

1 A complex interplay between balancing selection and introgression maintains a  
2 genus-wide alternative life history strategy

3 **Authors:**

4 Kalle Tunström<sup>1,10\*</sup>, Alyssa Woronik<sup>1,2,10</sup>, Joseph J. Hanly<sup>3</sup>, Pasi Rastas<sup>4</sup>, Anton Chichvarkhin<sup>5</sup>, Andrew D Warren<sup>6</sup>,  
5 Akito Kawahara<sup>6</sup>, Sean D. Schoville<sup>7</sup>, Vincent Ficarrotta<sup>3</sup>, Adam H. Porter<sup>8</sup>, Ward B. Watt<sup>9</sup>, Arnaud Martin<sup>3</sup>,  
6 Christopher W. Wheat<sup>1\*</sup>

7 10: these authors contributed equally

8 \*Corresponding authors, email: [kalle.tunstrom@gmail.com](mailto:kalle.tunstrom@gmail.com), [chris.wheat@zoologi.su.se](mailto:chris.wheat@zoologi.su.se)

9 **Affiliations:**

10 1: Department of Zoology, Stockholm University, Stockholm, Sweden

11 2: Department of Biology, Sacred Heart University, Fairfield, CT, United States

12 3: Department of Biological Sciences, The George Washington University, Washington, DC 20052, USA

13 4: Institute of Biotechnology, University of Helsinki, 00014, Finland

14 5: National Scientific Center of Marine Biology, Far Eastern Branch of Russian Academy of Sciences Palchevskogo  
15 17, Vladivostok 690022

16 5: McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida,  
17 Gainesville, FL 32611, USA.

18 6: Department of Entomology, University of Wisconsin-Madison, Madison, WI, United States

19 7: Department of Biology, University of Massachusetts Amherst, Amherst, MA 01003, USA.

20 8: Department of Biology, University of South Carolina, Columbia, SC 29208, USA & Rocky Mountain Biological  
21 Laboratory, Crested Butte, CO 81224, USA

22 **Abstract:**

23 Alternative life-history strategies (ALHS) are genetic polymorphisms generating phenotypes differing in life  
24 histories that generally arise due to metabolic resource allocation tradeoffs. Although ALHS are often limited  
25 to a single sex or populations of a species, they can, in rare cases, be found among several species across a  
26 genus. In the butterfly genus *Colias*, at least a third of the species have a female limited ALHS called Alba. While  
27 many females develop brightly pigmented wings, Alba females reallocate nitrogen resources used in pigment  
28 synthesis to reproductive development, producing white-winged, more fecund females. Whether this ALHS  
29 evolved once or many times, and whether it has moved among species via introgression or been maintained via  
30 long-term balancing selection, has not been established. Answering these questions presents an opportunity to  
31 investigate the genetic basis and evolutionary forces acting upon ALHS, which have rarely been studied at a  
32 genus level. Here we identify the genetic locus of *Alba* in a second *Colias* species, allowing us to compare this  
33 with previous results in a larger phylogenetic context. Our findings suggest *Alba* has a singular origin and has  
34 been maintained in *Colias* through a combination of balancing selection and introgression for nearly one million  
35 years and at least as many generations. Finally, using CRISPR/Cas9 deletions in the cis-regulatory region of the  
36 *Alba* allele, we demonstrate that the *Alba* allele is a modular enhancer for the *BarH1* gene and is necessary for  
37 the induction of the ALHS, which potentially facilitates its long-term persistence in the genus.

## 38 Introduction:

39 Species vary in their life histories, allocating resources in differing amounts to growth, maintenance, and  
40 reproduction in an attempt to maximize fitness (Rose and Mueller, 1993). Within species, individuals also vary,  
41 whether plastically or genetically, in how they allocate resources. When genetically determined and causing  
42 distinctly different phenotypes, such polymorphisms within species are called alternative life-history strategies  
43 (ALHS)(Gross, 1996). Examining why some species remain polymorphic, as opposed to becoming fixed for one  
44 strategy, will provide insight into the ecological and evolutionary forces that shape ALHS during the adaptive  
45 diversification of populations and species (Jamie and Meier, 2020), and in turn, will inform our understanding of  
46 how complex traits evolve (Zakas et al., 2018).

47 Alternative life-history strategies are unstable by nature, with a single phenotype predicted to fixate within  
48 species over time (Ford, 1945; Llaurens et al., 2017). Thus, when an ALHS is observed in multiple species within  
49 a genus, at least one of the following must be true: **a)** the ALHS arose once and has been maintained by some  
50 form of balancing selection (Mérot et al., 2020a), **b)** the ALHS evolved independently via novel mutations  
51 causing similar phenotypes in separate species (Blow et al., 2021; Yassin et al., 2016), or **c)** the ALHS moved  
52 between species via introgression (Dasmahapatra et al., 2012). Distinguishing the relative role these  
53 nonexclusive alternatives have played in the evolution of an ALHS is challenging, especially as the genetic basis  
54 of ALHS is poorly understood in most species, let alone among many related species. Thus, an understanding of  
55 how ALHS evolve remains elusive.

56 Although butterflies in the genus *Colias* are characterized by their yellow to orange wing coloration, a third of  
57 the approximately 90 butterfly species of this genus have a female limited ALHS called Alba, where female wings  
58 are white (Limeri and Morehouse, 2016; Remington, 1954). The remaining species appear to be fixed for Alba or  
59 colored wings. Alba females reallocate metabolic resources from wing pigmentation to reproductive  
60 development, resulting in white wings, rather than the colored wings shared by males and the remaining  
61 females (Descimon and Pennetier, 1989; Graham et al., 1980; Nielsen and Watt, 1998; Watt, 1973; Woronik et  
62 al., 2018). Genetic studies in six *Colias* species consistently found that Alba is caused by a single dominant,  
63 autosomal locus (Remington, 1954). In *Colias crocea*, a Eurasian species polymorphic for Alba, the locus of the  
64 ALHS has been mapped to a transposable element insertion downstream of the gene *BarH1* (Woronik et al.,  
65 2019). This insertion leads to a gain of function of the *BarH1* gene in the developing wings, resulting in a lack of  
66 pigment granules in Alba wing scales (Woronik et al., 2019), in addition to a previously described reduction of  
67 pterin pigments (Watt 1973). In contrast, white color in other Pieridae species is primarily due to white  
68 pteridine pigment granules (Giraldo and Stavenga, 2007; Watt and Bowden, 1966; Wijnen et al., 2007).

69 This detailed knowledge of the genetic basis of this ALHS presents an opportunity to investigate the  
70 evolutionary dynamics of a complex life-history trait among species. The Alba life-history tradeoff between color  
71 production and reproductive investment has been studied in detail in both North American *C. eurytheme* and  
72 Eurasian *C. crocea* (Descimon and Pennetier, 1989; Woronik et al., 2018). Although Alba in both species appears  
73 to be a similar ALHS, the genetic basis of the Alba phenotype in any North American species is unknown. Since  
74 these two species likely last shared an ancestor before the North American and Eurasian clades separated, and  
75 these two clades contain the vast majority of *Colias* species, resolving whether these two species have the same  
76 or independent Alba phenotypes is necessary in order to understand the prevalence of this ALHS across the  
77 genus. While balancing selection can maintain life history polymorphisms, there is very little evidence that such  
78 a mechanism alone can maintain ALHS over deep evolutionary time. Shared ancestry, however, does not  
79 necessarily require that a polymorphism be ancient. Among *Heliconius* butterflies, regulatory units affecting  
80 wing color patterning readily move among populations and species via introgression, followed by strong  
81 directional selection leading to fixation (Morris et al., 2020; Wallbank et al., 2016; Westerman et al., 2018a).  
82 Whether introgression could play a similar role for an ALHS like Alba is unknown.

83 Here we set out to test among the alternative (single or multiple origins) and complimentary (balancing  
84 selection, introgression) evolutionary mechanisms responsible for the prevalence of the Alba ALHS (being fixed,  
85 absent or polymorphic) across the *Colias* genus. Using a combination of phylogenetic analyses, GWAS, and  
86 genetic manipulation, we find evidence that 1) the Alba polymorphism arose once, likely at the root of the

87 genus, 2) it is maintained by balancing selection with introgression, and 3) that the *Alba* allele acts as a modular  
88 enhancer controlling this trans-specific ALHS.

## 89 Results

### 90 Phylogenomic analysis

91 A chromosome-level genome for *C. eurytheme* was generated and used as a reference for aligning whole-  
92 genome sequence data from 21 species representing the global distribution of *Colias* and diverse Alba  
93 phenotypes (Fig. 1; Supplementary Table 1). After haplomerging and polishing, the final haploid genome was  
94 328 MB, with an N50 of 5.2MB across 108 scaffolds and high gene completeness with low duplication (97.7% of  
95 expected single-copy genes were complete and unique, i.e., 5166 of 5286 BUSCO genes; Supplementary Table  
96 2). We then generated and used a linkage map to assemble chromosomes, finding evidence for 31, which is the  
97 expected number for most *Colias* (Maeki and Remington, 1960). The *C. eurytheme* chromosomal structure was  
98 highly syntenic with the standard organization of Lepidoptera chromosomes (Supplementary Fig. 1). Annotation  
99 of the resulting chromosome level assembly identified 18077 transcripts from 16842 genes.

100 We mined these alignments for single-copy orthologs in Lepidoptera, then used the longest exon per gene to  
101 estimate maximum-likelihood gene trees, followed by species tree estimation using ASTRAL. Although there was  
102 extensive conflict among gene trees (Fig 1a; Supplementary Fig. 2; (n=4,244)), the species tree (Fig. 1c) supports  
103 three conclusions: species from South America are sister to the rest of the *Colias* genus, the major divergence  
104 within *Colias* is between a North American and a Eurasian + African clade, and circumpolar taxa, which we refer  
105 to as Holarctic, fall between and among these two major clades (Fig. 1c). In order to further assess these  
106 relationships in a multispecies coalescent framework, and to estimate divergence time among the North  
107 American and the Eurasian + African clades, we used a Bayesian species-tree inference approach (SNAPP)  
108 calibrated on an age estimate of when *Colias* and its sister genus *Zerene* last shared a common ancestor (Chazot  
109 et al., 2019). SNAPP is computationally demanding, necessitating a down sampling of data and taxa. Following  
110 recommendations (Stange et al., 2018), we analyzed a random set of 1000 SNPs selected from a reduced set of  
111 taxa, selected to remove redundancy among closely related species while retaining regional diversity. The  
112 SNAPP estimated phylogeny was largely concordant with the Astral species tree, with strong support for nodes  
113 separating North American from Eurasian + African clades, while the vast majority of remaining nodes were  
114 poorly supported (Fig. 1d). The mean crown age of *Colias* was estimated at 2.66 million years old (2.10 – 3.34  
115 posterior distribution 95 % limits), while the mean age of the last common ancestor of the non-South American  
116 *Colias* at 0.98 million years ago (0.75-1.21). The SNAPP analysis also estimated extremely short branch lengths  
117 among the majority of species (Fig. 1d), which corresponds to the extensive conflict among gene trees (Fig. 1a)  
118 and previous single and multigene studies (Limeri and Morehouse, 2016; Pollock et al., 1998; Wheat and Watt,  
119 2008), in suggesting that non-South American *Colias* rapidly diversified into two regional clades in the past  
120 million years. Using these results, we can now view Alba and putative Alba phenotypes in a global phylogenetic  
121 context, revealing that Alba is in all the geographic clades of *Colias* (Fig. 1b, d). While much of the phylogenetic  
122 conflict observed arises from rapid speciation events and incomplete lineage sorting, introgression has likely  
123 been extensive during these events and potentially an important contributor to moving Alba among species and  
124 regions.

### 125 Genome-wide Introgression analysis

126 In order to assess to what extent introgression played in the current distribution of Alba, we first evaluated  
127 introgression among *Colias* species using D-statistics. All possible species trios were assessed for introgression,  
128 using the South American *C. lesbia* as an outgroup. Grouping results by region (North America, Eurasian,  
129 Holoartic), we found significant levels of introgression among non-South American species (Fig. 1e, Fig. 2).  
130 Additionally, the proportion of significant trios showing introgression, as well as the general level of  
131 introgression, increased when we combined taxa from either North America or the Eurasian region with the  
132 Holoartic taxa. To estimate when these introgression events among the non-South American species occurred,  
133 we combined our D-statistics with our species tree and used the F-branch metric (Malinsky et al., 2018), which  
134 differentiates signatures of historical introgression between internal branch nodes from introgression between  
135 extant species. This revealed introgression between an ancestor of the Holarctic *nastes* clade and the North

136 American species, as well as low levels of introgression among the Eurasian species (Fig. 2). Since species in the  
137 *nastes* clade are fixed for Alba, the Alba allele may have been transferred through introgression between this  
138 clade and an ancestor to the North American species. However, this analysis is unable to resolve the direction of  
139 introgression or capture localized intra-chromosomal introgression events. To specifically investigate the role of  
140 introgression around the Alba locus, sliding windows of 50 SNPs across the genome were analyzed and  $f_{dM}$   
141 calculated (an alternative statistic to D that is appropriate for windows-based analyses). However, no significant  
142 signatures of elevated introgression near *BarH1* were detected in any species trio (Supplementary Fig. 3). While  
143 this suggests no introgression of the *Alba* locus among *Colias* species, our results may arise due to a lack of  
144 sufficient informative SNPs to detect localized introgression. In *C. crocea*, Alba is caused by transposable  
145 element insertion not found in orange haplotypes. In our  $f_{dM}$  analysis, this Alba insertion is necessarily  
146 excluded when analyzing all species. Thus, our intra-chromosomal analysis of introgression is dependent on  
147 finding enough linked variants in the region surrounding the Alba locus rather than the locus itself, and this  
148 could reduce analysis power. Thus, our failure to detect local introgression dynamics near the identified Alba  
149 locus in *C. crocea* could be due to insufficient power resulting from the genomic architecture of the trait,  
150 introgression not moving the Alba locus extensively or recently among *Colias* species, or Alba having a different  
151 genetic basis among species.

#### 152 *Identification of the Alba locus in Colias eurytheme*

153 In order to thoroughly test whether Alba has a shared or *de novo* origin among species, we next mapped Alba in  
154 the North American species, *Colias eurytheme*. We first identified the chromosome carrying Alba, using a  
155 linkage map generated from a *Colias eurytheme* x *Colias philodice* F2 cross segregating the Alba phenotype,  
156 which allowed us to not only position our genome assembly scaffolds in chromosomal order but also to identify  
157 the chromosome bearing the Alba locus (Fig. 3a). This F2 cross showed Mendelian 1:1 segregation of the  
158 dominant Alba phenotypic state among the F2 females ( $\chi^2=0.006$ , d.f.=2, P-value>0.9), showing inheritance from  
159 an F1 single hybrid parent. The linkage map positioned the Alba locus on Chromosome 3 (Fig. 3b).  
160 Unfortunately, we could not further resolve the position of Alba within this chromosome since the Alba  
161 polymorphism was inherited from the maternal side of the F1 cross and was thus void of crossing-over events,  
162 due to the achiasmic mode of maternal meiosis in Lepidoptera (Traut et al., 2007).

163 A genome-wide association study (GWAS) was used to fine-map the Alba locus within *C. eurytheme*, using  
164 genomic data from 15 Alba and 14 orange wild-caught females mapped the reference genome. This identified  
165 two loci, the most significant of which was a single locus on Chr. 3 situated immediately downstream of the  
166 *BarH1* gene (Supplementary Fig. 4; but also Fig. 3c,d), which is concordant with both our previous mapping  
167 (Fig.3b) and the location of the Alba locus identified in *C. crocea* (Woronik et al., 2018). The second locus, which  
168 had less support, was located on a different chromosome between a PIFI-like helicase and a PiggyBac  
169 transposon (Supplementary Fig.5). We hypothesized that the second locus was an artifact arising from aligning  
170 reads to a reference genome lacking the Alba insertion since the reference was from an orange female  
171 individual. To test this, we generated an Alba genomic reference by combining a draft assembly made using  
172 linked read technology from an Alba *C. eurytheme* female and the reference genome (Supplementary Fig. 6),  
173 which added ~36kb of sequence downstream of *BarH1*. Repeating the GWAS using this synthetic Alba reference  
174 genome identified only the previous *BarH1*-associated locus, indicating reference bias as a likely cause of the  
175 second peak (Fig. 3c).

#### 176 *Investigating the Alba insertion*

177 To further investigate the Alba associated insertion region, which we expected to be composed of repeat  
178 content and regions unique to Alba, we conducted a read depth analysis by mapping the *C. eurytheme*  
179 individual genomes from the GWAS onto the Alba genome. By contrasting uniquely mapped reads that were at  
180 expected coverage depth to 1) reads mapping at higher-than-expected depth, or 2) reads not mapping uniquely,  
181 we could distinguish between unique Alba content and low complexity or repeat regions found in other parts of  
182 the genome. In the Alba associated insertion region, we identified an approximately 20 kb region containing two  
183 stretches of unique Alba content (where no orange reads mapped); data from orange females showed no  
184 unique content (Fig. 3d). Next, we similarly aligned reads from orange and Alba *C. crocea* females (n=15 each),  
185 revealing that only one of these two regions contained reads unique to Alba in both species (Fig. 3d), suggesting

186 a region of high sequence similarity between the two species, which is notable given their deep divergence (Fig.  
187 1d). We hypothesized that this shared region causes the Alba ALHS and hereafter refer to it as the Alba  
188 candidate locus (Fig. 3d). We further documented the uniqueness of the Alba candidate locus in *C. eurytheme*  
189 using PCR genotyping for an additional eight wild-caught females of each color morph (Supplementary Fig. 7).

190 ***Comparative analysis of Alba candidate locus***

191 To test the hypothesis that our identified Alba candidate locus is associated with Alba in additional *Colias*  
192 species, we generated an additional draft genome to resolve the Alba locus for *C. nastes*, a species fixed for the  
193 Alba color phenotype. In *C. nastes*, the Alba candidate locus and the *BarH1* gene assembled as a single contig,  
194 consistent with the hypothesis of the genetic basis for Alba having a shared ancestry among *Colias* species  
195 (Supplementary Fig. 8-9). In the aforementioned second region, which was unique to Alba only in *C. eurytheme*,  
196 we observed a dramatic increase in read depth and nucleotide diversity in *C. crocea*. This suggests that this  
197 segment is found in more than one copy in both orange and Alba *C. crocea* individuals and highlights the  
198 complex nature and evolutionary history of the locus.

199 Next, we formally tested the hypothesis of a shared origin of Alba by quantifying the association between having  
200 the Alba candidate locus and the Alba phenotype. All species with white wings had reads covering the entire  
201 Alba candidate locus except for *C. phicomone*, where reads covered only approximately half the candidate locus.  
202 In contrast, none of the samples from females with colored wings had reads covering the insertion region,  
203 instead reads piled up in low complexity areas flanking the locus. Thus, we observed a perfect correlation  
204 among species between having the insertion and the Alba phenotype. To estimate the likelihood of such a  
205 correlation on a genomic scale across species, we performed a window-based analysis of read coverage across  
206 the genome for all species (n=546,228 windows, 600bp in length each). A window located in the Alba candidate  
207 locus was the only region in the entire genome where the presence of coverage segregated with female wing  
208 color (Supplementary Fig. 10), suggesting the Alba insertion causes Alba across *Colias*.

209 ***Phylogenetic analysis of the Alba locus***

210 To further investigate the evolution of the Alba candidate locus, we constructed a phylogenetic tree for this  
211 region, which revealed a grouping of species discordant with the species tree (Fig. 5). While the *nastes* clade,  
212 European *C. crocea* and *C. erate* were all placed concordantly with the species tree, North American samples  
213 were not. Instead, they showed a pattern suggestive of separate introgression events. *C. eurytheme* grouped  
214 with *C. philodice* originating from the Maryland hybrid population and *C. nastes* clade as the closest outgroup,  
215 suggesting that Alba has been exchanged between the two former species during their ongoing hybridization  
216 and that this allele might have originated from an ancient introgression event with the Holarctic *C. nastes* clade.  
217 Meanwhile the *C. philodice* originating from British Columbia grouped with *C. canadensis* and *C. pelidne*,  
218 suggesting an independent introgression event of Alba entering *C. philodice* (though whether this sample is *C.*  
219 *philodice* or a subspecies thereof will require more regional sampling). Within *C. eurytheme*, our sampling of  
220 many Alba females reveals branch length differences among individual alleles, which suggest that the Alba allele  
221 has been maintained within *C. eurytheme* long enough for mutations to accumulate (Fig. 5). In contrast, the  
222 multiple alleles sampled from European *C. crocea* form a polytomy with *C. erate*, with the *C. erate* allele  
223 identical to several *C. crocea* samples, consistent with documented ongoing hybridization between *C. crocea*  
224 and *C. erate* (Descimon and Mallet, 2009). Together, these observations suggest that, while some regional  
225 hybridization has resulted in a reticulate phylogeny of the Alba locus, balancing selection has also played a role  
226 in maintaining Alba within species.

227 ***Functional validation of insertion.***

228 Our comparative analyses of the Alba insertion locus strongly suggested that the conserved Alba locus has a  
229 functional role in regulating the expression of *BarH1* in females' wing scales, so we tested this hypothesis by  
230 generating a somatic deletion mosaic of the Alba candidate locus using CRISPR/Cas9 in *C. crocea*. Along with  
231 Cas9, four gRNAs all targeting different parts of the locus were injected individually and together as a cocktail to  
232 generate multiple cuts and remove a significant portion within its 1.2 kb length. While injections with single  
233 gRNA did not produce any phenotypic changes (Supplementary Table. 7), of the forty eggs injected with the  
234 four-gRNA cocktail, unfortunately, due to bad rearing, many died before we were able to place the hatched

235 larvae on hostplant, and we were only able to rear five individuals, two of which were males. From the  
236 remaining three females, we had two mosaic mutant individuals that exhibited extensive wing clones where  
237 scales recovered the orange pigmentation of non-Alba females (Fig. 4, Supplementary Fig. 11). Both females  
238 were genetically Alba, as validated by PCR (Supplementary Fig. 12, and the remaining orange female showed no  
239 abnormalities. Successful mutagenesis was confirmed by PCR fragment size polymorphism relative to uninjected  
240 Alba females. The observed wing phenotypes are reminiscent of the effects seen in previous mosaic knock-outs  
241 within the coding region of *BarH1* (Woronik et al., 2019). While the previous coding knock-outs produced eye  
242 color phenotypes in both sexes (Woronik et al., 2019), consistent with the known role of BarH1 in insect eye  
243 development (Hayashi et al., 1998; Kojima et al., 2000; Woronik et al., 2019), our work here targeting the Alba  
244 candidate locus did not produce any noticeable eye aberrations (Supplementary Table. 7), indicating it functions  
245 not only as a regulatory region but as a modular enhancer region required for female and scale-specific  
246 expression of *BarH1*.

## 247 Discussion

248 The distribution of the Alba ALHS within the *Colias* genus, wherein some species are polymorphic for Alba or  
249 fixed for either Alba or the color phenotype, has been facilitated by a complex interplay between balancing  
250 selection and introgression. By identifying and comparing the genetic basis of Alba in two species that last  
251 shared a common ancestor at the base of the main species radiation in *Colias*, we confirm Alba's single ancestral  
252 origin for most, if not all *Colias*. Our comparative analysis of this trans-specific polymorphism allowed us to  
253 hypothesize about the location of Alba's alternate *cis*-regulatory module of *BarH1*, which we were able to  
254 confirm by genetic manipulation. The resulting indel-associated regulatory sequence, which has a wing-specific  
255 effect on the function of the transcription factor *BarH1*, is in concordance with the hypothesis that discrete  
256 regulatory modules with spatially discrete functions might be common in morphological evolution (Prud'homme  
257 et al., 2007). Our findings extend this hypothesis to an ALHS and constitute the first functional identification of a  
258 modular enhancer in a butterfly.

259 For the majority of *Colias* species, we are able to reject the hypothesis of multiple independent origins of Alba.  
260 While we cannot identify the maximum age of Alba at this time, we show it evolved prior to the separation of  
261 the North American and Eurasian clades approx. one million years ago. Additionally, as most *Colias* have at least  
262 one generation per year, this polymorphism has been maintained in these divergent clades since they shared a  
263 common ancestor at least one million generations ago. However, given the presence of Alba in South American  
264 taxa (Hovanitz, 1945), Alba is likely to be much older (e.g., age of *Colias*). Thus, the Alba ALHS is an ancient trans-  
265 specific polymorphism. The larger challenge is determining the relative role ancient balancing selection and  
266 introgression have played in its maintenance over time (Gao et al., 2015). Within-species diversity of the *Alba*  
267 allele is consistent with balancing selection (Fig. 5). While we do find evidence of historical introgression events  
268 (Fig 1c, Fig. 2), we did not detect any significant localized introgression around the *Alba* locus (Supplementary  
269 Fig. 3), potentially due to low power arising from the genetic architecture of *Alba*. However, the *Alba* locus  
270 phylogeny (Fig. 5) suggests several recent introgression events that include the *Alba* allele (Fig. 1).

271 The study of the genus-wide maintenance of an ALHS presented us with many inherent challenges that may  
272 account for the paucity of other well-characterized examples in the literature. Some ALHS may be exclusively  
273 physiological and lack a clear visual manifestation, which makes tracking the presence or absence of the ALHS  
274 across taxa particularly challenging. Our study here relied upon the large body of work linking resource  
275 allocation and performance with wing color (Gilchrist and Rutowski, 1986; Graham et al., 1980; Limeri and  
276 Morehouse, 2016; Nielsen and Watt, 1998; Remington, 1954; Watt, 1973; Woronik et al., 2018, 2019), allowing  
277 us to include additional species where the specific metabolic resource allocation tradeoffs of Alba have yet to be  
278 established. Furthermore, to persist through time, polymorphisms of complex traits, such as ALHS, are  
279 expected to be the result of structural variants, as these are less likely to be broken up by recombination (Llaurens  
280 et al., 2017). However, the genetic tools traditionally used for the identification of genetic associations  
281 generally exclude, and lack power to accurately detect, structural variants (Chaisson et al., 2019; Kishikawa et  
282 al., 2019; Merot et al., 2020b). If the causal locus is absent in the genomic reference, as was the case when we  
283 used our orange reference genome for our GWAS analysis, causally associated reads cannot map to the  
284 genome, while repeat content within the causal locus will map to non-orthologous regions producing spurious

285 associations (Supplementary Fig. 4, 5). Additionally, since downstream analyses and phylogenetic tools generally  
286 require discarding indel variants (Alonge et al., 2020), this likely decreased our power to detect localized  
287 introgression near the *Alba* locus. Fortunately, individual resequencing using long reads is becoming more  
288 accessible and has already been essential for the discovery of novel genomic regions associated with  
289 adaptation, such as premating isolation and speciation in crows (Weissensteiner et al., 2020). Further  
290 development of tools and pipelines for the downstream analysis of structural variants via long read data at the  
291 individual level will hopefully remove the aforementioned biases.

292 Finally, while the use of CRISPR/Cas9 for the functional validation of candidate genes is now accessible to a wide  
293 range of taxa (Sun et al., 2017), when candidate gene KOs have lethal effects the generation of transgenic lines  
294 are impossible. Hence, many such studies have been limited to assessing effects in low penetrance F0 mosaic  
295 phenotypes (e.g. Woronik et al., 2019; Zhang and Reed, 2017). While such mosaic phenotypes have greatly  
296 advanced the study of morphological traits (Burg et al., 2020; Perry et al., 2016; Westerman et al., 2018b), such  
297 mosaics of physiological phenotypes are nearly impossible to interpret, leaving the study of non-morphological  
298 phenotypes unable to use these new advances. While targeting the cis-regulatory regions of genes isn't a novel  
299 idea, the inherent difficulty in identifying such candidate regions have limited this approach. Here, the  
300 conserved nature of the *Alba* allele allowed us to leverage a phylogenetic foot printing approach (Tomoyasu and  
301 Halfon, 2020) to localize the locus and conduct the first targeted mutagenesis of a regulatory enhancer in  
302 Lepidoptera (but see (Murugesan et al., 2021)). Finally, while induced mutations in the coding exons of *BarH1*  
303 had a high lethality and pleiotropic effects (e.g. eye phenotypes), mutations in the CRE region of *Alba* were not  
304 lethal and appear to lack such negative pleiotropic effects. This is consistent with the CRE being a modular  
305 enhancer, only necessary for regulating *BarH1* expression in scale cell precursors. Facilitated by its modularity,  
306 introgression of this segment alone could be sufficient to (re)introduce *Alba* in a species lacking the Alba ALHS.  
307 Future studies in the system will be able to take advantage of these advances, especially those allowing the  
308 establishment of germline mutations, which will enable further detailed studies not only of the *Alba* CRE region,  
309 but direct investigation of the physiological aspects of the Alba ALHS.

## 310 References:

311 Aljanabi, S.M., and Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-  
312 based techniques. *Nucleic Acids Res.* **25**, 4692–4693.

313 Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Cirene,  
314 D., et al. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement  
315 in Tomato. *Cell* **182**, 145–161.e23.

316 Blow, R., Willink, B., and Svensson, E.I. (2021). A molecular phylogeny of forktail damselflies (genus  
317 *Ischnura*) reveals a dynamic macroevolutionary history of female colour polymorphisms. *Mol. Phylogenet. Evol.*  
318 **160**, 107134.

319 Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones,  
320 G., Kühnert, D., Maio, N.D., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary  
321 analysis. *PLOS Comput. Biol.* **15**, e1006650.

322 Brúna, T., Hoff, K.J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2020). BRAKER2: Automatic Eukaryotic  
323 Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. *BioRxiv*  
324 2020.08.10.245134.

325 Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A., and RoyChoudhury, A. (2012). Inferring Species Trees  
326 Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Mol. Biol. Evol.* **29**,  
327 1917–1932.

328 Burg, K.R.L. van der, Lewis, J.J., Brack, B.J., Fandino, R.A., Mazo-Vargas, A., and Reed, R.D. (2020). Genomic  
329 architecture of a genetically assimilated seasonal color pattern. *Science* **370**, 721–725.

330 Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L.,  
331 Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human  
332 genomes. *Nat. Commun.* **10**, 1784.

333 Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK:  
334 rising to the challenge of larger and richer datasets. *GigaScience* **4**.

335 Chazot, N., Wahlberg, N., Freitas, A.V.L., Mitter, C., Labandeira, C., Sohn, J.-C., Sahoo, R.K., Seraphim, N., de  
336 Jong, R., and Heikkilä, M. (2019). Priors and Posteriors in Bayesian Timing of Divergence Analyses: The Age of  
337 Butterflies Revisited. *Syst. Biol.* **68**, 797–813.

338 Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O’Malley, R., Figueroa-  
339 Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time  
340 sequencing. *Nat. Methods* **13**, 1050–1054.

341 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth,  
342 G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158.

343 Dasmahapatra, K.K., Walters, J.R., Briscoe, A.D., Davey, J.W., Whibley, A., Nadeau, N.J., Zimin, A.V., Hughes,  
344 D.S.T., Ferguson, L.C., Martin, S.H., et al. (2012). Butterfly genome reveals promiscuous exchange of mimicry  
345 adaptations among species. *Nature* **487**, 94–98.

346 Descimon, H., and Mallet, J. (2009). Bad species. *Ecol. Butterflies Eur.* **219**–249.

347 Descimon, H., and Pennetier, J.-L. (1989). Nitrogen metabolism in *Colias croceus* (Linné) and its “Alba” mutant  
348 (Lepidoptera Pieridae). *J. Insect Physiol.* **35**, 881–885.

349 Farré, D., Roset, R., Huerta, M., Adsúara, J.E., Roselló, L., Albà, M.M., and Messeguer, X. (2003). Identification of  
350 patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. *Nucleic Acids Res.* **31**, 3651–  
351 3653.

352 Ford, E.B. (1945). Polymorphism. *Biol. Rev.* **20**, 73–88.

353 Gao, Z., Przeworski, M., and Sella, G. (2015). Footprints of ancient-balanced polymorphisms in genetic variation  
354 data from closely related species. *Evolution* **69**, 431–446.

355 Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing.  
356 ArXiv12073907 Q-Bio.

357 Gilchrist, G.W., and Rutowski, R.L. (1986). Adaptive and incidental consequences of the alba polymorphism in an  
358 agricultural population of *Colias* butterflies: female size, fecundity, and differential dispersion. *Oecologia* **68**,  
359 235–240.

360 Giraldo, M.A., and Stavenga, D.G. (2007). Sexual dichroism and pigment localization in the wing scales of *Pieris*  
361 *rapae* butterflies. *Proc. R. Soc. B Biol. Sci.* **274**, 97–102.

362 Girgis, H.Z. (2015). Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale.  
363 *BMC Bioinformatics* **16**, 227.

364 Graham, S.M., Watt, W.B., and Gall, L.F. (1980). Metabolic resource allocation vs. mating attractiveness:  
365 Adaptive pressures on the “alba” polymorphism of *Colias* butterflies. *Proc. Natl. Acad. Sci.* **77**, 3615–3619.

366 Gross, M.R. (1996). Alternative reproductive strategies and tactics: diversity within sexes. *Trends Ecol. Evol.* **11**,  
367 92–98.

368 Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances circular visualization  
369 in R. *Bioinformatics* **30**, 2811–2812.

370 Hayashi, T., Kojima, T., and Saigo, K. (1998). Specification of Primary Pigment Cell and Outer Photoreceptor  
371 Fates by BarH1 Homeobox Gene in the Developing *Drosophila* Eye. *Dev. Biol.* **200**, 131–145.

372 Hovanitz, W. (1945). The distribution of *Colias* in the equatorial Andes. *Caldasia* **3**, 283–300.

373 Huang, S., Kang, M., and Xu, A. (2017). Haplomerger2: rebuilding both haploid sub-assemblies from high-  
374 heterozygosity diploid genome assembly. *Bioinformatics* **33**, 2577–2579.

375 Jamie, G.A., and Meier, J.I. (2020). The Persistence of Polymorphisms across Species Radiations. *Trends Ecol.*  
376 *Evol.* **35**, 795–808.

377 Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and  
378 genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915.

379 Kishikawa, T., Momozawa, Y., Ozeki, T., Mushiroda, T., Inohara, H., Kamatani, Y., Kubo, M., and Okada, Y. (2019).  
380 Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci. Rep.* **9**,  
381 1784.

382 Kojima, T., Sato, M., and Saigo, K. (2000). Formation and specification of distal leg segments in *Drosophila* by  
383 dual Bar homeobox genes, BarH1 and BarH2. *Development* **127**, 769–778.

384 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009).  
385 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

386 Limeri, L.B., and Morehouse, N.I. (2016). The evolutionary history of the ‘alba’ polymorphism in the butterfly  
387 subfamily Coliadinae (Lepidoptera: Pieridae). *Biol. J. Linn. Soc.* **117**, 716–724.

388 Llaurens, V., Whibley, A., and Joron, M. (2017). Genetic architecture and balancing selection: the life and death  
389 of differentiated variants. *Mol. Ecol.* **26**, 2430–2448.

390 Maeki, K., and Remington, C.L. (1960). Studies of the Chromosomes of North American Thopalocera 2.  
391 *Hesperide*, Megathymidie, and Pieridae. *J. Lepidopterists Soc.* **14**, 21.

392 Malinsky, M., Svardal, H., Tyers, A.M., Miska, E.A., Genner, M.J., Turner, G.F., and Durbin, R. (2018). Whole-  
393 genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat. Ecol. Evol.* **2**,  
394 1940–1955.

395 Malinsky, M., Matschiner, M., and Svardal, H. (2020). Dsuite - fast D-statistics and related admixture evidence  
396 from VCF files. *BioRxiv* 634477.

397 Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: A fast and  
398 versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944.

399 Martin, S.H., Davey, J.W., and Jiggins, C.D. (2015). Evaluating the use of ABBA-BABA statistics to locate  
400 introgressed loci. *Mol. Biol. Evol.* **32**, 244–257.

401 Merot, C., Llaurens, V., Normandea, E., Bernatchez, L., and Wellenreuther, M. (2020a). Balancing selection via  
402 life-history trade-offs maintains an inversion polymorphism in a seaweed fly. *Nat. Commun.* **11**, 670.

403 Merot, C., Oomen, R.A., Tigano, A., and Wellenreuther, M. (2020b). A Roadmap for Understanding the  
404 Evolutionary Significance of Structural Genomic Variation. *Trends Ecol. Evol.* **35**, 561–572.

405 Messeguer, X., Escudero, R., Farré, D., Núñez, O., Martínez, J., and Albà, M.M. (2002). PROMO: detection of  
406 known transcription regulatory elements using species-tailored searches. *Bioinforma. Oxf. Engl.* **18**, 333–334.

407 Morris, J., Hanly, J.J., Martin, S.H., Belleghem, S.M.V., Salazar, C., Jiggins, C.D., and Dasmahapatra, K.K. (2020).  
408 Deep Convergence, Shared Ancestry, and Evolutionary Novelty in the Genetic Architecture of *Heliconius*  
409 Mimicry. *Genetics* 216, 765–780.

410 Murugesan, S.N., Connahs, H., Matsuoka, Y., Gupta, M. das, Huq, M., Gowri, V., Monroe, S., Deem, K.D., Werner,  
411 T., Tomoyasu, Y., et al. (2021). Butterfly eyespots evolved via co-option of the antennal gene-regulatory  
412 network. *BioRxiv* 2021.03.01.429915.

413 Nallu, S., Hill, J.A., Don, K., Sahagun, C., Zhang, W., Meslin, C., Snell-Rood, E., Clark, N.L., Morehouse, N.I.,  
414 Bergelson, J., et al. (2018). The molecular genetic basis of herbivory between butterflies and their host plants.  
415 *Nat. Ecol. Evol.* 2, 1418–1427.

416 Nielsen, M.G., and Watt, W.B. (1998). Behavioural fitness component effects of the alba polymorphism of *Colias*  
417 (Lepidoptera, Pieridae): resource and time budget analysis. *Funct. Ecol.* 12, 149–158.

418 Page, J.T., Liechty, Z.S., Huynh, M.D., and Udall, J.A. (2014). BamBam: genome sequence analysis tools for  
419 biologists. *BMC Res. Notes* 7, 829.

420 Perry, M., Kinoshita, M., Saldi, G., Huo, L., Arikawa, K., and Desplan, C. (2016). Molecular logic behind the three-  
421 way stochastic choices that expand butterfly colour vision. *Nature* 535, 280–284.

422 Pollock, D.D., Watt, W.B., Rashbrook, V.K., and Iyengar, E.V. (1998). Molecular Phylogeny for *Colias* Butterflies  
423 and Their Relatives (Lepidoptera: Pieridae). *Ann. Entomol. Soc. Am.* 91, 524–531.

424 Prud'homme, B., Gompel, N., and Carroll, S.B. (2007). Emerging principles of regulatory evolution. *Proc. Natl.*  
425 *Acad. Sci.* 104, 8605–8612.

426 Remington, C.L. (1954). The Genetics of *Colias* (Lepidoptera). In *Advances in Genetics*, M. Demerec, ed.  
427 (Academic Press), pp. 403–450.

428 Rodriguez-Caro, L., Fenner, J., Benson, C., Van Belleghem, S.M., and Counterman, B.A. (2020). Genome  
429 Assembly of the Dogface Butterfly *Zerene cesonia*. *Genome Biol. Evol.* 12, 3580–3585.

430 Rose, M.R., and Mueller, L.D. (1993). Stearns, Stephen C., 1992. *The Evolution of Life Histories*. Oxford  
431 University Press, London xii + 249 pp., £16.95. *J. Evol. Biol.* 6, 304–306.

432 Sedlazeck, F.J., Rescheneder, P., and von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in  
433 highly polymorphic genomes. *Bioinforma. Oxf. Engl.* 29, 2790–2791.

434 Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing Genome Assembly and Annotation  
435 Completeness. *Methods Mol. Biol.* Clifton NJ 1962, 227–245.

436 Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing  
437 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212.

438 Stange, M., Sánchez-Villagra, M.R., Salzburger, W., and Matschiner, M. (2018). Bayesian Divergence-Time  
439 Estimation with Genome-Wide Single-Nucleotide Polymorphism Data of Sea Catfishes (Ariidae) Supports  
440 Miocene Closure of the Panamanian Isthmus. *Syst. Biol.* 67, 681–699.

441 Sun, D., Guo, Z., Liu, Y., and Zhang, Y. (2017). Progress and Prospects of CRISPR/Cas Systems in Insects and Other  
442 Arthropods. *Front. Physiol.* 8.

443 Tomoyasu, Y., and Halfon, M.S. (2020). How to study enhancers in non-traditional insect models. *J. Exp. Biol.*  
444 223, jeb212241.

445 Traut, W., Sahara, K., and Marec, F. (2007). Sex chromosomes and sex determination in Lepidoptera. *Sex. Dev.* 1,  
446 332–346.

447 Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J.,  
448 Young, S.K., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome  
449 Assembly Improvement. *PLOS ONE* 9, e112963.

450 Wallbank, R.W.R., Baxter, S.W., Pardo-Diaz, C., Hanly, J.J., Martin, S.H., Mallet, J., Dasmahapatra, K.K., Salazar, C.,  
451 Joron, M., Nadeau, N., et al. (2016). Evolutionary Novelty in a Butterfly Wing Pattern through Enhancer  
452 Shuffling. *PLOS Biol.* 14, e1002353.

453 Wang, S., Meyer, E., McKay, J.K., and Matz, M.V. (2012). 2b-RAD: a simple and flexible method for genome-wide  
454 genotyping. *Nat. Methods* 9, 808–810.

455 Watt, W.B. (1973). Adaptive Significance of Pigment Polymorphisms in *Colias* Butterflies. iii. Progress in the  
456 Study of the “Alba” Variant. *Evolution* 27, 537–548.

457 Watt, W.B., and Bowden, S.R. (1966). Chemical Phenotypes of Pteridine Colour Forms in *Pieris* Butterflies.  
458 *Nature* 210, 304–306.

459 Weissensteiner, M.H., Bunkis, I., Catalán, A., Francoijis, K.-J., Krief, U., Heim, W., Peona, V., Pophaly, S.D.,  
460 Sedlazeck, F.J., Suh, A., et al. (2020). Discovery and population genomics of structural variation in a songbird  
461 genus. *Nat. Commun.* 11, 3403.

462 Westerman, E.L., Letchinger, R., Tenger-Trolander, A., Massardo, D., Palmer, D., and Kronforst, M.R. (2018a).  
463 Does male preference play a role in maintaining female limited polymorphism in a Batesian mimetic butterfly?  
464 *Behav. Processes* 150, 47–58.

465 Westerman, E.L., VanKuren, N.W., Massardo, D., Tenger-Trolander, A., Zhang, W., Hill, R.I., Perry, M., Bayala, E.,  
466 Barr, K., Chamberlain, N., et al. (2018b). Aristaless Controls Butterfly Wing Color Variation Used in Mimicry and  
467 Mate Choice. *Curr. Biol.* 28, 3469–3474.e4.

468 Wheat, C.W., and Watt, W.B. (2008). A mitochondrial-DNA-based phylogeny for some evolutionary-genetic  
469 model species of *Colias* butterflies (Lepidoptera, Pieridae). *Mol. Phylogenet. Evol.* 47, 893–902.

470 Wijnen, B., Leertouwer, H.L., and Stavenga, D.G. (2007). Colors and pterin pigmentation of pierid butterfly  
471 wings. *J. Insect Physiol.* 53, 1206–1217.

472 Woronik, A., Stefanescu, C., Käkelä, R., Wheat, C.W., and Lehmann, P. (2018). Physiological differences between  
473 female limited, alternative life history strategies: The Alba phenotype in the butterfly *Colias croceus*. *J. Insect  
474 Physiol.* 107, 257–264.

475 Woronik, A., Tunström, K., Perry, M.W., Neethiraj, R., Stefanescu, C., Celorio-Mancera, M. de la P., Brattström,  
476 O., Hill, J., Lehmann, P., Käkelä, R., et al. (2019). A transposable element insertion is associated with an  
477 alternative life history strategy. *Nat. Commun.* 10, 1–11.

478 Yassin, A., Delaney, E.K., Reddiex, A.J., Seher, T.D., Bastide, H., Appleton, N.C., Lack, J.B., David, J.R., Chenoweth,  
479 S.F., Pool, J.E., et al. (2016). The pdm3 Locus Is a Hotspot for Recurrent Evolution of Female-Limited Color  
480 Dimorphism in *Drosophila*. *Curr. Biol.* 26, 2412–2422.

481 Zakas, C., Deutscher, J.M., Kay, A.D., and Rockman, M.V. (2018). Decoupled maternal and zygotic genetic effects  
482 shape the evolution of development. *ELife* 7, e37143.

483 Zhang, L., and Reed, R.D. (2017). A Practical Guide to CRISPR/Cas9 Genome Editing in Lepidoptera. In *Diversity  
484 and Evolution of Butterfly Wing Patterns*, (Springer, Singapore), pp. 155–172.

485 Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction  
486 from partially resolved gene trees. *BMC Bioinformatics* 19, 153.

487 **Methods:**

488 ***C. eurytheme* genome**

489 High molecular weight DNA from six female pupae originating from Davis (CA, USA) and reared in the lab for  
490 several generations was sequenced on PacBio Sequel v1 at the U. Maryland - Baltimore Institute of Genomic  
491 Sciences. Assembly followed the Falcon/Falcon-Unzip/Arrow assembly pipeline (Chin et al., 2016) and led to a  
492 diploid genome length of 583Mb, with N50 of 2.7Mb, haploidization was performed using Haplomerger2  
493 leading to a haploid genome of 364.5Mb and 123 scaffolds (Huang et al., 2017).

494 ***C. eurytheme* genome polishing, quality control, and annotation**

495 Pilon v.1.22 (Walker et al., 2014) was used to polish the genome, using 150bp PE reads (350bp insert, Illumina  
496 HiseqX) aligned with NextGenMap v.0.5.2 (Sedlazeck et al., 2013). Genome quality before and after polishing  
497 was assessed using Busco v1.1b1 with OrthoDBs Lepidoptera v10, as well as N50 (Seppey et al., 2019; Simão et  
498 al., 2015). Repetitive regions were softmasked using RED v:05/22/2015 (Girgis, 2015). The genome was  
499 annotated using the Braker2 (Brúna et al., 2020) pipeline with transcriptome data generated in a previous study  
500 (Nallu et al., 2018) aligned with Hisat2 v2.2.1 (Kim et al., 2019) and protein data from OrthoDBs Arthropod  
501 database (V10).

502 **Synteny comparative analysis**

503 To assess our genome assembly and check for any large-scale structural changes compared to other sequenced  
504 Lepidopteran genomes, we compared our *C. eurytheme* chromosome to one from the sister genus, *Zerene*  
505 *cesonia* (Rodriguez-Caro et al., 2020). Whole-genome alignments were performed using nucmer v4.0 (Marçais  
506 et al., 2018) followed by circos plotting using the R package circlize v.0.4.9 (Gu et al., 2014).

507 ***Colias* phylogenetic analysis**

508 For each individual, whole-genome sequencing reads were generated via DNA extracted from thorax and/or  
509 abdomen via a salting out method (Aljanabi and Martinez, 1997). DNA quality was evaluated using a 260/280  
510 ratio (Nanodrop 8000 spectrophotometer; Thermo Scientific, Waltham, MA, USA). The library preparation and  
511 short read paired end sequencing (500Bp insert) for all individuals was performed at BGI China. Reads were  
512 filtered for adapters and trimmed at the 5' and 3' end based on a PHRED quality score >20. Reads were aligned  
513 to the *C. eurytheme* reference genome using NextGenMap v0.5.5 (Sedlazeck et al., 2013). Using these bam files  
514 after MapQ > 20 filtering via Samtools v.1.9 (Li et al., 2009), the longest exon per gene for each individual was  
515 obtained the CDS annotation for *C. eurytheme* via bam2fasta script from the package bambam v1.4 tool-kit  
516 (Supplementary methods)(Page et al., 2014). *Zerene cesonia* (Rodriguez-Caro et al., 2020) was used as an  
517 outgroup, the dataset of which was generated by aligning Illumina sequencing reads from *Z. cesonia*  
518 ([SRR11021459](https://sra.ncbi.nlm.nih.gov/sra/SRR11021459)) to the *C. eurytheme* genome, as per the bam2fasta pipeline outlined above. Individual gene  
519 trees were then estimated using iQTree v. 2.0.6, which were then used to estimate a species tree via ASTRAL  
520 v.5.7.3 (Zhang et al., 2018). Gene tree support for the species tree was assessed using Phyparts  
521 (<https://bitbucket.org/blackrim/phyparts/src/master/>). Species trees for each chromosome were generated  
522 using all genes trees of a given chromosomes to generate an Astral species tree. Analyses were repeated after  
523 higher quality filtering, with identical species tree results (Supplemental methods). SNAPP v.1.3.0 (Bryant et al.,  
524 2012) analysis, implemented in BEAST2 v. 2.6.3 (Bouckaert et al., 2019), followed previous extensive analyses  
525 for optimal analysis settings (Stange et al., 2018), with dataset construction using snapp\_prep.rb  
526 ([https://github.com/mmatschiner/snapp\\_prep](https://github.com/mmatschiner/snapp_prep)). Calibration for the timing of the split between *Zerene* and  
527 *Colias* used a secondary calibration of 10.9 million years ago (Chazot et al., 2019), along with two monophyletic  
528 constraints set to increase run speed (SA taxa, non-SA taxa). Please see Supplementary methods for more  
529 details.

530 **Introgression analysis**

531 Introgression between different species was estimated using D-statistics calculated from ABBA-BABA between  
532 all possible trio combinations using the Dsuite software package (v0.3)(Malinsky et al., 2020). Using Dsuite we  
533 also calculated an f-branch metric, a statistic related to the f-4 statistics, which allows you to summarize the

534 amount of shared introgressed material on a branch and infer past gene-flow (Malinsky et al., 2018). Using the  
535 Alba reference genome, we aligned the reads of each species using NextGenMapper, and then called variants  
536 using Freebayes (Garrison and Marth, 2012). The resulting vcf-file was filtered (see SM for details) using vcftools  
537 (Danecek et al., 2011).

538 Using the Dinvestigate tool part of the Dsuite tool-kit *f\_d* and *f\_dM* was calculated in windows non-overlapping  
539 windows of 15 and 100 informative SNPs to look for signals of adaptive introgression along the chromosomes.  
540 (more details in Supplementary Methods)(Martin et al., 2015).

541 ***C. eurytheme* x *C. philodice* 2b-RADseq genotyping and linkage map**

542 By generating a linkage map, using the 2b-RADseq (Wang et al., 2012)whole-genome genotypes of the F2 brood  
543 from a *C. eurytheme* x *C. philodice* hybrid cross, we turned the haploid genome scaffolds into chromosome-wide  
544 super-scaffolds. Linkage mapping followed the basic LepMap3 with some exceptions (explained in the SM) and  
545 resulted in 31 linkage groups, with one short unplaced scaffold. The linkage map was output as a 4-way cross,  
546 which was imported into R package *r/qtl* using custom code (contributed by Karl Broman). We performed a  
547 genome scan with a single QTL binary model and ran a permutation test (n=1000) to determine a 5%  
548 significance threshold.

549 ***GWAS of Alba in C. eurytheme***

550 Individuals used in the genome resequencing were from 15 Alba and 14 orange *C. eurytheme* females caught in  
551 Tracy, California, stored at -20 in 95% ethanol. For DNA extraction through to read cleaning, see Supplemental  
552 Materials. Cleaned reads were mapped to the *C. eurytheme* reference genome using NextGenMap v0.5.2,  
553 followed by duplicate marking, and then Freebayes v1.3.1-16-g85d7bfc for variant calling. The variants were  
554 filtered using VCFTOOLS v0.1.13 (Danecek et al., 2011). Variants were associated with the Alba phenotype using  
555 PLINK v1.9 (Chang et al., 2015). Two separate sets of filters were used, one with stronger priors, where the  
556 nature of the inheritance pattern was taken into account, and one weaker where sites were filtered quality and  
557 depth mainly; for more detailed information on the filters, please refer to the SM.

558 ***Generation of an Alba specific reference genomes***

559 To get an idea of what the sequence and structure of the Alba insertion was, we generated an Alba reference  
560 genome. First, we generated a draft genome using a 10X Chromium library, sequenced on a NovaSeq S4,  
561 2x150bp PE reads, followed by assembly with Supernova v2.1.1 (performed by SciLifeLab). In addition to *C.*  
562 *eurytheme*, we also generated draft genomes for *C. nastes*, a species fixed for Alba as well as *C. crocea* (used as  
563 a control to compare against the previous genome), using the same protocol.

564 ***Characterizing the Alba insertion in C. eurytheme***

565 We identified the scaffold containing *BarH1* by using tBLASTn in the *C. eurytheme* Alba Supernova-assembly. We  
566 then aligned all the resequencing data from the GWAS to this contig. Read depth along the contig was analyzed  
567 visually in IGV, and differences between the Orange and Alba morph noted. Regions where no orange reads  
568 aligned, but Alba did, were extracted and blasted back against the *C. crocea* reference genome (Woronik 2019),  
569 to assess whether this was the previously identified Alba insertion region. To identify borders of the Alba  
570 insertion in *C. eurytheme*, we aligned the Alba contig against the orange *C. eurytheme* reference genome using  
571 BLASTn. This then provided the boundaries of the Alba insertion region for *C. eurytheme*, which was then used  
572 to place this haplotype into the orange *C. eurytheme* reference genome assembly, creating what we refer to as  
573 the *C. eurytheme* Alba reference genome.

574 ***GWAS using the Alba reference genome***

575 Using the new Alba reference genome, we repeated the steps done in the initial GWAS, seeing if the alternative  
576 loci disappeared with new targets to map against.

577 ***PCR-based validation of insertion***

578 The presence and uniqueness of the insertion to Alba individuals were validated using PCR-based markers with  
579 primers designed to bind within the insertion region. Primers were designed using primer3 software (libprimer3

580 release 2.5.0, Untergasser et al. - 2012). DNA from 8 orange and 8 Alba females independent females were used  
581 in the analysis, and CytC was used as a positive control in each reaction. For more information about primer  
582 design and the reaction, see the SM.

583 ***Identification of Alba insertion in C. nastes***

584 The genome assembly were scanned for the presence of the *C. crocea* insertion sequence, as well as the  
585 sequence identified in the *C. eurytheme* Chromium assembly, and whether it was found in linkage with the  
586 *BarH1* gene using the same combination of tBLASTN and BLASTN as we used in *C. eurytheme*.

587 ***Alignment and assessment of the Alba insertion across species***

588 Resequence-data generated for the phylogeny was aligned to the Alba reference using NextGenMap, filtered  
589 (MAPQ 20 and proper pairs), and had coverage across the Alba insertion region visually inspected in IGV (v2.7).  
590 We assessed whether the presence of read-coverage segregated with female wing color. Regions that were  
591 unique to only white-colored species (putative Alba ALHS species) were considered to be conserved regions of  
592 the Alba insertion and likely causal for the phenotype. The likelihood of coverage segregating between the two  
593 color-morphs to this degree was estimated using a window-based analysis (See SM).

594 ***Phylogenetic relationship of Alba insertion***

595 We extracted the consensus sequence of this using the bam2fasta tool (part of the bambamv1.4 tool-kit). We  
596 then assessed their phylogenetic relationships using IQtree with the same settings as in the primary analysis  
597 (Supplemental Methods).

598 ***CRISPR/Cas9 targeted mutagenesis of the Alba insertion.***

599 We used the PROMO tool (Farré et al., 2003; Messeguer et al., 2002) to scan the conserved Alba locus for  
600 potential transcription factors, comparing the motifs against version 8.3 of the TRANSFAC database. The output  
601 of this scan was analyzed in three ways, 1) relevant candidate transcription factors were identified, such as  
602 *Doublesex*; 2) sites with a high density of potential TF-binding sites were recorded; and 3) sites that were highly  
603 conserved between different species were preferentially selected. We designed four sgRNAs seeking to produce  
604 multiple cuts and produce a large >100 bp deletion. gRNA design and injection were performed following the  
605 steps outlined in Woronik et al. 2019 and injected as a cocktail together with Cas9 at a 500 ng/ul concentration.  
606 *C. crocea* Alba females (n = 6) from Aiguamolls de l'Empordà, Spain, were captured and transported to  
607 Stockholm alive and allowed to oviposit on *Vicia villosa*. Eggs were collected three times daily for injection,  
608 ensuring that they were not more than 4 h old at the time of injection. Injected eggs were kept on glass slides  
609 inside sealed Petri dishes, together with moist paper. Hatched larvae were transferred to fresh *Vicia villosa* and  
610 kept in feeding cups with no more than five larvae at 23°C until pupation. Once pupated, the pupae were  
611 transferred to a climate cabinet kept at 16°C until eclosion. Mutated individuals had the CRISPR cut site  
612 validated using PCR-based assay.

613 **Figures**

614 **Figure 1.** Evolutionary relationships among *Colias* butterfly species, their distribution, Alba phenotypes, and  
615 introgression. **a.** Thirty-one species trees are shown overlapping, one from each chromosome generated using  
616 gene trees based upon a single exon per gene (on average 129 genes / chromosome; n=4011 genes). **b.** For  
617 each specimen, first column indicates wing color morph (blue = Alba, orange = colored), while second column  
618 Alba morph type (black boxes = species is fixed for one morph, white = species is polymorphic). **c.** Species tree  
619 generated from all BUSCO exon trees, with branches color-coded by their samples regional distribution (Blue =  
620 North America, Orange = Holarctic, Green=Eurasia and Africa, Purple = South America). **d.** SNAPP tree generated  
621 using 1000 SNPs with secondary calibration, with millions of years on X axis. Dots upon nodes indicate posterior  
622 support, with values > 0.92 indicate by large blue circles. **e.** Distribution of minimal D-statistic of all species-trios  
623 that showed significant levels of introgression (BH corrected p < 0.05). Trios are grouped by Eurasian or North  
624 American regions, and then combined with the Holarctic species.

625 **Figure 2.** Signature of historical introgression across the *Colias* species tree phylogeny. Each cell in the grid  
626 indicates the F-branch statistic, identifying excess sharing of derived alleles between branch nodes on the y-axis

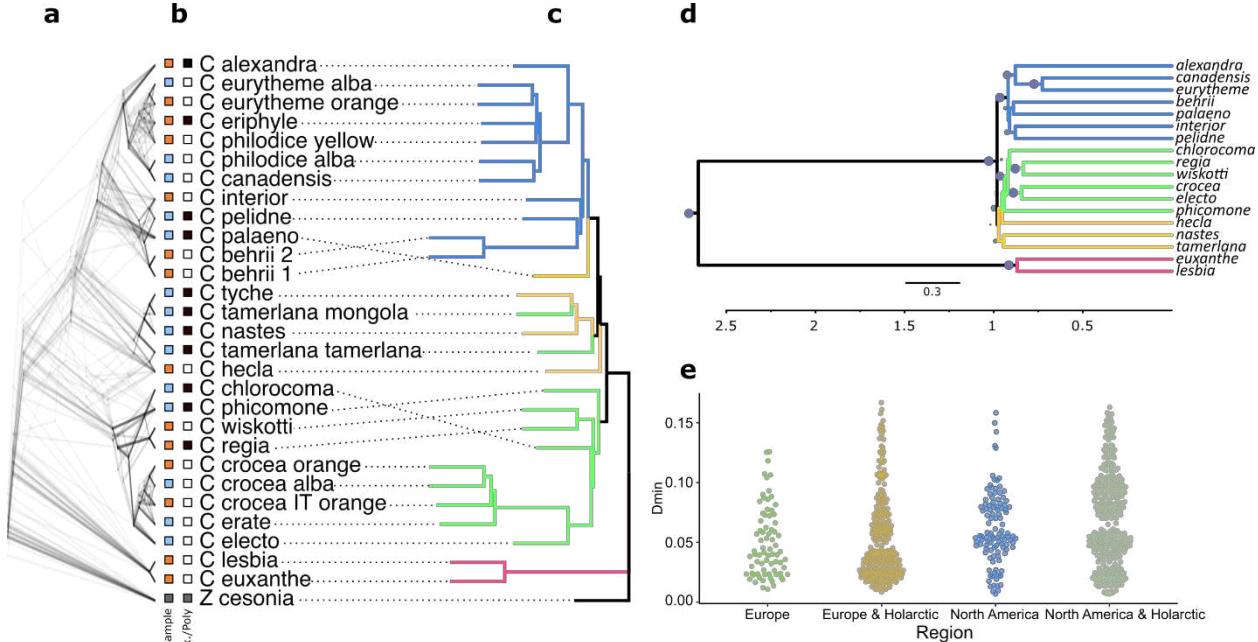
627 and individual species on the x-axis. A darker color in the heatmap indicates higher  $F_B$  and suggests gene flow  
628 between that branch and species. Allele sharing between an internal node and current species can be used to  
629 infer past gene flow events between an ancestor of a clade A and a species B. Results indicate a strong signal of  
630 introgression between an ancestor of the *nastes* clade and the North American clade, as well as between *C.*  
631 *phicomone* and the Eurasian species.

632 **Figure 3.** Identification of the *Alba* locus in *Colias eurytheme*. **a.** *Colias eurytheme* specimens and schematic of  
633 female informative crosses used in the linkage analysis. **b.** Linkage mapping of the *Alba* trait generated from  
634 female informative hybrid crosses of *C. eurytheme* and *C. philodice* revealed a single autosomal locus on Chr. 3  
635 that was associated with *Alba*. **c.** GWAS using WGS data from 14 and 15 orange and *Alba* *C. eurytheme* females.  
636 Blocks in the manhattan-plot are colored by chromosome, ordered 1:31 (y-axis represent negative log value  
637 Bonferroni-Holm corrected false discovery rate (-log10(FDR\_BH)) of each variant, the x-axis is scaffold position,  
638 the second row is a close up of scaffold 2, which makes up part of Chr. 3 as well as containing the *Alba* locus. **d.**  
639 Detailed view of the 58kb region identified from the 10X Chromium assembly showing WGS mapping coverage  
640 of an *Alba* (blue) and orange (orange) female from *C. eurytheme* (top) and *C. crocea* (bottom), respectively. The  
641 reads were aligned to the entire *Alba* reference genome (filtered to MAPQ > 20 and good pairs). Colored boxes  
642 highlight the *Alba* insertion (green), repetitive content within the insertion (grey), and a region of high sequence  
643 similarity with *C. crocea* *Alba* insertion, which is referred to as the *Alba* candidate locus (red). The grey region  
644 unique to *Alba* individuals in *C. eurytheme* is not unique to *Alba* individuals in *C. crocea* and is found at elevated  
645 coverage in all Eurasian species. Location and direction of the *BarH1* gene is indicated by the blue gene model.

646 **Figure 4.** Association of the *Alba* candidate locus with wing color across a species tree of *Colias* butterflies. **a.**  
647 Species tree of *Colias* colored by geographic region, with purple = South America, blue = North America, orange  
648 = Holarctic, green = Eurasia and Northern Africa. **b.** Read depth coverage across the *Alba* candidate locus for the  
649 sequenced samples, with separate columns depicting coverage plots according to female wing color. In cases  
650 where we have sequence data for both morphs, both are shown next to each other. **c.** Location of the *Alba*  
651 candidate locus in *C. crocea* (blue square, in the *Alba* insertion (red) and location of CRISPR cut sites (blue  
652 arrows). **d.** Images of dorsal front- and hind-wings wild-type *C. crocea* females; from left to right: orange, *Alba*,  
653 and one of the successful *Alba*-CRE mutants. Orange areas on the wings of the on the *Alba*-CRE KO individual  
654 are somatic CRISPR mosaic KO cells showing a return of color production.

655 **Figure 5** Phylogeny of the 1.2 Kbp long *Alba* candidate locus. Branch color corresponds to geographic region  
656 (blue = North America, yellow = Holarctic, green = Eurasia and Northern Africa), and branch length corresponds  
657 to substitution differences. The branching structure reveals shared alleles among *C. eurytheme* and the *C.*  
658 *philodice* originating from the Maryland hybrid population. Note the distinct variation among alleles in *C.*  
659 *eurytheme*, which suggests these alleles have been within this species long enough to accumulate variation.  
660 While similar diversity is seen within *C. crocea*, the shared allele with *C. erate* suggests ongoing *Alba* allelic  
661 exchange between these species, similar to that between *C. eurytheme* and *C. philodice*. The large divergence  
662 between the alleles found in the Maryland *C. philodice* and the *C. philodice* used in the phylogenetic analysis

663 further supports the reticulate evolution of Alba among North American *Colias*.

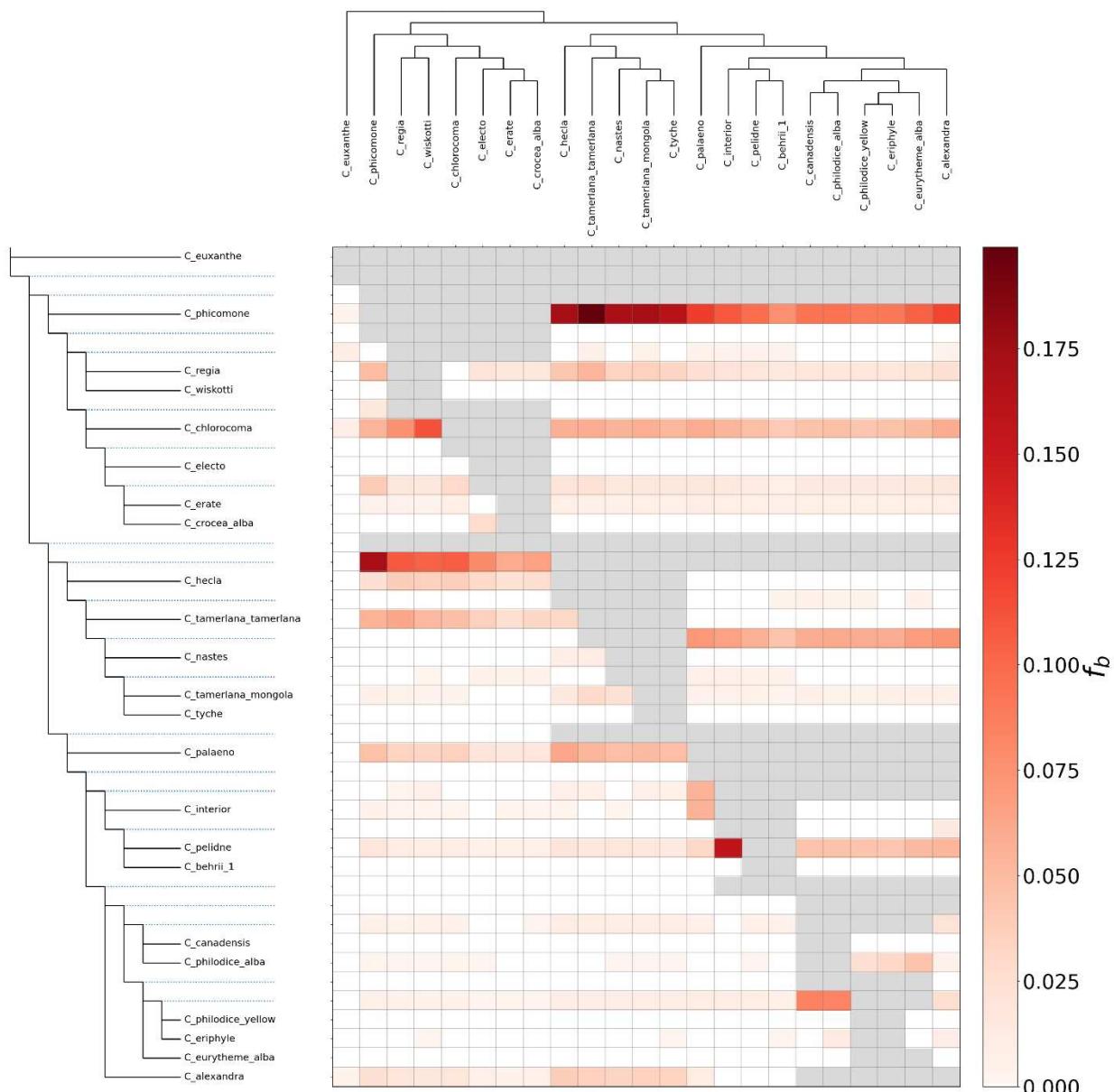


664

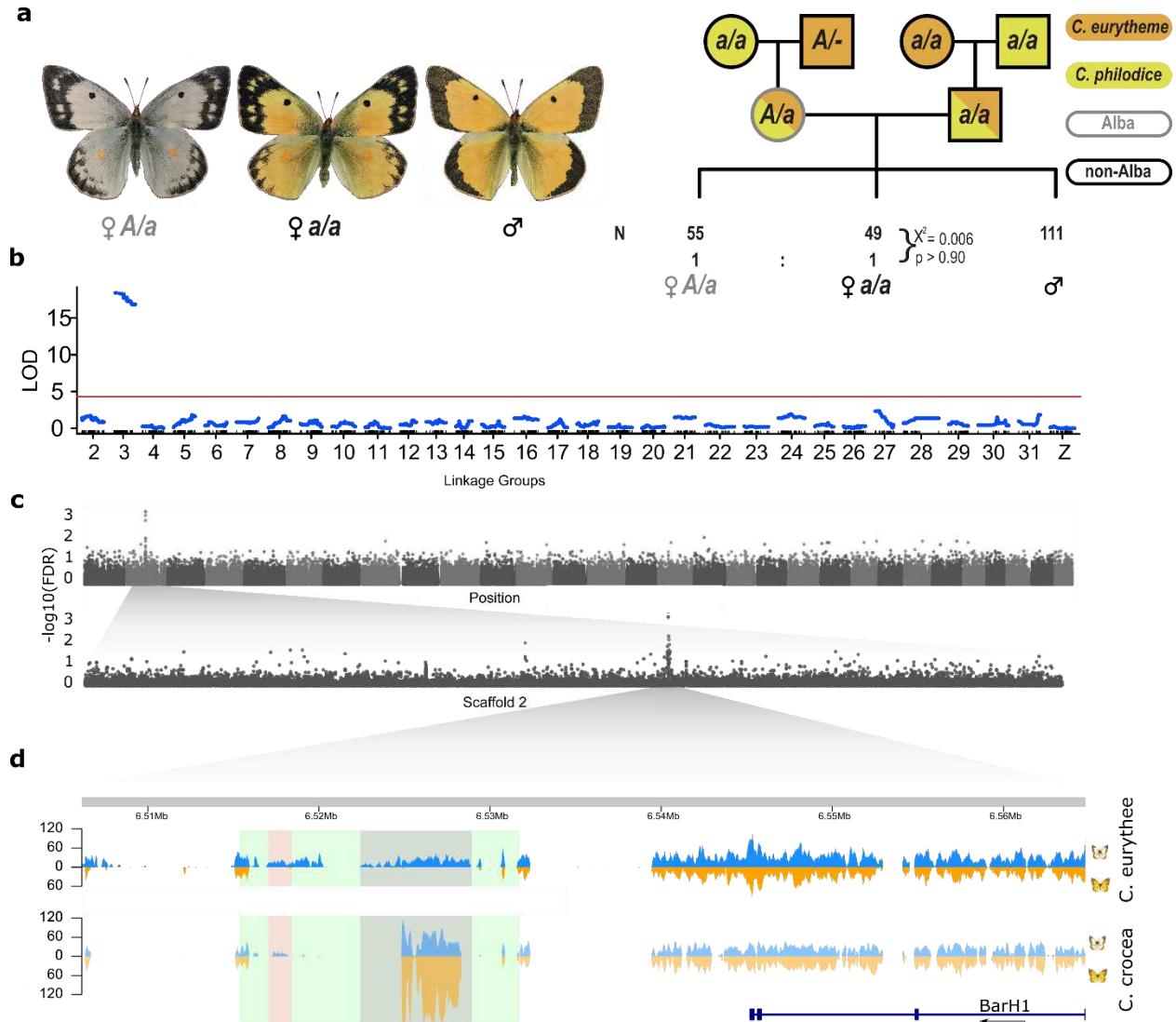
665 **Figure 1**

666

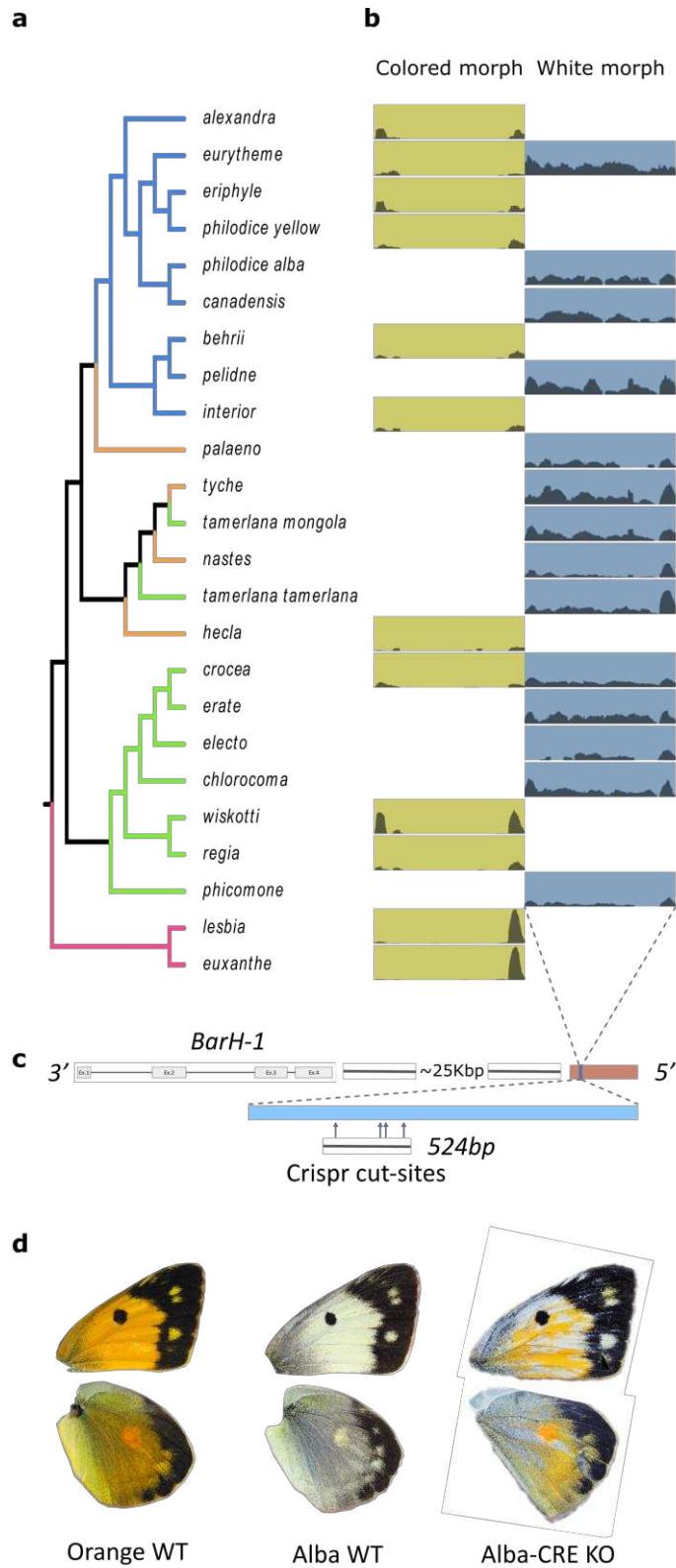
667



668 **Figure 2**

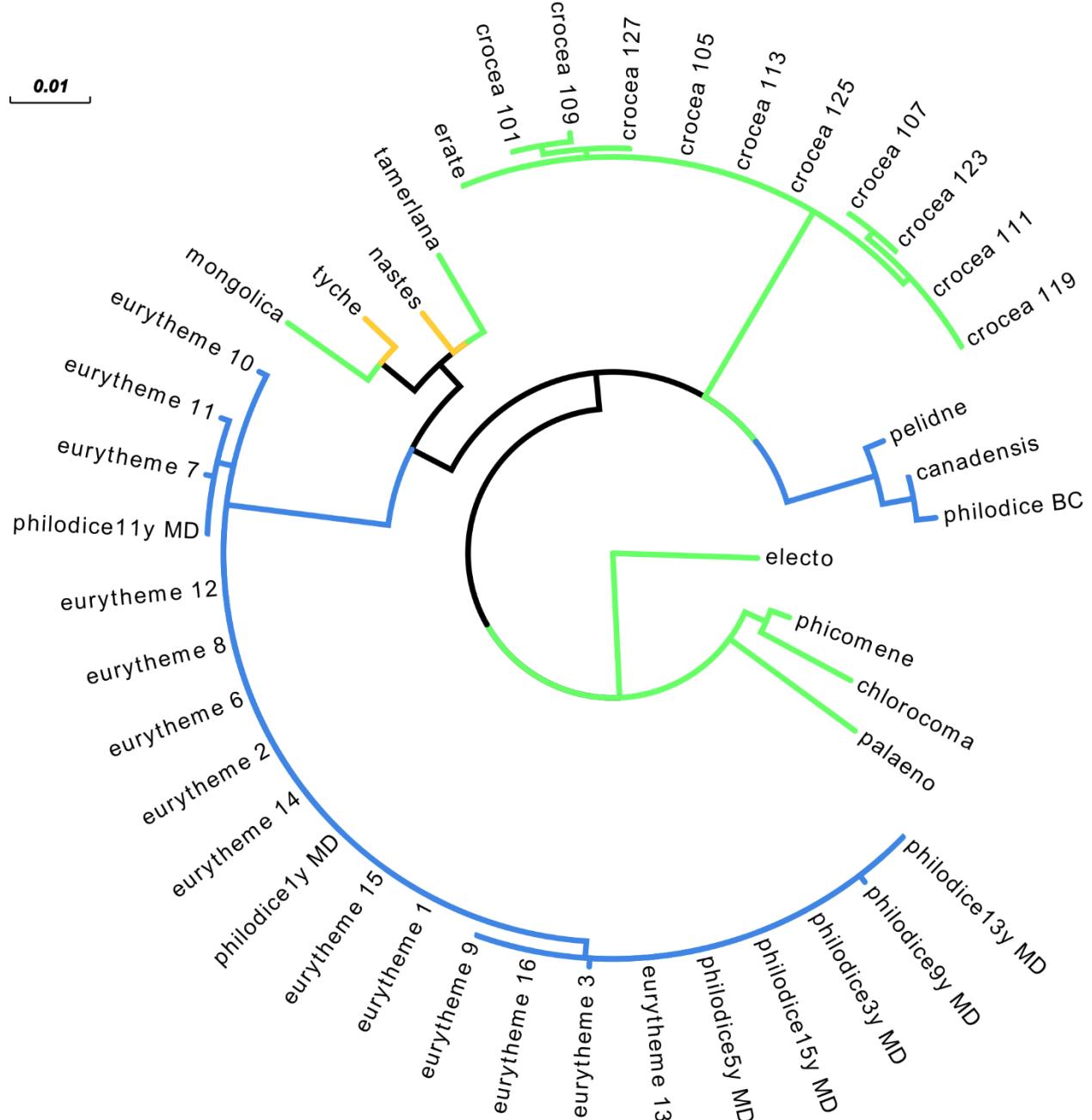


669 **Figure 3**



670    Figure 4

671



# 673 Supplemental Information

674 A complex interplay between balancing selection and introgression maintains a  
675 genus-wide alternative life history strategy

676 **Authors:**

677 Kalle Tunström<sup>1,10\*</sup>, Alyssa Woronik<sup>1,2,10</sup>, Joseph J. Hanly<sup>3</sup>, Pasi Rastas<sup>4</sup>, Anton Chichvarkhin<sup>5</sup>, Andy Warren<sup>6</sup>, Akito  
678 Kawahara<sup>6</sup>, Sean D. Schoville<sup>7</sup>, Vincent Ficarotta<sup>3</sup>, Adam Porter<sup>8</sup>, Ward B. Watt<sup>9</sup>, Arnaud Martin<sup>3</sup>, Christopher  
679 W. Wheat<sup>1\*</sup>

680 10: these authors contributed equally

681 \* email: [kalle.tunstrom@gmail.com](mailto:kalle.tunstrom@gmail.com), [chris.wheat@zoologi.su.se](mailto:chris.wheat@zoologi.su.se)

682 **Affiliations:**

683 1: Department of Zoology, Stockholm University, Stockholm, Sweden

684 2: Department of Biology, Sacred Heart University, Fairfield, CT, United States

685 3: Department of Biological Sciences, The George Washington University, Washington, DC 20052, USA

686 4: Institute of Biotechnology, University of Helsinki, 00014, Finland

687 5:

688 6: McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida,  
689 Gainesville, FL 32611, USA.

690 7: Department of Entomology, University of Wisconsin-Madison, Madison, WI, United States

691 8: Department of Biology, University of Massachusetts Amherst, Amherst, MA 01003, USA.

692 9: Department of Biology, University of South Carolina, Columbia, SC 29208, USA & Rocky Mountain Biological  
693 Laboratory, Crested Butte, CO 81224, USA

694

## 695 Supplementary Methods

### 696 *Bioinformatic Scripts*

697 Detailed scripts for bioinformatic analysis, if not specified in the relevant section, be found on the projects  
698 github.

### 699 *Colias eurytheme DNA extraction and genome sequencing:*

700 A *Colias eurytheme* stock originating from Davis (CA, USA) was maintained in a laboratory setting for several  
701 generations. High molecular weight DNA from six female pupae was isolated using the Qiagen Genomic-tip  
702 100/G. The specimen yielding the most DNA (574.4 nanograms per microliter) was submitted for quality  
703 control, BluePippin extraction fragments > 15kb, PacBio SMRTBell Express library preparation, and PacBio  
704 Sequel v1 sequencing at the U. Maryland – Baltimore Institute of Genomic Sciences. Six PacBio Sequel v1 cells  
705 were sequenced and yielded a total of 50.46 Gb of sequence data with an average sub-read length of 10kb, *i.e.*,  
706 a coverage of 144x assuming a genome size of 350 Mb based on fluocytometry. Assembly of the data using the  
707 Falcon/Falcon-Unzip/Arrow assembly pipeline (Chin et al. 2016) was outsourced to DNAexus (Mountainview,  
708 CA), which led to a diploid genome length of 583 Mb with an N50 of 2.7 Mb. Haplodization of the genome was  
709 performed using Haplomerger2, leading to a haplogenome of 364.4 Mb assembled into 123 scaffolds with a  
710 scaffold N50 = 4.82 Mb.

### 711 *C. eurytheme x C. philodice 2b-RADseq genotyping and linkage map*

712 To further improve the assembly into chromosome-wide super-scaffolds, a linkage map of the genome was  
713 generated using the 2b-RADseq ((S. Wang et al. 2012)) whole-genome genotypes of the F2 brood from a *C.*  
714 *eurytheme* x *C. philodice* hybrid cross (B. Wang and Porter 2004). Briefly, we extracted DNA from the thorax of  
715 frozen individuals using a bead-shaker and the Quick-DNA 96 Kit (Zymo Research), digested 400 ng of DNA per  
716 individual with the *BcgI* enzyme (New England Biolabs), purified the 36 bp restriction fragments using the [ZR-96](#)  
717 [Oligo Clean & Concentrator](#) kit (Zymo Research), added barcoded adapters corresponding to a 16 reduced tag  
718 representation (S. Wang et al. 2012) before ligation with the NEBNext Multiplex Primers Set 1-4 (New England  
719 Biolabs). The pooled library was enriched for a 155-175 bp (target inserts of 166bp) using a BluePippin  
720 instrument and sequenced using an Illumina HiSeq4000 SR50 run.

721 We investigated whether the haplogenome still contained some haplotype scaffolds/contigs. To achieve this,  
722 the two first steps (hm.batchA1.initiation\_and\_all\_lastz + hm.batchA2.chainNet\_and\_netToMaf) of  
723 Haplomerger2 (Huang, Kang, and Xu 2017) were run on the haplogenome to create the alignment chain  
724 (all.chain.gz). The alignments of the remaining 123 scaffolds revealed that there were clearly some haplotypes  
725 left. We manually classified all scaffolds into full, partial, or unique scaffolds; there were 16 full and 11 partial  
726 haplotypes. Using a custom script, the remaining haplotypes and partial haplotypes were removed from the  
727 assembly to construct the final scaffold assembly with 108 scaffolds (one scaffold was cut to two, as likely  
728 chimera).

729 The genotype data for the linkage map for the *C. eurytheme* genome was obtained by Lep-MAP3 (LM3)(Rastas  
730 2017) pipeline. First, the individual fastq files were mapped to the scaffold assembly using bwa mem (H. Li  
731 2013), and using LM3 pipeline (pileupParser.awk, pileup2posterior.awk) and SAMtools mpileup (H. Li et al.  
732 2009), we obtained the input genotype likelihoods.

733 Linkage mapping followed the basic LM3 pipeline as follows (non-default parameters inside parenthesis):

- 734 1) ParentCall2(ZLimit=2, removeNonInformative=1)
- 735 2) Filtering2 (dataTolerance=0.0001)
- 736 3) SeparateChromosomes2(lodLimit=14.5 maleTheta=0.5 femaleTheta=0.0001 distortionLod=1  
737 sizeLimit=4)
- 738 4) 2 x JoinSingles2All (lodLimit=10 lodDifference=2 maleTheta=0.05 femaleTheta=0.0001  
739 distortionLod=1)
- 740 5) OrderMarkers2 (chromosome=1..31 recombination2=0 informativeMask=13 useMorgan=1).

741 This yielded 31 linkage groups but after inspecting the markers occurring in the same scaffolds, two linkage  
742 groups were joined (27+30 and 29+25). Moreover, two linkage groups had very long maps (>100cM) and these  
743 groups were split, group 1 with SeparateChromosomes2 (map=map14.5.txt maleTheta=0.5 femaleTheta=0.0001  
744 distortionLod=1 lg=1 renameLGs=0 lodLimit=16) and group 10 based on sex/autosome markers (indicated by \*  
745 in the output of ParentCall2).

746 After these splits and joins, the JoinSingles2All and OrderMarkers2 were run again to obtain final (*de novo*)  
747 linkage maps with 31 linkage groups. The splits and joins were necessary due to complex family structure  
748 (multiple families) and low marker density.

749 With the help of the linkage map, the scaffolds (all except Sc0000116) were manually put together into these 31  
750 linkage groups. Linkage groups were named based on chromosome numbers in *Melitea cinxia* (Ahola et al.  
751 2014). The linkage map was re-evaluated in this physical order with OrderMarkers2 (recombination2=0  
752 chromosome=1..31 evaluateOrder=phys\_order improveOrder=0 hyperPhaser=1 phasingIterations=3), put into  
753 grand parental phase (phasematch.awk) and this map was used for QTL mapping.

#### 754 ***C. eurytheme* genome polishing, quality control, and annotation**

755 Polishing of the genome was performed using Pilon v1.2.2 ( Walker et al. - 2014), using data from a single  
756 orange female, with a Illumina TruSeq Nano library prep and sequenced (150 bp PE reads with 350bp350 bp  
757 insert size, Illumina HiSeqX) to provide ~30X genome coverage, aligned using NextGenMap v0.5.2. The assembly  
758 quality was assessed using custom scripts for basic length metrics and BUSCO v1.1b1 before and after polishing  
759 to evaluate the difference, using the insecta\_odb9 dataset. The genome was softmasked for repetitive regions  
760 using RED (Version: 05/22/2015, Girgis, H.Z 2015 (Girgis 2015)).

761 Genome annotation was generated using BRAKER2 (v2.1) trained on data from *C. eurytheme* transcriptome and  
762 proteins. RNA-seq data generated in a previous study (Nallu et al.?), consisting or transcriptome data from  
763 several developmental life stages. We used a reference protein dataset from the arthropoda section of OrthoDB  
764 (v10). Transcriptome reads were aligned using HISAT2 v.2.1.0 (Kim et al. 2019, 2), against the unmasked  
765 genome, and the alignment was then filtered, sorted and indexed using SAMTOOLS v.1.7. Braker2 was run using  
766 the ETP mode and set to take softmasking into account. Quality of the annotation was then assessed by  
767 counting the number of good transcripts (genes containing both start and stop codon). Further assessment was  
768 done using OHR analysis (modified from (O’Neil et al. 2010))), reads transcripts with 0.9 identity or higher were  
769 clustered using CDHit v.4.8.1 (W. Li and Godzik 2006) and then compared against a protein set from *B. mori*.  
770 This gives an estimate of the fraction of full-length transcripts in the annotation, as well as completeness.  
771 Annotation of the resulting chromosome level assembly identified 18,460 genes, 18,081 of which had a correct  
772 start and stop codon. Clustering these "good genes" at 90% identity resulted in 16,352 genes. When we  
773 compared our annotation with the *B. mori* proteome, our set covered 11,623 of the 14,052 *B. mori* proteins  
774 with at least 80% of the *B. mori* length, indicating a high-quality annotation.

#### 775 **Synteny comparative analysis**

776 To assess our genome assembly and check for any large-scale structural changes compared to other sequenced  
777 Lepidopteran genomes, we compared our *C. eurytheme* chromonome to one from the sister genus, *Zerene*  
778 *cesonia* (Rodriguez-Caro et al. 2020). Whole-genome alignments were performed using nucmer followed by  
779 circos plotting using the R package circlize v.0.4.9  
780 (<https://academic.oup.com/bioinformatics/article/30/19/2811/2422259>). Bioinformatic details can be found on  
781 our github page.

#### 782 **Phylogenetic analyses**

783 The longest exon per gene dataset was generated by running BUSCO upon the protein dataset from our  
784 annotation of the *C. eurytheme* genome, with the protein dataset generated using our GFF annotation and the  
785 genome as inputs for the gffread script from cufflinks v.2.2.1 (Trapnell et al. 2010). Of the total lepidopteran  
786 BUSCO genes searched (n= 5286), 4476 were found complete and single copy in our protein dataset for *C.*  
787 *eurytheme*. Among the BUSCO outputs is a table where for each annotated protein identified, its BUSCO status  
788 is indicated (e.g. as single and complete, duplicated, etc). Using this table, the exons of the complete and single

789 copy proteins in our annotation were extracted and converted to bed. Then the length of each exon was  
790 calculated, allowing for the longest exon per protein ID to be selected using custom scripts, and the resulting  
791 bed file of these longest exons was then used as input for the bam2fasta script from the package bambam v1.4  
792 tool-kit (Page et al. 2014), along with all the bam alignment files and a minimum depth requirement of 5 reads  
793 per base pair. The resulting set of fasta files (busco\_exons) was then used to generate gene trees using iQtree,  
794 with each gene tree using extended model selection, a random starting tree, 1000 ultrafast bootstraps and  
795 optimization, and *Z. cesonia* set as an outgroup (-m MFP -t RANDOM -bb 1000 -alrt 1000 -bnni -o Z\_cesonia). A  
796 total of 4244 gene trees were generated. These were then used as input for species tree estimate by Astral  
797 using default settings. Using the genome annotation, gene trees were also grouped by chromosome, which  
798 allowed for species tree for each chromosome to be similarly estimated.

799 We also generated a filtered set of the busco\_exon dataset, using AMAS V.1.0 (Borowiec 2016) to remove *Z.*  
800 *cesonia* from all fasta files, and then generate a summary table of all files, which was then parsed to produce a  
801 set of files filtered to remove those with missing content > 1 %, < 5 % variable sites), and length < 300 bp and >  
802 2000 bp. With this filtered set of fasts file IDs (n=1400), these iQtree gene trees were then selected for species  
803 tree estimation, to assess whether dataset quality had any effect on species tree topology. The resulting species  
804 tree from these filtered fasta files was identical to the full busco\_exon analysis. Additional analyses using full  
805 CDS for ~9000 genes, or their 2<sup>nd</sup> exon, produced essentially identical species tree results (data not shown).

806 Gene tree concordance with the species tree was assessed using Phyplots (Smith et al. 2015), to calculate the  
807 number of gene trees concordant and discordant with the species tree topology, per node. Phyplots output was  
808 further parsed to distinguish among gene trees discordant with the species tree, into those supporting a main  
809 alternative tree vs. many alternative trees, as well as those gene trees having less than 50% bootstrap support  
810 at the node in question. We used pieplots to represent these proportions  
([https://github.com/mossmatters/MJPythonNotebooks/blob/master/PhyParts\\_PieCharts.ipynb](https://github.com/mossmatters/MJPythonNotebooks/blob/master/PhyParts_PieCharts.ipynb)). Importantly,  
812 these pieplots results were concordant with estimated gene concordance factors via iQtree (data not shown).

813 SNAPP analyses (Bryant et al. 2012), which are run as an addon to the BEAST2 software program, were used for  
814 generating multi-species-coalescent analyses, but instead of using whole genes as in StarBEAST2, SNAPP uses  
815 SNPs. A further difference to StarBEAST2 is that gene trees are not estimated for each SNP, although SNPs are  
816 considered unique markers SNAPP; SNAPP estimate the species tree probability by integrating over all the  
817 possible gene trees observed among the SNPs. Despite this approach dramatically decreasing parameter space  
818 for analyses, this approach remains computationally demanding. Recent work has investigated the number of  
819 SNPs for optimal inference while minimizing computation demands, as well determined accurate SNAPP settings  
820 for time calibration, which uses a strick-clock model using fossil calibrations and the linking of all population  
821 sizes during analysis (Stange et al. 2018). Thus, in order to stay within a multi-species-coalescent framework for  
822 divergence time and phylogenetic relationship estimation, we followed these recommendations (Stange et al.  
823 2018) and down sampled our taxa to remove redundancy among closely related species while retaining regional  
824 diversity. This took our full dataset of 29 species down to 18. Using AMAS, exon fasta files were subsampled to  
825 these 18 species, exons concatenated and then converted to phylip file format. A ruby script was used to  
826 generate an input XML file for SNAPP via snapp\_prep.rb ([https://github.com/mmatschiner/snapp\\_prep](https://github.com/mmatschiner/snapp_prep)), which  
827 takes a phylip formatted sequence dataset, allows for specifications of run iterations, inter-SNP intervals, total  
828 SNPs, a starting tree, and diverse constraints to incorporated. Constraints included a temporal calibration for  
829 the timing of the split between *Zerene* and *Colias* using a secondary calibration of 10.9 million years ago with  
830 sigma=1, along with two monophyletic constraints well supported by previous Astral analyses (one for South  
831 American taxa, one for the remaining *Colias* species). Two SNP datasets were constructed, each containing 1000  
832 random SNPs: 1) SNPs sampled from among the concatenated BUSCO exons with at least 300 bp between each  
833 SNP, 2) SNPs from among concatenated filtered BUSCO exons, with at least 100 bp between each SNP. Each  
834 dataset was run twice, for between 2 to 3 million iterations, which returned effective sample sizes (ESS) > 200  
835 and were convergent within dataset with stable likelihood values, which was assessed using Tracer in the  
836 BEAST2 package. Posterior estimates used mean tree credibility after a 10% of data was discarded as burn n, as  
837 implemented in Treeannotator in the BEAST2 package. The resulting phylogenetic relationships and divergence  
838 estimates for *Colias* and non-South American *Colias* crown groups we nearly identical in their results, which was  
839 visualized using Figtree v.1.4.4 (Rambaut and Drummond 2012).

840 ***Introgression analysis***

841 In order to assess the level of gene flow and introgression among *Colias* species, we calculated the D-statistic  
842 between all possible trios of species for which we had sequence data. Using the software Dsuite (ref) we were  
843 able to analyze this jointly and infer between which, past and present taxa, we've likely had introgression  
844 (Malinsky, Matschiner, and Svardal 2020). By calculating a F-branch statistic and analyzing it together with the  
845 phylogeny, we can infer between which taxa and which nodes in the phylogeny gene flow is most likely.

846 For this analysis, we used the sequence data for different *Colias* species mentioned previously and aligned them  
847 to the manually curated Alba-reference genome using NGM (same settings as previously). No filtering was done  
848 before using Freebayes to call variants. The resulting vcf-file filtered for strand-bias, quality score of 30, 90% of  
849 species sharing the site, minimum sample depth of 5, and minor allele frequency of 5%. In the analysis, only  
850 biallelic SNPs are included. D statistic between all our trios was calculated with the phylogeny as a reference  
851 tree to guide the analysis using the Dtrios tool in the Dsuite toolkit (v.0.4). This compares all possible trios of  
852 species and calculates a genome-wide minimum D, as well as a significance value of each trial, it will also  
853 calculate F4 statistics for each trio that we used to calculate Fbranch statistic that we used to infer signals of  
854 past introgression. We used the DInvestigate tool from Dsuite to generate window-based measures of f\_dM, as  
855 a measure of introgression for each trio that showed significant (BH corrected p < 0.05) introgression in the  
856 genome-wide analysis (Fig. 1 e) in overlapping windows of 50 SNPs (25SNP overlap). We were interested to see  
857 if there were any trios that showed an elevated signal of introgression around the BarH1 locus. To this end, we  
858 filtered out all trios that showed an increase in f\_dM in the 600Kb region surrounding BarH1, this revealed 3  
859 species trios (Supplementary figure 3). However, when investigated in detail, they were found to represent  
860 increased haplotype similarity between 2 fixed alba species in comparison to fixed orange species; additionally,  
861 none of these had any increase in f\_dM across the actual BarH1 genic region or across the insertion region.

862

863 ***GWAS of Alba in C. eurytheme***

864 DNA was extracted using KingFisher Cell and Tissue DNA Kit from ThermoFisher Scientific (N11997) and the  
865 robotic Kingfisher Duo Prime purification system. DNA purity was assessed using 260/280 ratio (Nanodrop 8000  
866 spectrophotometer; Thermo Scientific, MA, USA) and concentration was quantified on a Qubit 2.0 Fluorometer  
867 (dsDNA BR; Invitrogen, Carlsbad, CA, USA). DNA was sent to the Science for Life Laboratory (Stockholm, Sweden)  
868 for library preparation and sequencing (Rubicon, 150bp PE reads with 350bp insert size, Illumina HiSeqX).  
869 Sequencing libraries were sequenced twice, to a predicted depth of 10x each time. Raw reads were clone  
870 filtered, had adaptors trimmed and low-quality bases (PHRED 20) removed using the BBduk tool from the  
871 BBmap software package v34.86 (Bushnell B. sourceforge.net/projects/bbmap/). Cleaned reads were mapped  
872 to the *C. eurytheme* reference genome using NextGenMap v0.5.2. SAMTOOLS was used to filter out unmapped  
873 reads, sort, duplicate marking of reads and indexing. PICARD-tools v1.139 was used to add Readgroups and then  
874 again to merge the separate sequencing runs of each sample, before a final round of duplicate marking using  
875 SAMTOOLS. Variants were called using Freebayes v1.3.1-16-g85d7bfc (Garrison E -2012 arXiv), and the resulting  
876 VCF-file was filtered using a mix of VCFTOOLS v0.1.13 Danecek, (Petr, et al. - 2011) and custom awk scripts. The  
877 final GWAS was run using PLINK v1.9 (Chang et al. - 2015) to identify associated loci. We filtered the GWAS  
878 based on two separate sets of filters were used. First, we filtered the VCF file to remove low quality SNPs or  
879 sites that did not match our depth or frequency criteria (minimum depth 3, minQ 20, max-missing 0.95, and  
880 minor allele cutoff of 0.05) . Second, we added an additional filter to this set added the prior criteria to include  
881 only sites that were unique to the Alba females. Additionally, we also filtered the output using the information  
882 gained from the QTL analysis, and only kept SNPs on the scaffolds that make up Chromosome 3.

883 The same filtering approach was also applied to the GWAS done using the Alba reference genome.

884

885 ***Generation of draft genomes alba individuals of C. eurytheme, C. nastes and C. crocea***

886 DNA from wild caught females was extracted from the thorax of samples stored frozen in 95% ethanol using the  
887 same protocol as in the resequencing done for the GWAS. Prior to library preparation an additional analysis of

888 molecular wight of the DNA was performed using gel electrophoresis (0.5% agarose LE). DNA was extracted  
889 from 2 individuals of each species and the ones with the highest quality DNA, as determined by gel  
890 electrophoresis, nanodrop measurement and Qbit were selected for library prep. Sequencing and assembly with  
891 Supernova v2.1.1 at SciLifeLab.

892 ***PCR-based validation of insertion***

893 The presence and uniqueness of the insertion to Alba individuals were further validated using PCR-based  
894 markers. The primers were designed to bind within the insertion region. that was found to be unique to Alba.  
895 Optimal primer binding sites were identified using the primer3 software (libprimer3 release 2.5.0, Untergasser  
896 et al. - 2012). PCR-reactions were run on DNA extracted from 8 orange and 8 Alba females, that had not yet  
897 been sequenced or used in the GWAS. Positive controls were run using previously validated primers binding to  
898 mitochondrial cytochrome oxidase I gene (Brower et al. 2006). The reactions were run using Invitrogen Platinum  
899 Taq in a Veriti 96-well thermocycler (Applied Biosystems, Foster City, CA, USA) using the recommended settings  
900 for the polymerase (72C x 2min followed 35 cycles of 94C x 30sec + 54C x 30sec + 72C x 15 sec followed by 72C  
901 x 5min). The PCR product was visualized by agarose gel electrophoresis in a 1% gel (Supplementary figure 7).

902 ***Generation of Alba reference genome***

903 We identified the scaffold containing *BarH1* in the supernova assembly of the *C. eurytheme* using tBLASTn. We  
904 then aligned all the resequencing data from the GWAS to this contig alone and evaluated along the contig  
905 visually in IGV. Regions where no orange reads aligned, but Alba did, were extracted and blasted back against  
906 the *C. crocea* reference genome (Woronik 2019) to assess whether this was the previously identified Alba  
907 insertion region. Once we had established the Contig containing both *BarH1* and the Alba insertion, we blasted  
908 the entire contig back against the Orange *C. eurytheme* reference genome to establish the edges of the contig in  
909 the reference genome as well as to establish orthology. Once edges were established, we manually inserted the  
910 entire contig, making sure to remove the overlapping sequence, generating what we refer to as the Alba  
911 reference genome.

912 We decided to generate the additional *C. crocea* assembly to make sure that we would be able to assemble and  
913 subsequently detect the alba insertion sequence. This was done using the same method as described in the  
914 main paper for the detection of the Alba allele in the *C. nastus*, and *C. eurytheme* assemblies, where we blasted  
915 the previously known alba allele detected in the *C. crocea* assembly from Woronik et al. 2019, and if this  
916 insertion was co-located with the *BarH1* gene. We also assessed sequence similarity and synteny using Blastn,  
917 which we visualized using Kablamm (supplementary figure 8 & 9).

918

919 ***Alignment and assessment of the Alba insertion across species***

920 The sequence data generated for the phylogeny was aligned to the Alba reference-genome using NextGenMap,  
921 filtered for mapQ 20 and being in proper pairs. Mapability of the data, and confirmation of sex in cases where it  
922 was uncertain was done using read coverage analysis using goleft indexcov (see SM\_CovStats). This confirmed  
923 that 7 samples in our analysis were indeed male (*Colias erate* *poligraphus*, *Colias tam.* *mogola*, *Colias*  
924 *phicomone*, *Colias* *Tyche*, *Colias behri*(B033), *Colias tamerlana* (B037) and *Colias wiscotti*) based on the  
925 increased coverage on the scaffolds that make up the sex-chromosome (9, 19 &52). It also emphasizes the  
926 divergence between the South American species *C. lesbia* and *C. euxanthe*, as those species consistently showed  
927 reduced coverage on the terminal ends of the shorted chromosomes in the assembly (SM\_CovStats).

928 Coverage across the insertion region was then visually inspected visually in IGV. We especially assessed the  
929 regions where we had observed a difference in coverage between Orange and Alba eurytheme individuals and  
930 whether this extended to other species where we had sampled either Orange or Alba individuals. The sequence  
931 that was unique to only white-colored species (putative Alba species) was the conserved Alba region and likely  
932 causal for the phenotype across *Colias*.

933 To establish a null expectation for how often a region of the same size as the conserved Alba region would  
934 segregate in this manner between white and colored species (e. g orange; yellow; red) we performed a read

935 depth analysis of all the sequenced *Colias* species. Reads were aligned using Nextgenmapper and filtered for  
936 being good pairs. We then used goleft v.0.2.1 (<https://github.com/brentp/goleft>) to evaluate the average read  
937 depth of 600bp windows across all scaffolds. We selected 600 basepairs rather than the full 1200bp of the  
938 Candidate alba locus to ensure that at least one window ended up inside the insertion region, this also made  
939 the analysis more sensitive compared to using a larger window size. Read depth in the windows was then binary  
940 classified as either having read coverage or not, with the latter classification assigned if they had less than 25%  
941 of the read depth compared to the scaffold average. Thus, for each window, each species had a value of 0 or 1.  
942 Then, using these values of either having reads covering the window or not, we calculated the mean value per  
943 window for the white and colored groups of species, then calculated the white minus colored value. For the  
944 Alba identified region in white-colored butterflies, this value was 1, which was then compared to the rest of the  
945 windows across the genome, which served as a genome scale control.

946 ***Phylogenetic analysis of the Alba insertion.***

947 We extracted the consensus sequence of the region from all 43 Alba samples in our analysis (14 *erytheme*, ten  
948 *crocea*, seven *philodice* (Maryland), 1 *philodice* (from British Columbia) and one from each remaining alba  
949 species in the phylogeny) using the bam2fasta 1.4 tool-kit. We keep all sites with at least one read covering it  
950 and selected the most common allele at polymorphic sites. We then ran IQtree to generate a phylogenetic tree  
951 of the sequences. IQtree was run with model finder plus enabled to allow it to find the most parsimonious  
952 model.

953 ***CRISPR/Cas9 targeted mutagenesis of the Alba insertion.***

954 We used the PROMO tool to narrow down regions of interest in the Conserved Alba locus, while we did keep an  
955 out for candidate genes that have been previously shown to either interact with BarH1 or known to be involved  
956 in sexual dimorphisms, such as *Doublesex*, we also used it to get an idea of sites that showed a higher density of  
957 many different potential transcription factors. In the end, we selected four target sites, two that targeted the  
958 only putative binding site for the *Doublesex* transcription factor and two that targeted upstream and  
959 downstream of this position. Together, if all cuts would be successful, this would remove ~500basepairs of  
960 sequence and remove a large portion of the potential binding sites. We did four types of cocktails of gRNA/Cas9  
961 mixture that we injected;

962 gRNA\_P(doublesex binding site)

963 gRNA\_P and gRNA\_U(3bp upstream of doublesex binding site)

964 gRNA\_2 (~250bp upstream of gRNA1) and gRNA\_5 (~150 bp downstream of gRNA1).

965 And all four together.

966 Unfortunately, due to poor egg-laying of the alba females, we could not inject an equal number of eggs with  
967 each combination, instead primarily injecting eggs with either gRNA\_P alone, or gRNA\_U and gRNA\_P.

968 The only combination we saw any phenotype from was when we injected all 4 together. But as this was only  
969 done once the females had almost stopped to inject eggs, we didn't inject more than 40 eggs with this  
970 combination. The major cause of death among our injections was the larvae getting stuck in the double-sided  
971 tape that was used to fixate the eggs to a glass slide during injection. Out of 200 injected eggs, we were only  
972 able to transfer 39 larvae to fresh hostplant (5 of the 40 eggs we injected with all four gRNAs).

973

974 ***Validation of CRISPR/Cas9 targeted mutagenesis.***

975 Using Primer3 we designed PCR primer pairs that bind upstream and downstream of gRNA\_2 and gRNA\_5. Due  
976 to a large amount of repetitive DNA in the region, we were forced to have the forward primer bind to a non-  
977 unique location, leading to a risk of having some off-target bind sites. The reverse primer was unique, though,  
978 and there were no alternative bind-sites for the forward primer within 200Kb of the intended location. The  
979 reactions were run using Invitrogen Platinum Taq in a Veriti 96-well thermocycler (Applied Biosystems, Foster

980 City, CA, USA) using the recommended settings for the polymerase (72C x 2min followed 35 cycles of 94C x  
981 30sec + 54C x 30sec + 72C x 15 sec followed by 72C x 5min). The PCR product was visualized by agarose gel  
982 electrophoresis in a 1% gel (Supplementary figure 12). We used the primer pair developed by Woronik. Et al  
983 2019. to validate the Alba status of the sample and as a positive control for the reaction, this primer pair does  
984 not bind in the vicinity of each other.

985

986

987

988 *Tables.*

989

990 **Table1. Quality metrics of the assembly**

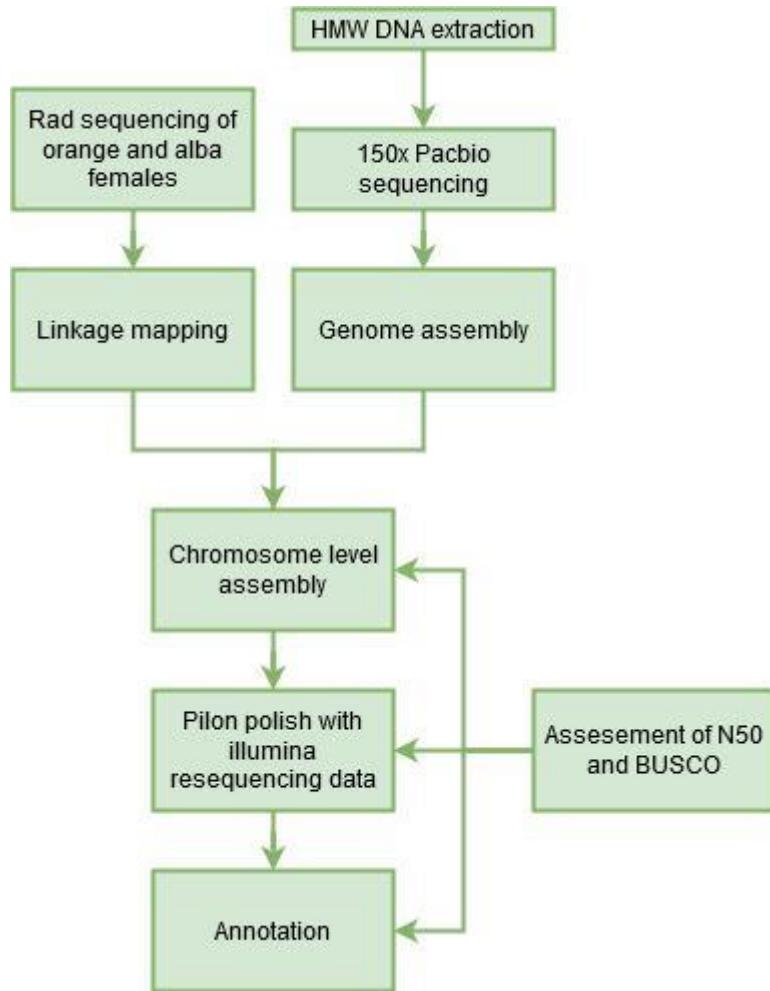
<b>Size</b>	
Contigs Generated :	108
Maximum Contig Length (Mb) :	13
Minimum Contig Length :	38,228
Average Contig Length (Mb) :	3.0 ± 2.9
Total Contigs Length (Mb):	327
N50 value MB :	5.2
<b>Completeness</b>	
BUSCO summary	C:98.5%[S:97.7%,D:0.8%],F:0.2%,M:1.3%,n:52 86
Complete BUSCOs	5209
Complete and single-copy BUSCOs	5166
Complete and duplicated BUSCOs	43
Fragmented BUSCOs	12
Missing BUSCOs	65
Total BUSCO groups searched	5286
<b>Annotation</b>	
Transcripts	16842

991

992 ST1. Metadata table of sequenced individuals, including capture site, sequencing depth, and sample name.  
993 ST2. Top hits of GWAS after light filtering of SNPs, using the orange reference.  
994 ST3. Top hits using informed priors, using orange reference  
995 ST4. Top hits using QTL informed priors, orange reference  
996 ST5. Top hits using Alba reference genome and light filters. Note that the hits not on the alba locus are on the  
997 sex chromosome.  
998 ST6. gRNA sequence  
999 ST7. Injection statistics and survival.  
1000 ST8. PCR primers.  
1001 ST\_covstats. Normalized coverage of reads across scaffolds of all samples.  
1002  
1003  
1004  
1005  
1006

1007 **Figures**

1008



1009

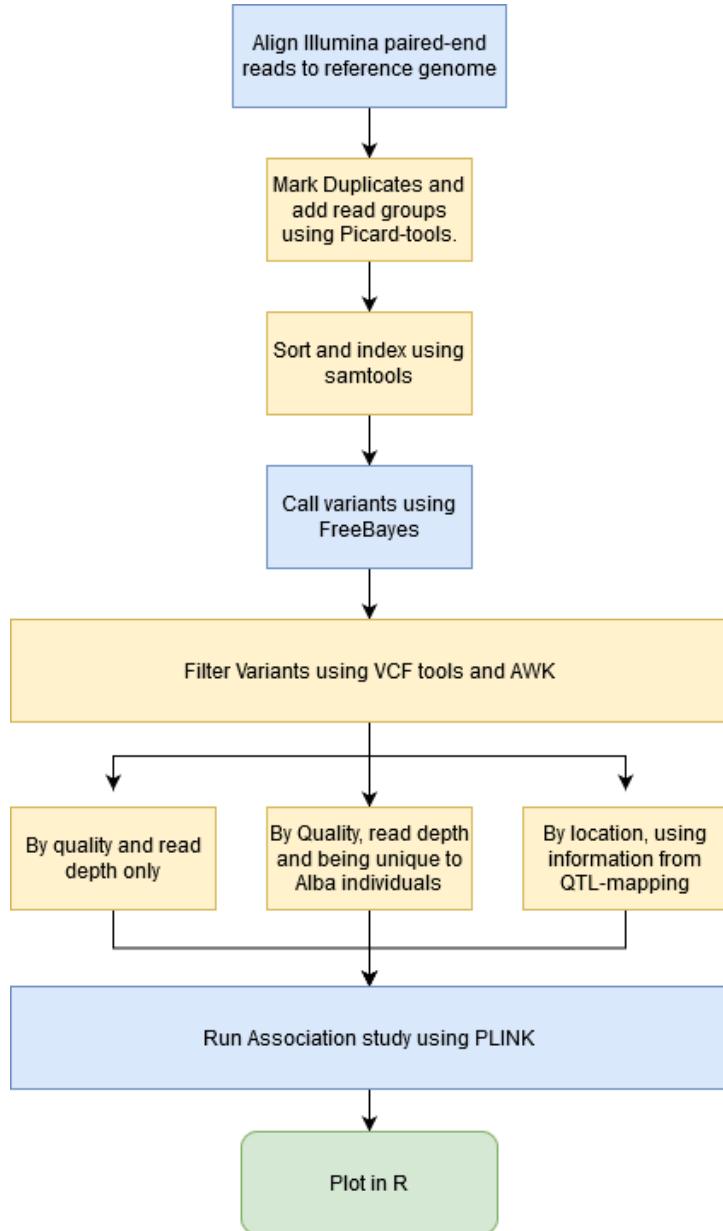
1010

1011 **Supplementary flowchart 1. Assembly pipeline.** Flowchart illustrating the genome assembly, annotation, and  
1012 evaluation process. Generation of raw contigs was done using 150x Pacbio sequence data coming from a single  
1013 female pupa. Linkage mapping of the contig data was done using Rad sequencing from multiple individuals of *C.*  
1014 *erytheme* X *C. philodice* hybrid crosses. The draft assembly was then polished using Pilon to reduce indel errors  
1015 introduced by the pacbio sequencing, and finally annotated using BRAKER2. Each step of the assembly process  
1016 was analyzed using N50 and BUSCO.

1017

1018

1019

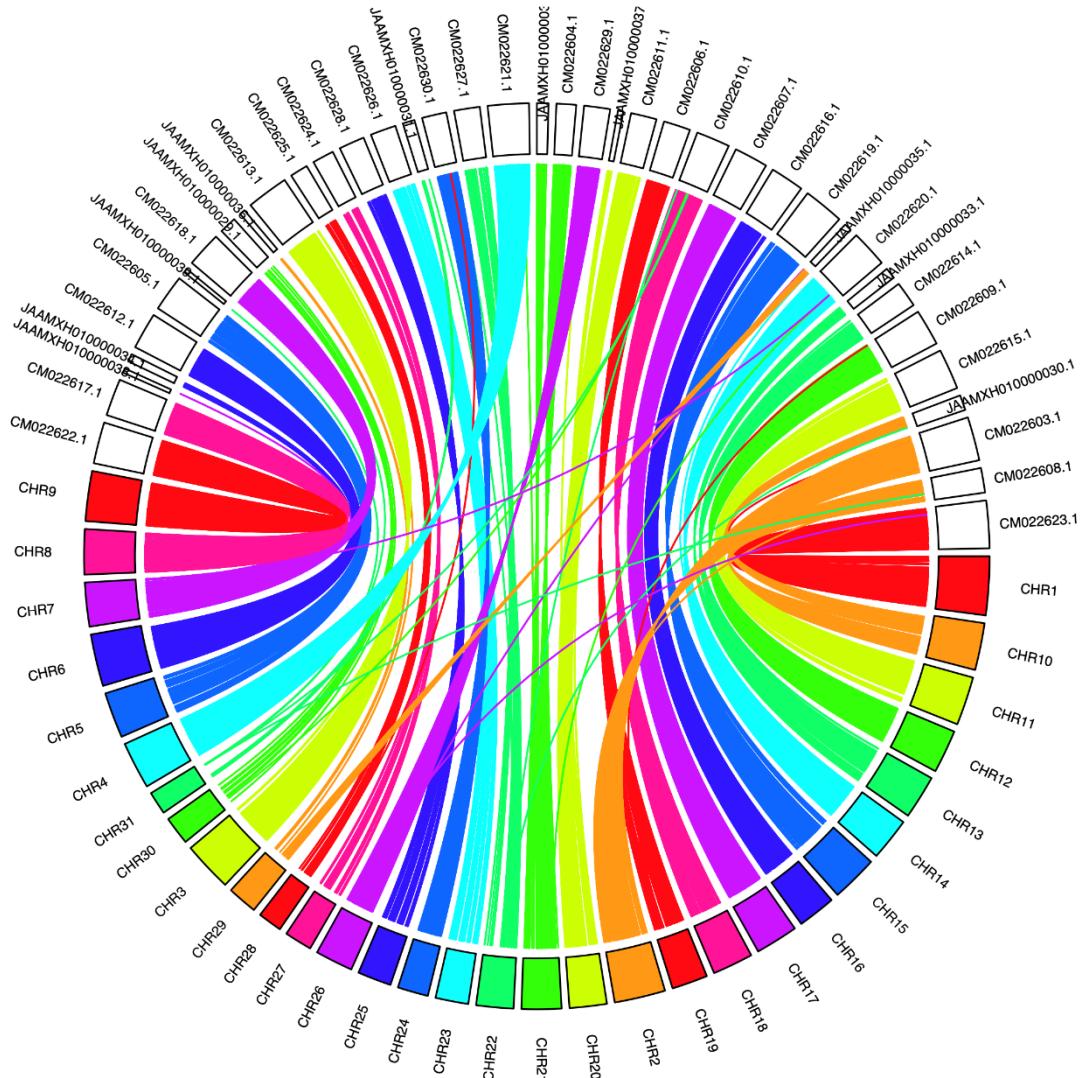


1020

1021

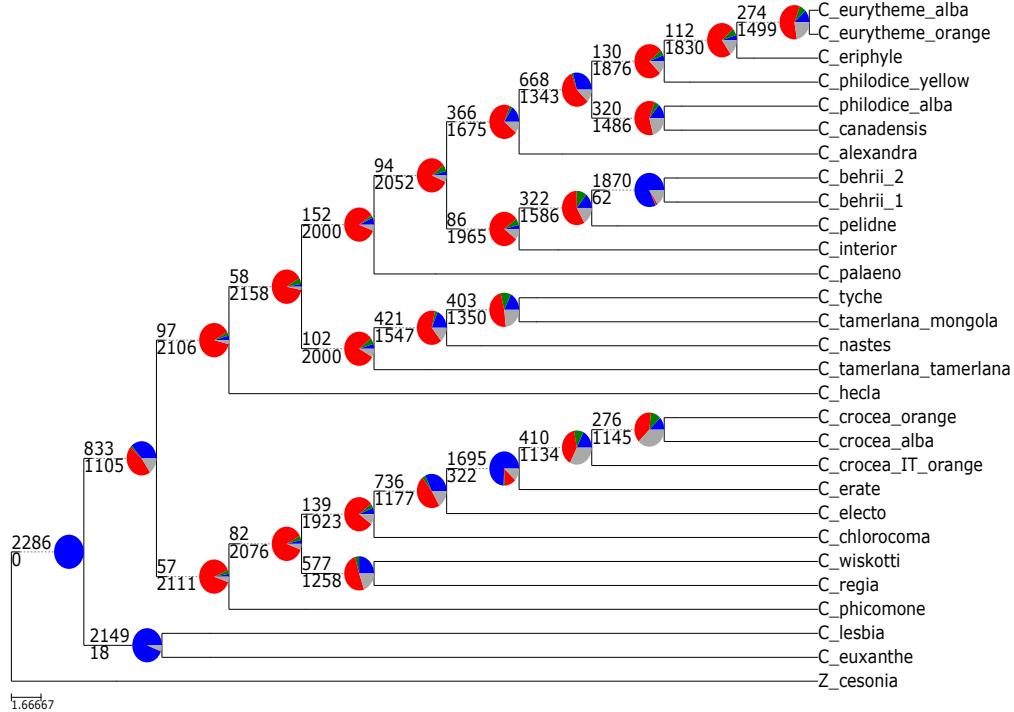
1022 **Supplementary flowchart 2. Variant calling pipeline.** Flowchart describing the steps involved in the Genome  
1023 Wide Association study (GWAS). Blue boxed represent the generation of Data, Yellow boxed filtering of data and  
1024 green boxes visualization. In total we ran 3 levels of filtering of the raw VCF-file with different level of priors.

1025



1026 **Supplementary figure 1. Syntheny plot illustrating the conserved chromosomal structural comparison**  
1027 **between the *Colias eurytheme* genome and *Zerene cesonia*.** Shown are the 31 lineage groups identified in *C.*  
1028 *eurytheme* colored by chromosome, while the scaffolds of *Z. cesonia* are uncolored. Each line represents an  
1029 inferred orthologous region. Only scaffolds > 1 Mb of *Z. cesonia* were included, with nucleotide alignment  
1030 identity > 88 %. Note how several *Z. cesonia* scaffolds are brought together within *C. eurytheme* chromosomes  
1031 (e.g. Chr 6 and Chr 29). Given the high synteny between *Z. cesonia* and *Heliconius erato* (Rodriguez-Caro et al.  
1032 2020), and *Heliconius* to other butterflies and moths (Ahola et al. 2014), we infer that *Colias* chromosomal  
1033 structure adheres to the standard Lepidoptera chromosome structure (Hill et al. 2019).

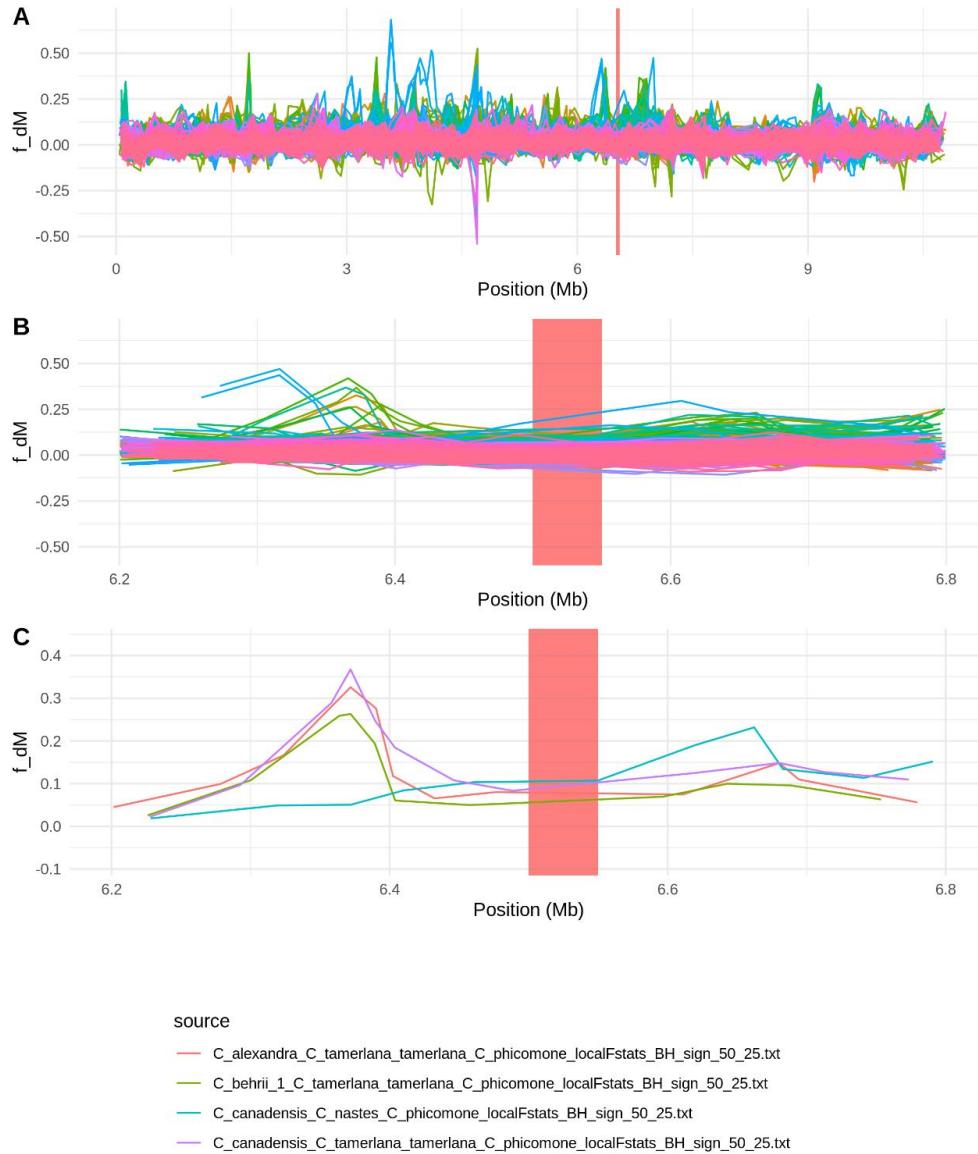
1034



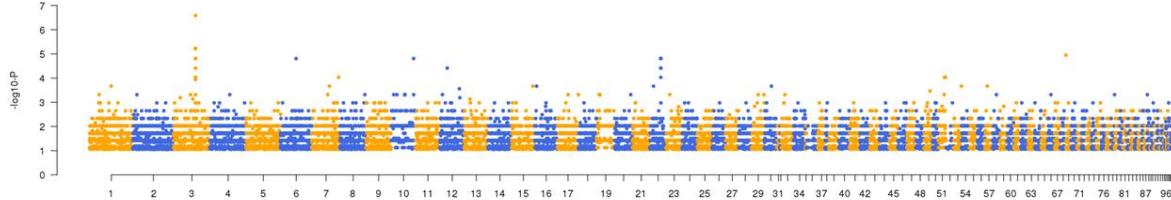
1035

1036 **Supplemental figure 2. Pie charts of gene tree concordance**, conflict, and lack of signal compared to the Astral  
1037 species tree. While traditional node support values, whether bootstraps or posterior probability, generally  
1038 overstate gene tree support, here we present direct quantification per node of gene tree topology with the  
1039 species tree topology. At each node, the number of gene trees concordant (top number) and in conflict (bottom  
1040 number) is shown. Pie charts at each node give a further breakdown of gene tree proportions by those  
1041 concordant with species tree (blue), those that support a common alternative topology (green), those that  
1042 support the remaining low-frequency alternatives (red), and those lacking robust information as they have less  
1043 than 50% bootstrap support (gray). Here, the South American taxa are very well supported, as are the clades  
1044 containing: *C. crocea*, *C. erate*, *C. electo*; *C. tyche*, *C. tamerlana mongola*, *C. nastes*; *C. pelidne*, *C. behrii*; *C.*  
1045 *alexandra*, *C. canadensis*, *C. philodice*, *C. eurytheme*, *C. eriphyle*. Note how the vast majority of nodes are red,  
1046 indicating extensive gene tree conflict rather than lack of phylogenetic signal (gray).

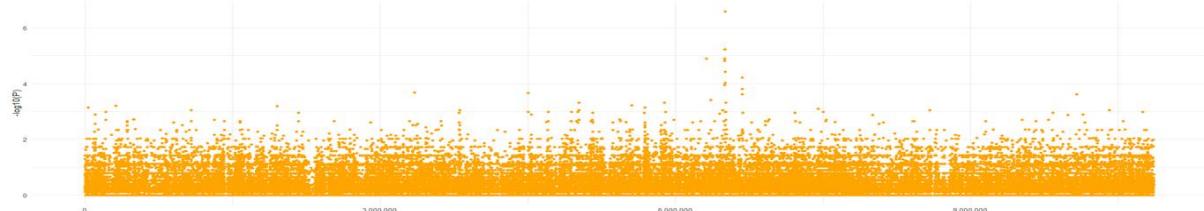
1047



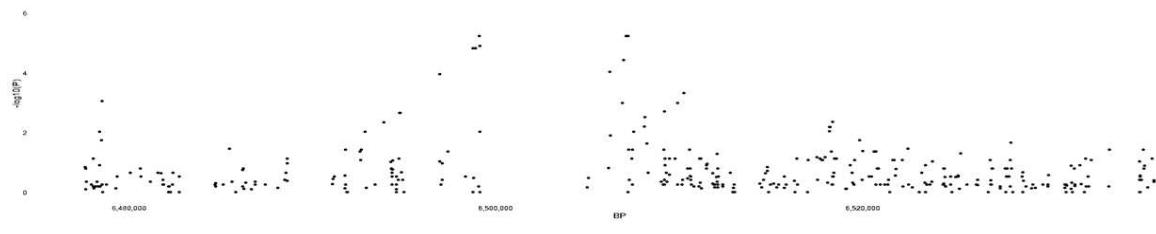
1048 **Supplementary figure 3. Assessment of signals of localized introgression.** By looking at  $f_dM$  in sliding windows  
1049 of 50 SNPs across the genome for all species trios showing significant levels of introgression in the genome-wide  
1050 Introgression analysis looking at A. Whole scaffold containing BarH1, B. 600Kb region surrounding BarH1 and C.  
1051 The four species trios with an  $f_dM$  higher than 0.2 anywhere in the 600Kb region surrounding BarH1. The  
1052 BarH1 region is highlighted red in all three plots.



1053



1054

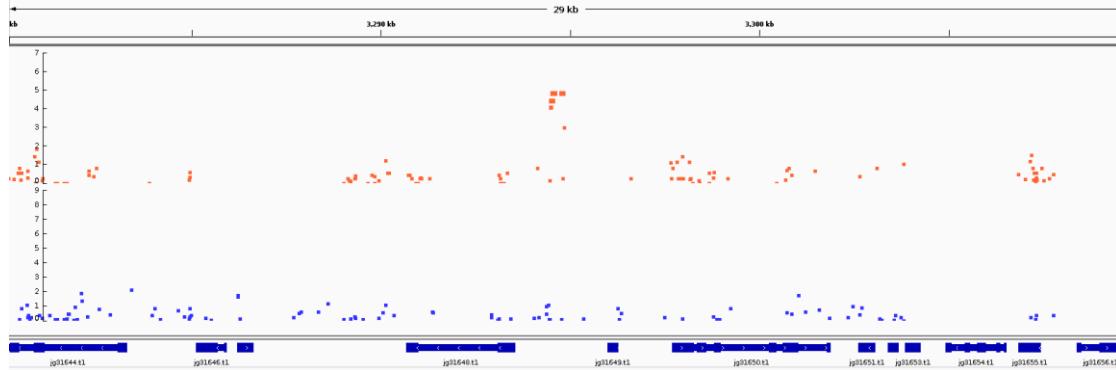


1055

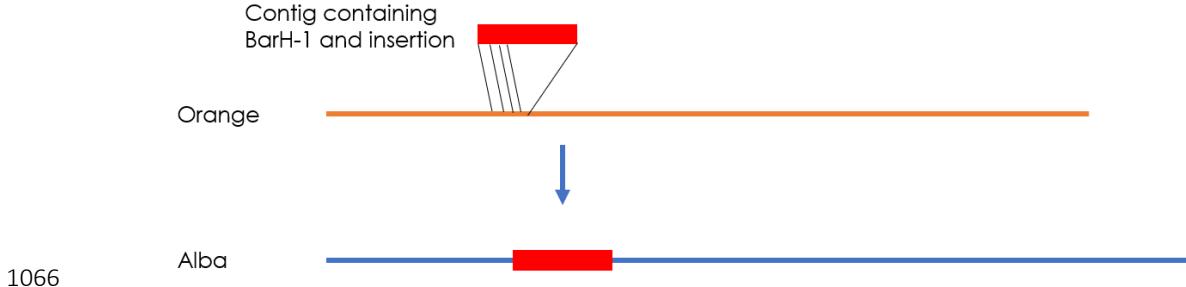
1056

1057

1058 **Supplementary figure 4 GWAS results.** GWAS against the Orange reference genome using the light filters of  
1059 quality and depth only.  $-\log_{10}(p\text{-value})$  for SNP correlating with alba color against the position in the genome.  
1060 Panel A: Genome-wide, B: Sc0000002, C: 60kb around the highest peak.

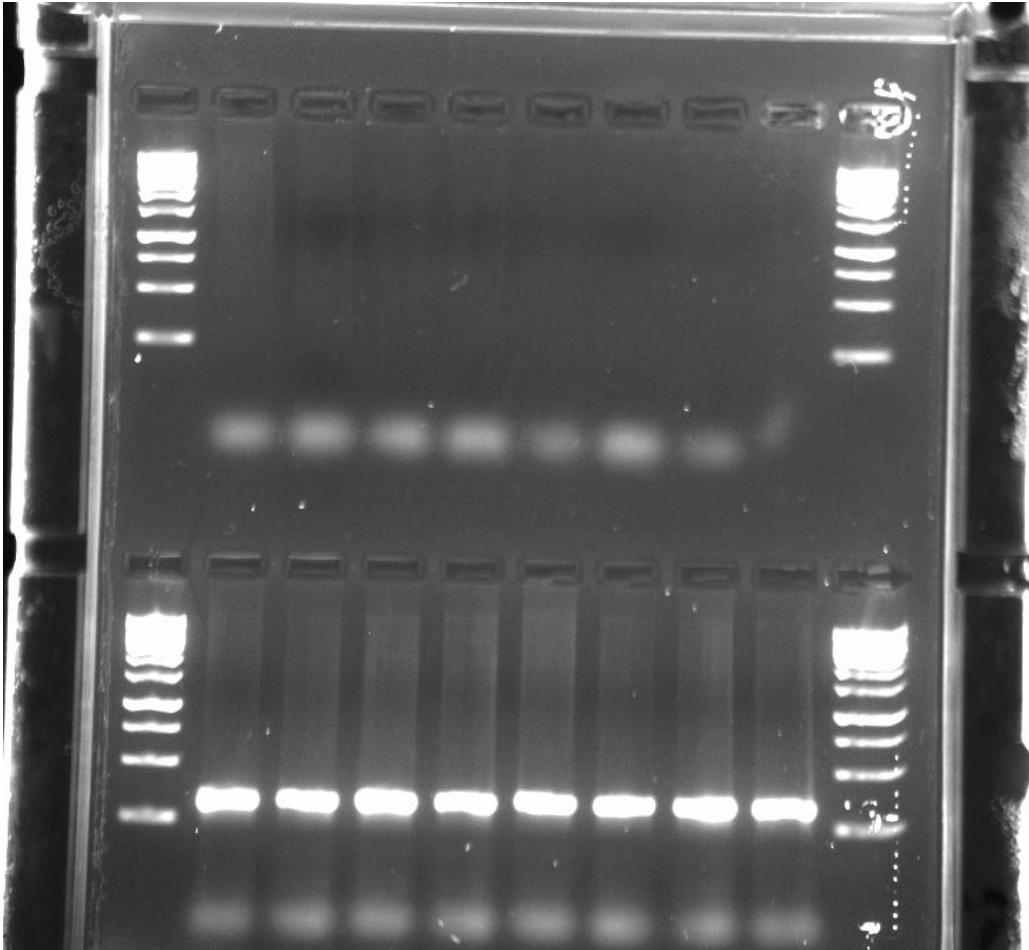


1061 *Supplementary figure 5. Close up on the second GWAS locus identified on Scaffold 22 when the orange*  
1062 *reference genome was used. The top row colored in orange represents sites identified against the orange*  
1063 *reference genome, while the blue are from the synthetic Alba reference genome. The gene to the right of the*  
1064 *highly associated loci is similar PIFI-like transposase when blasted against NCBI, and the gene to the right is*  
1065 *similar to a PiggyBac transposon.*



1066      Alba

1067 **Supplementary figure 6. Illustration of the generation of the Alba reference genome.** The contig containing  
1068 *BarH1* and the insertion was identified using blast and read depth analysis. Overlapping sequences between the  
1069 contig and the orange reference genome was used to define the edges at which to insert the sequence. The  
1070 sequence was then inserted and overlapping sequences removed, favoring the Alba Contig sequence leading to  
1071 the generation of the Alba reference genome.

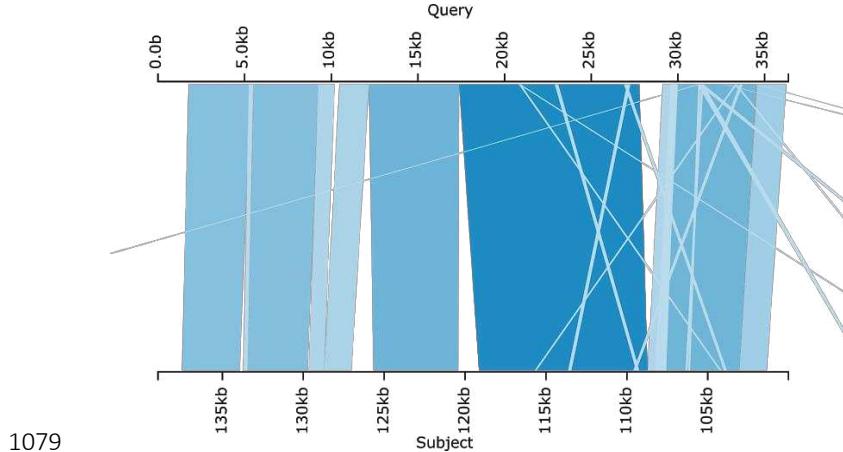


1072

1073 **Supplementary figure 7. Verification of the Alba locus using PCR-based markers.** 8 orange and 8 Alba C.  
1074 *cerytheme* individuals, independent from the GWAS analysis, had DNA extracted and then genotyped for the  
1075 insertion. PCR products were visualized on a 1% agarose gel. The top row shows negative results in orange  
1076 females, and the bottom row shows positive results from alba females.

1077

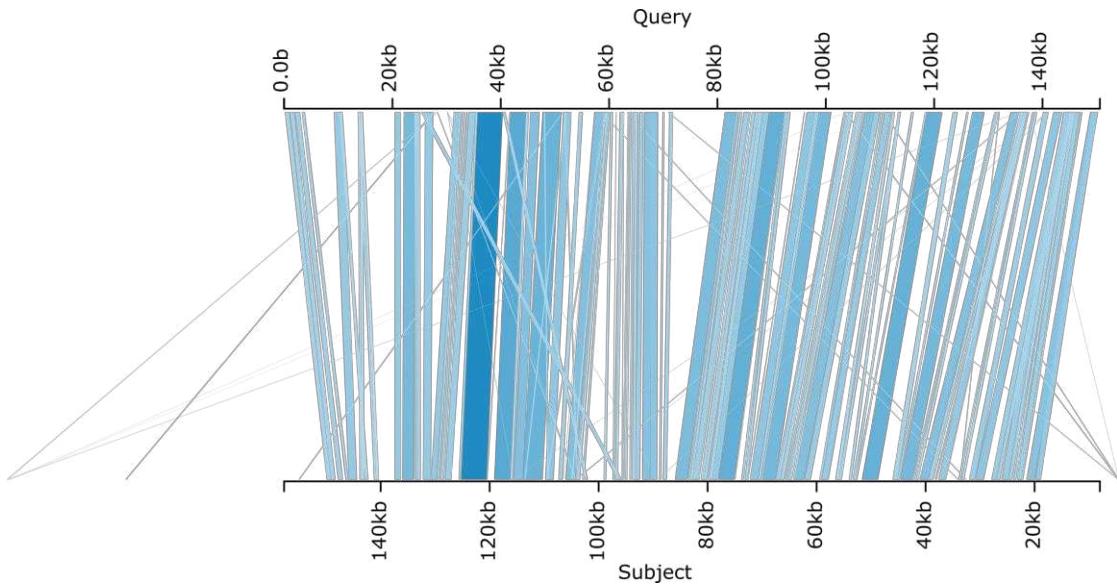
1078



1080 **Supplementary figure 8. Orthology assessment of the *C. crocea alba* scaffold identified in the Chromium 10X**  
1081 **assembly** (query), identified by blasting the *C. eurytheme* BarH1 gene and insertion sequence against the  
1082 assembly, compared to the one found in the *C. crocea* reference genome(subject). Darker blue hue means  
1083 higher similarity.

1084

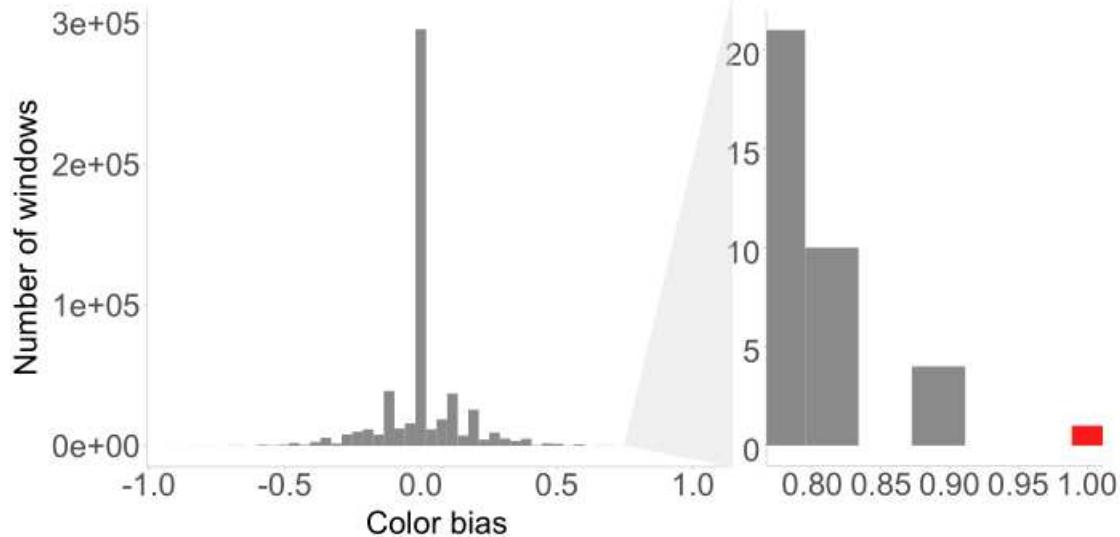
1085



1086

1087 **Supplementary figure 9. Orthology assessment of the Alba insertion in C. nastes.** The alba scaffold, as  
1088 identified using the *BarH1* gene and insertion identified in *C. eurytheme*, from the *C. nastes* Chromium 10X  
1089 assembly(query), aligned against the *Colias crocea* reference genome(subject). Darker blue hue means higher  
1090 similarity

1091



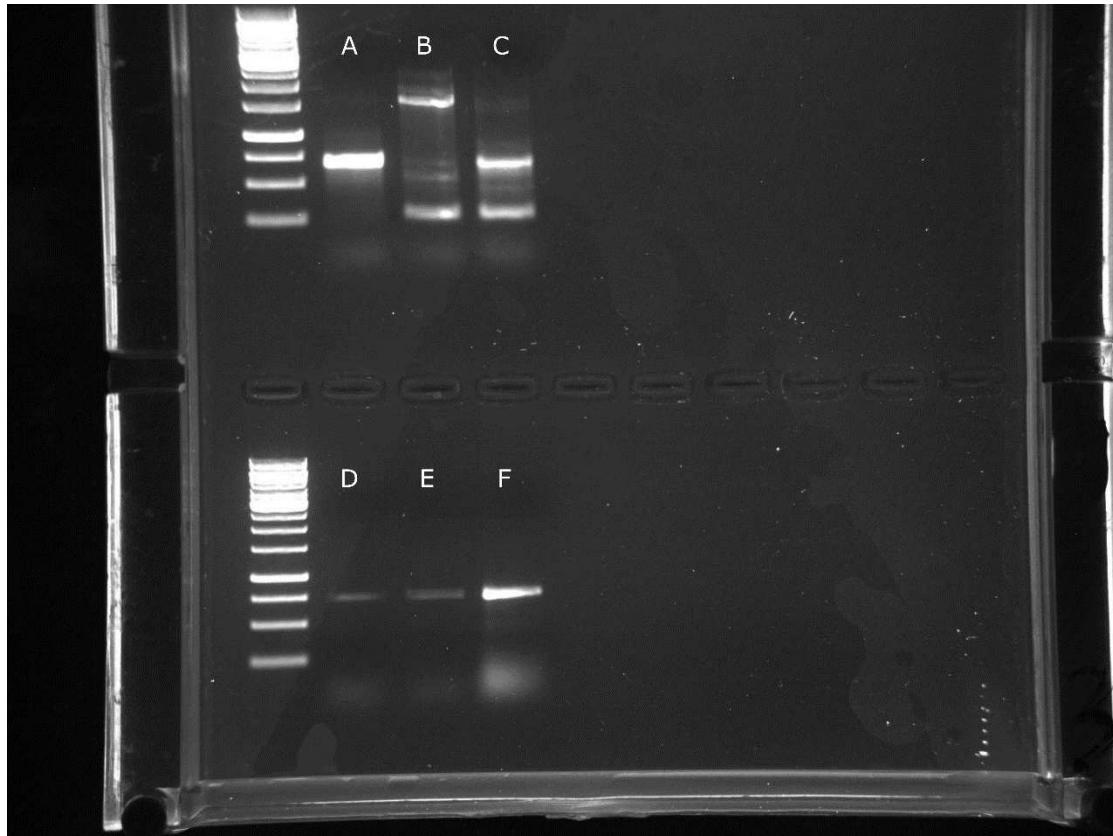
1092 **Supplementary figure 10. Read depth analysis between Orange and Alba species in 600bp windows.** Read  
1093 depth was calculated for each species in 600bp windows when the reads were aligned against the synthetic alba  
1094 reference genome. Each window was then classified as either having reads covering or not if the read coverage  
1095 was less or more than 25% of the mean scaffold depth for the species. A mean was then calculated for all the  
1096 alba species and all the Orange species, and finally we subtracted the orange mean from the Alba beam. Only 1  
1097 window, the conserved alba locus, had zero coverage in the Orange individuals, and complete coverage in the  
1098 Alba. No window showed the opposite pattern.

1099



1100

1101 **Supplementary figure 11. Alba CRE-KO.** Images of both successful “conserved Alba region” CRISPR-KOs with  
1102 their phenotypes: ind. Alba CRE-KO 1 wing (top left), ind. Alba CRE-KO 2 wing (bottom left), eye ind. Alba CRE-KO  
1103 1 (top right), eye ind. Alba CRE-KO 2 (bottom right)



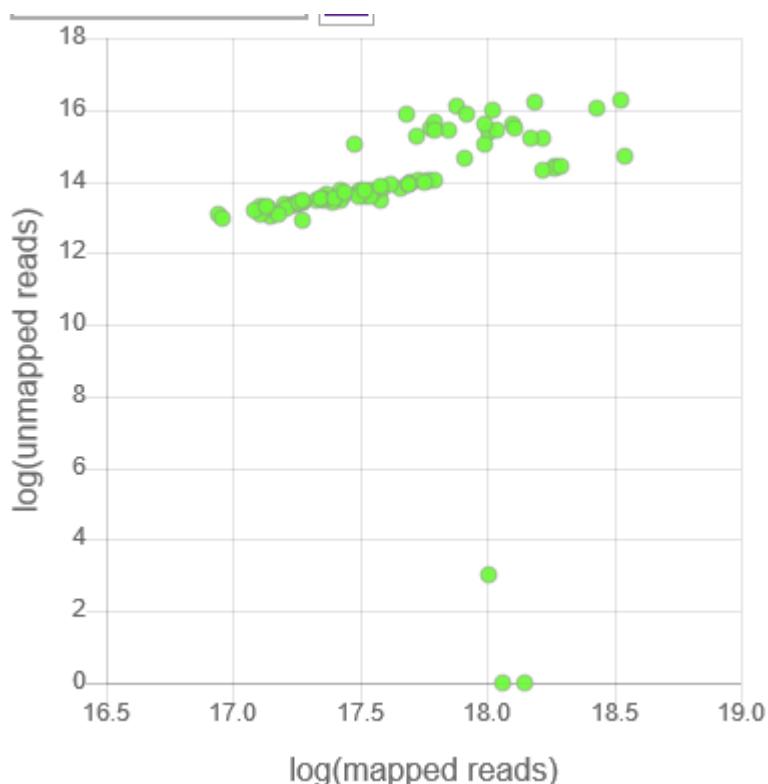
1104

1105 *Supplementary figure 12. PCR validation of CRISPR-induced deletions .*

1106 Gel verifying CRSPR KO as well as alba status for the two successful mutants. From left to right the well show, **A**:  
1107 Alba-CRE WT, **B**: Alba-CRE KO-1, **C**: Alba-CRE KO-2, **D:F**: Alba control validation primer on the sample in the  
1108 column above. Not the variation in band sizes caused by the Cas9 in the two knockout samples.

1109

1110



1111 *Supplementary figure 13. Mappability of short read datasets.* Scatterplot showing counts of mapped reads to  
1112 unmapped reads. Note that the samples with very few unmapped reads had been filtered for mapped reads  
1113 prior to the generation of this figure.

1114 Ahola, Virpi, Rainer Lehtonen, Panu Somervuo, Leena Salmela, Patrik Koskinen, Pasi Rastas, Niko Välimäki, et al.  
1115 2014. 'The Glanville Fritillary Genome Retains an Ancient Karyotype and Reveals Selective Chromosomal Fusions  
1116 in Lepidoptera'. *Nature Communications* 5 (1): 4737. <https://doi.org/10.1038/ncomms5737>.

1117 Borowiec, Marek L. 2016. 'AMAS: A Fast Tool for Alignment Manipulation and Computing of Summary Statistics'.  
1118 *PeerJ* 4 (January). <https://doi.org/10.7717/peerj.1660>.

1119 Bryant, David, Remco Bouckaert, Joseph Felsenstein, Noah A. Rosenberg, and Arindam RoyChoudhury. 2012.  
1120 'Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent  
1121 Analysis'. *Molecular Biology and Evolution* 29 (8): 1917–32. <https://doi.org/10.1093/molbev/mss086>.

1122 Chin, Chen-Shan, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum,  
1123 Christopher Dunn, et al. 2016. 'Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing'.  
1124 *Nature Methods* 13 (12): 1050–54. <https://doi.org/10.1038/nmeth.4035>.

1125 Gligis, Hani Z. 2015. 'Red: An Intelligent, Rapid, Accurate Tool for Detecting Repeats de-Novo on the Genomic  
1126 Scale'. *BMC Bioinformatics* 16 (1): 227. <https://doi.org/10.1186/s12859-015-0654-5>.

1127 Huang, Shengfeng, Mingjing Kang, and Anlong Xu. 2017. 'HaploMerger2: Rebuilding Both Haploid Sub-  
1128 Assemblies from High-Heterozygosity Diploid Genome Assembly'. *Bioinformatics* 33 (16): 2577–79.  
1129 <https://doi.org/10.1093/bioinformatics/btx220>.

1130 Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. 'Graph-  
1131 Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype'. *Nature Biotechnology* 37 (8):  
1132 907–15. <https://doi.org/10.1038/s41587-019-0201-4>.

1133 Li, Heng. 2013. 'Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM'.  
1134 *ArXiv:1303.3997 [q-Bio]*, May. <http://arxiv.org/abs/1303.3997>.

1135 Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and  
1136 Richard Durbin. 2009. 'The Sequence Alignment/Map Format and SAMtools'. *Bioinformatics* 25 (16): 2078–79.  
1137 <https://doi.org/10.1093/bioinformatics/btp352>.

1138 Li, Weizhong, and Adam Godzik. 2006. 'Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of  
1139 Protein or Nucleotide Sequences'. *Bioinformatics* 22 (13): 1658–59.  
1140 <https://doi.org/10.1093/bioinformatics/btl158>.

1141 Malinsky, Milan, Michael Matschiner, and Hannes Svardal. 2020. 'Dsuite - Fast D-Statistics and Related  
1142 Admixture Evidence from VCF Files'. *BioRxiv*, February, 634477. <https://doi.org/10.1101/634477>.

1143 O'Neil, Shawn T., Jason DK Dzurisin, Rory D. Carmichael, Neil F. Lobo, Scott J. Emrich, and Jessica J. Hellmann.  
1144 2010. 'Population-Level Transcriptome Sequencing of Nonmodel Organisms Erynnis Propertius and Papilio  
1145 Zelicaon'. *BMC Genomics* 11 (1): 1–15. <https://doi.org/10.1186/1471-2164-11-310>.

1146 Page, Justin T., Zachary S. Liechty, Mark D. Huynh, and Joshua A. Udall. 2014. 'BamBam: Genome Sequence  
1147 Analysis Tools for Biologists'. *BMC Research Notes* 7 (1): 829. <https://doi.org/10.1186/1756-0500-7-829>.

1148 Rambaut, Andrew, and AJ Drummond. 2012. *FigTree Version 1.4. 0*.

1149 Rastas, Pasi. 2017. 'Lep-MAP3: Robust Linkage Mapping Even for Low-Coverage Whole Genome Sequencing  
1150 Data'. *Bioinformatics* 33 (23): 3726–32. <https://doi.org/10.1093/bioinformatics/btx494>.

1151 Rodriguez-Caro, Luis, Jennifer Fenner, Caleb Benson, Steven M. Van Belleghem, and Brian A. Counterman. 2020.  
1152 'Genome Assembly of the Dogface Butterfly Zerene Cesonia'. *Genome Biology and Evolution* 12 (1): 3580–85.  
1153 <https://doi.org/10.1093/gbe/evz254>.

1154 Smith, Stephen A., Michael J. Moore, Joseph W. Brown, and Ya Yang. 2015. 'Analysis of Phylogenomic Datasets  
1155 Reveals Conflict, Concordance, and Gene Duplications with Examples from Animals and Plants'. *BMC*  
1156 *Evolutionary Biology* 15 (1): 150. <https://doi.org/10.1186/s12862-015-0423-0>.

1157 Stange, Madlen, Marcelo R Sánchez-Villagra, Walter Salzburger, and Michael Matschiner. 2018. 'Bayesian  
1158 Divergence-Time Estimation with Genome-Wide Single-Nucleotide Polymorphism Data of Sea Catfishes (Ariidae)  
1159 Supports Miocene Closure of the Panamanian Isthmus'. *Systematic Biology* 67 (4): 681–99.  
1160 <https://doi.org/10.1093/sysbio/syy006>.

1161 Trapnell, Cole, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L.  
1162 Salzberg, Barbara J. Wold, and Lior Pachter. 2010. 'Transcript Assembly and Quantification by RNA-Seq Reveals  
1163 Unannotated Transcripts and Isoform Switching during Cell Differentiation'. *Nature Biotechnology* 28 (5): 511–  
1164 15. <https://doi.org/10.1038/nbt.1621>.

1165 Wang, Baiqing, and Adam H. Porter. 2004. 'An AFLP-Based Interspecific Linkage Map of Sympatric, Hybridizing  
1166 Colias Butterflies'. *Genetics* 168 (1): 215–25. <https://doi.org/10.1534/genetics.104.028118>.

1167 Wang, Shi, Eli Meyer, John K. McKay, and Mikhail V. Matz. 2012. '2b-RAD: A Simple and Flexible Method for  
1168 Genome-Wide Genotyping'. *Nature Methods* 9 (8): 808–10. <https://doi.org/10.1038/nmeth.2023>.

1169