

1 **Title:** Synteny-based genome assembly for 16 species of *Heliconius* butterflies, and an assessment of
2 structural variation across the genus

3

4 **Running Title:**

5 Synteny and structural genomics in *Heliconius*

6

7 **Authors:** Fernando A. Seixas^{1*}, Nathaniel B. Edelman^{1,2}, James Mallet^{1*}

8 ¹Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

9 ²Yale Institute for Biospheric Studies, 170 Whitney Avenue, RM 213A New Haven, CT 06511 USA

10 **Corresponding authors:** fernando_seixas@g.harvard.edu, jmallet@oeb.harvard.edu

11

12 **Abstract**

13 *Heliconius* butterflies (Lepidoptera: Nymphalidae) are a group of 48 neotropical species widely studied in
14 evolutionary research. Despite the wealth of genomic data generated in past years, chromosomal level
15 genome assemblies currently exist for only two species, *Heliconius melpomene* and *H. erato*, each a
16 representative of one of the two major clades of the genus. Here, we use these reference genomes to
17 improve the contiguity of previously published draft genome assemblies of 16 *Heliconius* species. Using
18 a reference-assisted scaffolding approach, we place and order the scaffolds of these genomes onto
19 chromosomes, resulting in 95.7-99.9% of their genomes anchored to chromosomes. Genome sizes are
20 somewhat variable among species (270-422 Mb) and in one small group of species (*H. hecale*, *H.*
21 *elevatus* and *H. pardalinus*) differences in genome size are mainly driven by a few restricted repetitive
22 regions. Genes within these repeat regions show an increase in exon copy number, an absence of internal
23 stop codons, evidence of constraint on non-synonymous changes, and increased expression, all of which

24 suggest that the extra copies are functional. Finally, we conducted a systematic search for inversions and
25 identified five moderately large inversions fixed between the two major *Heliconius* clades. We infer that
26 one of these inversions was transferred by introgression between the lineages leading to the *erato/sara*
27 and *burneyi/doris* clades. These reference-guided assemblies represent a major improvement in
28 *Heliconius* genomic resources that should aid further genetic and evolutionary studies in this genus.

29 **Keywords:** *Heliconius*, Genome Assembly, Structural Variation, Inversions, Copy Number Variation

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46 **Introduction**

47 Advances in sequencing technology have revolutionized the field of evolutionary biology. Generating
48 short-read genomic datasets is now common practice, enabling investigation of fundamental evolutionary
49 processes including the genetic basis of adaptive traits, dynamics of selection on particular alleles, and
50 demographic histories of populations. In order to exploit the power of low-cost short-read data, one must
51 usually align reads to a reference genome.

52 The availability of high-quality reference genomes can determine the breadth and power of comparative
53 and population genomic analyses in evolutionary studies. For instance, placing genome scaffolds on
54 chromosomes allows one to contrast patterns between autosomes and sex chromosomes which has been
55 important for understanding speciation (Coyne and Orr 1989; Coyne 2018; Prowell 1998; Masly and
56 Presgraves 2007; Fontaine et al. 2015; Ellegren et al. 2012; Seixas et al. 2018; Martin et al. 2019).

57 Anchoring scaffolds to chromosomes can also enable discovery of divergence and gene flow along
58 chromosomes and how it is modified by recombination rate variation (Schumer et al. 2018; Martin et al.
59 2019). Furthermore, chromosome-level assemblies have been shown to greatly improve the power and
60 resolution of genome-wide association and QTL studies (Benevenuto et al. 2019; Markelz et al. 2017).

61 However, high-quality, chromosome-level, contiguous reference genome assemblies are often limited to
62 one or a few species in many groups of taxa, especially in non-model organisms. This is partly due to the
63 fact that generating near-complete chromosome-level assemblies normally requires integrating a mixture
64 of high fidelity short-read sequencing data (today typically Illumina), and more costly long-read
65 sequencing data (such as PacBio or Nanopore), genetic linkage mapping, optical (restriction site)
66 mapping, and/or chromatin interaction frequency data (Hi-C) (Rice and Green 2019; Ghurye and Pop
67 2019; Yang et al. 2020; Wei et al. 2020; Deschamps et al. 2018; Yu et al. 2019). These methods can be
68 prohibitively expensive and time consuming, especially for entire clades.

69 With 48 described species, *Heliconius* butterflies are a prime example of an adaptive radiation where
70 multiple chromosome-level reference assemblies could improve evolutionary analyses. Currently,

71 published high-contiguity genome assemblies (hereafter, reference genomes) exist for only two species –
72 *H. melpomene* (Davey et al. 2017) and *H. erato* (*H. erato lativitta* – Lewis et al., 2016; *H. erato*
73 *demophoon* – Van Belleghem et al., 2017). While these chromosome-level reference assemblies are
74 essential tools for genomic studies in *Heliconius*, each has limitations. At 275 Mb, *H. melpomene* has the
75 smallest *Heliconius* genome assembled to date (Edelman et al. 2019). Mapping short-read sequencing
76 data from other species with larger genomes to this reference genome is likely to result both in the loss of
77 information, due to loss of ancestral orthologous sequence in the *H. melpomene* genome, and spurious
78 read mapping to similar but non-orthologous regions. In contrast, the two *H. erato* reference genomes
79 (383 and 418 Mb) are among the largest *Heliconius* genomes assembled to date. However, while these
80 might be appropriate for studies focusing on closely related species (e.g. species within the *erato* clade),
81 mapping accuracy decreases in more divergent species (Prüfer et al. 2010) and better results are obtained
82 when mapping to closer reference genomes (Gopalakrishnan et al. 2017). Also, as we move from
83 comparative (e.g. phylogenomic) towards more functional genetics studies (Lewis et al. 2016; Lewis and
84 Reed 2019; Pinharanda et al. 2019), this genus could benefit greatly from higher-quality species-specific
85 genomic resources.

86 Recently, *de novo* draft genomes of 16 *Heliconius* species (Supplemental Table S1) have been assembled
87 (Edelman et al. 2019). These genomes were generated from Illumina PCR-free libraries sequenced at
88 deep coverage (at least 60X coverage) using paired-end 250-bp reads on the Illumina Hi-Seq 2500 and
89 assembled using *w2rap* (Clavijo et al. 2017), an extension of the DISCOVAR *de novo* genome assembly
90 method (<https://software.broadinstitute.org/software/discovar/blog/>; Love et al. 2016; Weisenfeld et al.
91 2014). This strategy results in high-quality genomes in terms of read accuracy, contiguity within
92 scaffolds, and genome completeness (87.5-97.3% complete single copy core BUSCO genes present;
93 Edelman et al. 2019). Nonetheless, because these assemblies (hereafter, *w2rap* assemblies) used only
94 short-read data, they were considerably more fragmented (scaffold N50 = 23-106 kb) than the *Heliconius*
95 reference genomes. Furthermore, scaffolds were not assigned to chromosomes.

96 A cost-effective approach for improving the contiguity of existing draft genomes is to use synteny-based
97 methods that identify potentially adjacent scaffolds from multi-species alignments. Such methods are
98 particularly efficient if high-quality reference genome assemblies of closely related species are available,
99 and especially if there is high synteny between the genomes of the draft and reference assemblies (Alonge
100 et al. 2019), as in *Heliconius*. While a limited number of genomic rearrangements have been identified in
101 *Heliconius* (Davey et al. 2017; Jay et al. 2018; Edelman et al. 2019; Meier et al. 2020), even species as
102 divergent as *H. melpomene* and *H. erato*, which last shared a common ancestor over 10 million years ago,
103 remain highly collinear (Davey et al. 2017). Synteny-based assembly should thus be especially effective
104 within this genus.

105 Here, we exploit the chromosome-mapped assemblies of the *H. melpomene melpomene* and *H. erato*
106 *demophoon* reference genomes to guide improvement of contiguity of the *w2rap* draft genome assemblies
107 of 16 *Heliconius* species. The *w2rap* scaffolds were ordered, oriented and anchored onto chromosomes,
108 resulting in a level of completeness of the scaffolded *w2rap* assemblies similar to that of reference
109 genomes. A potential weakness of our synteny-based assembly method is that it can miss structural
110 variation among species where it occurs. However, we use these scaffolded *w2rap* assemblies (hereafter,
111 reference-guided assemblies) to identify clade-specific local genomic expansions due to local duplications
112 with potentially functional consequences. To estimate how much structural variation we might be
113 missing, we also carry out a systematic search for candidate inversions in the genus using the original
114 *w2rap* scaffolds to detect break-points, and demonstrate that the results can be used to investigate
115 phylogenetic uncertainty and gene flow deep in the tree of *Heliconius* species.

116

117 **Results**

118 *Reference-guided genome assemblies and annotation*

119 Alternative haplotype scaffolds in the *w2rap* assemblies were first merged using HaploMerger2, reducing
120 the numbers of scaffolds by 31.3-64.6% and total assembly length by 3.4-25.9% (Supplemental Table
121 S2). These haplotype-merged scaffolds were then assembled using our reference guided approach.
122 Standard metrics for the resulting assemblies can be found in Supplemental Table S2. Contiguity of all
123 assemblies was considerably improved, with a reduction in the numbers of scaffolds to 0.9-16.8% of the
124 original *w2rap* assemblies (Supplemental Table S2; Supplementary Figs. S1-2). N50 length values were
125 14.2-20.0 Mb when using the *H. melpomene* genome as reference (the N50 of the *H. melpomene*
126 reference genome is *ca.* 14.3 Mb) and 7.1-11.5 Mb when using the *H. erato demophoon* genome (*H. erato*
127 *demophoon* reference genome N50 is *ca.* 10.7Mb). In general, scaffolds in the reference anchored to
128 chromosomes have a single corresponding scaffold in each of our reference-guided assemblies
129 (Supplemental Figs. S3-S36). Overall, 94.5-99.5% and 91.6-99.7% of bases in each reference-guided
130 assembly were anchored to chromosomes using the *H. melpomene* and *H. erato* references, respectively
131 (Supplemental Table S2; Figure 1B; Supplemental Fig. S37). For each species, the proportion of
132 reference-guided assembly length anchored to chromosomes was higher in assemblies guided by the
133 genome of the phylogenetically closest species, *H. doris* being the only exception. This species is distant
134 from both reference genomes, but has been inferred to be phylogenetically closer to *H. melpomene*
135 (Edelman et al., 2019; Kozak et al., 2018; Kozak et al., 2015). However, it shows a 0.2% higher
136 proportion of the assembly length included in scaffolds anchored to chromosomes using *H. erato*
137 *demophoon* as the reference, likely because the larger genome of *H. erato* contains ancestral sequence
138 that was lost by the smaller *H. melpomene* genome but retained in the early branching *H. doris*.

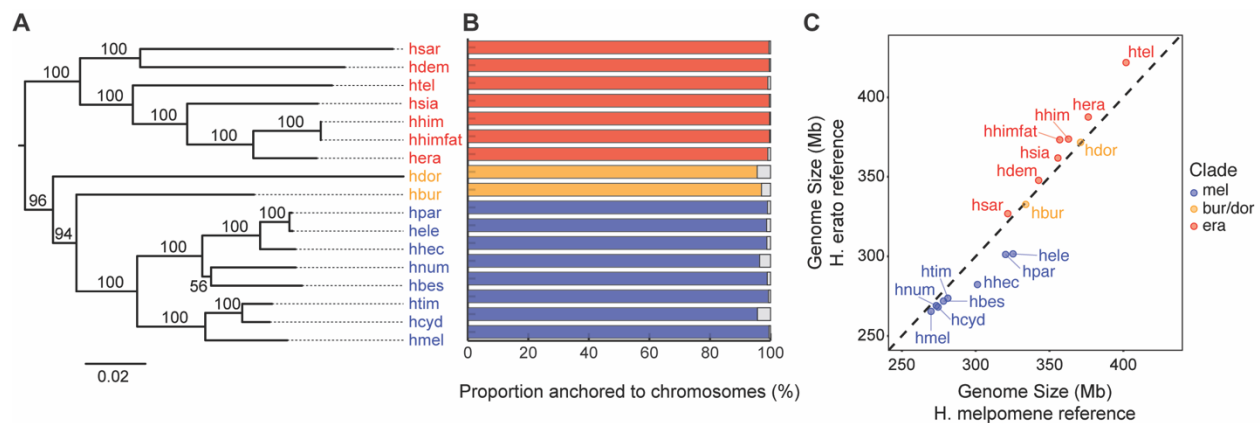
139 Genome sizes, considering only scaffolds anchored to chromosomes, varied between *ca.* 270-422 Mb
140 (Figure 1C; Supplemental Table S2). Phylogeny is a predictor of genome size: species within the
141 *erato/sara* clade have larger genomes (327-422 Mb) than species in the *melpomene/silvaniform* group
142 (270-325 Mb; Figure 1C), while genome sizes of *H. burneyi* and *H. doris* (334 and 371 Mb, respectively)
143 are more typical of those of the *erato/sara* group. The genome size of the *H. melpomene* reference-guided

144 assembly (270 Mb; 271 Mb including all scaffolds) is similar to that of the reference assembly (273 Mb;
145 275 Mb total; Davey et al., 2017), but both are smaller than estimates based on flow cytometry (292 Mb
146 +/- 2.4 Mb; Jiggins et al., 2005). The genome size of the *H. erato demophoon* reference-guided assembly
147 (388 Mb; 391 Mb total) is a little larger than that of the reference assembly (383 Mb; Van Belleghem et
148 al. 2017) but both are smaller than flow cytometry estimates (396-397 Mb, misnamed as *H. e. petiverana*;
149 Tobler et al., 2005). Despite the difference in genome sizes of the reference genome used to guide
150 scaffolding, genome sizes of our assemblies (considering only scaffolds anchored to chromosomes) did
151 not depend strongly on which reference genome was used (Spearman's Rank correlation test $\rho = 0.99$; P
152 $\ll 0.01$; linear regression slope=0.81; Figure 1C). Likewise, individual chromosome lengths of the
153 species assemblies scaffolded using the two different references differed little and were highly correlated
154 (Spearman's Rank correlation coefficient, $\rho = 0.94-0.99$; $P \ll 0.01$; linear regression slope = 0.80-1.05;
155 Supplemental Fig. S38; Supplemental Table S3).

156 Assembly completeness was evaluated by the presence of core arthropod genes in BUSCO. The
157 proportion of detected orthologs varied between 98.6 and 99.6%, values similar to those reported by
158 Edelman et al. (2019) for the original *w2rap* genomes (Supplemental Fig. S39; Supplemental Table S4).
159 There are however improvements (1-10% increase) in terms of the percentage of complete single copy
160 BUSCOs and a reduction in complete duplicated, fragmented and missing BUSCOs. These
161 improvements are likely a consequence of the increased contiguity and decreased scaffold redundancy
162 (due to the collapsing of alternative haplotype scaffolds) in the reference-guided assemblies, which allows
163 for better mapping of the core genes.

164 Gene annotation of *H. melpomene* and *H. erato demophoon* reference genomes was mapped onto the
165 reference-guided assemblies using the annotation lift-over tool LiftOff (Shumate and Salzberg 2020). We
166 considered only transcripts with ORFs (i.e. start and stop codon, no frame-shift mutation and no internal
167 stop codons) as successful mappings. Out of the 21,656 transcripts from 20,096 *H. melpomene* annotated
168 genes and 20,118 transcripts from 13,676 *H. erato demophoon* annotated genes, we were able to

169 successfully map 5,817-14,838 *H. melpomene* genes (6,217-16,007 transcripts) and 4,530-9,780 *H. erato*
 170 *demophoon* genes (6,139-14,472 transcripts) - Supplemental Table S5. The success of the gene annotation
 171 lift-over approach decreased with phylogenetic distance to the reference. While some of the genes that
 172 were not successfully lifted-over could potentially represent mis-annotations in the reference, this could
 173 also reflect differences in the structure of these genes or differences in gene composition between species.
 174 In fact, Liftoff is designed to map annotations between assemblies of the same or closely-related species
 175 and assumes gene structure is conserved between target and reference assemblies. Species-specific *de*-
 176 *novo* gene annotation using transcriptome data would be needed to obtain a more comprehensive
 177 annotation for all species.
 178



179
 180 **Figure 1 – Reference-guided assemblies.** **A** – Maximum-likelihood tree from whole mitochondrial genomes assembled here.
 181 Bootstrap values are shown next to the branches. The tree was rooted using the *E. tales* mitochondrial genome. **B** – Proportion of
 182 the reference-scaffolded assemblies length anchored to chromosomes. The results are shown for the reference-guided assemblies
 183 mapped to the closest reference (either *H. melpomene* or *H. erato demophoon*). For a complete report of the results see
 184 Supplemental Fig. S37. **C** – Reference-scaffolded assemblies nuclear genome sizes using either *H. melpomene* (x-axis) or *H.*
 185 *erato demophoon* (y-axis) as the reference. The dashed line represents the expectation if there was a 1:1 correspondence. In all
 186 panels, subclade memberships are represented by different colours – *melpomene/silvaniform* (blue), *burneyi+doris* (yellow),
 187 *erato/sara* (red). Species codes for all the new reference-guided assemblies are as follows: hmel – *H. melpomene*; hcyd – *H.*
 188 *cydno*; htim – *H. timareta*; hbes – *H. besckei*; hnum – *H. numata*; hhec – *H. hecale*; hele – *H. elevatus*; hpar – *H. pardalinus*; hbur

189 – *H. burneyi*; hdor – *H. doris*; hera – *H. erato*; hhimfat – *H. himera*; hhim – *H. himera*; hsia – *H. hecalesia*; htel – *H. telesiphe*;
190 hdem – *H. demeter*; hsar – *H. sara*.

191

192 ***Whole mitochondrial genome assemblies***

193 The *de novo* assembly of *Heliconius* mitochondrial genomes allowed the recovery of partially complete
194 mitochondrial sequences (*ca.* 15-kb, typical of *Heliconius*) for all 16 species, including part of the
195 mitochondrial DNA control region. A genealogy based on these mitochondrial genomes (Figure 1A) did
196 not differ from that for the mitochondrial genomes assembled using reference-aided approaches (Kozak et
197 al. 2015; Massardo et al. 2020), thereby validating our *de novo* approach.

198

199 ***Improved mapping efficiency using the reference-guided assemblies***

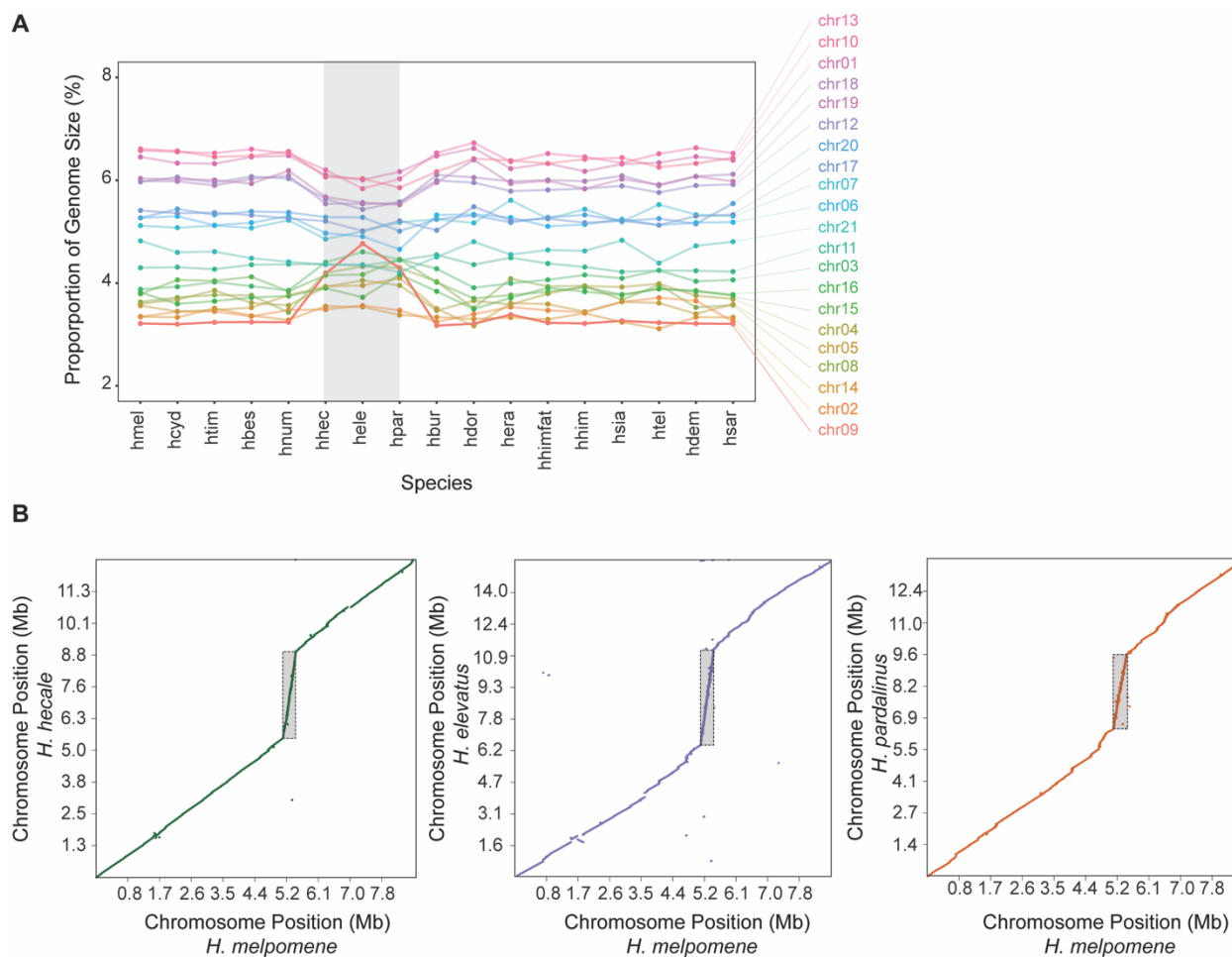
200 Mapping the original *w2rap* Illumina short read sequence data to the reference-guided genome assemblies
201 of their own species resulted in 0.45-12.77% more mapped reads and 0.60-40.77% more properly paired
202 reads than when mapping to the closest reference genome (Supplemental Table S6). These mappings also
203 show an increase of the depth of coverage (1.02-2.29 times the coverage obtained when mapped to the
204 closest reference; Supplemental Table S6), which is also more uniform along chromosomes
205 (Supplemental Figs. S40-S41). The largest increases in depth of coverage were observed for *H. burneyi*
206 and *H. doris*, which are the two sequenced *Heliconius* species phylogenetically most distant to either
207 reference genome. Increases in depth of coverage tend to be larger in species in the *erato/sara* clade (1.06
208 to 1.97 times more coverage) than in species in the *melpomene/silvaniform* clade (1.02 to 1.35 times more
209 coverage). This is expected since species sampled within the *erato/sara* clade were typically more
210 divergent from *H. erato* than species in the *melpomene/silvaniform* group are from *H. melpomene*.
211 Importantly, these results show how studies focusing on *Heliconius* species with deeper divergence to
212 both *H. melpomene* and *H. erato* will benefit from mapping re-sequence data to the reference-guided

213 assemblies generated here. Also, the greater uniformity of coverage along chromosomes when mapping
214 reads to the reference-guided assemblies suggests that they should better capture fine-scale structural
215 variation. This likely reflects the ability of the high sequencing fidelity of the original *w2rap* assemblies
216 to resolve short imperfect repeats (< 500 bp long) (Love et al. 2016; Edelman et al. 2019) that differ
217 between species.

218

219 ***Genome expansions and gene duplications***

220 Although genome sizes vary among *Heliconius* species, the relative but not absolute sizes of
221 chromosomes were generally conserved (Figure 2A, Supplemental Fig. S42). The three closely related
222 species with the largest genomes in the *melpomene*/silvaniform group (*H. hecale*, *H. elevatus* and *H.*
223 *pardalinus*) are exceptions. Upon closer inspection, the variation in chromosome size in these three
224 species is particularly accentuated on chromosome 9 (Figure 2A; Supplemental Fig. S42). Alignment of
225 reference-guided assemblies of these three species to the *H. melpomene* reference genome suggests that
226 the increase in size of chromosome 9 mainly corresponds to a single genomic region in *H. melpomene*
227 (Hmel209001o:5125000-5450000, Figure 2B). This region is *ca.* 325 kb long in *H. melpomene* but the
228 scaffolds that map to it total over 10x as long (3.350-4.125 Mb) in the *hecale/elevatus/pardalinus* trio.



229

230 **Figure 2 – Chromosome size variation and local genomic expansions.** **A** – Chromosome sizes in proportion to the genome
 231 size across the different species for the reference-guided assemblies mapped to the *H. melpomene* reference genome.
 232 Chromosome relative sizes are generally similar across species, with the exception of *H. hecale*, *H. elevatus*, and *H. pardalinus*,
 233 particularly chromosome 9. **B** – Genome to genome alignment showing the repeat region on chromosome 9 (highlighted by the
 234 grey rectangles) in the species trio: *H. hecale*, *H. elevatus* and *H. pardalinus*.

235

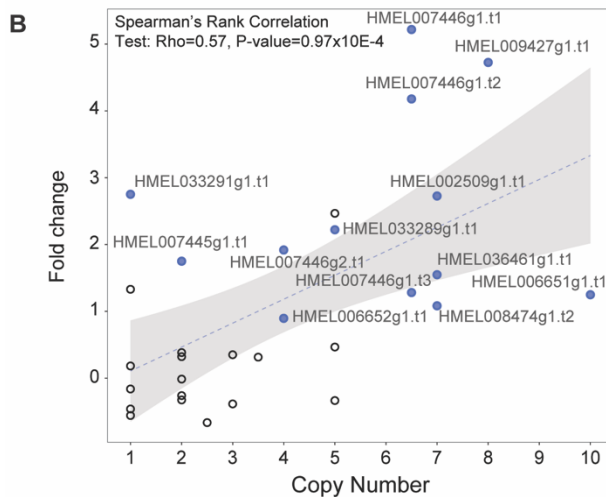
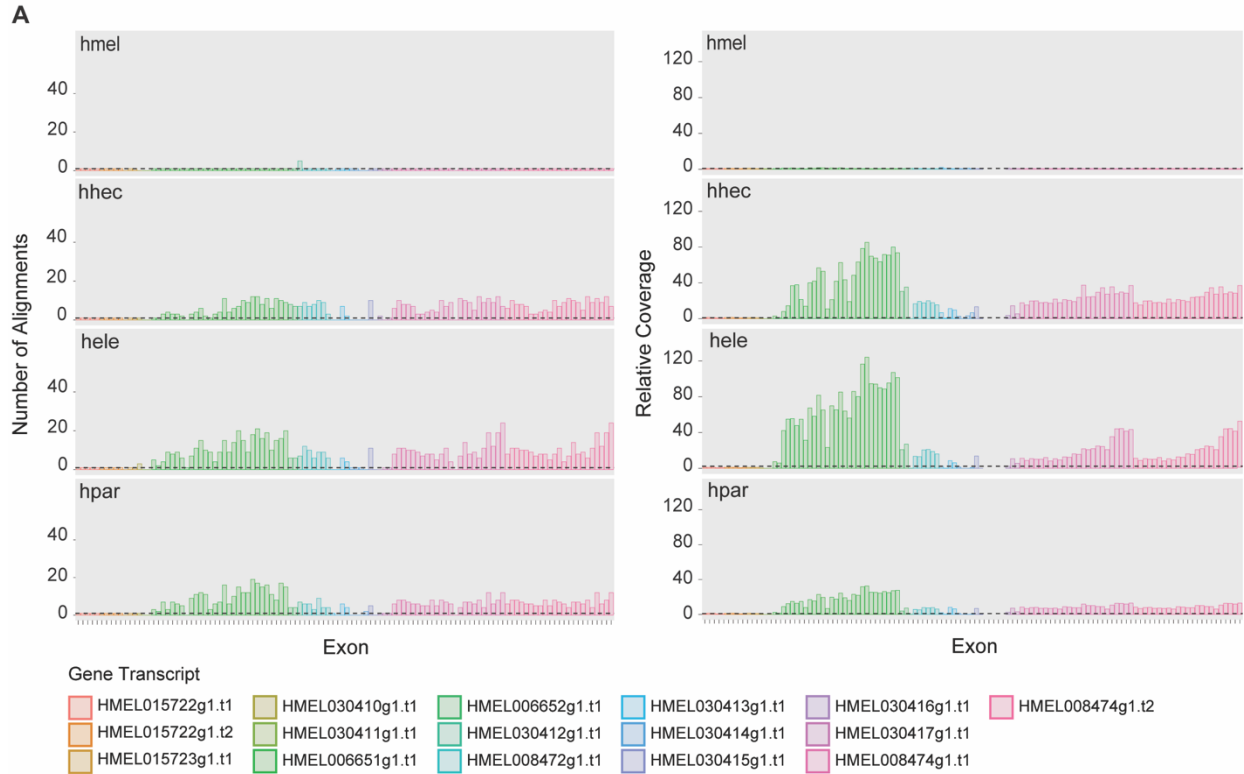
236 We investigated whether other genomic regions also underwent an increase in size specifically in these
 237 three species. There are four regions that show exceptionally high coverage in these three species (at least
 238 5-fold local increase in the *hecale/elevatus/pardalinus* trio and less than 2-fold local increase in every
 239 other species, in at least two consecutive 25 kb windows). These included the region on chromosome 9
 240 discussed above and three other regions on chromosome 2 (Hmel202001o:4075000-4125000),

241 chromosome 4 (Hmel204001o:5650000-5875000), and chromosome 8 (Hmel208001o:3300000-3475000)
242 (Supplemental Fig. S43, Supplemental Table S7). In contrast, mapping reads onto the reference-guided
243 assemblies resulted in more uniform coverage in these regions (Supplemental Figs. S44-S47). This
244 suggests the repeats are divergent enough so that they could be at least partially resolved in the *w2rap*
245 assemblies.

246 All repeat regions harbor protein coding genes (Supplemental Table S7), as annotated in the *H.*
247 *melpomene* reference genome, and thus structural variation in these regions could have resulted in gene
248 copy number variations with potential functional consequences. To address this, we first estimated the
249 copy number of each exon based both on i) the number of valid alignments of *H. melpomene* exon
250 sequences onto the reference-guided assemblies and ii) the normalized mean per base coverage for each
251 exon, mapping *H. hecale*, *H. elevatus* and *H. pardalinus* re-sequencing data to the *H. melpomene*
252 reference. Copy number estimates based on number of exon sequence alignments are generally lower than
253 estimates based on read coverage (Figure 2C; Supplemental Figs. S48-S50). Nevertheless, for both
254 measures, exonic copy number is much larger in *hecale/elevatus/pardalinus* trio than in *H. melpomene*,
255 suggesting duplications of the corresponding genes. It should also be noted that copy number is variable
256 between exons of the same gene and, and while it can probably be attributed to different alignment
257 efficiency due to variation in exons sequence length (Li 2018), it might also be due to partial duplications
258 of some of these genes. However, given the fragmented nature of the *w2rap* genomes, we could not assess
259 whether genes were wholly or partially duplicated, nor whether the duplications were translocated
260 elsewhere in the genome or are located in the same region as in *H. melpomene*. Long read sequencing
261 data would be required to resolve this.

262 These gene duplications could result in pseudogenes, in which case we might expect to find stop codons
263 within exons and a relaxation of selection. In general, we find high exon copy numbers even after
264 excluding exon copies with stop codons (10-14% exon copies have a stop codon; Supplemental Figs.
265 S51-S54). Also, dN/dS estimates are overall close to zero, suggestive of purifying selection

266 (Supplemental Figs. S55-S58). RNA-Seq shows a significant correlation between gene copy number and
 267 expression levels and that many of these genes have significantly higher expression in *H. pardalinus* than
 268 in *H. melpomene* (Figure 2D). Together, these results suggest that many of the gene copies are functional
 269 and that CNV at these genes resulted in altered gene dosage.
 270



271

272 **Figure 3 - Copy number variation and increased expression levels of genes in the repeat region on chromosome 9. A –**
273 Exon copy number variation for genes in the chromosome 9 repeat region. The number of alignments (left panel) and relative
274 coverage (right panel) were used as proxies of copy number. Relative coverage was calculated by dividing exon coverage by the
275 median genomic coverage, based on mappings to the *H. melpomene* reference. On the left panel, coloured bars depict the number
276 of alignments to the expected chromosome. Dashed horizontal lines on both plots represent a copy number of one. *Our new H.*
277 *melpomene* assembly was also included as a control. **B –** Fold change in expression level in *H. pardalinus* compared to *H.*
278 *melpomene* (y-axis) as a function of *H. pardalinus* transcript copy number (x-axis). For each transcript, copy number was
279 calculated as the median number of alignments across exons for the *H. pardalinus* sample. Full blue circles represent transcripts
280 for which the levels of expression in *H. pardalinus* were significantly higher than in *H. melpomene*. The best fit linear model
281 regression line and confidence intervals are depicted by the dashed line and grey band, respectively. Species codes are as in
282 Figure 1.

283

284 ***Inversions fixed between the two Heliconius major clades***

285 Reference-guided assemblies will inevitably be ineffective at detecting inversions or translocated regions,
286 so it seems important to quantify potential drawbacks of our approach. Previous studies showed that some
287 regions of the genome with unusual phylogenomic patterns in the *erato/sara* clade were associated with
288 inversions (Edelman et al. 2019). Here, we make a systematic search for small to medium sized inversion
289 differences among *Heliconius* species, focusing on those 50 kb - 2 Mb long. At the broad scale, the
290 genome structure of the reference-guided assemblies is constrained by the reference genome, so we
291 returned to the *w2rap* scaffolds (after collapsing alternative haplotypes with HaploMerger2), mapping
292 these to the *H. melpomene* and the *H. erato* reference genomes to infer inversion breakpoints. In total, and
293 after filtering, we found 2560 and 3829 scaffolds for which one end aligns to the positive strand of the
294 reference genome and the other end maps to the negative strand, using the *H. melpomene* and *H. erato*,
295 respectively. Of these, 900 and 1786 support inversions 50 kb - 2 Mb long, yielding 345 and 741 unique
296 candidate inversions across all species (mapping to *H. melpomene* and *H. erato demophaon*,
297 respectively), supported by at least one scaffold per species, some of which were shared by multiple
298 species (Supplemental Table S8).

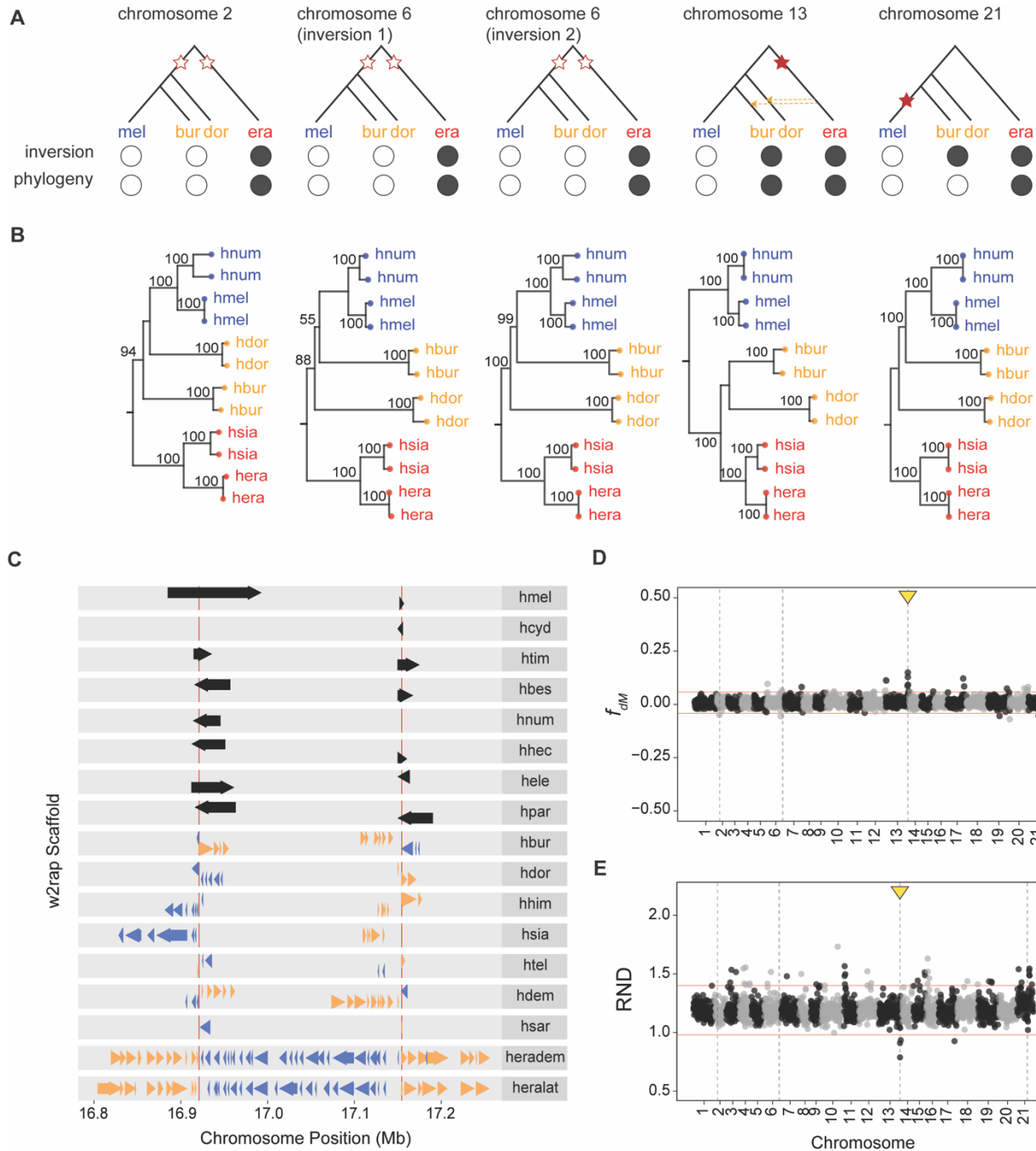
299 Our systematic search confirmed previous findings of two independent but overlapping introgressed
300 inversions around a color patterning locus on chromosome 15 (one shared by *H. sara*, *H. demeter*, *H.*
301 *telesiphe* and *H. hecalesia* and the other shared by *H. pardalinus* and *H. numata*) and another inversion
302 on chromosome 2 (shared by *H. erato* and *H. telesiphe*) (Edelman et al. 2019). In addition, we found five
303 moderately large inversions, previously identified as inversion candidates based on alignments between
304 *H. melpomene* and *H. erato* reference genomes (Davey et al. 2017), to be fixed between major branches
305 of the *Heliconius* phylogeny (Figure 4A). Such shared inversions occur on chromosome 2 (Supplemental
306 Fig. S59), chromosome 6 (Supplemental Fig. S60), chromosome 13 (Figure 2C; Supplemental Fig. S61)
307 and the Z chromosome, chromosome 21 (Supplemental Fig. S62; Supplemental Table S9). The two
308 inversions on chromosome 6 occur in tandem and are further supported by linkage maps in *H. melpomene*
309 and *H. erato* (Davey et al. 2017).

310 The placement of *H. doris* and *H. burneyi* in the *Heliconius* phylogeny remains contentious. These two
311 species have been inferred to be more closely related to the *melpomene*/silvaniform clade than to the
312 *erato/sara* clade (see also the mitochondrial tree of Figure 1A), but node supports are relatively weak and
313 the internal branches leading to *H. doris* and *H. burneyi* are short (Kozak et al. 2015). We here test
314 whether homologous inversions can be used as a phylogenetic character to resolve their placement. Figure
315 1A). Both *H. burneyi* and *H. doris* group with the *melpomene*/silvaniform group based on the orientation
316 of three inversions on chromosomes 2 and 6, but with the *erato/sara* clade based on the of homologous
317 inversions on chromosomes 13 and Z. These groupings are further confirmed based on maximum-
318 likelihood (ML) phylogenetic analysis of the inversion regions using a subset of species (Figure 4B). The
319 only exception is that *H. doris* and *H. burneyi* both group with *melpomene*/silvaniform species for the
320 inversion on the Z chromosome in the ML phylogeny, rather than with *erato/sara* (Figure 4B), as might
321 be expected solely based on presence/absence of the inversion. (Supplemental Fig. S61). This apparent
322 contradiction can be reconciled if the *melpomene*/silvaniform clade is sister both to *H. doris* and to *H.*
323 *burneyi*, but the Z chromosome inversion was derived in the *melpomene*/silvaniform ancestor after it split

324 from the *burneyi* and *doris* lineages. In this scenario, the sharing of the inversion between *burneyi/doris*
325 and the *erato/sara clade* on chromosome 13 must be explained by secondary transfer *via* introgression,
326 perhaps soon after the initial separation of the two major clades, or through incomplete lineage sorting of
327 an inversion polymorphism at the base of *Heliconius*. Previously, reticulation involving *H. burneyi* and *H.*
328 *doris* and the *erato/sara* group had been hypothesized, but different phylogenomic methods gave different
329 results (Kozak et al. 2018).

330 To test for introgression genome-wide, we used Patterson's *D*-statistic (Green et al. 2010; Durand et al.
331 2011). Specifically, we calculated *D* for all possible topologies of the triplets (*H. erato* – *H. melpomene* –
332 *H. doris*) and (*H. erato* – *H. melpomene* – *H. burneyi*), in each case using *Eueides tales* as an outgroup.
333 For a given triplet of species, the minimum absolute whole genome Patterson's *D*-statistic should result
334 for the topology that best describes the relationships between species. We found that this is the case when
335 *H. erato* is the inner outgroup in both triplets, implying that *H. burneyi* and *H. doris* are more closely
336 related to *H. melpomene*. Yet, Patterson's *D*-statistics are still significantly different from zero
337 (Patterson's *D* = 0.037 and 0.060 for *H. doris* and *H. burneyi*, respectively) based on block-jackknifing,
338 providing evidence of introgression among lineages leading to *H. burneyi*, *H. doris*, and *H. erato*. We
339 also used an alternative branch length-based approach, QuIBL (Quantifying Introgression via Branch
340 Lengths; Edelman et al. 2019), which further corroborated these results (Supplemental Table S10). To
341 understand which specific genomic regions were shared by introgression between these species, we
342 estimated the excess of shared derived mutations between *H. doris* and *H. burneyi* with either *H.*
343 *melpomene* or *H. erato*, using the f_{dM} statistic (Malinsky et al. 2015). The f_{dM} estimates in windows
344 overlapping the chromosome 13 inversion show a significant deviation from the genomic average, with
345 an excess of shared variation between *H. erato* and both *H. doris* and *H. burneyi* (Figure 4D;
346 Supplemental Fig. S63). Likewise, relative divergence between *H. erato* to both *H. burneyi* and *H. doris*
347 is significantly reduced in the inversion region (Figure 4D; Supplemental Fig. S64). We also used QuIBL
348 with the triplets (*H. erato* – *H. melpomene* – *H. doris*) and (*H. erato* – *H. melpomene* – *H. burneyi*) to

349 calculate the likelihood that the discordant phylogenies at the chromosome 13 inversion were due to
350 introgression. For both triplets, the average internal branch of gene trees within the chromosome 13
351 inversion is larger than the genome-wide average, corresponding to a 90.1% and 86.7% probability of
352 introgression, respectively (Supplemental Fig. S65) We found no significant f_{dM} or relative divergence
353 estimates for any of the other four inversions, including the Z chromosome inversion. These results
354 strongly support the argument that the chromosome 13 inversion of *H. doris* and *H. burneyi* results from
355 introgression from the common ancestor of the *erato/sara* clade.



356

357 **Figure 4 – Large Inversions fixed between the *melpomene*/silvaniform and *erato/sara* clades.** **A** – Possible scenario for the
 358 origin and sharing of the inversion. Stars represent inversions and the branch in which these likely took place. Empty stars are
 359 used when the inversion is equally likely to have occurred in two branches. Introgression between branches are represented by
 360 arrows, the direction of the arrow indicating directionality. Full and empty circles represent the orientation of the inversion and
 361 the groupings based on the phylogeny. **B** – ML phylogenies of each major inversion estimated using IQ-TREE, based on
 362 mapping of re-sequence data to the *H. melpomene* reference genome. **C** – Alignments of scaffolds to the *H. melpomene* reference genome

363 supporting the inversion on chromosome 13. Inversion breakpoints are depicted by the vertical red lines. Scaffold alignments are
364 shown represented by the arrows, the direction and colour of the arrows representing whether the alignments are to the forward
365 strand (blue rightwards arrows) or the reverse strand (yellow leftwards arrows). Black arrows represent alignments spanning the
366 inversion breakpoints. **D,E** f_{dM} and relative node depth (RND) statistics along the genome. Both statistics were calculated in 25
367 kb non-overlapping windows across the genome, based on mapping of re-sequencing data to the *H. melpomene*. Chromosomes
368 are shown with alternating grey and black colours. The location of inversions is given by the dashed vertical lines while
369 horizontal red lines represent ± 3 SD from the mean f_{dM} and RND values. Outlier windows overlapping the chromosome 13
370 inversion are indicated by the yellow arrows. Positive f_{dM} values (and lowered RND) indicate an excess of shared variation
371 between *H. burneyi* with *H. erato* and negative values of f_{dM} represent an excess of shared with *H. melpomene*. In this test, *H.*
372 *melpomene* and *H. burneyi* were considered to be the ingroup species and *H. erato* the inner outgroup. Derived alleles were
373 determined using *E. tales*. Species codes are as in Figure 1.

374

375 **Discussion**

376 ***Genome assembly improvements and limitations***

377 Here, we implement a purely *in silico* reference-guided scaffolding approach to improve draft genome
378 assemblies of 16 species from across the genus *Heliconius*. The contiguity of our new assemblies is
379 similar to that of the reference genomes. For instance, the *H. melpomene* reference genome assembly has
380 38 scaffolds anchored to chromosomes (99.1% of the assembly length), and the reference-guided
381 assemblies scaffolded based on this reference have 31-36 scaffolds anchored to chromosomes
382 representing 83.8-99.1% of the total assembly. Similarly, the *H. erato* reference has 195 scaffolds
383 anchored to chromosomes (100% of the assembly length), and the reference-guided assemblies scaffolded
384 based on this reference have 94-168 scaffolds anchored to chromosomes representing 83.2-99.9% of the
385 total assembly.

386 Our reference-guided assembly strategy assumes that the orientation and order of the new scaffolds in our
387 genomes is the same as the reference. Clearly, it may not fully represent the structure of these genomes.
388 While small genomic rearrangements spanned by the original scaffolds (i.e. rearrangements in relation to

389 the reference present within *w2rap* scaffolds) are recovered in our reference-guided assemblies, larger
390 genomic rearrangements relative to the reference not spanned by a single *w2rap* scaffold can be missed.
391 One such example is the case of the known *ca.* 400 kb inversion around a color pattern locus known from
392 *H. numata* and *H. pardalinus* on chromosome 15 (Jay et al. 2018) which we do not recover in the
393 reference-guided assemblies, in either species. This is also the case for the five large inversions we
394 discovered that are fixed between the two *Heliconius* major clades, depending on the reference genome
395 used to guide scaffolding. For instance, for species in the *melpomene*/silvaniform group, all reference-
396 guided assemblies mapped to the *H. melpomene* reference have the correct orientation for all five
397 inversions, but not when mapped to the *H. erato* reference. The same logic applies for species in the
398 *erato/sara* group, when mapped to different references. For *H. burneyi* and *H. doris* however, neither of
399 the two alternative reference-guided assemblies recovers the correct orientation of all five inversions,
400 since these two species share the same orientation as *H. melpomene* for the inversions on chromosome 2
401 and 6, but not for chromosomes 13 and 21 (for which they have the same orientation as *H. erato*). Long-
402 read sequence data and/or linkage mapping could better resolve the genome structure of species-specific
403 assemblies. Nevertheless, our reference-guided assemblies represent a major improvement over mapping
404 short read data directly to existing reference genomes, and researchers that use these and other reference-
405 guided assemblies for this purpose will see marked improvement in their data quality.

406 Mapping the original *w2rap* Illumina reads back to the reference-guided assembly of their own species
407 resulted in more than doubling of the median genomic coverage in some species and in a more uniform
408 depth of coverage along the genome than when mapping to the closest reference genome. Mapping
409 efficiency improves in all species studied here (Supplemental Table S6), but we see the greatest benefits
410 in *H. burneyi* and *H. doris*, the two *Heliconius* species studied here that are most divergent from either
411 reference genome assembly. In these two species, the proportion of properly mapped reads increases from
412 53.6% and 49.9% (for *H. burneyi* and *H. doris*, respectively) when mapped to the *H. melpomene*
413 reference genome, to 90.7% and 90.6% when mapped to their own reference-guided assembly. In another

414 study (Rosser et al., in preparation), a linkage map produced from backcrosses of F1 male hybrids,
415 between *H. pardalinus butleri* and *H. p. sergestus*, to the parental *H. p. butleri* population contained ca.
416 29% more markers when RADseq data was mapped to the new *H. pardalinus* reference-guided assembly
417 than to the *H. melpomene* reference. The use of reference-guided assemblies of the closest species thus
418 greatly improves the efficiency of mapping resequencing data over mapping to the currently available
419 reference genomes.

420 The more uniform depth of coverage when mapping to reference-guided assemblies also leads to
421 improvements in discovery of species-specific genomic variation and in resolving imperfect repeat
422 regions. Indeed, given variation in genome sizes among *Heliconius* species (275-418 Mb), the new
423 genomes are helpful in mapping variation that is otherwise lost or mapped to similar but non-orthologous
424 regions of more divergent reference genomes. Variations in depth of coverage along the genome, if not
425 properly filtered, could lead to biased estimates of diversity and divergence. For example, partially
426 divergent repeats mapping to the same region in the reference genome (resulting in unusually high
427 coverage) could inflate local estimates of diversity and thus be spuriously implicated as important sites
428 for species divergence. This is especially likely in studies focusing on *Heliconius* species with larger
429 genomes when mapping reads to the *H. melpomene* reference, the smallest genome assembled here. On
430 the other hand, if regions with abnormal coverage are filtered out, information could be lost by discarding
431 genomic regions with potentially relevant biological signals. For example, highly divergent regions may
432 result in abnormally low coverage, even though such regions could be important for diversification of the
433 group.

434 Overall, our reference-guided assemblies extend the number of applications for which these genomes can
435 be used. By ordering, orienting and anchoring scaffolds onto chromosomes, the new reference-guided
436 assemblies enable improved chromosome-scale analyses and genome scans.

437

438 ***Prevalence of structural variants in Heliconius butterflies***

439 Chromosomal rearrangements can play a major role in adaptation and speciation (Wellenreuther and
440 Bernatchez 2018; Feulner and De-Kayne 2017). By reducing recombination, inversions can facilitate the
441 build-up of associations between loci involved in traits responsible for reproductive isolation, and thus
442 could play a role in establishing or reinforcing species barriers (Noor et al. 2001). Inversions can also be
443 favored by selection by maintaining adaptive combinations of locally adapted alleles (Todesco et al.
444 2020; Faria et al. 2019; Christmas et al. 2019).

445 In *Heliconius*, a previous study focusing on two closely related species (*H. melpomene* and *H. cydno*)
446 found no evidence for major inversions that might have aided speciation (Davey et al. 2017). Thus,
447 *Heliconius* appeared to have low rates of chromosomal rearrangement, and selection without the help of
448 chromosomal rearrangements was believed to maintain the differences between these two species. In
449 another species, *H. numata*, the tandem inversion complex that forms the supergene locus *P* allows the
450 maintenance of a multi-allele color pattern polymorphism of mimicry morphs (Joron et al. 2011). The
451 first inversion in the tandem supergene was most likely transferred to *H. numata* via introgression from
452 *H. pardalinus* (Jay et al. 2018). An independently derived inversion has since been found for the same
453 colour pattern determination region in four species in the *erato/sara* clade (*H. telesiphe*, *H. hecalesia*, *H.*
454 *demeter* and *H. sara*). This inversion was also inferred to have been shared via introgression, this time
455 between *H. telesiphe* and *H. sara* sub-clades (Edelman et al. 2019). In parallel hybrid zones of *H. erato*
456 and *H. melpomene*, 14 and 19 polymorphic inversions were detected within each species, respectively.
457 Most of these inversion polymorphisms did not differ across the hybrid zones of either species. The
458 frequency of only one inversion on chromosome 2 (different to the inversion on chromosome 2 reported
459 here) differed strongly across the hybrid zone between highland *H. e. notabilis* and lowland *H. e. lativitta*
460 races, and may be associated with ecological adaptation to altitude (Meier et al. 2020).

461 In the 16 species studied here, we systematically searched for inversions. We found several candidates in
462 all 16 species (17-61 and 40-126 inversions per species, compared with *H. melpomene* and *H. erato*,
463 respectively), including some previously described (Davey et al. 2017; Jay et al. 2018; Edelman et al.

464 2019). However, the strategy we implemented to search for inversions, i.e. split alignment of *w2rap*
465 scaffolds to forward and reverse strands of the reference genomes, is liable to false positives because
466 small interspersed duplications and translocations (for example due to transposable element activity)
467 might generate a similar signal. This is particularly likely in highly repetitive regions where we find many
468 different, partially overlapping candidate inversions in many or all species (Supplemental Fig. S66). It is
469 thus difficult to assess, solely based on these results, how pervasive inversions are among *Heliconius*
470 species. While it is possible that inversions in this group occur more frequently than earlier studies
471 indicated (Heliconius Genome Consortium 2012; Davey et al. 2017), long-read or linked-read
472 sequencing, preferably with a larger set of individuals per species, will ultimately be needed to answer
473 this question.

474 However, by focusing on phylogenetically informative inversions, we were able to verify five candidate
475 inversions that occurred deep in the *Heliconius* phylogeny. We searched for inversions fixed between the
476 *melpomene/silvaniform* and *erato/sara* groups. We are confident that these were correctly identified for
477 two reasons. First, the inversions are supported in multiple species, with breakpoint coordinates consistent
478 among species. Second, while a mis-assembly in the reference genome could generate a misleading signal
479 of inversion, this is unlikely to happen for the same candidate inversion when mapping to two or more
480 different genomes. All five of these inversions were supported in multiple species when mapping
481 scaffolds to either reference genome, the orientation of the inversion being mirrored depending on the
482 reference used. Furthermore, the inversion orientation shows a phylogenetic signal (fixed between clades)
483 that is unexpected if due to mis-assembly in one of the reference genomes.

484 The most parsimonious scenario that explains both the orientation and the phylogenetic pattern, taking all
485 five inversions into account, supports the hypothesis that *H. burneyi* and *H. doris* are more closely related
486 to the *melpomene/silvaniform* group than to the *erato/sara* group (Figure 4), in line with previous studies
487 (Edelman et al. 2019; Kozak et al. 2018). The relationships of the inversion on chromosome 13, which
488 groups *H. burneyi*, *H. doris* and the *erato/sara* group, is then explained by introgression between the

489 ancestor of the latter group and both *H. burneyi* and *H. doris* (Supplemental Figs. S63-S65). Introgression
490 almost certainly occurred from the *erato/sara* clade into *H. burneyi* and *H. doris*, since the relative
491 divergence between *H. erato* and both *H. burneyi* and *H. doris* is reduced at the chromosome 13 inversion
492 when compared to the rest of the genome (Figure 4E), but not between *H. erato* and *H. melpomene* as
493 expected if introgression took place in the other direction (Supplemental Fig. S67). Interestingly, *H.*
494 *burneyi* has been inferred to be on a separate branch from *H. doris*, although the two branches were
495 connected by introgression (Kozak et al. 2015, 2018). This suggests that introgression of the chromosome
496 13 inversion occurred twice. Either there were two separate introgression events from the *erato/sara*
497 ancestor to *H. burneyi* and to *H. doris*, or the inversion first passed from the *erato/sara* ancestor to one of
498 these two species which then passed it to the other. Altogether, and in line with previous studies (Edelman
499 et al. 2019; Kozak et al. 2018), this inversion supports a hypothesis that hybridization and introgression
500 among species occurred early in the radiation of *Heliconius*, as well as later, between more closely related
501 species extant within each major subgroup. Alignment issues have previously made it hard to interpret
502 evidence for introgression so deep in the phylogeny. Although we still do not know whether it has
503 functional implications, our finding of transfer of this chromosome 13 inversion provides stronger support
504 for introgression deeper in the Heliconiini tree than hitherto.

505 Species may also differ in gene copy number. Copy number can affect the phenotype by altering gene
506 dosage, altering the protein sequence, or by creating paralogs that can diverge and gain new functions
507 (Iskow et al. 2012). Copy number variation has been implicated in ecological adaptation – e.g. insecticide
508 resistance in *Anopheles* mosquitoes (Lucas et al. 2019), climate adaptation in white spruce (Prunier et al.
509 2017) and polar bears (Rinker et al. 2019), and resistance to malaria in humans (Leffler et al. 2017). Gene
510 copy number may also be involved in reproductive barriers among species – e.g. hybrid lethality in
511 *Mimulus* sympatric species (Zuellig and Sweigart 2018). Gene duplications within specific gene families
512 in the branch leading to *Heliconius* have been linked to evolution of visual complexity, development,
513 immunity (Heliconius Genome Consortium 2012) and female oviposition behavior (Briscoe et al. 2013).

514 Within the genus, gene copy number variation is plausibly associated with species divergence between *H.*
515 *melpomene* and *H. cydno* (Pinharanda et al. 2017).

516 Here we show that the genomes of different *Heliconius* species vary in size, with each chromosome
517 typically showing similar directional changes in size between species. Thus, genome expansions and
518 reductions in size seem typically to involve all chromosomes, so that the relative sizes of chromosomes
519 are conserved. Our study of the *Heliconius* butterfly radiation conforms, on a much more restricted
520 phylogenetic scale, to the pattern of relative chromosome size across eukaryotes: across many orders of
521 magnitude of genome size, relative chromosome sizes can be predicted based on chromosome number
522 and are almost always between $\sim 0.4x$ and $\sim 1.9x$ the mean (Li et al. 2011).

523 We find that, in *Heliconius*, genomic expansion is at least partially driven by small genomic regions that
524 became hotspots of repeat accumulation. Amplified regions tend to be conserved among closely related
525 species and are more frequent towards chromosome ends (Supplemental Fig. S68). However, in a
526 subclade of three closely related species (*H. hecale*, *H. elevatus* and *H. pardalinus*), we found four small
527 genomic regions with highly aberrant increases in size and exon copy number compared to related
528 species. These three are therefore exceptions to more or less orderly pattern across chromosomes in the
529 rest of the genus. Our approach for detecting exceptional repeat regions relies on the *H. melpomene*
530 genomic arrangement as a backbone. Hence, we do not know whether the additional copies we found
531 were translocated to other regions of the genome of these three species, or whether they remained
532 clustered as tandem copies at a single genomic location. By aligning the reference guided-assemblies to
533 the *H. erato demophoon* reference, we found a signal of local expansion in chromosome 9 (Supplemental
534 Figures S25-27) which would support that the repeats occur in tandem. However, we could not assess
535 whether this was also the case for the repeat regions in the three other chromosomes. Transposable
536 element activity is one possible mechanism responsible for these repeats (Bourque et al. 2018), and rapid
537 divergent transposable element evolution has already been found among *Heliconius* species (Ray et al.
538 2019). Hybridization could also spread variation in copy number among the species. *H. hecale*, *H.*

539 *elevatus* and *H. pardalinus* are sympatric in the Amazon where they are known to hybridize occasionally
540 (Mallet et al. 2007; Rosser et al. 2019). We found significantly higher copy numbers in the Amazon than
541 in extra-Amazonian populations of these species (Supplemental Fig. S69). The correlations of copy
542 number among species in an area suggests that hybridization might indeed have been involved.

543 Genes within highly amplified regions had significantly higher expression levels in *H. pardalinus* than in
544 *H. melpomene* (Figure 3B), which suggests that this gene copy variation could have functional
545 significance. An examination of genes within these regions shows that orthologs of these genes in
546 *Drosophila* are involved in important functions such as cytoskeletal processes and oogenesis (i.e.
547 *Dhc64C*, *sima*, *shotgun*, and *capicua*; Supplemental Table S7). Evaluating how variation in these critical
548 genes impacts phenotypes in *H. pardalinus*, *H. elevatus*, and *H. hecale* will advance our understanding of
549 the role of copy number variation in evolution.

550 The full extent to which inversions and copy number variation play a role in the evolution of *Heliconius*
551 butterflies remains to be examined. However, the current work suggests that the types of structural
552 variation examined here could be relevant to diversification. The characterization of intra- and
553 interspecific structural variation in this group could thus be an especially promising avenue for future
554 studies particularly now that improvements in sequencing technology allow for more detailed, rigorous
555 and cost-effective detection of structural variants (Wellenreuther et al. 2019; Logsdon et al. 2020).

556

557 **Methods**

558 ***Genome merging and scaffolding***

559 We used the draft genome scaffolder MEDUSA (Bosi et al. 2015) for reference-aided assembly of the
560 existing DISCOVAR *de novo/w2rap* genomes (Edelman et al. 2019). MEDUSA relies on reference
561 genomes from closely related species to determine the correct order and orientation of the draft genome
562 scaffolds, assuming collinearity between reference and the lower contiguity genome. The *w2rap* genome

563 assemblies of 16 *Heliconius* species produced by Edelman et al. (2019) - Supplemental Table S1 - and
564 high-quality reference genome assemblies of two *Heliconius* species - *H. melpomene* (Hmel2.5) and *H.*
565 *erato demophoon* (*Heliconius_erato_demophoon_v1*) - were downloaded from lepbases.org. Before the
566 reference-scaffolding step, alternative haplotypes present in the *w2rap* assemblies were collapsed using
567 the HaploMerger2 pipeline (version 20180603; Huang, Kang, & Xu, 2017). Repetitive elements and low
568 complexity regions in the *w2rap* assemblies were first soft-masked using WindowMasker (Morgulis et al.
569 2006) with default settings. A score matrix for LASTZ (used within HaploMerger) was generated for each
570 *w2rap* assembly. This was done using the `lastz_D_Wrapper.pl` script with `identity = 90` and splitting the
571 *w2rap* assemblies into two sets of scaffolds (scaffolds greater or smaller than 150kb). HaploMerger2
572 batch scripts A and B were then run using default settings. Finally, MEDUSA was used with default
573 parameters to place and orient the *w2rap* assembly scaffolds based on either of the two reference
574 genomes, placing 100 Ns between adjacent pairs of scaffolds mapping to the same reference
575 chromosome/scaffold. This resulted in two scaffolded assemblies per species (one based on mapping to
576 *H. melpomene* and another based on mapping to *H. erato demophoon* reference genomes).
577 Reference-guided assemblies were then re-aligned to the *H. melpomene* and *H. erato* reference genomes
578 using the Mashmap aligner as implemented in D-GENIES v1.2.0 online tool (Cabanettes and Klopp
579 2018) to assess collinearity. Scaffolds in the reference-guided assemblies aligning to reference assembly
580 scaffolds anchored to chromosomes were renamed to reflect their association to chromosomes and order
581 within chromosomes (as in the reference genomes). Also, when necessary, scaffold sequences were
582 reverse-complemented to maintain the same orientation as in the reference.

583

584 ***Mitochondrial genome assembly***

585 To assemble the mitochondrial genomes the 16 *Heliconius* species analyzed here, we first subsampled 1
586 million read pairs from the original reads used to produce the *w2rap* assemblies. We then used ABySS
587 2.0 (Jackman et al. 2017) to assemble the reads, using 5 different *k*-mer sizes (64, 80, 96, 112 and 128-bp)

588 and requiring a minimum mean unitig k -mer coverage of 10. All other parameters were left as default.
589 Because of the higher number of mtDNA copies relative to nuclear DNA, resulting in higher mtDNA
590 coverage, we were able to recover the mitochondrial genome as a single large contig (about the size of the
591 complete mitogenome) while any nuclear contigs should be small. In *Heliconius*, the sizes of the
592 mitogenomes sequenced so far are *ca.* 15,300-bp, thus only contigs larger than 15 kb were retained. These
593 were then blasted to the NCBI Nucleotide collection (nr/nt) to confirm that they corresponded to the
594 mitochondrial genome. Finally, for each species, only the largest contig (after removing Ns) was retained.
595 The mitochondrial sequences were aligned using MAFFT v7.407 (Katoh and Standley 2013), with default
596 parameters and a maximum-likelihood (ML) tree was estimated using IQ-TREE v1.6.10 (Nguyen et al.
597 2015) – Figure 1A. Model selection was performed using ModelFinder (Kalyaanamoorthy et al. 2017)
598 and branch support was assessed with 1000 ultra-fast bootstraps (Hoang et al. 2018), as implemented in
599 IQ-TREE. We also used this approach to recover the mitogenome of *Eueides tales* (Accession number:
600 SRS4612550) to use as an outgroup.

601

602 ***Scaffolded assemblies quality assessment***

603 Basic statistics (e.g. scaffold N50, cumulative length, proportion of missing sequence) of the reference-
604 guided scaffolded genome assemblies were calculated using QUAST v5.0.2 (Gurevich et al. 2013).
605 Assembly completeness was assessed using BUSCO_V3 (Simão et al. 2015), which looks for the
606 presence (complete, partial or duplicated) or absence (missing) of core arthropod genes (arthropoda-odb9
607 dataset; https://busco.ezlab.org/datasets/arthropoda_odb9.tar.gz).

608

609 ***Gene annotation***

610 We used the Liftoff tool (Shumate and Salzberg 2020) to lift gene annotations from the reference
611 genomes to the new reference-guided assemblies. We used either the *H. melpomene* (Hmel2.5.gff3) or *H.*

612 *erato demophoon* (*Heliconius_erato_v1_-_genes.gff.gz*) gene annotations (downloaded from
613 www.butterflygenome.org and lepbase.org, respectively), depending on the reference genome used for the
614 scaffolding of the reference-guided assemblies. We ran Liftoff setting the maximum distance between two
615 nodes to be either i) twice the distance between two nodes in the reference genome (i.e. distance scaling
616 factor of 2) or ii) 20 kb distance between in the target, depending on which of these distances is greater. In
617 order to improve mapping of exons at the ends of genes we extended gene sequences by 20% of the gene
618 length, to include flanking sequences on each side (-flank 0.2). Given the *w2rap* scaffolds were ordered,
619 oriented and anchored to chromosomes using the reference genomes as the backbone, and thus we know
620 the association between scaffolds in the reference genomes and in the reference-guided assemblies, we
621 have also enabled the option to first align genes chromosome by chromosome. All other parameters were
622 set as default.

623

624 ***Mapping and genotype calling of re-sequencing data***

625 Mapping efficiency of the original *w2rap* reads to the reference-guided assemblies was compared with
626 mapping efficiency of the same reads to the reference genomes. Reads were first filtered for Illumina
627 adapters using cutadapt v1.8.1 (Martin 2011) and then mapped to their respective reference-guided
628 genome assemblies, the *H. melpomene* and *H. erato demophoon* reference genomes using BWA mem
629 v0.7.15 (Li 2013), with default parameters and marking short split hits as secondary. Mapped reads were
630 sorted and duplicate reads removed using sambamba v0.6.8 (Tarasov et al. 2015). Realignment around
631 indels was performed with the Genome Analysis Toolkit (GATK) v3.8 RealignerTargetCreator and
632 IndelRealigner modules (McKenna et al. 2010; DePristo et al. 2011), in order to reduce the number of
633 indel miscalls. Mapping statistics and mean read depth were calculated in non-overlapping sliding
634 windows of 25 kb using the *flagstat* and *depth* modules implemented in sambamba v0.6.8, respectively.
635 Genotype calling was also performed for reads mapped to either of the two reference genomes and for
636 each individual separately with bcftools v1.5 (Li et al. 2009) *mpileup* and *call* modules (Li 2011), using

637 the multiallelic-caller model (call -m) and requiring a minimum base and mapping qualities of 20.
638 Genotypes were filtered using the bcftools *filter* module. Both invariant and variant sites were required to
639 have a minimum quality score (QUAL) of 20. Furthermore, individual genotypes were filtered to have a
640 depth of coverage (DP) ≥ 8 (except for the Z-chromosome of females for which the minimum required
641 depth was 4) and genotype quality (GQ) ≥ 20 . All genotypes not fulfilling these requirements or within
642 5-bp of an indel (--SnpGap) were recoded as missing data.

643

644 ***Copy number variation and selection tests***

645 Copy number variation (CNV) of genes within repeat regions of interest was estimated using two
646 different approaches. The first relies on mapping exonic sequences of genes annotated in the *H.*
647 *melpomene* reference within regions of interest onto the reference-guided assemblies. The reference-
648 guided assemblies were split back into the original scaffolds by breaking apart regions separated by 100
649 consecutive Ns, in order to avoid potential mis-mappings over scaffold breakpoints. Exon sequences were
650 mapped to these scaffolds using minimap2 v2.9 (Li 2018), with default settings (except that, as we were
651 interested in repeats, we allowed a much larger threshold of up to 1000 different alignments). Only
652 alignments for which $\geq 50\%$ of the length of the exon was mapped were considered. Copy number of
653 each exon was then estimated based on the number of alignments to these genomes. The second approach
654 is based on read coverage of the original *w2rap* read data, mapped to the *H. melpomene* reference genome
655 using BWA as described above. For each species, the mean read coverage within an exon (based on the
656 coordinates of exons as annotated in *H. melpomene*) was calculated using the sambamba v0.6.8 ‘*depth*’
657 module (Tarasov et al. 2015). Exon coverage was then normalized dividing by the median genomic
658 coverage (calculated in non-overlapping windows of 25 kb along the genome as described above) to
659 estimate copy number. This second approach was also used to estimate CNV in Amazon and extra-
660 Amazonian populations of *H. hecale*, *H. elevatus* and *H. pardalinus* (Supplemental Table S11).

661 We further investigated whether CNV in specific genes resulted in potentially functional copies or
662 pseudogenization by analyzing signals of codon-based selection and looking for the presence or absence
663 of stop codons. For each gene we examine each exon independently since different exons can show
664 different copy number. Sequences of the different putative copies were extracted from the reference-
665 guided assemblies, based on the coordinates obtained by aligning the reference *H. melpomene* exon
666 sequences to the reference-guided assemblies (as described above in this section). When shorter than the
667 exon length, coordinates were extended to match the total exon length. Exon sequences including 10
668 consecutive Ns (introduced during the *w2rap* assembly process) were excluded from this analysis to avoid
669 artificial sequence frameshifts. The remaining exonic sequences of all species were then aligned to the *H.*
670 *melpomene* reference genome using MAFFT v7.407 (Katoh and Standley 2013), with default parameters
671 and allowing reverse complementing of sequences when necessary. Bases before the start and after the
672 end of the *H. melpomene* reference sequence were removed from the alignment since these could have
673 been erroneously included when extending sequences to match the total exon length (see above). Also,
674 alignments including frameshift mutations (determined based on the *H. melpomene* sequence) were
675 excluded. We then calculated the ratio of non-synonymous versus synonymous changes (dN/dS) for each
676 pairwise comparison between exon copies detected in the reference-guided assemblies and the reference
677 *H. melpomene* sequence, using Li's (1993) method implemented in the 'seqinr' package in R. Finally, we
678 checked for the presence of stop codons using a custom script.

679

680 ***Detection of inversions in the w2rap assemblies***

681 In order to detect potential inversions in relation to the reference genomes, we mapped the *w2rap*
682 scaffolds (after filtering with HaploMerger2; see above) onto the reference genomes. Scaffolds of at least
683 5 kb were mapped to the *H. m. melpomene* and the *H. erato demophoon* reference genomes using
684 minimap2 (Li 2018) with default settings. Only primary alignments (tp:A:P), at least 1 kb long, with
685 mapping quality ≥ 60 and with less than 25% approximate per-base sequence divergence (dv) to the

686 reference were kept. Mappings of scaffolds spanning inversion breakpoints in the reference genome
687 should result in split alignments to different strands. We thus considered scaffolds as potentially
688 informative for inversions if they had at least two alignments to the same chromosome (split-alignments)
689 and at least one alignment to each strand as potentially informative for inversions. Same-scaffold
690 alignments mapping to the same strand, partially overlapping or not more than 50 kb apart were
691 concatenated. If less than 20% of the length of the scaffold aligned to the reference, the scaffold was
692 excluded. Furthermore, any scaffolds for which both forward and reverse alignments to the reference i)
693 come from overlapping scaffold regions (overlap greater than 5 kb), ii) overlap in the reference by more
694 than 5 kb or iii) in which the alignment in one strand is completely within the alignment to the other
695 strand, were removed as these likely represent spurious alignments, perhaps due to repeats. Candidate
696 inversions less than 50 kb from scaffold boundaries within chromosomes of the reference genome were
697 also excluded. Finally, we considered any two informative scaffolds to support the same candidate
698 inversion if they overlapped by at least 75% of the maximum length of the two. We also mapped the two
699 reference genomes against each other (and also the *H. erato lativitta* onto both) using minimap2 and
700 inferred candidate inversions by looking for alignments, within a scaffold, to the reverse strand. Only
701 alignments with a MQ \geq 10 and to the same chromosome in the reference were considered. Entire
702 scaffolds aligning to the reverse strand are possibly mis-oriented and were not considered to be
703 inversions.

704 For each candidate inversion we made sequence alignments for a subset of species (*H. melpomene*, *H.*
705 *numata*, *H. doris*, *H. burneyi*, *H. erato* and *H. hecalesia*, using *Eueides tales* as an outgroup) based on the
706 original *w2rap* sequencing data mapped to both *H. melpomene* and *H. erato* reference genomes. We then
707 estimated maximum-likelihood (ML) trees for these candidate regions using IQ-TREE v1.6.10 (Nguyen
708 et al. 2015). Model selection was performed using ModelFinder (Kalyaanamoorthy et al. 2017) and
709 branch support was assessed with 1000 ultra-fast bootstraps (Hoang et al. 2018), as implemented in IQ-
710 TREE.

711 We used the Patterson's D statistic (Green et al. 2010; Durand et al. 2011) to test i) which branching
712 pattern best describes the relationships between *H. doris*, *H. burney*, the *erato/sara* and the
713 *melpomene/silvaniform* groups and ii) whether the alternative clustering of *H. doris* and *H. burney* with
714 either of the two groups (both patterns were observed in the inversions) could be explained by
715 introgression. We used the ABBABABAwindows.py script (available from
716 github.com/simonhmartin/genomics_general) to estimate the D statistic in non-overlapping windows of 1
717 Mb, discarding all windows with fewer than 100 informative sites. The mean and variance of the D
718 *statistic* were calculated using a 1-Mb block jackknifing approach, allowing to test whether D differed
719 significantly from zero. We have also used the internal branch length based approach QuIBL (Edelman et
720 al. 2019), which uses the distribution of internal branch lengths and calculates the likelihood that the
721 triplet topologies discordant from the species tree are due to introgression rather than ILS alone. For this
722 analysis, we sampled 10 kb windows along the genome (50 kb apart) and for each we estimated
723 maximum-likelihood trees using the phylml_sliding_windows.py (available from
724 github.com/simonhmartin/genomics_general). Only alignments with less than 5% of the sites genotyped
725 were discarded. We then ran QuIBL on the filtered dataset with default parameters and adjusting the
726 number of steps to 50. In both Patterson's D and QuIBL analyses, *Eueides tales* was used as outgroup.
727 In order to detect local signals of introgression we also calculated the f_{dM} statistic (Malinsky et al. 2015),
728 which, like the f_d statistic (Martin et al. 2015), checks for imbalance in the number of shared variants
729 between the inner outgroup population and one of two ingroup populations, and was developed
730 specifically to investigate introgression of small genomic regions. Unlike the f_d statistic, it simultaneously
731 tests for an excess of shared variation between the inner outgroup population and either ingroup
732 population, at each genomic window. Again, we used the ABBABABAwindows.py script (available from
733 github.com/simonhmartin/genomics_general) to estimate the f_{dM} in non-overlapping windows of 100 kb,
734 discarding all windows with fewer than 100 informative sites. Because a local excess of derived alleles
735 could also be explained by retention of ancestral polymorphism (incomplete lineage sorting - ILS), we

736 calculated the divergence (D_{XY}) between both *H. doris* and *H. burneyi* to *H. erato*, normalized by
737 divergence to *H. melpomene* (i.e. Relative Node Depth, RND), to control for variation in substitution rate
738 across the genome. D_{XY} was calculated in 100 kb non-overlapping windows using the popgenWindows.py
739 script (available from github.com/simonhmartin/genomics_general). Finally, we also used QuIBL to
740 estimate the probability that gene trees within the chromosome 13 inversion were generated by
741 introgression.

742

743 ***Gene expression analyses***

744 Ovaries were dissected from adult females of *H. melpomene rosina* and *H. pardalinus butleri* at two
745 weeks post-eclosion, divided into developmental stages, and stored in RNALater. Ovaries were blotted
746 dry with kimWipes to remove excess RNALater solution. Tissue was then transferred to TRIZOL and
747 homogenized with the PRO200 tissue homogenizer (PRO Scientific). RNA was extracted with the Direct-
748 zol RNA miniprep kit (Zymo R2051). mRNA libraries were prepared by the Harvard University Bauer
749 Core with the KAPA mRNA HyperPrep kit, with mean fragment insert sizes of 200-300bp. mRNA was
750 sequenced with the NovaSeq S2, producing an average of 49 million paired-end, 50 bp reads.

751 RNASeq reads were mapped to the *H. melpomene* v2.5 transcriptome (Pinharanda et al. 2019) using
752 kallisto (Bray et al. 2016). Analysis was carried out in R using the Sleuth package (Pimentel et al. 2017).
753 Significant differences in expression levels between *H. melpomene* and *H. pardalinus* were assessed with
754 a likelihood ratio test, comparing expression as a function of developmental stage to expression as a
755 function of developmental stage + species identity.

756

757 **Data access**

758 On publication, the reference-guided assemblies and gene annotations generated in this study will have
759 been made available in Zenodo and all custom scripts used in this study will be made available on the
760 GitHub repository <https://github.com/FernandoSeixas/HeliconiusReferenceGuidedAssemblies>.

761

762 **Competing interest statement**

763 The authors declare no competing interests.

764

765 **Acknowledgements**

766 We thank the Harvard FAS Research Computing team for their support, A. Shumate for her guidance
767 using the Liftoff software, J. Davey and D. Ray for their valuable inputs to our thinking around genome
768 structural variation, and N. Rosser for the helpful discussions on *Heliconius*. This project was funded by
769 a SPARC Grant from the Broad Institute of Harvard and MIT and funds from Harvard University

770

771 **Bibliography**

- 772 Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019.
773 RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* **20**: 224.
- 774 Benevenuto J, Ferrão LF V., Amadeu RR, Munoz P. 2019. How can a high-quality genome assembly
775 help plant breeders? *Gigascience* **8**: 1–4.
- 776 Bosi E, Donati B, Galardini M, Brunetti S, Sagot M-F, Lió P, Crescenzi P, Fani R, Fondi M. 2015.
777 MeDuSa: a multi-draft based scaffolder. *Bioinformatics* **31**: 2443–2451.
- 778 Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin
779 HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. *Genome*
780 *Biol* **19**: 199.
- 781 Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification.
782 *Nat Biotechnol* **34**: 525–527.
- 783 Briscoe AD, Macias-Muñoz A, Kozak KM, Walters JR, Yuan F, Jamie GA, Martin SH, Dasmahapatra
784 KK, Ferguson LC, Mallet J, et al. 2013. Female Behaviour Drives Expression and Evolution of
785 Gustatory Receptors in Butterflies. *PLoS Genet* **9**: e1003620.
- 786 Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple

- 787 way. *PeerJ* **6**: e4958.
- 788 Christmas MJ, Wallberg A, Bunikis I, Olsson A, Wallerman O, Webster MT. 2019. Chromosomal
789 inversions associated with environmental adaptation in honeybees. *Mol Ecol* **28**: 1358–1374.
- 790 Clavijo B, Accinelli GG, Wright J, Heavens D, Barr K, Yanes L, Di-Palma F. 2017. W2RAP: a pipeline
791 for high quality, robust assemblies of large complex genomes from short read data. *bioRxiv*.
- 792 Coyne JA. 2018. “Two Rules of Speciation” revisited. *Mol Ecol* **27**: 3749–3752.
- 793 Coyne JA, Orr AH. 1989. Two rules of speciation. In *Speciation and its Consequences* (eds. D. Otte and
794 J.A. Endler), pp. 180–207, Sinauer Associates, Sunderland, MA.
- 795 Davey JW, Barker SL, Rastas PM, Pinharanda A, Martin SH, Durbin R, McMillan WO, Merrill RM,
796 Jiggins CD. 2017. No evidence for maintenance of a sympatric *Heliconius* species barrier by
797 chromosomal inversions. *Evol Lett* **1**: 138–154.
- 798 DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G,
799 Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-
800 generation DNA sequencing data. *Nat Genet* **43**: 491–8.
- 801 Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin H. 2018. A chromosome-scale
802 assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat Commun* **9**:
803 4844.
- 804 Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for Ancient Admixture between Closely
805 Related Populations. *Mol Biol Evol* **28**: 2239–2252.
- 806 Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey JW, Dikow RB, García-Accinelli G, Van
807 Belleghem SM, Patterson N, Neafsey DE, et al. 2019. Genomic architecture and introgression shape
808 a butterfly radiation. *Science* **366**: 594–599.
- 809 Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H,
810 Nadachowska-Brzyska K, Qvarnström A, et al. 2012. The genomic landscape of species divergence
811 in *Ficedula* flycatchers. *Nature* **491**: 756–760.
- 812 Faria R, Chaube P, Morales HE, Larsson T, Lemmon AR, Lemmon EM, Rafajlović M, Panova M,
813 Ravinet M, Johannesson K, et al. 2019. Multiple chromosomal rearrangements in a hybrid zone
814 between *Littorina saxatilis* ecotypes. *Mol Ecol* **28**: 1375–1393.
- 815 Feulner PGD, De-Kayne R. 2017. Genome evolution, structural rearrangements and speciation. *J Evol*
816 *Biol* **30**: 1488–1490.
- 817 Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov I V, Jiang X, Hall AB,
818 Catteruccia F, Kakani E, et al. 2015. Extensive introgression in a malaria vector species complex
819 revealed by phylogenomics. *Science* **347**: 1258524.
- 820 Ghurye J, Pop M. 2019. Modern technologies and algorithms for scaffolding assembled genomes. *PLOS*
821 *Comput Biol* **15**: e1006994.
- 822 Gopalakrishnan S, Samaniego Castruita JA, Sinding M-HS, Kuderna LFK, Räikkönen J, Petersen B,
823 Sicheritz-Ponten T, Larson G, Orlando L, Marques-Bonet T, et al. 2017. The wolf reference genome
824 sequence (*Canis lupus lupus*) and its implications for *Canis* spp. population genomics. *BMC*
825 *Genomics* **18**: 495.
- 826 Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-
827 Y, et al. 2010. A draft sequence of the Neandertal genome. *Science* **328**: 710–22.
- 828 Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome
829 assemblies. *Bioinformatics* **29**: 1072–5.

- 830 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast
831 Bootstrap Approximation. *Mol Biol Evol* **35**: 518–522.
- 832 Huang S, Kang M, Xu A. 2017. HaploMerger2: rebuilding both haploid sub-assemblies from high-
833 heterozygosity diploid genome assembly. *Bioinformatics* **33**: 2577–2579.
- 834 Iskow RC, Gokcumen O, Lee C. 2012. Exploring the role of copy number variants in human adaptation.
835 *Trends Genet* **28**: 245–257.
- 836 Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L,
837 Warren RL, et al. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom
838 filter. *Genome Res* **27**: 768–777.
- 839 Jay P, Whibley A, Frézal L, Rodríguez de Cara MÁ, Nowell RW, Mallet J, Dasmahapatra KK, Joron M.
840 2018. Supergene Evolution Triggered by the Introgression of a Chromosomal Inversion. *Curr Biol*
841 **28**: 1839–1845.
- 842 Jiggins CD, Mavarez J, Beltrán M, McMillan WO, Johnston JS, Bermingham E. 2005. A genetic linkage
843 map of the mimetic butterfly *Heliconius melpomene*. *Genetics* **171**: 557–570.
- 844 Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW,
845 Ferguson L, et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene
846 controlling butterfly mimicry. *Nature* **477**: 203–206.
- 847 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermini LS. 2017. ModelFinder: fast model
848 selection for accurate phylogenetic estimates. *Nat Methods* **14**: 587–589.
- 849 Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements
850 in Performance and Usability. *Mol Biol Evol* **30**: 772–780.
- 851 Kozak KM, McMillan O, Joron M, Jiggins CD. 2018. Genome-wide admixture is common across the
852 *Heliconius* radiation. *bioRxiv* doi: 10.1101/414201.
- 853 Kozak KM, Wahlberg N, Neild AFE, Dasmahapatra KK, Mallet J, Jiggins CD. 2015. Multilocus Species
854 Trees Show the Recent Adaptive Radiation of the Mimetic *Heliconius* Butterflies. *Syst Biol* **64**: 505–
855 524.
- 856 Leffler EM, Band G, Busby GBJ, Kivinen K, Le QS, Clarke GM, Bojang KA, Conway DJ, Jallow M,
857 Sisay-Joof F, et al. 2017. Resistance to malaria through structural variation of red blood cell
858 invasion receptors. *Science* **356**: 1140–1152.
- 859 Lewis JJ, Reed RD. 2019. Genome-Wide Regulatory Adaptation Shapes Population-Level Genomic
860 Landscapes in *Heliconius*. *Mol Biol Evol* **36**: 159–173.
- 861 Lewis JJ, van der Burg KRL, Mazo-Vargas A, Reed RD. 2016. ChIP-Seq-Annotated *Heliconius erato*
862 Genome Highlights Patterns of cis-Regulatory Evolution in Lepidoptera. *Cell Rep* **16**: 2855–2863.
- 863 Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and
864 population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993.
- 865 Li H. 2013. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv
866 Prepr. arXiv **0**: 3.
- 867 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.
- 868 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The
869 Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- 870 Li W-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol*
871 *Evol* **36**: 96–99.

- 872 Li X, Zhu C, Lin Z, Wu Y, Zhang D, Bai G, Song W, Ma J, Muehlbauer GJ, Scanlon MJ, et al. 2011.
873 Chromosome Size in Diploid Eukaryotic Species Centers on the Average Length with a Conserved
874 Boundary. *Mol Biol Evol* **28**: 1901–1911.
- 875 Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications.
876 *Nat Rev Genet* **21**: 597–614.
- 877 Love RR, Weisenfeld NI, Jaffe DB, Besansky NJ, Neafsey DE. 2016. Evaluation of DISCOVAR de novo
878 using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics* **17**: 187.
- 879 Lucas ER, Miles A, Harding NJ, Clarkson CS, Lawniczak MKN, Kwiatkowski DP, Weetman D,
880 Donnelly MJ. 2019. Whole-genome sequencing reveals high complexity of copy number variation
881 at insecticide resistance loci in malaria mosquitoes. *Genome Res* **29**: 1250–1261.
- 882 Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner
883 MJ, Turner GF. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African
884 crater lake. *Science* **350**: 1493–1498.
- 885 Mallet J, Beltrán M, Neukirchen W, Linares M. 2007. Natural hybridization in heliconiine butterflies: the
886 species boundary as a continuum. *BMC Evol Biol* **7**: 28.
- 887 Markelz RJC, Covington MF, Brock MT, Devisetty UK, Kliebenstein DJ, Weinig C, Maloof JN. 2017.
888 Using RNA-Seq for genomic scaffold placement, correcting assemblies, and genetic map creation in
889 a common *Brassica rapa* mapping population. *G3 Genes, Genomes, Genet* **7**: 2259–2270.
- 890 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.
891 *EMBnet.journal* **17**: 10.
- 892 Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the Use of ABBA–BABA Statistics to Locate
893 Introgressed Loci. *Mol Biol Evol* **32**: 244–257.
- 894 Martin SH, Davey JW, Salazar C, Jiggins CD. 2019. Recombination rate variation shapes barriers to
895 introgression across butterfly genomes. *PLOS Biol* **17**: e2006288.
- 896 Masly JP, Presgraves DC. 2007. High-Resolution Genome-Wide Dissection of the Two Rules of
897 Speciation in *Drosophila*. *PLoS Biol* **5**: e243.
- 898 Massardo D, VanKuren NW, Nallu S, Ramos RR, Ribeiro PG, Silva-Brandão KL, Brandão MM, Lion
899 MB, Freitas AVL, Cardoso MZ, et al. 2020. The roles of hybridization and habitat fragmentation in
900 the evolution of Brazil’s enigmatic longwing butterflies, *Heliconius nattereri* and *H. hermathena*.
901 *BMC Biol* **18**: 84.
- 902 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D,
903 Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for
904 analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.
- 905 Meier JI, Salazar PA, Ku M, Davies RW, Dréau A, Aldás I, Power OB, Nadeau NJ, Bridle JR, Rolian C,
906 et al. 2020. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species.
907 *bioRxiv* doi: 10.1101/2020.05.25.113688.
- 908 Morgulis A, Michael Gertz E, Schä AA, Agarwala R. 2006. WindowMasker: window-based masker for
909 sequenced genomes. **22**: 134–141.
- 910 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic
911 Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**: 268–274.
- 912 Noor MAF, Grams KL, Bertucci LA, Reiland J. 2001. Chromosomal inversions and the reproductive
913 isolation of species. *Proc Natl Acad Sci* **98**: 12084–12088.
- 914 Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. 2017. Differential analysis of RNA-seq

- 915 incorporating quantification uncertainty. *Nat Methods* **14**: 687–690.
- 916 Pinharanda A, Martin SH, Barker SL, Davey JW, Jiggins CD. 2017. The comparative landscape of
917 duplications in *Heliconius melpomene* and *Heliconius cydno*. *Heredity* **118**: 78–87.
- 918 Pinharanda A, Rousselle M, Martin SH, Hanly JJ, Davey JW, Kumar S, Galtier N, Jiggins CD. 2019.
919 Sexually dimorphic gene expression and transcriptome evolution provide mixed evidence for a fast-
920 Z effect in *Heliconius*. *J Evol Biol* **32**: 194–204.
- 921 Prowell PD. 1998. Sex linkage and speciation in Lepidoptera. In *Endless Forms. Species and Speciation*
922 (eds. D.J. Howard and S.H. Berlocher), pp. 309–319, Oxford University Press, New York.
- 923 Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE. 2010. Computational challenges in the
924 analysis of ancient DNA. *Genome Biol* **11**: R47.
- 925 Prunier J, Caron S, Lamothe M, Blais S, Bousquet J, Isabel N, MacKay J. 2017. Gene copy number
926 variations in adaptive evolution: The genomic distribution of gene copy number variations revealed
927 by genetic mapping and their adaptive role in an undomesticated species, white spruce (*Picea*
928 *glauca*). *Mol Ecol* **26**: 5989–6001.
- 929 Ray DA, Grimshaw JR, Halsey MK, Korstian JM, Osmanski AB, Sullivan KAM, Wolf KA, Reddy H,
930 Foley N, Stevens RD, et al. 2019. Simultaneous TE Analysis of 19 Heliconiine Butterflies Yields
931 Novel Insights into Rapid TE-Based Genome Diversification and Multiple SINE Births and Deaths.
932 *Genome Biol Evol* **11**: 2162–2177.
- 933 Rice ES, Green RE. 2019. New Approaches for Genome Assembly and Scaffolding. *Annu Rev Anim*
934 *Biosci* **7**: 17–40.
- 935 Rinker DC, Specian NK, Zhao S, Gibbons JG. 2019. Polar bear evolution is marked by rapid changes in
936 gene copy number in response to dietary shift. *Proc Natl Acad Sci* **116**: 13446–13451.
- 937 Rosser N, Queste LM, Cama B, Edelman NB, Mann F, Mori Pezo R, Morris J, Segami C, Velado P,
938 Schulz S, et al. 2019. Geographic contrasts between pre- and postzygotic barriers are consistent with
939 reinforcement in *Heliconius* butterflies. *Evolution* **73**: 1821–1838.
- 940 Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, Blazier JC, Sankararaman S, Andolfatto
941 P, Rosenthal GG, et al. 2018. Natural selection interacts with recombination to shape the evolution
942 of hybrid genomes. *Science* **368**: eaar3684.
- 943 Seixas FA, Boursot P, Melo-Ferreira J. 2018. The genomic impact of historical hybridization with
944 massive mitochondrial DNA introgression. *Genome Biol* **19**: 91.
- 945 Shumate A, Salzberg SL. 2020. Liftoff: an accurate gene annotation mapping tool. *bioRxiv* doi:
946 2020.06.24.169680.
- 947 Simão FA, Waterhouse R, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: Assessing
948 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:
949 3210–3212.
- 950 Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Genome analysis Sambamba : fast processing
951 of NGS alignment formats. *Bioinformatics* **31**: 2032–2034.
- 952 The *Heliconius* Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry
953 adaptations among species. *Nature* **487**: 94–98.
- 954 Tobler A, Kapan D, Flanagan NS, Gonzalez C, Peterson E, Jiggins CD, Johnstson JS, Heckel DG,
955 McMillan WO. 2005. First-generation linkage map of the warningly colored butterfly *Heliconius*
956 *erato*. *Heredity* **94**: 408–417.
- 957 Todesco M, Owens GL, Bercovich N, Légaré JS, Soudi S, Burge DO, Huang K, Ostevik KL, Drummond

- 958 EBM, Imerovski I, et al. 2020. Massive haplotypes underlie ecotypic differentiation in sunflowers.
959 *Nature*.
- 960 Van Belleghem SM, Rastas P, Papanicolaou A, Martin SH, Arias CF, Supple MA, Hanly JJ, Mallet J,
961 Lewis JJ, Hines HM, et al. 2017. Complex modular architecture around a simple toolkit of wing
962 pattern genes. *Nat Ecol Evol* **1**: 0052.
- 963 Wei S, Yang Y, Yin T. 2020. The chromosome-scale assembly of the willow genome provides insight
964 into Salicaceae genome evolution. *Hortic Res* **7**: 45.
- 965 Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ
966 C, et al. 2014. Comprehensive variation discovery in single human genomes. *Nat Genet* **46**: 1350–
967 1355.
- 968 Wellenreuther M, Bernatchez L. 2018. Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends*
969 *Ecol Evol* **33**: 427–440.
- 970 Wellenreuther M, Mérot C, Berdan E, Bernatchez L. 2019. Going beyond SNPs: The role of structural
971 genomic variants in adaptive evolution and species diversification. *Mol Ecol* **28**: 1203–1209.
- 972 Yang J, Wan W, Xie M, Mao J, Dong Z, Lu S, He J, Xie F, Liu G, Dai X, et al. 2020. Chromosome-level
973 reference genome assembly and gene editing of the dead-leaf butterfly *Kallima inachus*. *Mol Ecol*
974 *Resour* **20**: 1080–1092.
- 975 Yu A, Li F, Xu W, Wang Z, Sun C, Han B, Wang Y, Wang B, Cheng X, Liu A. 2019. Application of a
976 high-resolution genetic map for chromosome-scale genome assembly and fine QTLs mapping of
977 seed size and weight traits in castor bean. *Sci Rep* **9**: 1–11.
- 978 Zuellig MP, Sweigart AL. 2018. Gene duplicates cause hybrid lethality between sympatric species of
979 *Mimulus*. *PLoS Genet* **14**: 1–20.
- 980