewing with ~50% penetrance. (**B**) Model showing how fusion of the BC autosome to the W chromosome has produced a neo-W (purple) in contact zone females (top left), while males have two autosomal copies of the BC chromosome (top right). Daughters inherit the neo-W, while sons inherit the other BC chromosome haplotype from their mother. The latter allele is then lost due to male killing by *Spiroplasma*.

Results and Discussion

Identification of the BC supergene

We assembled a high quality draft genome for *D. chrysippus,* with a total length of 322 Megabases (Mb), a scaffold N50 length of 0.63 Mb, and a BUSCO [28] completeness score of 94% (S1 Table-S8 Table). We then further scaffolded the genome into a pseudo-chromosomal assembly based on homology with the *Heliconius melpomene* genome [29–31] accounting for known fusions that differentiate these species [9,30,32] (Fig. S1). We also re-sequenced 42 individuals representing monomorphic populations of each of the four subspecies and a polymorphic population from a known male-killing hotspot near Nairobi, in the contact zone (Fig 1A, S9 Table).

111 To identify the putative BC supergene, we scanned for genomic regions showing high
112 differentiation between the subspecies and an association with colour pattern. Genetic
113 differentiation (F_{ST}) and excessive divergence (d_{XY}) is largely restricted to a handful of broad
114 peaks, with a background level of approximately zero (Fig. 2A, S2 Fig, S3 Fig). This low
115 background level implies a nearly panmictic population across the continent. The effective
116 population size appears to be very large, as average genome-wide diversity at putatively neutral
117 4-fold degenerate 3rd codon positions is 0.042, which is among the highest values reported for
118 animals [33,34]. The islands of differentiation that stand out from this background imply
119 selection for local adaptation maintaining particular differences between the subspecies, similar
120 to patterns seen between geographic races of *Heliconius* butterflies [35]. However, here the
121 peaks of differentiation are broad, covering several Mb, implying some mechanism of
122 recombination suppression such as inversions that differentiate the subspecies.

123 The inclusion of the polymorphic contact-zone samples, and the fact that three of the
124 subspecies each carry a unique colour pattern allele (Fig 1A), allowed us to identify particular
125 differentiated regions associated with the three major colour pattern traits. A ~3 Mb region on
126 chromosome 4 is associated with the white hindwing patch (A locus) and a ~5 Mb region on
127 chromosome 15 (hereafter chr15) is associated with both background orange/brown (B locus)
128 and the forewing black tip (C locus) (Fig 2A, S2 Fig). Below, we refer to this region on chr15,
129 which spans over 200 protein-coding genes, as the BC supergene [19], although we note that
130 additional associated SNPs on chromosome 22 suggest that background wing melanism may also
131 be influenced by other loci.

132 Clustering analysis based on genetic distances reveals three clearly distinct alleles at the
133 BC supergene (Fig 2C). This further supports the hypothesis of recombination suppression,
134 although a number of individuals show mosaic ancestry consistent with occasional
135 recombination (S4 Fig). The three main alleles correspond to the three common forewing
136 phenotypes, so we term these $BC^{chrysippus}$ (orange background with black forewing tip, formerly
137 *bbcc*), $BC^{dorippus}$ (orange without black tip, formerly *bbCC*), and $BC^{orientis}$ (brown background with
138 black forewing tip, formerly *BBcc*) (Fig 2C). Fifteen of the twenty contact zone individuals are
139 heterozygous, carrying two distinct BC alleles, and a few carry putative recombinant alleles, as
140 do some of the southern African *orientis* individuals (S4 Fig). As shown previously, $BC^{dorippus}$

(which includes the dominant *C* allele) and $BC^{orientis}$ (which includes the dominant *B* allele) are both dominant over the recessive $BC^{chrysippus}$ (Fig. 2C, S4 Fig).

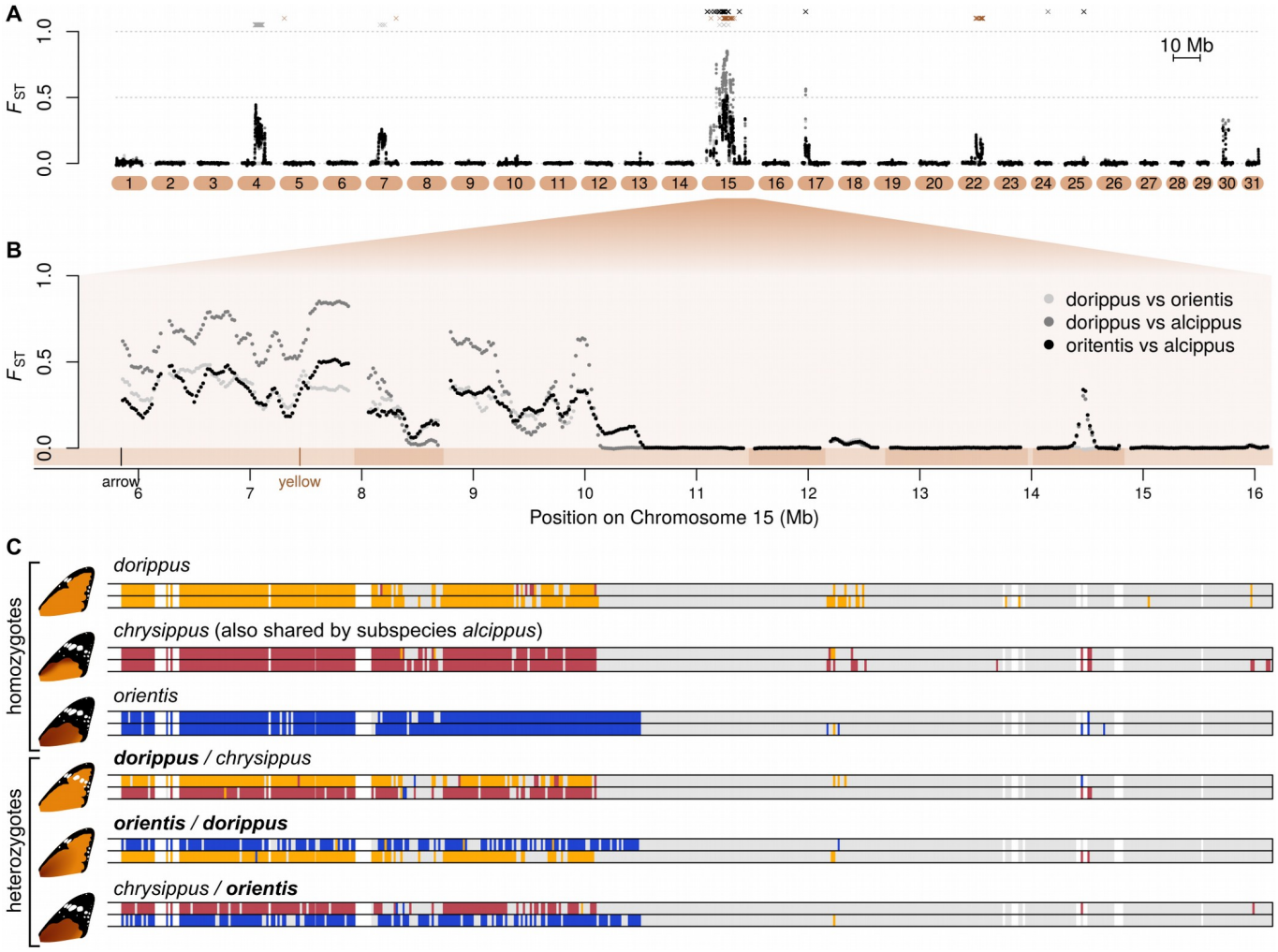


Fig 2. Identification of the BC supergene on chromosome 15. (**A**) Pairwise genetic differentiation (F_{ST}), plotted in 100 kb sliding windows with a step size of 20 kb across all chromosomes. Three subspecies pairs with sample sizes ≥ 6 are shown (see legend in panel B), revealing strong overlap in patterns of differentiation. Locations of SNPs most strongly associated with the A, B and C loci (Wald test, 99.99% quantile) are plotted above in grey, brown and black, respectively. See S2 Fig for a more detailed plot. (**B**) Expanded plot of F_{ST} across a 10 Mb portion of chromosome 15 (chr15). Note that the first ~6 Mb of the chromosome is not shown due to complex structural variation (see main text). Scaffolds are indicated below the plot in alternating shades. The locations of our most likely candidate genes for B (*yellow*) and C (*arrow*) are indicated. (**C**) Allelic clustering on chr15 in six representative individuals: three homozygotes and three heterozygotes (see S4 Fig for all individuals, and see panel B for chromosome positions). Coloured blocks indicate 20 kb windows in which sequence haplotypes could be assigned to one of three clusters based on pairwise genetic distances (see Methods for details). Windows in grey indicate insufficient relative divergence to be assigned to a cluster and white indicates missing data. Allele names reflect the subspecies from which they come. In heterozygotes, the name of the dominant allele is in bold.

159 Although it can be challenging to identify particular functional mutations in regions of
160 suppressed recombination, the presence of some recombinant individuals allowed us to narrow
161 down candidate regions for the B and C loci. A cluster of SNPs most strongly associated with
162 background colour (B locus) is found just upstream of the gene *yellow*, and a phylogenetic
163 network for a 30 kb region around *yellow* groups individuals nearly perfectly by phenotype,
164 although some individuals classed as heterozygous were intermingled with homozygotes (S5
165 Fig). Differential expression of Yellow in *Drosophila* is associated with different levels of
166 melanism [36] and *yellow* knockouts in other butterflies show reduced melanin pigmentation
167 [37], making this a compelling candidate for the background colour polymorphism. The
168 strongest associations with forewing tip (C locus) occur at the gene *arrow*, and a phylogenetic
169 network for a 100 kb region around this gene similarly clusters individuals by phenotype (S5
170 Fig). In *Drosophila*, Arrow is essential for Wnt signalling in wing development [38]. Wnt
171 signalling is known to underlie variation in colour pattern in *Heliconius* butterflies [39] and
172 knock-out mutants for the Wnt ligand gene *WntA* in *D. plexippus* show a loss of pigmentation
173 [40]. This makes *arrow* a promising candidate for the C locus gene. While these genes represent
174 our best candidates, numerous strongly-associated SNPs occurred closer to other genes in this
175 region (S10 Table). Future studies will aim to narrow down the associated regions and
176 performing functional validation.

177 Irrespective of their precise mode of action, the patterns of association imply that the B and
178 C loci are ~1.6 Mb apart (S5 Fig), and would therefore be fairly loosely linked under normal
179 recombination. This physical distance translates to around 7.6 cM, assuming crossover rates
180 similar to those in *Heliconius* [31,41], whereas the estimated recombination distance between B
181 and C based on crosses is 1.9 cM [42]. Theory predicts that recombination suppression can be
182 favoured if it maintains linkage disequilibrium (LD) between co-adapted alleles in the face of
183 gene flow [1–4]. Our study is one of only a few cases in which it can be shown that alleles at
184 distinct loci that each influence a component of a complex trait are maintained in LD by
185 suppressed recombination [43,44].

186 It is likely that chromosomal rearrangements contribute to recombination suppression at
187 the BC supergene. Although our short-read data do not allow us to test directly for inversions,
188 they do reveal dramatic variation in sequencing coverage over the proximal end of the

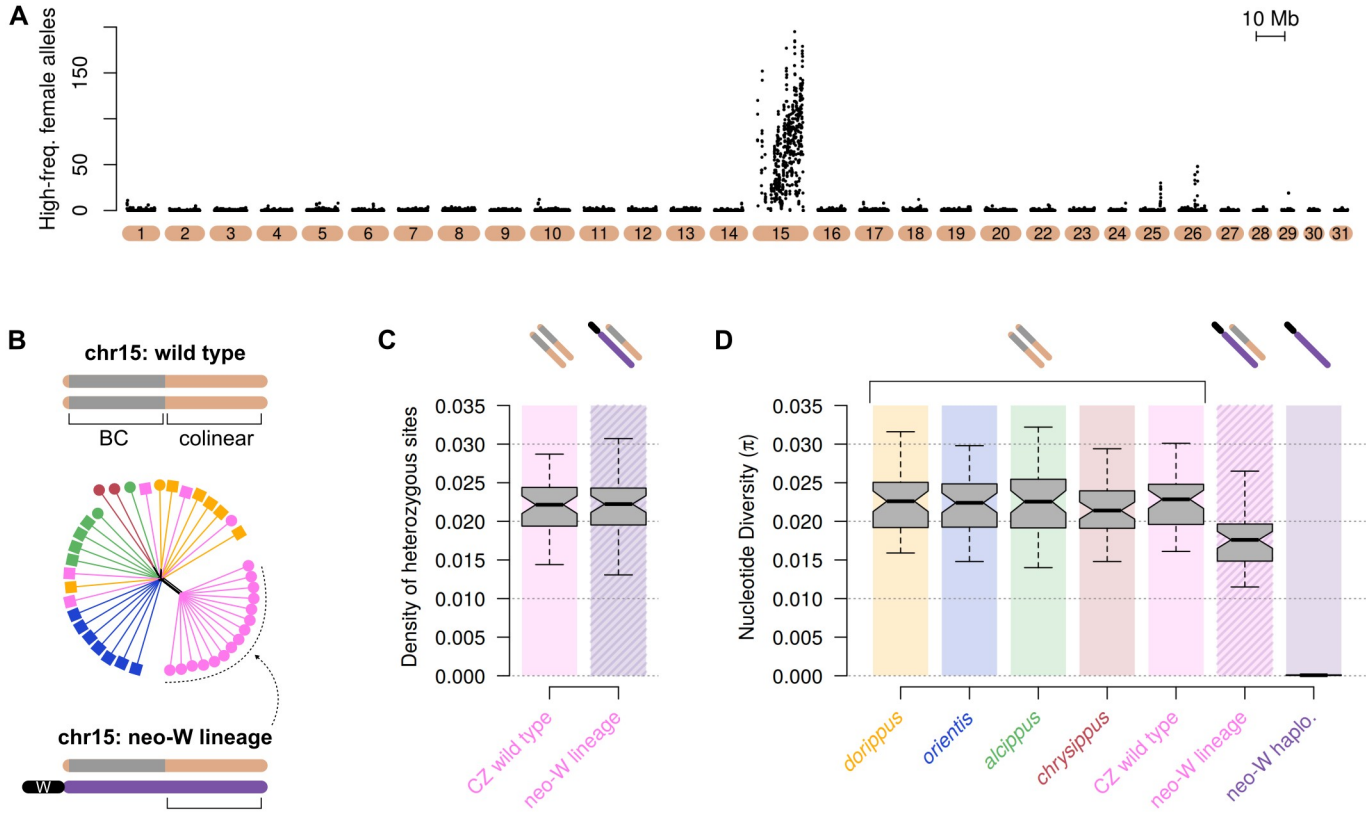
189 chromosome. Comparison of coverage among individuals suggests a large (~5 Mb) polymorphic
190 insertion in this region that tends to occur in individuals carrying the $BC^{dorippus}$ allele (S6 Fig).
191 Synteny comparison with *H. melpomene* reveals that this insertion involves an expansion in copy
192 number of a region of several hundred kb. Comparison of copy numbers for two of the genes in
193 the expansion with several other species confirms that it is derived in *D. chrysippus* (S7 Fig).
194 The expansion appears to occur just a few kb from the coding region of *arrow* (S6 Fig), and is
195 also perfectly associated with the presence of the dominant *dorippus* phenotype (absence of
196 black forewing tip) (S7 Fig). It is possible that it has a causal effect on the phenotype by
197 influencing the expression of *arrow*, but it might also be simply linked to the causative mutation.
198 Either way, we suggest that this large structural change, which increases the length of the
199 chromosome by nearly a third, contributes to recombination suppression between the $BC^{dorippus}$
200 allele and other supergene alleles by interfering with chromosome pairing in heterozygotes.

201 **A neo-W chromosome traps a single haplotype of chromosome 15 in contact zone females**

202 Previous crossing experiments indicated that the BC chromosome has become sex-linked
203 in contact zone females [22]. To confirm this hypothesis using genetic tools, we created a 'cured
204 line' by treating a female from an all-female brood with tetracycline to eliminate *Spiroplasma*
205 and allow the survival of male offspring [23]. A cross using this female confirms perfect sex-
206 linkage of forewing phenotype (n=22, chi-square test p=0.00002; S8 Fig). We then used PCR
207 assays on a subsequent sibling cross from the cured line to confirm that maternal alleles for
208 chr15 segregate with sex (n=22, p<0.00003) whereas paternal alleles segregate randomly (n=22,
209 p=0.36; S8 Fig). These results exactly match the model (Fig. 1B) in which the BC supergene has
210 become linked to the W chromosome in females, but continues to segregate as an autosome in
211 males.

212 Although we were unable to definitively identify any scaffolds from the ancestral W
213 chromosome, which is likely to be highly repetitive, we can test whether chr15 shows the
214 expected hallmarks of a young neo-W, hypothesised to have formed through fusion to the
215 ancestral W [22]. Due to the complete absence of recombination in females, we expect that a
216 single fused haplotype of chr15 would be spreading in the population. Any unique mutations
217 specific to this haplotype should therefore occur at high frequency in females and be absent in
218 males. We scanned for such high-frequency female-specific mutations, and found them to be
219 abundant across the entire length of chr15 and nearly absent throughout the rest of the genome

217 (Fig 3A). At the individual level, we can clearly identify 15 females (14 collected in the contact
218 zone and the single 'cured line' female) that consistently share these high-frequency mutations
219 (S9 Fig). Genetic distance among these females in the colinear region of chr15 (outside the BC
220 supergene) is reduced, indicating that they all share a similar haplotype of the fused chromosome
221 (Fig. 3B).



222 **Fig 3. Recent sweep of a young neo-W.** (**A**) The number of high-frequency female-specific
223 mutations (> 20% in females and absent in males) in 100 kb sliding windows (20 kb step size).
224 **(B)** Distance-based phylogenetic network for the distal colinear region of chr15 outside of the
225 BC supergene reveals that most contact zone females carry the conserved neo-W haplotype.
226 Cartoons show how the colinear region of chr15 is outside of the BC supergene but would still
227 capture reduced divergence among individuals carrying a shared non-recombining neo-W. (**C**)
228 Boxplot comparing the density of heterozygous sites in 100 kb windows in the colinear region of
229 chr15 between wild type individuals from the contact zone (CZ) and those carrying the neo-W.
230 Cartoon chromosomes above the plot match those shown in panel B. A relative lack of elevated
231 heterozygosity in the neo-W lineage indicates a lack of divergence of the fused neo-W haplotype,
232 consistent with the fusion being recent. (**D**) Boxplot of nucleotide diversity (π) within each
233 population for the same colinear region of chr15. On the far right, π is shown for the haploid neo-
234 W haplotype specifically, based on partial sequences isolated from this haplotype (see Methods
235 and S10 Fig for details). The near absence of genetic diversity implies a very rapid spread of the
236 neo-W through the population.

237 **The neo-W formed recently and spread rapidly**

238 Genetic variation accumulated in the neo-W lineage since its formation can tell us about its
239 age. Sequence divergence between the neo-W and autosomal copies of chr15 (inferred from the
240 density of heterozygous sites in the colinear region of chr15 in females carrying the neo-W) is
241 not significantly different from that between the autosomal copies in 'wild type' individuals that
242 lack the fusion (Fig. 3C, Wilcoxon signed rank test, p=0.36, n=48 100 kb windows). This implies
243 that insufficient time has passed since the fusion event for significant accumulation of new
244 mutations. The limited divergence of the neo-W haplotype from the autosomal copy of chr15 in
245 each female makes it challenging to isolate. Nonetheless, by using diagnostic mutations that are
246 unique to, and fixed in the neo-W lineage, we were able to identify sequencing reads from the
247 shared haplotype and reconstruct a partial neo-W sequence for each female (S10 Fig). A dated
248 genealogy based on these sequences places the root of the neo-W lineage at ~2200 years (26,400
249 generations) ago (posterior mean = 2201, std. dev. = 318).

250 The neo-W is present in all but one of the contact zone females, implying a rapid spread
251 since its formation. This process is similar to a selective sweep of a beneficial mutation, except
252 that complete recombination suppression in females means that the sweep affects the entire
253 chromosome equally. Unlike a conventional sweep, it is not expected to eliminate genetic
254 diversity from the population as these females will also carry an autosomal copy of chr15
255 inherited from their father (Fig 1B). Indeed, we see a 20% reduction in overall nucleotide
256 diversity (π) on chr15 in females of the neo-W lineage (Fig 3D). However, when we consider
257 only the neo-W haplotype in each of these females we see a nearly complete absence of genetic
258 variation, with a π of 0.00007, two orders of magnitude lower than for autosomal copies of chr15
259 (0.0228) (Fig. 3D). These results further support a very recent and rapid spread of the neo-W.

260 The neo-W haplotype carries the recessive $BC^{chrysippus}$ allele at the BC supergene (S4 Fig).
261 However, previous work [22] shows that at the the focal sampling site in the contact zone, most
262 males are immigrants homozygous for the dominant $BC^{dorippus}$ allele, and the vast majority of
263 females (84%) are heterozygous $BC^{dorippus}/BC^{chrysippus}$, as expected if most inherit $BC^{dorippus}$ from
264 their father and $BC^{chrysippus}$ (on the neo-W) from their mother. The dominant *dorippus* phenotype is
265 therefore by far the most abundant in this population. Since aposematic colouration should be
266 under positive frequency dependent selection, it is highly unlikely that the spread of the neo-W

267 can be explained by selection on colour pattern, highlighting the question of what else might
268 have driven its spread.

269 **Hitchhiking between the neo-W and *Spiroplama***

270 We hypothesised that the neo-W has spread as a result of co-inheritance with the male-
271 killing *Spiroplasma*, which is itself spreading through the population as a selfish element.
272 Experiments have suggested that infected female offspring have increased survival relative to
273 those from uninfected broods, possibly due to reduced competition for resources, although other
274 factors such as improved immunity have not been tested [45]. For this to drive the spread of the
275 neo-W, there would also need to be strict vertical inheritance of *Spiroplasma* down the female
276 line, such that it is always co-inherited with the neo-W.

277 We identified nine scaffolds making up the 1.75 Mb *Spiroplasma* genome in our *D.*
278 *chrysippus* assembly (S11 Fig). Infected individuals are clearly identifiable by mapping
279 resequencing reads to the *Spiroplasma* scaffolds (S11 Fig), and this was confirmed by PCR. As
280 predicted, all females in the neo-W lineage are infected (with the exception of the cured line
281 female, in which *Spiroplasma* had been eliminated). Moreover, all infected females fall into the
282 same mitochondrial clade (Fig 4A), consistent with matrilineal inheritance. To confirm that the
283 *Spiroplasma* is strictly vertically inherited and always associated with a single female lineage,
284 we used PCR assays for *Spiroplasma* and mitochondrial haplotype and expanded our sample size
285 to 158 individuals, including samples used in previous studies going back two decades
286 [19,23] (S12 Table, S12 Fig). This confirms the perfect association: 100% of infected individuals
287 (n=42) carry the same mitochondrial haplotype, and this haplotype is otherwise rare, occurring in
288 8% of uninfected individuals (n=116) (S12 Fig).

289 Like the neo-W, the *Spiroplasma* genomes carry limited variation among individuals (π =
290 0.0005), consistent with a single and recent outbreak of the endosymbiont. Although the lack of
291 variation makes it challenging to infer genealogies, our inferred maximum likelihood
292 genealogies for the neo-W and *Spiroplasma* are strikingly congruent (Fig 4B). The low bootstrap
293 support for multiple nodes is unsurprising given that these sequences descend from a recent
294 common ancestor, such that most nodes will be defined by only a few informative sites. This
295 does not weaken the support for congruence however, as the probability of two incorrectly
296 inferred topologies matching by chance is infinitesimally small. In a permutation test for
297 congruence between the two distance matrices [46], the observed level of congruence exceeds all

100,000 random permutations. There is therefore strong support for co-inheritance of the neo-W and *Spiroplasma* [47].

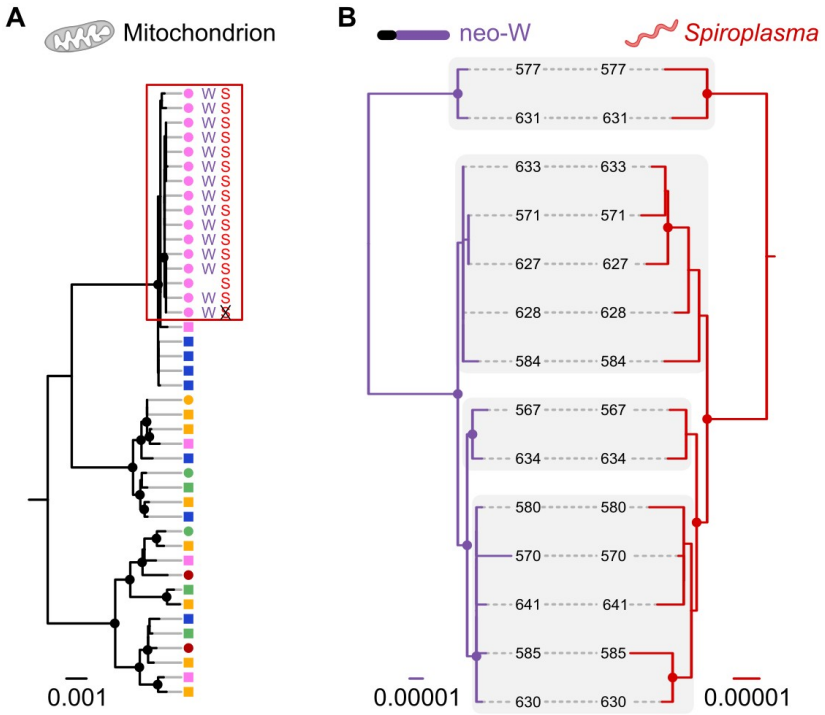


Fig 4. Matrilineal inheritance causes coupling between neo-W and *Spiroplasma*. (**A**) Maximum likelihood phylogeny for the whole mitochondrial genome. Individuals are coloured according to population of origin (see Fig 1A), and those carrying the neo-W ('W') and *Spiroplasma* ('S') are indicated (including one cured individual in which *Spiroplasma* was eliminated). Females are indicated by circles and males by squares. (**B**) Maximum likelihood phylogenies for the neo-W haplotype and *Spiroplasma* genome isolated from infected females. Corresponding clades are shaded to indicate congruence. Note that two samples are excluded in panel B: the cured sample, which lacked *Spiroplasma* due to tetracycline treatment, and one infected female found to lack the neo-W. Whether the latter represents an ancestral state or secondary loss requires further investigation. In all trees, nodes supported by more than 70 of 100 bootstrap replicates are indicated by circles.

The combined spread of three physically unlinked DNA molecules – the mitochondrial genome, neo-W and *Spiroplasma* genome – constitutes a form of genetic hitchhiking, but is facilitated by their strict matrilineal inheritance rather than physical linkage. We cannot entirely rule out the possibility that the neo-W is contributing to this spread, or even driving it entirely, through direct selection or meiotic drive. In theory this is testable by examining broods that carry the neo-W but lack *Spiroplasma*. We raised 11 such broods in our cured line, and Smith

319 [48] reported 10 natural broods that showed sex-linked colour pattern and no male killing.
320 Across these 21 broods, totalling 528 adult offspring, 51% were female. This is far from
321 significantly different from the null expectation of 50% (binomial test p=0.7). However, we note
322 that in order to detect meiotic drive causing a 1% female bias with good power would require a
323 far larger samples size of >15,000. Given the evidence from other work showing that male
324 killers can spread rapidly as selfish elements [49], and the endosymbionts can convey other
325 fitness benefits [50], this remains the most parsimonious explanation in this case. Selfish
326 elements have been shown to drive hitchhiking of the mitochondrial genome or a portion of a
327 chromosome through a population and even across species boundaries [51–53]. Our findings
328 show how an entire chromosome can be captured in the same way. Hitchhiking may therefore be
329 of general importance in driving the spread of neo-sex chromosomes.

330 In *D. chrysippus*, it is currently unclear whether the neo-W or male-killer emerged first. It
331 is also unclear whether their co-occurrence in a single ancestor was simply a coincidence or
332 instead reflects some functional connection, such as the suggestion that the neo-W might confer
333 susceptibility to the male killer [22]. It is important to note that this is not the first time a neo-sex
334 chromosome has formed in this taxon. A fusion of chromosome 21 to the ancestral Z
335 chromosome occurred in an ancestor of all *Danaus* species, producing a neo-Z [9,32,54]. It is
336 speculated that a complementary fusion of chromosome 21 to the ancestral W also occurred
337 [9,54], but this is difficult to conclusively verify due to degradation of the W chromosome over
338 longer timescales. If this hypothesis of an ancient neo-W is correct, then the neo-W we describe
339 (W-chr15) might in fact be better described as a neo-neo-W (W-chr21-chr15). It is possible that
340 the spread of the original W-chr21 was also driven by hitchhiking with a selfish endosymbiont.

341 **Genetic and phenotypic consequences of recombination suppression**

342 Sex chromosome evolution in many other taxa involves the progressive spread of
343 recombination suppression outward from the a sex-determining locus [55]. By contrast, the
344 absence of crossing over in female meiosis means that a lepidopteran neo-W experiences
345 complete and immediate recombination suppression over its entire length. Butterfly W
346 chromosomes are therefore thought to be highly degenerated and repetitive, and to our
347 knowledge none have been successfully assembled to date. The young age of the *D. chrysippus*
348 neo-W therefore provides a rare opportunity to study the early evolutionary consequences of
349 recombination suppression across an entire chromosome. Two related processes could shape its

evolution: hitchhiking of pre-existing deleterious mutations that were initially rare in the population [6], and accumulation of novel deleterious mutations due to reduced purging through recombination and selection (i.e. Muller's Ratchet) [7].

As a proxy for the 'genetic load' of deleterious mutations in the population, we considered P_n/P_s, the normalised ratio of non-synonymous to synonymous polymorphisms. Due to purifying selection, non-synonymous polymorphisms are typically rare, and where they do occur the mutant allele typically occurs at low frequency in the population [56]. When considering all polymorphisms in the neo-W lineage, P_n/P_s for chr15 (excluding the BC supergene, to avoid bias) is very slightly (~5%) higher than for other autosomes (Fig S13). Of 1000 bootstrap replicates, 916 reproduced this bias, corresponding to a p-value of 0.084. However, when we partition polymorphisms by allele frequency, we see that chr15 carries a large excess of non-synonymous polymorphisms in the highest frequency class (i.e. minor allele at 50%), with a P_n/P_s ratio >3 times larger than on other autosomes (S13 Fig). This holds across all 1000 bootstrap replicates (i.e. $p<0.001$). A change in the frequency distribution of non-synonymous variants, without a significant change in their abundance, is best explained by hitchhiking of pre-existing mildly-deleterious alleles that were initially rare in the population but were inadvertently carried to high frequency along with the neo-W haplotype, and are therefore now found in all females in this lineage. In fact, P_n/P_s for high-frequency polymorphisms on chr15 is higher than that for singletons on autosomes ($p=0.044$), suggesting that accumulation of additional mildly-deleterious alleles on the neo-W might have occurred early during its spread through the population.

At the phenotypic level, perhaps counter-intuitively, the spread of a single supergene allele on the neo-W has not caused homogenization of warning pattern among contact zone females, and might in fact have the opposite effect. In locations where the neo-W and *Spiroplasma* are nearly fixed, such as our sampling site near Nairobi, the high incidence of male killing implies that the population is strongly shaped by immigrant males. Since the $BC^{chrysippus}$ allele on the neo-W is universally recessive, daughters will tend to match the phenotype of their immigrant father. However, because the neo-W is always transmitted to daughters, the paternal chr15 copy will be lost to male killing after one generation, creating a genetic sink for immigrant male genes [22] (S14 Fig). This combination of processes results in a female population that is highly sensitive to the source of immigrants, which is known to fluctuate seasonally with monsoon

381 winds [16,57] (S14 Fig). This model leads to the testable prediction that seasonal fluctuations in
382 female phenotypes should be most dramatic where male killing is most abundant.

383 **Future evolutionary trajectories**

384 The future of the neo-W and *Spiroplasma* outbreak is uncertain. A lack of males could lead
385 to local extinctions [27], but extinction of the entire infected lineage is unlikely given the high
386 dispersal ability and seasonal influxes of males in the contact zone. Indeed, it is notable that
387 *Spiroplasma* infection has only been recorded within the contact zone population (with the
388 exception of a single South African female reported here, S12 Table), especially given theory
389 showing that male killers should spread very rapidly across the geographical range of a
390 panmictic population if they provide even a very weak selective advantage [49]. Future work
391 will investigate whether its spread might be curtailed by environmental factors, for example if
392 oviposition behaviour or host plant availability only leads to sibling competition (and consequent
393 benefits for all-female broods) under certain conditions [45]. An alternative and non-mutually-
394 exclusive hypothesis is that dispersal rates of infected females are strongly reduced. In other
395 systems, sex-ratio distortion has driven adaptive responses by the host, including changes to the
396 mating system [58] and the evolution of resistance to male-killing [59,60]. The absence of
397 evidence for these phenomena in *D. chrysippus* might simply reflect the recency of the male
398 killing outbreak. Eventually, we also expect the non-recombining neo-W to begin to degenerate
399 through further hitchhiking, gene loss and the spread of repetitive elements [8,55]. This young
400 system provides a rare opportunity to study how these phenomena unfold through time and
401 space.

402 ## Methods

403 **Reference genome sequencing, assembly and annotation**

404 Detailed methods for generation of the *D. chrysippus* reference genome are provided in S1
405 Text. Briefly, a draft assembly was generated using SPAdes [61] from a combination of paired-
406 end and mate-pair libraries of various insert sizes. Scaffolding and resolution of haplotypes was
407 performed using Redundans [62] and Haplomerger2 [63]. The assembly was annotated using a
408 combination of de-novo gene predictors yielding 16,654 protein coding genes. Mitochondrial
409 genomes were assembled using NOVOplasty [64].

410 Although we currently lack linkage information for further scaffolding, we generated a
411 pseudo-chromosomal assembly based on homology with the highly contiguous *Heliconius*
412 *melpomene* genome [30,31,65], adjusted for known karyotypic differences [9,30–32,54].
413 Although these genomes are diverged by approximately 90 million years, this homology-based
414 approach has been shown previously to be successful for reconstruct chromosomes in a
415 fragmented *D. plexippus* genome [9]. In total, 282 Mb (87% of the genome) could be confidently
416 assigned to chromosomes (S1 Fig).

417 Scaffolds representing the *Spiroplasma* genome were identified based on read depth of re-
418 mapped reads (S11 Fig) and homology to other available *Spirolasma* genomes. Annotation was
419 performed using the RAST server pipeline [66,67].

420 **Population sample resequencing and genotyping**

421 This study made use of 42 newly sequenced *D. chrysippus* individuals, as well as
422 previously sequenced individuals of the sister species, *D. petilia* (n=1), and next closest outgroup
423 *D. gilippus* (n=2) [68] (S9 Table). Details of DNA extraction, sequencing and genotyping are
424 provided in S1 Text. Briefly, DNA was extracted from thorax tissue and sequenced (paired-end,
425 150 bp) to a mean depth of coverage 20x or greater. Reads were mapped to the *D. chrysippus*
426 reference assembly using Stampy [69] v1.0.31 and genotyping was performed using GATK
427 version 3 [70,71]. Genotype calls were required to have an individual read depth ≥ 8, and
428 heterozygous and alternate allele calls were further required to have an individual genotype
429 quality (GQ) ≥ 20 for downstream analyses.

430 **Genomic differentiation and associations with wing pattern**

431 We used the fixation index (F_{ST}) to examine genetic differentiation across the genome
432 among the three subspecies for which we had six or more individuals sequenced. F_{ST} was
433 computed using the script popgenWindows.py (github.com/simonhmartin/genomics_general
434 release 0.2) with a sliding windows of 100 kb, stepping in increments of 20 kb. Windows with
435 fewer than 20,000 genotyped sites after filtering (see above) were ignored.

436 To identify SNPs associated with the three Mendelian colour pattern traits (i.e. the A B and
437 C loci) (Fig 1A), we used PLINK v1.9 [72] with the '--assoc' option, and provided quantitative
438 phenotypes of 0, 1 or 0.5 for assumed heterozygotes, which causes PLINK to use the Wald test
439 for quantitative traits. In addition to the quality and depth filters above, SNPs used for this

440 analysis were required to have genotypes for at least 40 individuals, a minor allele count of at
441 least 2, and to be heterozygous in no more than 75% of individuals. SNPs were also thinned to a
442 minimum distance of 100 bp.

443 To examine relationships among diploid individuals in specific regions of interest, we
444 constructed phylogenetic networks using the Neighbor-Net [73] algorithm, implemented in
445 SplitsTree [74]. Pairwise distances used for input were computed using the script distMat.py
446 (github.com/simonhmartin/genomics_general release 0.2).

447 **Haplotype cluster assignment**

448 To assign haplotypes to clusters in the BC supergene region, we first phased genotypes
449 using SHAPEIT2 [75,76] using SNPs filtered as for association analysis above, except with a
450 minor allele count of at least 4, and no thinning. Default parameters were used for phasing except
451 that the effective population size was set to $3x10^6$. To minimise phasing switch errors, we
452 analysed each 20 kb window separately. Cluster assignment for both haplotypes from each
453 individual was based on average genetic distance to all haplotypes from each of three reference
454 groups: *D. c. dorippus*, *D. c. orientis*, or *D. c. alcippus* (which is representative of *chrysippus* as
455 they share the same alleles at the BC supergene). A haplotype was assigned to one of the three
456 groups if its average genetic distance to members of that group was less than 80% the average
457 distance to the other two groups, otherwise it was left as unassigned. Genetic distances were
458 computed using the script popgenWindows.py (github.com/simonhmartin/genomics_general
459 release 0.2).

460 **Identification of neo-W specific sequencing reads**

461 To identify females carrying the neo-W chromosome, we visualised the distribution of
462 female-specific derived mutations that occur at high-frequency. Allele frequencies were
463 computed using the script freq.py (github.com/simonhmartin/genomics_general release 0.2). Due
464 to the absence of female meiotic crossing over in Lepidoptera, all females carrying the neo-W
465 fusion should share a conserved chromosomal haplotype for the entire fused chromosome. To
466 isolate this shared fused haplotype from the autosomal copy, we first identified diagnostic
467 mutations as those that are present in a single copy in each member of the 'neo-W lineage' and
468 absent from all other individuals and outgroups. We then isolated the sequencing read pairs from
469 each of these females that carry the derived mutation (S10 Fig). This resulted in a patchy
470 alignment file, with a stack of read pairs over each diagnostic mutation. Based on these aligned

471 reads, we genotyped each individuals as described above, except here setting the ploidy level to
472 1, and requiring a minimum read depth of 3.

473 **Diversity and divergence of the neo-W**

474 The lack of recombination across the neo-W makes it possible to gain insights into its age.
475 Over time, mutations will arise that differentiate the neo-W from the recombining autosomal
476 copies of the chromosome. We estimated this divergence based on average heterozygosity in
477 females carrying the neo-W, and compared it to heterozygosity from contact-zone individuals
478 not carrying the neo-W. Heterozygosity was computed using the Python script
479 popgenWindows.py (github.com/simonhmartin/genomics_general release 0.2) focusing only on
480 the colinear portion of the chromosome (i.e. the distal portion from 11 Mb onwards), which is
481 outside of the BC supergene. Heterozygosity was computed in 100 kb windows, and windows
482 were discarded if they contained fewer than 20,000 sites genotyped in at least two individuals
483 from each population.

484 A recent spread of the neo-W through the population should also be detectable in the form
485 of strong conservation of the neo-W haplotype in all females that carry it (i.e. reduced genetic
486 diversity). We therefore computed nucleotide diversity (π) in 100 kb windows as above.
487 Reported values of π and heterozygosity represent the mean +- standard deviation across 100 kb
488 windows.

489 **Genealogical analyses**

490 We produced maximum likelihood trees for the mitochondrial genome, neo-W and
491 *Spiroplasma* genome, using PhyML v3 [77] with the GTR substitution model. Given the small
492 number of SNPs in both the neo-W and *Spiroplasma* genome, regions with inconsistent coverage
493 across individuals were excluded manually. Only sites with no missing genotypes were included.

494 We estimated the root node age for the neo-W using BEAST2 [78,79] version 2.5.1 with a
495 fixed clock model and an exponential population growth prior. For all other priors we used the
496 defaults as defined by BEAUti v2.5.1. We assumed a mutation rate of 2.9×10^{-9} per generation
497 based on a direct estimate for *Heliconius* butterflies [80] and 12 generations per year [81].
498 BEAST2 was run for 500,000,000 iterations, sampling every 50,000 generations, and we used
499 Tracer [82] version 1.7.1 to check for convergence of posterior distributions and compute the
500 root age after discarding a burn-in of 10%.

501 We tested for congruence between the neo-W and *Spiroplasma* trees using PACo [46],
502 which assesses the goodness-of-fit between host and parasite distance matrices, with 100,000
503 permutations. Distance matrices were computed using the script distMat.py
504 (github.com/simonhmartin/genomics_general release 0.2).

505 **Analysis of synonymous and non-synonymous polymorphism**

506 We computed P_n/P_s as as the ratio of non-synonymous polymorphisms per non-
507 synonymous site to synonymous polymorphisms per synonymous site. Synonymous and non-
508 synonymous sites were defined conservatively as 4-fold degenerate and 0-fold degenerate codon
509 positions, respectively, with the requirement that the other two codon positions are invariant
510 across the entire dataset. Only sites genotyped in all 15 females in the neo-W lineage were
511 considered, and counts were stratified by minor allele frequency using the script sfs.py
512 (github.com/simonhmartin/genomics_general release 0.1).

513 **Butterfly rearing and molecular diagnostics**

514 To generate a stock line that is cured of *Spiroplasma* infection, we treated caterpillars from
515 all-female broods with Tetracycline, following Jiggins et al. [23]. A 'cured line' was initiated
516 from a single treated female that had the heterozygous *Cc* '*transiens*' phenotype (Fig 1A). This
517 female was crossed to a wild male (homozygous *cc*) to test for sex linkage of phenotype. The
518 cured line was maintained through sibling crosses for six generations and the persistence of
519 males indicated that *Spiroplasma* had been eliminated.

520 We then applied a molecular test for sex linkage of chr15 using the F5 brood from the
521 cured line. We designed two separate PCR diagnostics based on SNPs segregating on chr15 to
522 distinguish between the two chromosomes of the male and the female parents (S11 Table). PCR
523 was performed using the Phusion HF Master Mix and HF Buffer (New England Biosciences).

524 To screen for *Spiroplasma* infection, we designed a PCR assay targeting the
525 glycerophosphoryl diester phosphodiesterase (GDP) gene (S11 Table). PCR was performed as
526 above. We confirmed the the sensitivity of this diagnostic by analyzing individuals of known
527 infection status based on whole genome sequencing (12 infected and 11 uninfected).

528 To investigate whether *Spiroplasma* infection was always associated with a single
529 mitochondrial haplotype, we designed a PCR RFLP for the Cytochrome Oxidase Subunit I (COI)
530 that differentiates the infected 'K' lineage from uninfected (S10 Table). PCR was performed as

above. A subset of products were verified by Sanger sequencing after purification using the QIAquick PCR Purification Kit (Qiagen, USA).

Ethics Statement

Butterfly collection was performed under permit where relevant: NACOSTI/P15/3290/3607, NACOSTI/P15/2403/3602 (National Commission for Science and Technology, Kenya), MINEDUC/S&T/459/2017 (Ministry of Education, Rwanda), EMDEP006/17 (Environmental Management Division, St Helena Government); and always with permission of the land owner and/or local authorities. We also worked with local researchers wherever possible, including authors DJM, KSS & IJG, and with the Lepidopterists Society of Africa.

Data Availability

Raw genomic data and assemblies are available from GenBank (accession numbers to be provided upon acceptance), all processed data files underlying all figures are available from the Dryad Repository (accession number available upon acceptance). Scripts used for data analysis are available from https://github.com/simonhmartin/genomics_general.

References

1. Kirkpatrick M, Barton N. Chromosome inversions, local adaptation and speciation. Genetics. 2006;173: 419–34. doi:10.1534/genetics.105.047985

2. Bürger R, Akerman A. The effects of linkage and gene flow on local adaptation: A two-locus continent–island model. Theor Popul Biol. Academic Press; 2011;80: 272–288. doi:10.1016/J.TPB.2011.07.002

3. Guerrero RF, Rousset F, Kirkpatrick M. Coalescent patterns for chromosomal inversions in divergent populations. Philos Trans R Soc B Biol Sci. 2012;367: 430–438. doi:10.1098/rstb.2011.0246

4. Charlesworth D, Charlesworth B. Selection on recombination in clines. Genetics. 1979;91: 575–580.

5. Charlesworth D. The status of supergenes in the 21st century: Recombination suppression in Batesian mimicry and sex chromosomes and other complex adaptations. Evol Appl. John Wiley & Sons, Ltd (10.1111); 2016;9: 74–90. doi:10.1111/eva.12291

6. Rice WR. Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. Genetics. 1987;116: 161–167.

7. Charlesworth B. Model for evolution of Y chromosomes and dosage compensation. Proc Natl Acad Sci U S A. 1978;75: 5618–22.

8. Bachtrog D, Charlesworth B. The temporal dynamics of processes underlying Y chromosome degeneration. Genetics. Genetics; 2008;179: 1513–25. doi:10.1534/genetics.107.084012

9. Mongue AJ, Nguyen P, Voleníková A, Walters JR. Neo-sex chromosomes in the monarch butterfly, Danaus plexippus. G3. 2017;7: 3281–3294. doi:10.1534/g3.117.300187

10. Bracewell RR, Bentz BJ, Sullivan BT, Good JM. Rapid neo-sex chromosome evolution and incipient speciation in a major forest pest. Nat Commun. Springer US; 2017;8. doi:10.1038/s41467-017-01761-4

11. Pala I, Naurin S, Stervander M, Hasselquist D, Bensch S, Hansson B. Evidence of a neo-sex chromosome in birds. Heredity. Nature Publishing Group; 2012;108: 264–272. doi:10.1038/hdy.2011.70

12. Kitano J, Ross JA, Mori S, Kume M, Jones FC, Chan YF, et al. A role for a neo-sex chromosome in stickleback speciation. Nature. Nature Publishing Group; 2009;461: 1079–1083. doi:10.1038/nature08441

13. Nguyen P, Sykorova M, Sichova J, Kuta V, Dalikova M, Capkova Frydrychova R, et al. Neo-sex chromosomes and adaptive potential in tortricid pests. Proc Natl Acad Sci. 2013;110: 6931–6936. doi:10.1073/pnas.1220372110

14. Carabajal Paladino LZ, Provazníková I, Berger M, Bass C, Aratchige NS, López SN, et al. Sex Chromosome Turnover in Moths of the Diverse Superfamily Gelechioidea. Barluenga M, editor. Genome Biol Evol. Narnia; 2019;11: 1307–1319. doi:10.1093/gbe/evz075

15. Steinemann M, Steinemann S. Enigma of Y chromosome degeneration: neo-Y and neo-X chromosomes of Drosophila miranda a model for sex chromosome evolution. Genetica. 1998;102–103: 409–20.

16. Smith DAS, Owen DF, Gordon IJ, Lowis NK. The butterfly Danaus chrysippus (L .) in East Afiica : polymorphism and morph-ratio clines within a complex , extensive and dynamic hybrid zone. Zool J Linn Soc. 1997;120: 51–78.

17. Smith DAS. Heterosis, epistasis and linkage disequilibrium in a wild population of the polymorphic butterfly Danaus chrysippus (L .). Zool J Linn Soc. 1980;69: 87–109.

18. Smith D a S, Owen DF, Gordon IJ, Owiny AM. Polymorphism and evolution in the butterfly Danaus chrysippus (L.) (Lepidoptera: Danainae). Heredity. 1993;71: 242–251. doi:10.1038/hdy.1993.132

595 19. Smith DAS, Gordon IJ, Allen JA. Reinforcement in hybrids among once isolated
596 semispecies of Danaus chrysippus (L.) and evidence for sex chromosome evolution. Ecol
597 Entomol. 2010;35: 77–89. doi:10.1111/j.1365-2311.2009.01143.x

598 20. Smith DAS. Genetics of Some Polymorphic Forms of the African Butterfly Danaus
599 chrysippus L. (Lepidoptera: Danaidae). Insect Syst Evol. Brill; 1975;6: 134–144.
600 doi:10.1163/187631275X00235

601 21. Clarke CA, Sheppard PM, Smith AG. genetics of fore and hindwing colour in crosses
602 between Danaus chrysippus from Australia and from Sierra Leone (Danaidae). Lepid Soc
603 J. 1973;

604 22. Smith DAS, Gordon IJ, Traut W, Herren J, Collins S, Martins DJ, et al. A neo-W
605 chromosome in a tropical butterfly links colour pattern, male-killing, and speciation. Proc
606 R Soc B Biol Sci. 2016;283: 20160821. doi:10.1098/rspb.2016.0821

607 23. Jiggins FM, Hurst GD, Jiggins CD, v d Schulenburg JH, Majerus ME. The butterfly
608 Danaus chrysippus is infected by a male-killing Spiroplasma bacterium. Parasitology.
609 2000;120 (Pt 5: 439–46. doi:10.1017/S0031182099005867

610 24. Herren JK, Gordon I, Holland PWH, Smith D. The butterfly Danaus chrysippus
611 (Lepidoptera: Nymphalidae) in Kenya is variably infected with respect to genotype and
612 body size by a maternally transmitted male-killing endosymbiont (Spiroplasma). Int J
613 Trop Insect Sci. 2007;27: 62. doi:10.1017/S1742758407818327

614 25. Gordon IJ, Ireri P, Smith DAS. Hologenomic speciation: Synergy between a male-killing
615 bacterium and sex-linkage creates a “magic trait” in a butterfly hybrid zone. Biol J Linn
616 Soc. 2014;111: 92–109. doi:10.1111/bij.12185

617 26. Lushai G, Allen JA, Goulson D, Maclean N, Smith DAS. The butterfly Danaus chrysippus
618 (L.) in East Africa comprises polyphyletic, sympatric lineages that are, despite behavioural
619 isolation, driven to hybridization by female-biased sex ratios. Biol J Linn Soc. 2005;86:
620 117–131. doi:10.1111/j.1095-8312.2005.00526.x

621 27. Idris E, Saeed M. Hassan S. Biased sex ratios and aposematic polymorphism in African
622 butterflies: A hypothesis. Ideas Ecol Evol. 2013;6: 5–16. doi:10.4033/iee.2013.6.2.n

623 28. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO:
624 assessing genome assembly and annotation completeness with single-copy orthologs.
625 Bioinformatics. Oxford University Press; 2015;31: 3210–3212.
626 doi:10.1093/bioinformatics/btv351

627 29. Heliconius Genome Consortium T. Butterfly genome reveals promiscuous exchange of
628 mimicry adaptations among species. Nature. 2012;487: 94–8. doi:10.1038/nature11041

30. Davey JW, Chouteau M, Barker SL, Maroja L, Baxter SW, Simpson F, et al. Major Improvements to the Heliconius melpomene Genome Assembly Used to Confirm 10 Chromosome Fusion Events in 6 Million Years of Butterfly Evolution. G3. 2016;6: 695–708. doi:10.1534/g3.115.023655

31. Davey JW, Barker SL, Rastas PM, Pinharanda A, Martin SH, Durbin R, et al. No evidence for maintenance of a sympatric Heliconius species barrier by chromosomal inversions. Evol Lett. 2017;1: 138–154. doi:10.1002/evl3.12

32. Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, et al. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat Commun. 2014;5: 1–9. doi:10.1038/ncomms5737

33. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, et al. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 2012;10: e1001388. doi:10.1371/journal.pbio.1001388

34. Mackintosh A, Laetsch DR, Hayward A, Charlesworth B, Waterfall M, Vila R, et al. The determinants of genetic diversity in butterflies. Nat Commun. Nature Publishing Group; 2019;10: 3466. doi:10.1038/s41467-019-11308-4

35. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, et al. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 2013;23: 1817–1828. doi:10.1101/gr.159426.113

36. Wittkopp PJ, Vaccaro K, Carroll SB. Evolution of yellow Gene Regulation and Pigmentation in Drosophila. Curr Biol. Cell Press; 2002;12: 1547–1556. doi:10.1016/S0960-9822(02)01113-2

37. Zhang L, Martin A, Perry MW, van der Burg KRL, Matsuoka Y, Monteiro A, et al. Genetic Basis of Melanin Pigmentation in Butterfly Wings. Genetics. Genetics; 2017;205: 1537–1550. doi:10.1534/genetics.116.196451

38. Rives AF, Rochlin KM, Wehrli M, Schwartz SL, DiNardo S. Endocytic trafficking of Wingless and its receptors, Arrow and DFrizzled-2, in the Drosophila wing. Dev Biol. Academic Press; 2006;293: 268–283. doi:10.1016/j.ydbio.2006.02.006

39. Martin A, Papa R, Nadeau NJ, Hill RI, Counterman BA, Halder G. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. Proc Natl Acad Sci U S A. 2012;109: 12632–12637. doi:doi.org/10.1073/pnas.1204800109

40. Mazo-Vargas A, Concha C, Livraghi L, Massardo D, Wallbank RWR, Zhang L, et al. Macroevolutionary shifts of WntA function potentiate butterfly wing-pattern diversity. Proc Natl Acad Sci. 2017;114: 10701 LP – 10706.

663 41. Van Belleghem SM, Rastas P, Papanicolaou A, Martin SH, Arias CF, Supple MA, et al.
664 Complex modular architecture around a simple toolkit of wing pattern genes. Nat Ecol
665 Evol. 2017;1: 0052. doi:10.1038/s41559-016-0052

666 42. Smith DAS. Evidence for autosomal meiotic drive in the butterfly Danaus chrysippus L.
667 Heredity. 1976;36: 139–142.

668 43. Coughlan JM, Willis JH. Dissecting the role of a large chromosomal inversion in life
669 history divergence throughout the *Mimulus guttatus* species complex. Mol Ecol. John
670 Wiley & Sons, Ltd (10.1111); 2019;28: 1343–1357. doi:10.1111/mec.14804

671 44. Lee CR, Wang B, Mojica JP, Mandáková T, Prasad KVSK, Goicoechea JL, et al. Young
672 inversion with multiple linked QTLs under selection in a hybrid zone. Nat Ecol Evol.
673 2017;1. doi:10.1038/s41559-017-0119

674 45. Gordon IJ, Ireri P, Smith DAS. Preference for isolated host plants facilitates invasion of
675 Danaus chrysippus (Linnaeus, 1758) (Lepidoptera: Nymphalidae) by a bacterial male-
676 killer Spiroplasma. Austral Entomol. 2015;54: 210–216. doi:10.1111/aen.12113

677 46. Balbuena JA, Míguez-Lozano R, Blasco-Costa I. PACo: A Novel Procrustes Application
678 to Cophylogenetic Analysis. Moreau CS, editor. PLoS One. Public Library of Science;
679 2013;8: e61048. doi:10.1371/journal.pone.0061048

680 47. Richardson MF, Weinert LA, Welch JJ, Linheiro RS, Magwire MM, Jiggins FM, et al.
681 Population Genomics of the Wolbachia Endosymbiont in Drosophila melanogaster. PLoS
682 Genet. 2012;8. doi:10.1371/journal.pgen.1003129

683 48. Smith DAS. African Queens and Their Kin: A Darwinian Odyssey. Taunton, UK:
684 Brambleby Books; 2014.

685 49. Duplouy A, O'Neill SL. Rapid spread of male-Killing Wolbachia in the butterfly
686 Hypolimnas bolina. Evol Biol - Concepts, Mol Morphol Evol. 2010;23: 209–227.
687 doi:10.1007/978-3-642-12340-5_13

688 50. Hurst GDD, Hutchence KJ. Host defence: Getting by with a little help from our friends.
689 Curr Biol. Elsevier Ltd; 2010;20: R806–R808. doi:10.1016/j.cub.2010.07.038

690 51. Jiggins FM. Male-killing Wolbachia and mitochondrial DNA: selective sweeps, hybrid
691 introgression and parasite population dynamics. Genetics. 2003;164: 5–12.

692 52. Hurst GDD, Jiggins FM. Problems with mitochondrial DNA as a marker in population,
693 phylogeographic and phylogenetic studies: The effects of inherited symbionts. Proc R Soc
694 B Biol Sci. 2005;272: 1525–1534. doi:10.1098/rspb.2005.3056

695 53. Palopoli MF, Wu CI. Rapid evolution of a coadapted gene complex: Evidence from the
696 segregation Distorter (SD) system of meiotic drive in Drosophila melanogaster. Genetics.
697 1996;143: 1675–1688.

698 54. Traut W, Ahola V, Smith DAS, Gordon IJ, Ffrench-Constant RH. Karyotypes versus
699 Genomes: The Nymphalid Butterflies Melitaea cinxia, Danaus plexippus, and D.
700 chrysippus. Cytogenet Genome Res. 2018;153: 46–53. doi:10.1159/000484032

701 55. Wright AE, Dean R, Zimmer F, Mank JE. How to make a sex chromosome. Nature
702 Communications. Nature Publishing Group; 2016. p. 12087. doi:10.1038/ncomms12087

703 56. Fay JC, Wyckoff GJ, Wu C-I. Positive and Negative Selection on the Human Genome.
704 Genetics. 2001;158: 1227–1234.

705 57. Smith DAS, Owen DF. Colour genes as markers for migratory activity: The butterfly
706 Danaus chrysippus in Africa. Oikos. 1997;78: 127–135.

707 58. Jiggins FM, Hurst GDD, Majerus MEN. Sex-ratio-distorting Wolbachia causes sex-role
708 reversal in its butterfly host. Proc R Soc B Biol Sci. 2000;267: 69–73.
709 doi:10.1098/rspb.2000.0968

710 59. Hornett EA, Charlat S, Duplouy AMR, Davies N, Roderick GK, Wedell N, et al.
711 Evolution of male-killer suppression in a natural population. PLoS Biol. 2006;4: 1643–
712 1648. doi:10.1371/journal.pbio.0040283

713 60. Wilfert L, Jiggins FM. The dynamics of reciprocal selective sweeps of host resistance and
714 a parasite counter-adaptation in Drosophila. Evolution. Wiley/Blackwell (10.1111);
715 2013;67: 761–773. doi:10.1111/j.1558-5646.2012.01832.x

716 61. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A
717 New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J
718 Comput Biol. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY
719 10801 USA ; 2012;19: 455–477. doi:10.1089/cmb.2012.0021

720 62. Pryszcz LP, Gabaldón T. Redundans: an assembly pipeline for highly heterozygous
721 genomes. Nucleic Acids Res. Oxford University Press; 2016;44: e113–e113.
722 doi:10.1093/nar/gkw294

723 63. Huang S, Kang M, Xu A. HaploMerger2: rebuilding both haploid sub-assemblies from
724 high-heterozygosity diploid genome assembly. Berger B, editor. Bioinformatics. Oxford
725 University Press; 2017;33: 2577–2579. doi:10.1093/bioinformatics/btx220

726 64. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: *de novo* assembly of organelle
727 genomes from whole genome data. Nucleic Acids Res. Oxford University Press; 2016;45:
728 gkw955. doi:10.1093/nar/gkw955

729 65. Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, et al.
730 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species.
731 Nature. Nature Publishing Group; 2012;487: 94–8. doi:10.1038/nature11041

732 66. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the
733 Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic
734 Acids Res. Oxford University Press; 2014;42: D206–D214. doi:10.1093/nar/gkt1226

735 67. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server:
736 Rapid Annotations using Subsystems Technology. BMC Genomics. BioMed Central;
737 2008;9: 75. doi:10.1186/1471-2164-9-75

738 68. Zhan S, Zhang W, Niitepõld K, Hsu J, Haeger JF, Zalucki MP, et al. The genetics of
739 monarch butterfly migration and warning colouration. Nature. 2014;514: 317–21.
740 doi:10.1038/nature13812

741 69. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of
742 Illumina sequence reads. Genome Res. 2011;21: 936–9. doi:10.1101/gr.111120.110

743 70. DePristo M a, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework
744 for variation discovery and genotyping using next-generation DNA sequencing data. Nat
745 Genet. 2011;43: 491–8. doi:10.1038/ng.806

746 71. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A,
747 et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best
748 practices pipeline. Curr Protoc Bioinforma. Hoboken, NJ, USA: John Wiley & Sons, Inc.;
749 2013;43: 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

750 72. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation
751 PLINK: rising to the challenge of larger and richer datasets. Gigascience. Narnia; 2015;4:
752 7. doi:10.1186/s13742-015-0047-8

753 73. Bryant D, Moulton V. Neighbor-Net: An Agglomerative Method for the Construction of
754 Phylogenetic Networks. Mol Biol Evol. 2003;21: 255–265. doi:10.1093/molbev/msh018

755 74. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Mol
756 Biol Evol. 2006;23: 254–267. doi:10.1093/molbev/msj030

757 75. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease
758 and population genetic studies. Nat Methods. 2013;10: 5–6. doi:10.1038/nmeth.2307

759 76. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of
760 genomes. Nat Methods. 2012;9: 179–181. doi:10.1038/nmeth.1785

761 77. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms
762 and methods to estimate maximum-likelihood phylogenies: assessing the performance of
763 PhyML 3.0. Syst Biol. 2010;59: 307–21. doi:10.1093/sysbio/syq010

764 78. Bouckaert RR. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics.
765 2010;26: 1372–3. doi:10.1093/bioinformatics/btq110

766 79. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: A
767 Software Platform for Bayesian Evolutionary Analysis. Prlic A, editor. PLoS Comput
768 Biol. Public Library of Science; 2014;10: e1003537. doi:10.1371/journal.pcbi.1003537

769 80. Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, et al.
770 Estimation of the Spontaneous Mutation Rate in Heliconius melpomene. Mol Biol Evol.
771 2015;32: 239–243. doi:10.1093/molbev/msu302

772 81. Owen DF, Chanter DO. Population biology of tropical African butterflies. Sex ratio and
773 genetic variation in Acraea encedon. J Zool. 1969;157: 345–374. doi:10.1111/j.1469-
774 7998.1969.tb01707.x

775 82. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in
776 Bayesian Phylogenetics Using Tracer 1.7. Susko E, editor. Syst Biol. 2018;67: 901–904.
777 doi:10.1093/sysbio/syy032

778 Acknowledgements

779 We are grateful to Godfrey Amoni Etelej, Laura Hebberecht-Lopez and Glennis Julian for
780 support with butterfly rearing. We thank Jenny York, Frank Jiggins and Deborah Charlesworth
781 helpful comments. This work was funded by European Research Council (ERC) European Union
782 Horizon 2020 research and innovation programme grant 646625 (CB), ERC grant 339873 (CDJ,
783 SHM), National Geographic Society Research Grant WW-138R-17 (IJG), and a Royal Society
784 University Research Fellowship (SHM). The funders had no role in study design, data collection
785 and analysis, decision to publish, or preparation of the manuscript.