



This is a repository copy of *The gene cortex controls mimicry and crypsis in butterflies and moths*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/100074/>

Version: Accepted Version

Article:

Nadeau, N.J. orcid.org/0000-0002-9319-921X, Pardo-Diaz, C., Whibley, A. et al. (16 more authors) (2016) The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature*, 534 (7605). pp. 106-110. ISSN 0028-0836

<https://doi.org/10.1038/nature17961>

Reuse

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

1 **The gene cortex controls mimicry and crypsis in butterflies and moths**

2 Nicola J. Nadeau^{1,2}, Carolina Pardo-Diaz³, Annabel Whibley^{4,5}, Megan Supple^{2,6}, Suzanne V.
3 Saenko⁴, Richard W. R. Wallbank^{2,7}, Grace C. Wu⁸, Luana Maroja⁹, Laura Ferguson¹⁰,
4 Joseph J. Hanly^{2,7}, Heather Hines¹¹, Camilo Salazar³, Richard Merrill^{2,7}, Andrea Dowling¹²,
5 Richard ffrench-Constant¹², Violaine Llaurens⁴, Mathieu Joron⁴, W. Owen McMillan², Chris
6 D. Jiggins^{7,2}

7 ¹Department of Animal and Plant Sciences, University of Sheffield, UK; ²Smithsonian
8 Tropical Research Institute, Panama; ³Biology Program, Faculty of Natural Sciences and
9 Mathematics. Universidad del Rosario. Cra. 24 No 63C-69, Bogotá D.C., 111221, Colombia;
10 ⁴Institut de Systématique, Evolution et Biodiversité (UMR 7205 CNRS, MNHN, UPMC,
11 EPHE, Sorbonne Université), Museum National d'Histoire Naturelle, CP50, 57 rue Cuvier,
12 75005 PARIS, France; ⁵Cell and Developmental Biology, John Innes Centre, Norwich, UK,
13 NR4 7UH, ⁶The Australian National University, ACT, Australia; ⁷Department of Zoology,
14 University of Cambridge, UK; ⁸Energy and Resources Group, University of California at
15 Berkeley, CA, USA; ⁹Department of Biology, Williams College, MA, USA; ¹⁰Department
16 of Zoology, University of Oxford, UK; ¹¹ Penn State University, 517 Mueller, University
17 Park, PA 16802; ¹²School of Biosciences, University of Exeter in Cornwall, Penryn, UK
18 TR10 9EZ
19

20 The wing patterns of butterflies and moths (Lepidoptera) are diverse and striking examples of
21 evolutionary diversification by natural selection^{1,2}. Lepidopteran wing colour patterns are a
22 key innovation, consisting of arrays of coloured scales. We still lack a general understanding
23 of how these patterns are controlled and if there is any commonality across the 160,000 moth
24 and 17,000 butterfly species. Here, we identify a gene, *cortex*, through fine-scale mapping
25 using population genomics and gene expression analyses, which regulates pattern switches in
26 multiple species across the mimetic radiation in *Heliconius* butterflies. *cortex* belongs to a
27 fast evolving subfamily of the otherwise highly conserved *fizzy* family of cell cycle
28 regulators³, suggesting that it most likely regulates pigmentation patterning through
29 regulation of scale cell development. In parallel with findings in the peppered moth (*Biston*
30 *betularia*)⁴, our results suggest that this mechanism is common within Lepidoptera and that
31 *cortex* has become a major target for natural selection acting on colour and pattern variation
32 in this group of insects.

33

34 In *Heliconius*, there is a major effect locus, *Yb*, that controls a diversity of colour pattern
35 elements across the genus. It is the only locus in *Heliconius* that regulates all scale types and
36 colours, including the diversity of white and yellow pattern elements in the two co-mimics *H.*
37 *melpomene* (*Hm*) and *H. erato* (*He*), but also whole wing variation in black, yellow, white,
38 and orange/red elements in *H. numata* (*Hn*)⁵⁻⁷. In addition, genetic variation underlying the
39 Bigeye wing pattern mutation in *Bicyclus anynana*, melanism in the peppered moth, *Biston*
40 *betularia*, and melanism and patterning differences in the silkworm, *Bombyx mori*, have all
41 been localised to homologous genomic regions⁸⁻¹⁰ (Fig 1). Therefore, this genomic region
42 appears to contain one or more genes that act as major regulators of wing pigmentation and
43 patterning across the Lepidoptera.

44 Previous mapping of this locus in He, Hm and Hn identified a genomic interval of ~1Mb¹¹⁻¹³
45 (Extended Data Table 1), which also overlaps with the 1.4Mb region containing the
46 carbonaria locus in *B. betularia*⁹ and a 100bp non-coding region containing the Ws mutation
47 in *B. mori*¹⁰ (Fig 1). We took a population genomics approach to identify single nucleotide
48 polymorphisms (SNPs) most strongly associated with phenotypic variation within the ~1Mb
49 Heliconius interval. The diversity of wing patterning in Heliconius arises from divergence at
50 wing pattern loci⁷, while convergent patterns generally involve the same loci and sometimes
51 even the same alleles¹⁴⁻¹⁶. We used this pattern of divergence and sharing to identify SNPs
52 associated with colour pattern elements across many individuals from a wide diversity of
53 colour pattern phenotypes (Fig 2).

54 In three separate Heliconius species, our analysis consistently implicated the gene cortex as
55 being involved in adaptive differences in wing colour pattern. In He the strongest associations
56 with the presence of a yellow hindwing bar were centred around the genomic region
57 containing cortex (Fig 2A). We identified 108 SNPs that were fixed for one allele in He
58 favorinus, and fixed for the alternative allele in all individuals lacking the yellow bar, the
59 majority of which were in introns of cortex (Extended Data Table 2). 15 SNPs showed a
60 similar fixed pattern for He demophoon, which also has a yellow bar. These were non-
61 overlapping with those in He favorinus, consistent with the hypothesis that this phenotype
62 evolved independently in the two disjunct populations¹⁷.

63 Previous work has suggested that alleles at the Yb locus are shared between Hm and the
64 closely related species *H. timareta*, and also the more distantly related species *H. elevatus*,
65 resulting in mimicry between these species¹⁸. Across these species, the strongest associations
66 with the yellow hindwing bar phenotype were again found at cortex (Fig 2D, Extended Data
67 Fig 1A and Table 3). Similarly, the strongest associations with the yellow forewing band
68 were found around the 5' UTRs of cortex and gene HM00036, an orthologue of *D.*

69 melanogaster washout gene. A single SNP ~17kb upstream of cortex (the closest gene) was
70 perfectly associated with the yellow forewing band across all Hm, *H. timareta* and *H.*
71 *elevatus* individuals (Extended Data Fig 1A, Fig 2 and Table 3). We found no fixed coding
72 sequence variants at cortex in a larger sample (43-61 individuals) of *Hm aglaope* and *Hm*
73 *amaryllis* (Extended Data Figure 3, Supplementary Information), which differ in Yb
74 controlled phenotypes¹⁹, suggesting that functional variants are likely to be regulatory rather
75 than coding. We found extensive transposable element variation around cortex but it is
76 unclear if any of these associate with phenotype (Extended Data Figure 3 and Table 4;
77 Supplementary Information).

78 Finally, in Hn large inversions at the P supergene locus (Fig 1) are associated with different
79 morphs¹³. There is a steep increase in genotype-by-phenotype association at the breakpoint of
80 inversion 1, consistent with the role of these inversions in reducing recombination (Fig 2E).
81 However, the *bicoloratus* morph can recombine with all other morphs across one or the other
82 inversion, permitting finer-scale association mapping of this region. As in He and Hm, this
83 analysis showed a narrow region of associated SNPs corresponding exactly to the cortex gene
84 (Fig 2E), again with the majority of SNPs in introns (Extended Data Table 2). This associated
85 region does not correspond to any other known genomic feature, such as an inversion or
86 inversion breakpoint.

87 To determine whether sequence variants around cortex were regulating its expression we
88 investigated gene expression across the Yb locus. We used a custom designed microarray
89 including probes from all predicted genes in the *H. melpomene* genome¹⁸, as well as probes
90 tiled across the central portion of the Yb locus, focussing on two naturally hybridising Hm
91 races (*plesseni* and *malleti*) that differ in Yb controlled phenotypes⁷. cortex was the only gene
92 across the entire interval to show significant expression differences both between races with
93 different wing patterns and between wing sections with different pattern elements (Fig 3).

94 This finding was reinforced in the tiled probe set, where we observed strong differences in
95 expression of cortex exons and introns but few differences outside this region (Extended Data
96 Table 2). cortex expression was higher in *Hm malleti* than *Hm plesseni* in all three wing
97 sections used (but not eyes) (Fig 3C; Extended Data Fig 4C). When different wing sections
98 were compared within each race, cortex expression in *Hm malleti* was higher in the distal
99 section that contains the Yb controlled yellow forewing band, consistent with cortex
100 producing this band. In contrast, *Hm plesseni*, which lacks the yellow band, had higher cortex
101 expression in the proximal forewing section (Fig 3F; Extended Data Fig 4J). Expression
102 differences were found only in day 1 and day 3 pupal wings rather than day 5 or day 7
103 (Extended Data Fig 4), similar to the pattern observed previously for the transcription factor
104 *optix*²⁰.

105 Differential expression was not confined to the exons of cortex; the majority of differentially
106 expressed probes in the tiling array corresponded to cortex introns (Fig 3). This does not
107 appear to be due to transposable element variation (Extended Data Table 2), but may be due
108 to elevated background transcription and unidentified splice variants. RT-PCR revealed a
109 diversity of splice variants (Extended Data Fig 5), and sequenced products revealed 8 non-
110 constitutive exons and 6 variable donor/acceptor sites, but this was not exhaustive
111 (Supplementary Information). We cannot rule out the possibility that some of the
112 differentially expressed intronic regions could be distinct non-coding RNAs. However, qRT-
113 PCR in other hybridising races with divergent Yb alleles (*aglaope/amaryllis* and
114 *rosina/melpomene*) also identified expression differences at cortex and allele-specific splicing
115 differences between both pairs of races (Extended Data Figs 1 and 5, Supplementary
116 Information).

117 Finally, in situ hybridisation of cortex in final instar larval hindwing discs showed expression
118 in wing regions fated to become black in the adult wing, most strikingly in their

119 correspondence to the black patterns on adult Hn wings (Fig 4). In contrast, the array results
120 from pupal wings were suggestive of higher expression in non-melanic regions. This may
121 suggest that cortex is upregulated at different time-points in wing regions fated to become
122 different colours.

123 Overall, cortex shows significant differential expression and is the only gene in the candidate
124 region to be consistently differentially expressed in multiple race comparisons and between
125 differently patterned wing regions. Coupled with the strong genotype-by-phenotype
126 associations across multiple independent lineages (Extended Data Table 1), this strongly
127 implicates cortex as a major regulator of colour and pattern. However, we have not excluded
128 the possibility that other genes in this region also influence pigmentation patterning. A
129 prominent role for cortex is also supported by studies in other taxa; our identification of
130 distant 5' untranslated exons of cortex (Supplementary Information) suggests that the 100bp
131 interval containing the Ws mutation in *B. mori* is likely to be within an intron of cortex and
132 not in intergenic space as previously thought¹⁰. In addition, fine-mapping and gene
133 expression also implicate cortex as controlling melanism in the peppered moth⁴.

134 It seems likely that cortex controls pigmentation patterning through control of scale cell
135 development. The cortex gene falls in an insect specific lineage within the fizzy/CDC20
136 family of cell cycle regulators (Extended Data Fig 6A). The phylogenetic tree of the gene
137 family highlighted three major orthologous groups, two of which have highly conserved
138 functions in cell cycle regulation mediated through interaction with the anaphase promoting
139 complex/cyclosome (APC/C)^{3,21}. The third group, cortex, is evolving rapidly, with low amino
140 acid identity between *D. melanogaster* and Hm cortex (14.1%), contrasting with much higher
141 identities for orthologues between these species in the other two groups (fzy, 47.8% and
142 rap/fzr, 47.2%, Extended Data Fig 6A). *Drosophila melanogaster* cortex acts through a

143 similar mechanism to *fzy* in order to control meiosis in the female germ line²²⁻²⁴. Hm cortex
144 also has some conservation of the fizzy family C-box and IR elements (Supplementary
145 Information) that mediate binding to the APC/C²³, suggesting that it may have retained a cell
146 cycle function, although we found that expressing Hm cortex in *D. melanogaster* wings
147 produced no detectable effect (Extended Data Fig 6, Supplementary Information).

148 Previously identified butterfly wing patterning genes have been transcription factors or
149 signalling molecules^{20,25}. Developmental rate has long been thought to play a role in
150 lepidopteran patterning^{26,27}, but cortex was not a likely a priori candidate, because its
151 *Drosophila* orthologue has a highly specific function in meiosis²³. The recruitment of cortex
152 to wing patterning appears to have occurred before the major diversification of the
153 Lepidoptera and this gene has repeatedly been targeted by natural selection^{1,7,9,28} to generate
154 both cryptic⁴ and aposematic patterns.

155 **References**

- 156 1. Cook, L. M., Grant, B. S., Saccheri, I. J. & Mallet, J. Selective bird predation on the
157 peppered moth: the last experiment of Michael Majerus. *Biol. Lett.* **8**, 609–612 (2012).
- 158 2. Jiggins, C. D. Ecological Speciation in Mimetic Butterflies. *BioScience* **58**, 541–548
159 (2008).
- 160 3. Dawson, I. A., Roth, S. & Artavanis-Tsakonas, S. The *Drosophila* Cell Cycle Gene fizzy
161 Is Required for Normal Degradation of Cyclins A and B during Mitosis and Has
162 Homology to the CDC20 Gene of *Saccharomyces cerevisiae*. *J. Cell Biol.* **129**, 725–737
163 (1995).
- 164 4. Van't Hof, A. E. et al. The industrial melanism mutation in British peppered moths is a
165 transposable element. *Nature* **This issue**,

- 166 5. Joron, M. et al. A Conserved Supergene Locus Controls Colour Pattern Diversity in
167 Heliconius Butterflies. *PLoS Biol.* **4**, (2006).
- 168 6. Sheppard, P. M., Turner, J. R. G., Brown, K. S., Benson, W. W. & Singer, M. C.
169 Genetics and the Evolution of Mullerian Mimicry in Heliconius Butterflies. *Philos.*
170 *Trans. R. Soc. Lond. B. Biol. Sci.* **308**, 433–610 (1985).
- 171 7. Nadeau, N. J. et al. Population genomics of parallel hybrid zones in the mimetic
172 butterflies, *H. melpomene* and *H. erato*. *Genome Res.* **24**, 1316–1333 (2014).
- 173 8. Beldade, P., Saenko, S. V., Pul, N. & Long, A. D. A Gene-Based Linkage Map for
174 *Bicyclus anynana* Butterflies Allows for a Comprehensive Analysis of Synteny with the
175 Lepidopteran Reference Genome. *PLoS Genet* **5**, e1000366 (2009).
- 176 9. van't Hof, A. E., Edmonds, N., Dalíková, M., Marec, F. & Saccheri, I. J. Industrial
177 Melanism in British Peppered Moths Has a Singular and Recent Mutational Origin.
178 *Science* **332**, 958–960 (2011).
- 179 10. Ito, K. et al. Mapping and recombination analysis of two moth colour mutations, Black
180 moth and Wild wing spot, in the silkworm *Bombyx mori*. *Heredity* (2015).
181 doi:10.1038/hdy.2015.69
- 182 11. Counterman, B. A. et al. Genomic Hotspots for Adaptation: The Population Genetics of
183 Müllerian Mimicry in *Heliconius erato*. *PLoS Genet.* **6**, e1000796 (2010).
- 184 12. Ferguson, L. et al. Characterization of a hotspot for mimicry: assembly of a butterfly
185 wing transcriptome to genomic sequence at the *HmYb/Sb* locus. *Mol. Ecol.* **19**, 240–254
186 (2010).
- 187 13. Joron, M. et al. Chromosomal rearrangements maintain a polymorphic supergene
188 controlling butterfly mimicry. *Nature* **477**, 203–206 (2011).
- 189 14. Hines, H. M. et al. Wing patterning gene redefines the mimetic history of *Heliconius*
190 butterflies. *Proc. Natl. Acad. Sci.* **108**, 19666–19671 (2011).

- 191 15. Pardo-Diaz, C. et al. Adaptive Introgression across Species Boundaries in Heliconius
192 Butterflies. *PLoS Genet* **8**, e1002752 (2012).
- 193 16. Wallbank, R. W. R. et al. Evolutionary Novelty in a Butterfly Wing Pattern through
194 Enhancer Shuffling. *PLoS Biol* **14**, e1002353 (2016).
- 195 17. Maroja, L. S., Alschuler, R., McMillan, W. O. & Jiggins, C. D. Partial Complementarity
196 of the Mimetic Yellow Bar Phenotype in Heliconius Butterflies. *PLoS ONE* **7**, e48627
197 (2012).
- 198 18. The Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of
199 mimicry adaptations among species. *Nature* **487**, 94–98 (2012).
- 200 19. Mallet, J. The Genetics of Warning Colour in Peruvian Hybrid Zones of *Heliconius erato*
201 and *H. melpomene*. *Proc. R. Soc. Lond. B Biol. Sci.* **236**, 163–185 (1989).
- 202 20. Reed, R. D. et al. optix Drives the Repeated Convergent Evolution of Butterfly Wing
203 Pattern Mimicry. *Science* **333**, 1137–1141 (2011).
- 204 21. Barford, D. Structural insights into anaphase-promoting complex function and
205 mechanism. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 3605–3624 (2011).
- 206 22. Chu, T., Henrion, G., Haegeli, V. & Strickland, S. Cortex, a *Drosophila* gene required to
207 complete oocyte meiosis, is a member of the Cdc20/fizzy protein family. *genesis* **29**,
208 141–152 (2001).
- 209 23. Pesin, J. A. & Orr-Weaver, T. L. Developmental Role and Regulation of cortex, a
210 Meiosis-Specific Anaphase-Promoting Complex/Cyclosome Activator. *PLoS Genet* **3**,
211 e202 (2007).
- 212 24. Swan, A. & Schüpbach, T. The Cdc20/Cdh1-related protein, Cort, cooperates with
213 Cdc20/Fzy in cyclin destruction and anaphase progression in meiosis I and II in
214 *Drosophila*. *Dev. Camb. Engl.* **134**, 891–899 (2007).

- 215 25. Martin, A. et al. Diversification of complex butterfly wing patterns by repeated
216 regulatory evolution of a Wnt ligand. *Proc. Natl. Acad. Sci.* **109**, 12632–12637 (2012).
- 217 26. Koch, P. B., Lorenz, U., Brakefield, P. M. & ffrench-Constant, R. H. Butterfly wing
218 pattern mutants: developmental heterochrony and co-ordinately regulated phenotypes.
219 *Dev. Genes Evol.* **210**, 536–544 (2000).
- 220 27. Gilbert, L. E., Forrest, H. S., Schultz, T. D. & Harvey, D. J. Correlations of ultrastructure
221 and pigmentation suggest how genes control development of wing scales of *Heliconius*
222 butterflies. *J. Res. Lepidoptera* **26**, 141–160 (1988).
- 223 28. Mallet, J. & Barton, N. H. Strong Natural Selection in a Warning-Color Hybrid Zone.
224 *Evolution* **43**, 421–431 (1989).
- 225 29. Wahlberg, N., Wheat, C. W. & Peña, C. Timing and Patterns in the Taxonomic
226 Diversification of Lepidoptera (Butterflies and Moths). *PLoS ONE* **8**, e80875 (2013).
- 227 30. Surridge, A. et al. Characterisation and expression of microRNAs in developing wings of
228 the neotropical butterfly *Heliconius melpomene*. *BMC Genomics* **12**, 62 (2011).

229

230 **Supplementary Information** is linked to the online version of the paper at

231 www.nature.com/nature.

232 **Acknowledgements** We thank Christopher Sasaki, Clemson University, for assembly of the
233 He BACs. Moises Abanto and Adriana Tapia assisted with raising butterflies. Thanks to
234 Mathieu Chouteau, Jake Morris and Kanchon Dasmahapatra for providing larvae for in situ
235 hybridisations. Anna Morrison, Robert Tetley, Sarah Carl and Hanna Wegener assisted with
236 lab work at the University of Cambridge. Simon Baxter made the Hm fosmid libraries. We
237 thank the governments of Colombia, Ecuador, Panama and Peru for permission to collect
238 butterflies. This work was funded by a Leverhulme Trust award and BBSRC grant
239 (H01439X/1) to CDJ, NSF grants (DEB 1257689, IOS 1052541) to WOM, an ERC starting

240 grant to MJ and a French National Agency for Research (ANR) grant to VL (ANR-13-JSV7-
241 0003-01). NJN is funded by a NERC fellowship (NE/K008498/1).

242

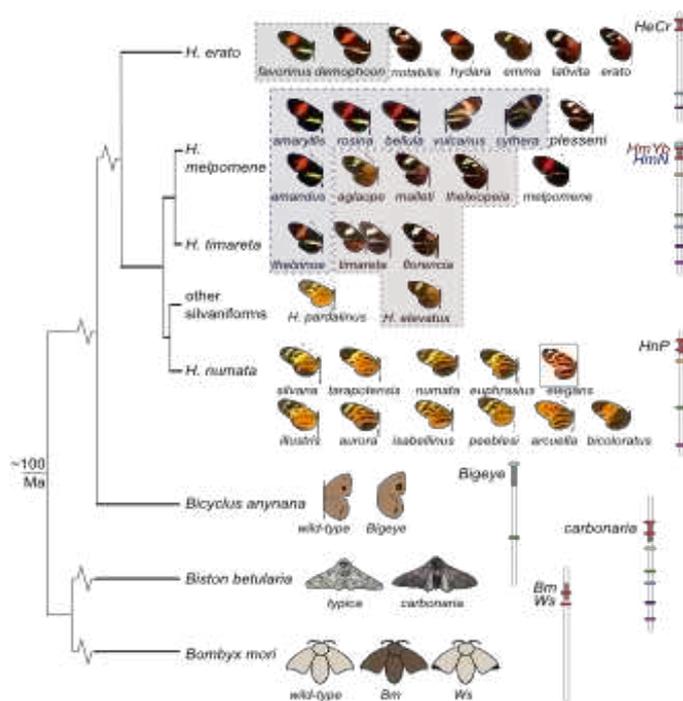
243 **Author Contributions** NJN performed the association analyses, 5' RACE, RT-PCR, qRT-
244 PCR and prepared the manuscript. NJN and CDJ co-ordinated the research. CP-D performed
245 and analysed the microarray and RNAseq experiments. AW performed the Hn association
246 analysis. MS assembled and annotated the HeCr BAC reference and the He alignments. SVS
247 performed in situ hybridizations. RWRW performed the transgenic experiments and analysis
248 of de novo assembled sequences and fosmids together with JJH. GW and LF initially
249 identified splicing variants of cortex. LM performed crosses between Hm races. HH screened
250 the HeCr BAC library. CS and RM provided samples. AD contributed to the Hm BAC
251 sequencing and annotation. R-fC, MJ, VL, WOM and CDJ are PIs who obtained funding and
252 led the project elements. All authors commented on the manuscript.

253

254 **Author Information** Short read sequence data generated for this study are available from
255 ENA (<http://www.ebi.ac.uk/ena>) under study accession PRJEB8011 and PRJEB12740 (see
256 Supplementary Table 1 for previously published data accessions). The updated Cr contig is
257 deposited in Genbank with accession KC469893. The assembled Hm fosmid sequences are
258 deposited in Genbank with accessions KU514430-KU514438. The microarray data are
259 deposited in GEO with accessions GSM1563402- GSM1563497. Reprints and permissions
260 information is available at www.nature.com/reprints. Correspondence and requests for
261 materials should be addressed to n.nadeau@sheffield.ac.uk or c.jiggins@zoo.cam.ac.uk

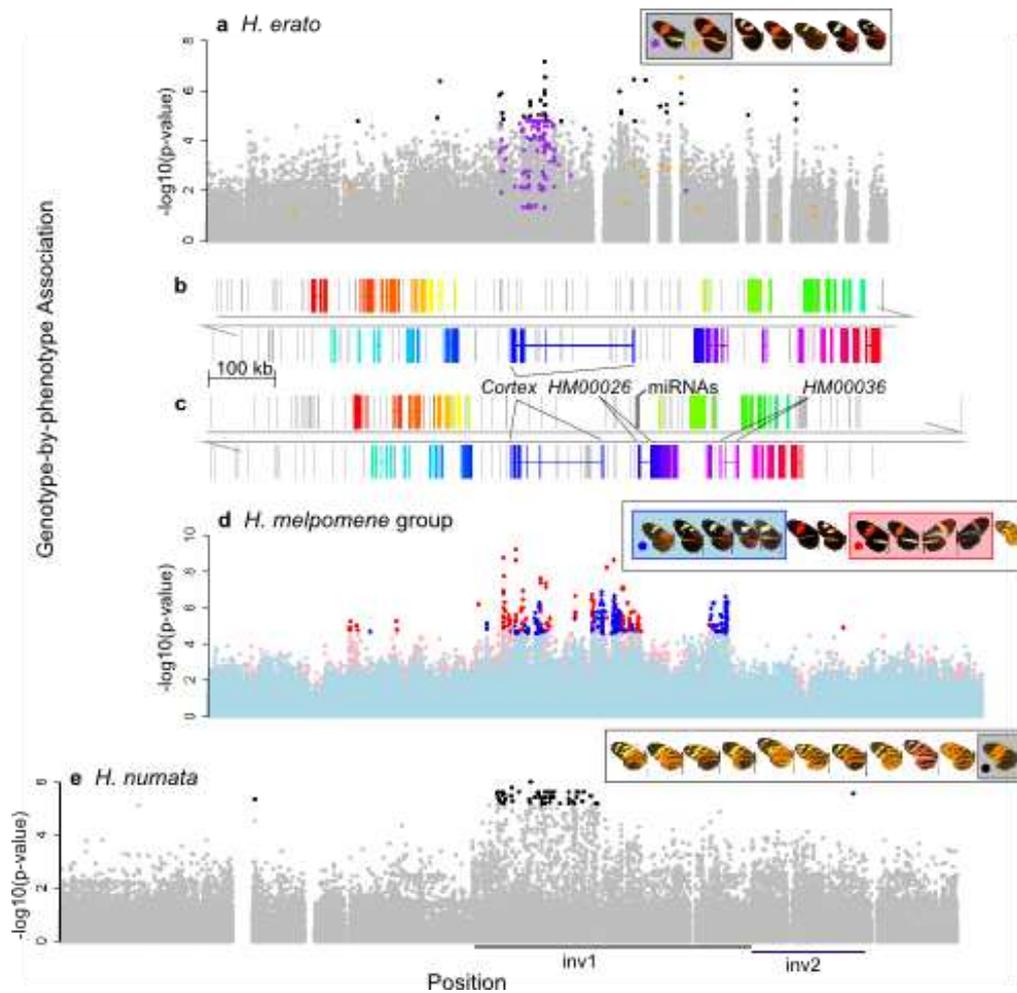
262

263



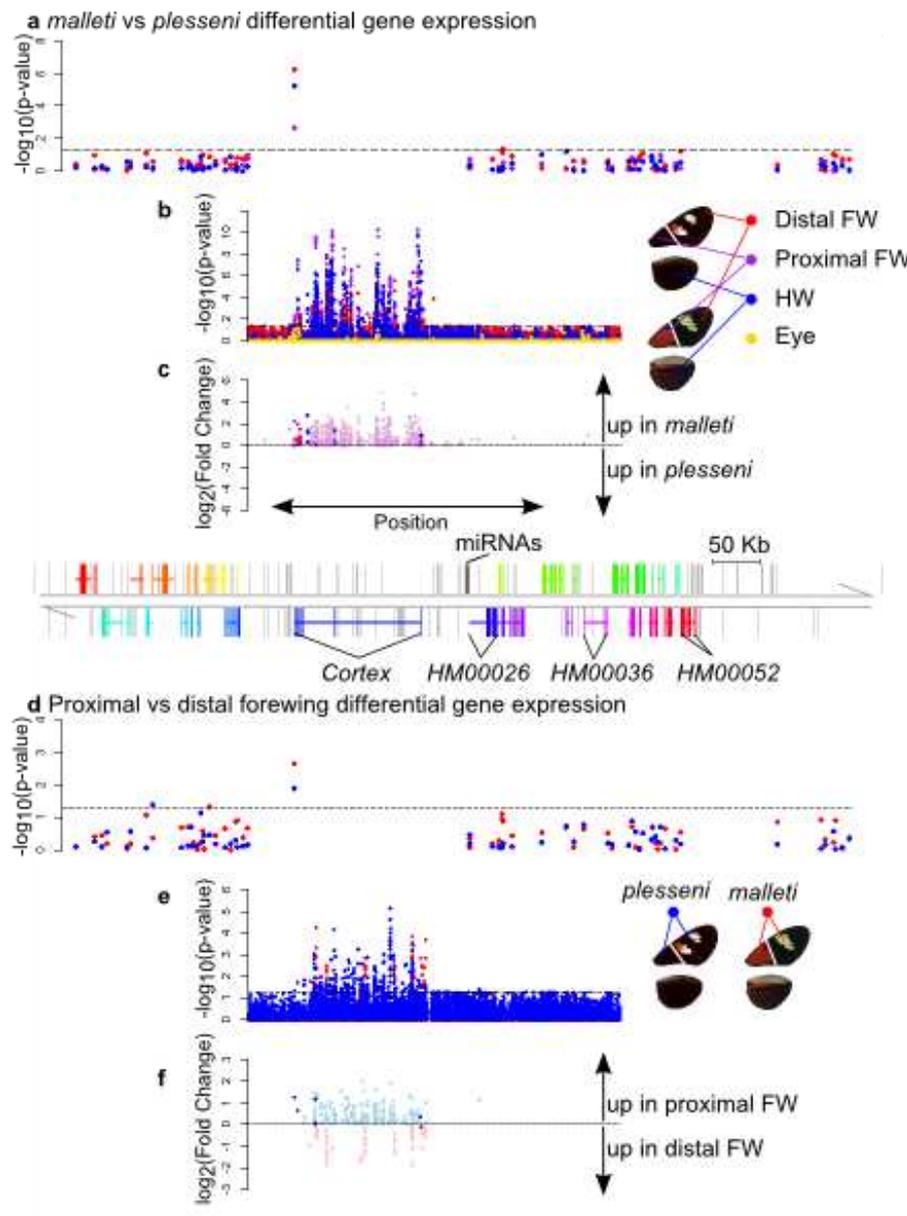
264

265 Figure 1. A homologous genomic region controls a diversity of phenotypes across the
 266 Lepidoptera. Left: phylogenetic relationships²⁹. Right: chromosome maps with colour pattern
 267 intervals in grey, coloured bars represent markers used to assign homology^{5,8-10}, the first and
 268 last genes from Fig 2 shown in red. In He the HeCr locus controls the yellow hind-wing bar
 269 phenotype (grey boxed races). In Hm it controls both the yellow hind-wing bar (HmYb, pink
 270 box) and the yellow forewing band (HmN, blue box). In Hn it modulates black, yellow and
 271 orange elements on both wings (HnP), producing phenotypes that mimic butterflies in the
 272 genus Melinaea. Morphs/races of Heliconius species included in this study are shown with
 273 names.



274

275 Figure 2. Association analyses across the genomic region known to contain major colour
 276 pattern loci in *Heliconius*. A) Association in He with the yellow hind-wing bar (n=45).
 277 Coloured SNPs are fixed for a unique state in He demophon (orange) or He favorinus
 278 (purple). B) Genes in He with direct homologs in Hm. Genes are in different colours with
 279 exons (coding and UTRs) connected by a line. Grey bars are transposable elements. C) Hm
 280 genes and transposable elements: colours correspond to homologous He genes; MicroRNAs³⁰
 281 in black. D) Association in the Hm/timareta/silvaniform group with the yellow hind-wing bar
 282 (red) and yellow forewing band (blue) (n=49). E) Association in Hn with the bicoloratus
 283 morph (n=26); inversion positions¹³ shown below. In all cases black/dark coloured points are
 284 above the strongest associations found outside the colour pattern scaffolds (He $p=1.63e-05$;
 285 Hm $p=2.03e-05$ and $p=2.58e-05$; Hn $p=6.81e-06$).

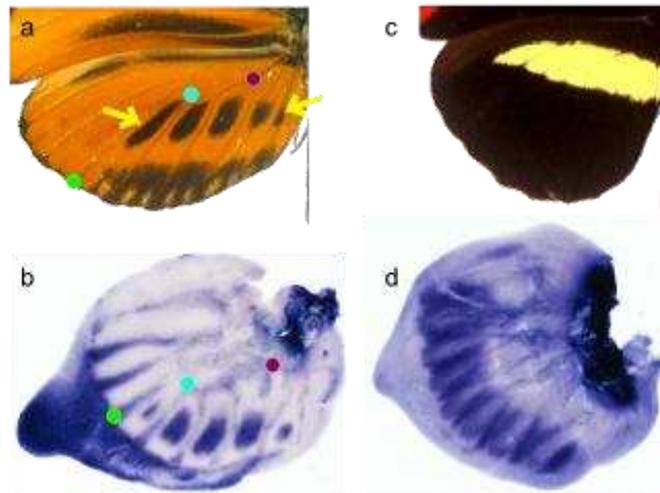


286

287 Figure 3. Differential gene expression across the genomic region known to contain major
 288 colour pattern loci in *Heliconius melpomene*. Expression differences in day 3 pupae, for all
 289 genes in the Yb interval (A,D) and tiling probes spanning the central portion of the interval
 290 (B,C,E,F). Expression is compared between races for each wing region (A,B,C) and between
 291 proximal and distal forewing sections for each race (D,E,F). C and F: magnitude and
 292 direction of expression difference (\log_2 fold-change) for tiling probes showing significant
 293 differences ($p \leq 0.05$); probes in known cortex exons shown in dark colours. Gene HM00052

294 was differentially expressed between other races in RNA sequence data (Supplementary
 295 Information) but is not differentially expressed here.

296



297

298 Figure 4. In situ hybridisations of cortex in hind-wings of final instar larvae. B) Hn
 299 tarapotensis; adult wing shown in A, coloured points indicate landmarks, yellow arrows
 300 highlight adult pattern elements corresponding to the cortex staining. D) Hm rosina; adult
 301 wing shown in C, staining patterns in other Hm races (meriana and aglaope) appeared
 302 similar. The probe used was complementary to the cortex isoform with the longest open
 303 reading frame (also the most common, Supplementary Information).

304

305 **Methods**

306 **He Cr reference**

307 Cr is the homologue of Yb in He (Fig 1). An existing reference for this region was available
 308 in 3 pieces (467,734bp, 114,741bp and 161,149bp, GenBank: KC469893.1)³¹. We screened
 309 the same BAC library used previously^{11,31} using described procedures¹¹ with probes designed

310 to the ends of the existing BAC sequences and the HmYb BAC reference sequence. Two
311 BACs (04B01 and 10B14) were identified as spanning one of the gaps and sequenced using
312 Illumina 2x250 bp paired-end reads collected on the Illumina MiSeq. The raw reads were
313 screened to remove vector and *E. coli* bases. The first 50k read pairs were taken for each
314 BAC and assembled individually with the Phrap³² software and manually edited with
315 *consed*³³. Contigs with discordant read pairs were manually broken and properly merged
316 using concordant read data. Gaps between contig ends were filled using an in-house
317 finishing technique where the terminal 200bp of the contig ends were extracted and queried
318 against the unused read data for spanning pairs, which were added using the
319 *addSolexaReads.perl* script in the *consed* package. Finally, a single reference contig was
320 generated by identifying and merging overlapping regions of the two consensus BAC
321 sequences.

322 In order to fill the remaining gap (between positions 800,387 and 848,446) we used the
323 overhanging ends to search the scaffolds from a preliminary He genome assembly of five
324 Illumina paired end libraries with different insert sizes (250, 500, 800, 4300 and 6500bp)
325 from two related He demophon individuals. We identified two scaffolds (*scf1869* and
326 *scf1510*) that overlapped and spanned the gap (using 12,257bp of the first scaffold and
327 35,803bp of the second).

328 The final contig was 1,009,595bp in length of which 2,281bp were unknown (N's). The HeCr
329 assembly was verified by aligning to the HmYb genome scaffold (HE667780) with *mummer*
330 and *blast*. The HeCr contig was annotated as described previously³², with some minor
331 modifications. Briefly this involved first generating a reference based transcriptome assembly
332 with existing *H. erato* RNA-seq wing tissue (GenBank accession SRA060220). We used
333 *Trimmomatic*³⁴ (v0.22), and *FLASH*³⁵ (v1.2.2) to prepare the raw sequencing reads, checking
334 the quality with *FastQC*³⁶ (v0.10.0). We then used the *Bowtie/TopHat/Cufflinks*³⁷⁻³⁹ pipeline

335 to generate transcripts for the unmasked reference sequence. We generated gene predictions
336 with the MAKER pipeline⁴⁰ (v2.31). Homology and synteny in gene content with the Hm Yb
337 reference were identified by aligning the Hm coding sequences to the He reference with
338 BLAST. Homologous genes were present in the same order and orientation in He and Hm
339 (Fig 2B,C). Annotations were manually adjusted if genes had clearly been merged or split in
340 comparison to *H. melpomene* (which has been extensively manually curated¹²). In addition
341 He cortex was manually curated from the RNA-seq data and using Exonerate⁴¹ alignments of
342 the *H. melpomene* protein and mRNA transcripts, including the 5' UTRs.

343 **Genotype-by-phenotype association analyses**

344 Information on the individuals used and ENA accessions for sequence data are given in
345 Supplementary Table 1. We used shotgun Illumina sequence reads from 45 He individuals
346 from 7 races that were generated as part of a previous study³¹ (Supplementary Information).
347 Reads were aligned to an He reference containing the Cr contig and other sequenced He
348 BACs^{11,31} with BWA⁴², which has previously been found to work better than Stampy⁴³
349 (which was used for the alignments in the other species) with an incomplete reference
350 sequence³¹. The parameters used were as follows: Maximum edit distance (n), 8; maximum
351 number of gap opens (o), 2; maximum number of gap extensions (e), 3; seed (l), 35;
352 maximum edit distance in seed (k), 2. We then used Picard tools to remove PCR and optical
353 duplicate sequence reads and GATK⁴⁴ to re-align indels and call SNPs using all individuals
354 as a single population. Expected heterozygosity was set to 0.2 in GATK. 132,397 SNPs were
355 present across Cr. A further 52,698 SNPs not linked to colour pattern loci were used to
356 establish background association levels.

357 For the Hm / Hn clade we used previously published sequence data from 19 individuals from
358 enrichment sequencing targeting of the Yb region, the unlinked HmB/D region that controls

359 the presence/absence of red colour pattern elements, and ~1.8Mb of non-colour pattern
360 genomic regions⁴⁵, as well as 9 whole genome shotgun sequenced individuals^{18,46}. We added
361 targeted sequencing and shotgun whole genome sequencing of an additional 47 individuals
362 (Supplementary Information). Alignments were performed using Stampy⁴³ with default
363 parameters except for substitution rate which was set to 0.01. We again removed duplicates
364 and used GATK to re-align indels and call SNPs with expected heterozygosity set to 0.1.

365 The analysis of the Hm/timareta/silvaniform included 49 individuals, which were aligned to
366 v1.1 of the Hm reference genome with the scaffolds containing Yb and HmB/D swapped with
367 reference BAC sequences¹⁸, which contained fewer gaps of unknown sequence than the
368 genome scaffolds. 232,631 SNPs were present in the Yb region and a further 370,079 SNPs
369 were used to establish background association levels.

370 The Hn analysis included 26 individuals aligned to unaltered v1.1 of the Hm reference
371 genome, because the genome scaffold containing Yb is longer than the BAC reference
372 making it easier to compare the inverted and non-inverted regions present in this species. We
373 tested for associations at 262,137 SNPs on the Yb scaffold with the Hn bicoloratus morph,
374 which had a sample of 5 individuals.

375 We measured associations between genotype and phenotype using a score test (qtscore) in the
376 GenABEL package in R⁴⁷. This was corrected for background population structure using a
377 test specific inflation factor, λ , calculated from the SNPs unlinked to the major colour pattern
378 controlling loci (described above), as the colour pattern loci are known to have different
379 population structure to the rest of the genome^{14,15,18}. We used a custom perl script to convert
380 GATK vcf files to Illumina SNP format for input to genABEL⁴⁷. genABEL does not accept
381 multiallelic sites, so the script also converted the genotype of any individuals for which a
382 third (or fourth) allele was present to a missing genotype (with these defined as the lowest

383 frequency alleles). Custom R scripts were used to identify sites showing perfect associations
384 with calls for >75% of individuals.

385 **Microarray Gene Expression Analyses**

386 We designed a Roche NimbleGen microarray (12x135K format) with probes for all annotated
387 Hm genes¹⁸ and tiling the central portion of the Yb BAC sequence contig that was previously
388 identified as showing the strongest differentiation between Hm races⁴⁵. In addition to the
389 HmYb tiling array probes there were 6,560 probes tiling HmAc (a third unlinked colour
390 pattern locus) and 10,716 probes tiling HmB/D, again distanced on average at 10bp intervals.
391 The whole-genome gene expression array contained 107,898 probes in total.

392 This was interrogated with Cy3 labelled double stranded cDNA generated from total RNA
393 (with a SuperScript double-stranded cDNA synthesis kit, Invitrogen, and a one-colour DNA
394 labelling kit, Niblegen) from four pupal developmental stages of Hm plesseni and malleti.
395 Pupae were from captive stocks maintained in insectary facilities in Gamboa, Panama. Tissue
396 was stored in RNA later at -80°C prior to RNA extraction. RNA was extracted using TRIZOL
397 (Invitrogen) followed by purification with RNeasy (Qiagen) and DNase treated with DNA-
398 free (Ambion). Quantification was performed using a Qubit 2.0 fluorometer (Invitrogen) and
399 purity and integrity assessed using a Bioanalyzer 2100 (Agilent). Samples were randomised
400 and each hybridised to a separate array. The HmYb probe array contained 9,979 probes
401 distanced on average at 10bp. The whole-genome expression array contained on average 9
402 probes per annotated gene in the genome (v1.1¹⁸) as well as any transcripts not annotated but
403 predicted from RNA-seq evidence.

404 Background corrected expression values for each probe were extracted using NimbleScan
405 software (version 2.3). Analyses were performed with the LIMMA package implemented in
406 R/Bioconductor⁴⁸. The tiling array and whole-genome data sets were analysed separately.

407 Expression values were extracted and quantile-normalised, log₂-transformed, quality
408 controlled and analysed for differences in expression between individuals and wing regions.
409 P-values were adjusted for multiple hypotheses testing using the False Discovery Rate (FDR)
410 method⁴⁹.

411 We detected isoform-specific expression differences between Hm aglaope/amaryllis and Hm
412 rosina/melpomene using RT-PCR and qRT-PCR on RNA extracted from developing hind-
413 wing tissue (further details in Supplementary Information). Previously published RNAseq
414 data was also used to assess gene expression differences between Hm aglaope and
415 amaryllis¹⁸ (further details in Supplementary Information).

416 **In situ hybridisations**

417 Hn and Hm larvae were reared in a greenhouse at 25-30°C and sampled at the last instar. In
418 situ hybridizations were performed according to previously described methods²⁵ with a cortex
419 riboprobe synthesized from a 831-bp cDNA amplicon from Hn. Wing discs were incubated in
420 a standard hybridization buffer containing the probe for 20-24 h at 60°C. For secondary
421 detection of the probe, wing discs were incubated in a 1:3000 dilution of anti-digoxigenin
422 alkaline phosphatase Fab fragments and stained with BM Purple for 3-6 h at room
423 temperature. Stained wing discs were photographed with a Leica DFC420 digital camera
424 mounted on a Leica Z6 APO stereomicroscope.

425 **De novo assembly of short read data in Hm and related taxa**

426 In order to better characterise indel variation from the short-read sequence data used for the
427 genotype-by-phenotype association analysis, we performed de novo assemblies of a subset of
428 Hm individuals and related taxa with a diversity of phenotypes (Extended Data Figure 2).
429 Assemblies were performed using the de novo assembly function of CLCGenomics
430 Workbench v.6.0 under default parameters. The assembled contigs were then BLASTed

431 against the Yb region of the Hm melpomene genome¹⁸, using Geneious v.8.0. The contigs
432 identified by BLAST were then concatenated to generate an allele sequence for each
433 individual. Occasionally two unphased alleles were generated when two contigs were
434 matched to a given region. If more than two contigs of equal length matched then this was
435 considered an unresolvable repeat region and replaced with Ns. The assembled alleles were
436 then aligned using the MAFFT alignment plugin in Geneious v.8.0.

437 **Long-range PCR targeted sequencing of cortex in Hm aglaope and Hm amaryllis**

438 We generated two long-range PCR products covering 88.8% of the 1,344bp coding region of
439 cortex (excluding 67bp at the 5' end and 83bp at the 3' end, further details in Supplementary
440 Information). A product spanning coding exons 5 to 9 (the final exon) was obtained from 29
441 Hm amaryllis individuals and 29 Hm aglaope individuals; a product spanning coding exons 2
442 to 5 was obtained from 32 Hm amaryllis individuals and 14 Hm aglaope. In addition, a
443 product spanning exons 4 to 6 was obtained from 6 Hm amaryllis and 5 Hm aglaope that
444 failed to amplify one or both of the larger products. Long-range PCR was performed using
445 Extensor long-range PCR mastermix (Thermo Scientific) following manufacturers guidelines
446 with a 60°C annealing temperature in a 10-20µl volume. The product spanning coding exons
447 5 to 9 was obtained with primers HM25_long_F1 and HM25_long_R4 (see Supplementary
448 Table 2 for primer sequences); the product spanning coding exons 2 to 5 was obtained with
449 primers HM25_long_F4 and HM25_long_R2; the product spanning exons 4 to 6 was
450 obtained with primers 25_ex5-ex7_r1 and 25_ex5-ex7_f1. Products were pooled for each
451 individual, including 5 additional products from the Yb locus and 7 products in the region of
452 the HmB/D locus. They were then cleaned using QIAquick PCR purification kit (QIAGEN)
453 before being quantified with a Qubit Fluorometer (Life Technologies) and pooled in
454 equimolar amounts for each individual, taking into account variation in the length and
455 number of PCR products included for each individual (because of some PCR failures, ie.

456 proportionally less DNA was included if some PCR products were absent for a given
457 individual).

458 Products were pooled within individuals (including additional products for other genes not
459 analysed here) and then quantified and pooled in equimolar amounts for each individual
460 within each race. The pooled products for each race (Hm *aglaope* and *amaryllis*) were then
461 prepared as two separate libraries with molecular identifiers and sequenced on a single lane
462 of an Illumina GAIIx. Analysis was performed using Galaxy and the history is available at
463 <https://usegalaxy.org/u/njnadeau/h/long-pcr-final>. Reads were quality filtered with a
464 minimum quality of 20 required over 90% of the read, which resulted in 5% of reads being
465 discarded. Reads were then quality trimmed to remove bases with quality less than 20 from
466 the ends. They were then aligned to the target regions using the fosmid sequences from
467 known races⁴⁵ with sequence from the Yb BAC walk¹² used to fill any gaps. Alignments were
468 performed with BWA v0.5.6⁴² and converted to pileup format using Samtools v0.1.12 before
469 being filtered based on quality (≥ 20) and coverage (≥ 10). BWA alignment parameters were
470 as follows: fraction of missing alignments given 2% uniform base error rate (aln -n) 0.01;
471 maximum number of gap opens (aln -o) 2; maximum number of gap extensions (aln -e) 12;
472 disallow long deletion within 12 bp towards the 3'-end (aln -d); number of first subsequences
473 to take as seed (aln -l) 100. We then calculated coverage and minor allele frequencies for
474 each race and the difference between these using custom scripts in R⁵⁰.

475 **Sequencing and analysis of Hm fosmid clones**

476 Fosmid libraries had previously been made from single individuals of 3 Hm races (*rosina*,
477 *amaryllis* and *aglaope*) and several clones overlapping the Yb interval had been sequenced⁴⁵.
478 We extended the sequencing of this region, particularly the region overlapping cortex by
479 sequencing an additional 4 clones from Hm *rosina* (1051_83D21, accession KU514430;

480 1051_97A3, accession KU514431; 1051_65N6, accession KU514432; 1051_93D23,
481 accession KU514433) 2 clones from Hm amaryllis (1051_13K4, accession KU514434;
482 1049_8P23, accession KU514435) and 3 clones from Hm aglaope (1048_80B22, accession
483 KU514437; 1049_19P15, accession KU514436; 1048_96A7, accession KU514438). These
484 were sequenced on a MiSeq 2000, and assembled using the de novo assembly function of
485 CLCGenomcs Workbench v.6.0. The individual clones (including existing clones 1051-
486 143B3, accession FP578990; 1049-27G11, accession FP700055; 1048-62H20, accession
487 FP565804) were then aligned to the BAC and genome scaffold¹⁸ references using the
488 MAFFT alignment plugin of Geneious v.8.0. Regions of general sequence similarity were
489 identified and visualised using MAUVE⁵¹. We merged overlapping clones from the same
490 individual if they showed no sequence differences, indicating that they came from the same
491 allele. We identified transposable elements (TEs) using nBLAST with an insect TE list
492 downloaded from Repbase Update⁵² including known *Heliconius* specific TEs⁵³.

493 **5' RACE, RT-PCR and qRT-PCR**

494 All tissues used for gene expression analyses were dissected from individuals from captive
495 stocks derived from wild caught individuals of various races of Hm (aglaope, amaryllis,
496 melpomene, rosina, plesseni, malleti) and F2 individuals from a Hm rosina (female) x Hm
497 melpomene (male) cross. Experimental individuals were reared at 28°C-31°C. Developing
498 wings were dissected and stored in RNAlater (Ambion Life Technologies). RNA was
499 extracted using a QIAGEN RNeasy Mini kit following the manufacturer's guidelines and
500 treated with TURBO DNA-free DNase kit (Ambion Life Technologies) to remove remaining
501 genomic DNA. RNA quantification was performed with a Nanodrop spectrophotometer, and
502 the RNA integrity was assessed using the Bioanalyzer 2100 system (Agilent).

503 Total RNA was thoroughly checked for DNA contamination by performing PCR for EF1 α
504 (using primers ef1-a_RT_for and ef1-a_RT_rev, Table S2) with 0.5 μ l of RNA extract (50ng-
505 1 μ g of RNA) in a 20 μ l reaction using a polymerase enzyme that is not functional with RNA
506 template (BioScript, Bioline Reagents Ltd.). If a product amplified within 45 cycles then the
507 RNA sample was re-treated with DNase.

508 Single stranded cDNA was synthesised using BioScript MMLV Reverse Transcriptase
509 (Bioline Reagents Ltd.) with random hexamer (N6) primers and 1 μ g of template RNA from
510 each sample in a 20 μ l reaction volume following the manufacturer's protocol. The resulting
511 cDNA samples were then diluted 1:1 with nuclease free water and stored at -80°C.

512 5' RACE was performed using RNA from hind-wing discs from one Hm aglaope and one
513 Hm amaryllis final instar larvae with a SMARTer RACE kit from Clontech (California,
514 USA). The gene specific primer used for the first round of amplification was anchored in
515 exon 4 (fzl_raceex5_R1, Supplementary Table 2). Secondary PCR of these products was then
516 performed using a primer in exon 2 (HM25_long_F2, Supplementary Table 2) and the nested
517 universal primer A. Other isoforms were detected by RT-PCR using primers within exons 2
518 and 9 (gene25_for_full1 and gene25_rev_ex3). We identified isoforms from 5' RACE and
519 RT-PCR products by cutting individual bands from agarose gels and if necessary by cloning
520 products before Sanger sequencing. Cloning of products was performed using TOPO TA
521 (Invitrogen) or pGEM-T (Promega) cloning kits. Sanger sequencing was performed using
522 BigDye terminator v3.1 (Applied Biosystems) run on an ABI13730 capillary sequencer.
523 Primers fzl_ex1a_F1 and fzl_ex4_R1 were used to confirm expression of the furthest 5'
524 UTR. For isoforms that appeared to show some degree of race specificity we designed
525 isoform specific PCR primers spanning specific exon junctions (Extended Data Fig 2, 4,
526 Supplementary Table 2) and used these to either qualitatively (RT-PCR) or quantitatively
527 (qRT-PCR) assess differences in expression between races.

528 We performed qRT-PCR using SensiMix SYBR green (Bioline Reagents Ltd.) with 0.2-
529 0.25 μ M of each primer and 1 μ l of the diluted product from the cDNA reactions. Reactions
530 were performed in an Opticon 2 DNA engine (MJ Research), with the following cycling
531 parameters: 95°C for 10min, 35-50 x: (95°C for 15sec, 55-60°C for 30sec, 72° for 30sec),
532 72°C for 5min. Melting curves were generated between 55°C and 90°C with readings taken
533 every 0.2°C for each of the products to check that a single product was generated. At least
534 one product from each set of primers was also run on a 1% agarose gel to check that a single
535 product of the expected size was produced and the identity of the product confirmed by direct
536 sequencing (See Supplementary Table 2 for details of primers for each gene). We used two
537 housekeeping genes (*EF1 α* and Ribosomal Protein S3A) for normalisation and all results
538 were taken as averages of triplicate PCR reactions for each sample.

539 C_t values were defined as the point at which fluorescence crossed a threshold (R_{Ct}) adjusted
540 manually to be the point at which fluorescence rose above the background level.

541 Amplification efficiencies (E) were calculated using a dilution series of clean PCR product.

542 Starting fluorescence, which is proportional to the starting template quantity, was calculated

543 as $R_0 = R_{Ct} (1+E)^{-Ct}$. Normalized values were then obtained by dividing R_0 values for the

544 target loci by R_0 values for *EF1 α* and *RPS3A*. Results from both of these controls were

545 always very similar, therefore the results presented are normalized to the mean of *EF1 α* and

546 *RPS3A*. All results were taken as averages of triplicate PCR reactions. If one of the triplicate

547 values was more than one cycle away from the mean then this replicate was excluded.

548 Similarly any individuals that were more than two standard deviations away from the mean of

549 all individuals for the target or normalization genes were excluded (these are not included in

550 the numbers of individuals reported). Statistical significance was assessed by Wilcoxon rank

551 sum tests performed in R⁵⁰.

552 **RNAseq analysis of *Hm amaryllis/aglaope***

553 RNA-seq data for hind-wings from three developmental stages had previously been obtained
554 for two individuals of each race at each stage (12 individuals in total) and used in the
555 annotation of the Hm genome¹⁸ (deposited in ENA under study accessions ERP000993 and
556 PRJEB7951). Four samples were multiplexed on each sequencing lane with the fifth instar
557 larval and day 2 pupal samples sequenced on a GAIIx sequencer and the day 3 pupal wings
558 sequenced on a HiSeq 2000 sequencer.

559 Two methods were used for alignment of reads to the reference genome and inferring read
560 counts, Stampy⁴³ and RSEM (RNAseq by Expectation Maximisation)⁵⁴. In addition we used
561 two different R/Bioconductor packages for estimation of differential gene expression,
562 DESeq⁵⁵ and BaySeq⁵⁶. Read bases with quality scores < 20 were trimmed with FASTX-
563 Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Stampy was run with default
564 parameters except for mean insert size, which was set to 500, SD 100 and substitution rate,
565 which was set to 0.01. Alignments were filtered to exclude reads with mapping quality <30
566 and sorted using Samtools⁵⁷. We used the HT seq-count script in with HTseq⁵⁸ to infer counts
567 per gene from the BAM files.

568 RSEM⁵⁴ was run with default parameters to infer a transcriptome and then map RNAseq
569 reads against this using Bowtie³⁷ as an aligner. This was run with default parameters except
570 maximum number of mismatches, which was set to 3.

571 **Annotation and alignment of fizzy family proteins**

572 In the arthropod genomes, some fizzy family proteins were found to be poorly annotated
573 based on alignments to other family members. In these cases annotations were improved
574 using well annotated proteins from other species as references in the program Exonerate⁴¹
575 and the outputs were manually curated. Specifically, the annotation of *B. mori* fzr was
576 extended based on alignment of *D. plexippus* fzr; the annotation of *B. mori* fzy was altered

577 based on alignment of *Drosophila melanogaster* and *D. plexippus* *fzy*; *H. melpomene* *fzy* was
578 identified as part of the annotated gene HMEL017486 on scaffold HE671623 (Hmel v1.1)
579 based on alignment of *D. plexippus* *fzy*; the *Apis mellifera* *fzr* annotation was altered based
580 on alignment of *D. melanogaster* *fzr*; the annotation of *Acyrtosiphon pisum* *fzr* was altered
581 based on alignment of *D. melanogaster* *fzr*. No one-to-one orthologues of *D. melanogaster*
582 *fzr2* were found in any of the other arthropod genera, suggesting that this gene is *Drosophila*
583 specific. Multiple sequence alignment of all the fizzy family proteins was then performed
584 using the Expresso server⁵⁹ within T-coffee⁶⁰, and this alignment was used to generate a
585 neighbour joining tree in Geneious v8.1.7.

586 **Expression of *H. melpomene* cortex in *D. melanogaster* wings**

587 *D. melanogaster* Cortex is known to generate an irregular microchaete phenotype when
588 ectopically expressed in the posterior compartment of the adult fly wing²⁴. We performed the
589 same assay using *H. melpomene* cortex in order to test if this functionality was conserved.
590 Following the methods of Swan and Schüpbach²⁴ a UAS-GAL4 construct was created using
591 the coding region for the long isoform of Hm cortex, plus a *Drosophila* cortex version to act
592 as positive control. The HA-tagged *H. melpomene* UAS-cortex expression construct was
593 generated using cDNA reverse transcribed (Revert-Aid, Thermo-Scientific) from RNA
594 extracted (Qiagen RNeasy) from pre-ommochrome pupal wing material. An HA-tagged
595 *D. melanogaster* UAS-cortex version was also constructed, following the methods of Swan
596 and Schüpbach, (2007). Expression was driven by *hsp70* promoter. Constructs were injected
597 into ϕ C31-attP40 flies (#25709, Bloomington stock centre, Indiana; Cambridge University
598 Genetics Department, UK, fly injection service) by site directed insertion into CII via an attB
599 site in the construct. Homozygous transgenic flies were crossed with *w,y'*;en-GAL4;UAS-
600 GFP (gift of M. Landgraf lab, Cambridge University Zoology Department) to drive

601 expression in the engrailed posterior domain of the wing, and adult offspring wings
602 photographed (Extended Data Fig 6B-D). Expression of the construct was confirmed by IHC
603 (standard *Drosophila* protocol) of final instar larval wing discs using mouse anti-HA and goat
604 anti-mouse alexa-fluor 568 secondary antibodies (Abcam), imaged by Leica SP5 confocal.
605 Successful expression of Hm_Cortex was confirmed by IHC against an HA tag inserted at the
606 N terminal of either protein (Extended Data Fig 6E).

607

608 **References**

- 609 31. Supple, M. A. et al. Genomic architecture of adaptive color pattern divergence and
610 convergence in *Heliconius* butterflies. *Genome Res.* **23**, 1248–1257 (2013).
- 611 32. de la Bastide, M. & McCombie, W. R. Assembling genomic DNA sequences with
612 PHRAP. *Curr. Protoc. Bioinforma.* Ed. Board Andreas Baxevanis Al **Chapter 11**,
613 Unit11.4 (2007).
- 614 33. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing.
615 *Genome Res.* **8**, 195–202 (1998).
- 616 34. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
617 sequence data. *Bioinformatics* btu170 (2014). doi:10.1093/bioinformatics/btu170
- 618 35. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve
619 genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
- 620 36. Andrews, S. FastQC. (2011).
- 621 37. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient
622 alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
- 623 38. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with
624 RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

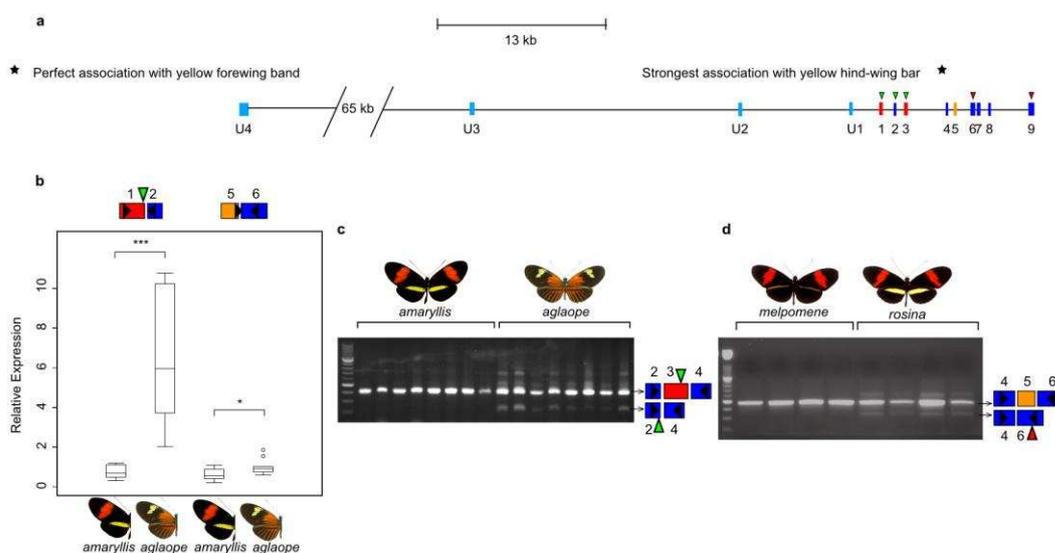
- 625 39. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals
626 unannotated transcripts and isoform switching during cell differentiation. *Nat.*
627 *Biotechnol.* **28**, 511–515 (2010).
- 628 40. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database
629 management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491
630 (2011).
- 631 41. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence
632 comparison. *BMC Bioinformatics* **6**, 31 (2005).
- 633 42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
634 transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).
- 635 43. Lunter, G. & Goodson, M. Stampy: A statistical algorithm for sensitive and fast mapping
636 of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
- 637 44. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-
638 generation DNA sequencing data. *Nat Genet* **43**, 491–498 (2011).
- 639 45. Nadeau, N. J. et al. Genomic islands of divergence in hybridizing *Heliconius* butterflies
640 identified by large-scale targeted sequencing. *Philos. Trans. R. Soc. B Biol. Sci.* **367**,
641 343–353 (2012).
- 642 46. Martin, S. H. et al. Genome-wide evidence for speciation with gene flow in *Heliconius*
643 butterflies. *Genome Res.* **23**, 1817–1828 (2013).
- 644 47. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for
645 genome-wide association analysis. *Bioinforma. Oxf. Engl.* **23**, 1294–1296 (2007).
- 646 48. Smyth, G. K. in *Bioinformatics and Computational Biology Solutions Using R and*
647 *Bioconductor* (eds. Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.)
648 397–420 (Springer New York, 2005).

- 649 49. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
650 Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300
651 (1995).
- 652 50. R Development Core Team. *R: A language and environment for statistical computing.*
653 (R Foundation for Statistical Computing, 2011).
- 654 51. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: Multiple Alignment of
655 Conserved Genomic Sequence With Rearrangements. *Genome Res.* **14**, 1394–1403
656 (2004).
- 657 52. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet.*
658 *Genome Res.* **110**, 462–467 (2005).
- 659 53. Lavoie, C. A., Platt, R. N., Novick, P. A., Counterman, B. A. & Ray, D. A. Transposable
660 element evolution in *Heliconius* suggests genome diversity within Lepidoptera. *Mob.*
661 *DNA* **4**, 21 (2013).
- 662 54. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data
663 with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
- 664 55. Anders, S. & Huber, W. Differential expression analysis for sequence count data.
665 *Genome Biol.* **11**, 1–12 (2010).
- 666 56. Hardcastle, T. J. & Kelly, K. A. baySeq: Empirical Bayesian methods for identifying
667 differential expression in sequence count data. *BMC Bioinformatics* **11**, 422 (2010).
- 668 57. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.*
669 **25**, 2078–2079 (2009).
- 670 58. Anders, S., Pyl, P. T. & Huber, W. HTSeq - A Python framework to work with high-
671 throughput sequencing data. *bioRxiv* (2014). doi:10.1101/002824

- 672 59. Armougom, F. et al. Espresso: automatic incorporation of structural information in
 673 multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* **34**, W604–608
 674 (2006).
- 675 60. Di Tommaso, P. et al. T-Coffee: a web server for the multiple sequence alignment of
 676 protein and RNA sequences using structural information and homology extension.
 677 *Nucleic Acids Res.* **39**, W13–17 (2011).

678

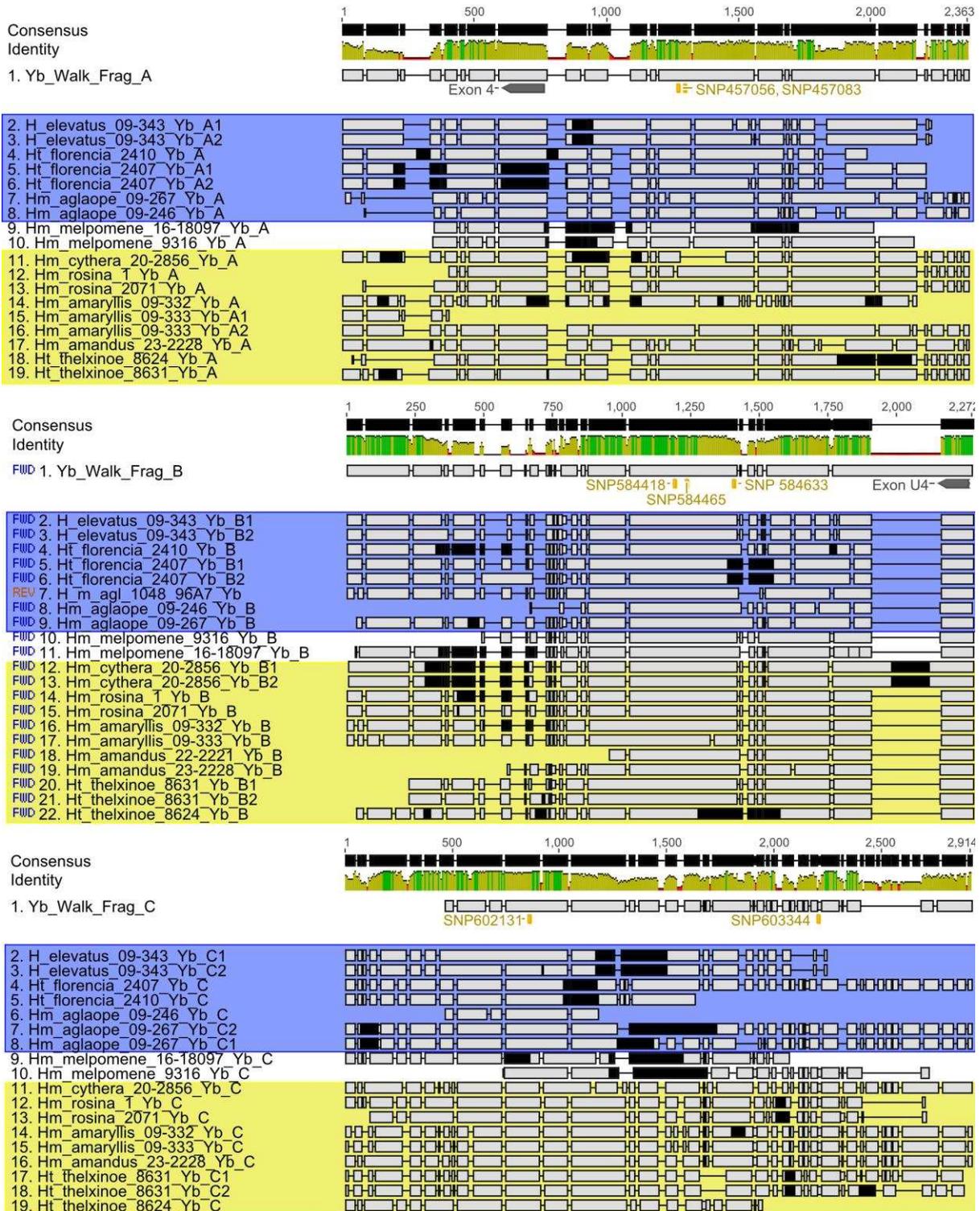
679

680 **Extended Data**

681

- 682 Extended Data Figure 1. A) Exons and splice variants of cortex in Hm. Orientation is
 683 reversed with respect to figures 2 and 4, with transcription going from left to right. SNPs
 684 showing the strongest associations with phenotype are shown with stars. B) Differential
 685 expression of two regions of cortex between Hm amaryllis and Hm aglaope whole hindwings
 686 (N=11 and N=10 respectively). Boxplots are standard (median; 75th and 25th percentiles;
 687 maximum and minimum excluding outliers – shown as discrete points) C) Expression of a

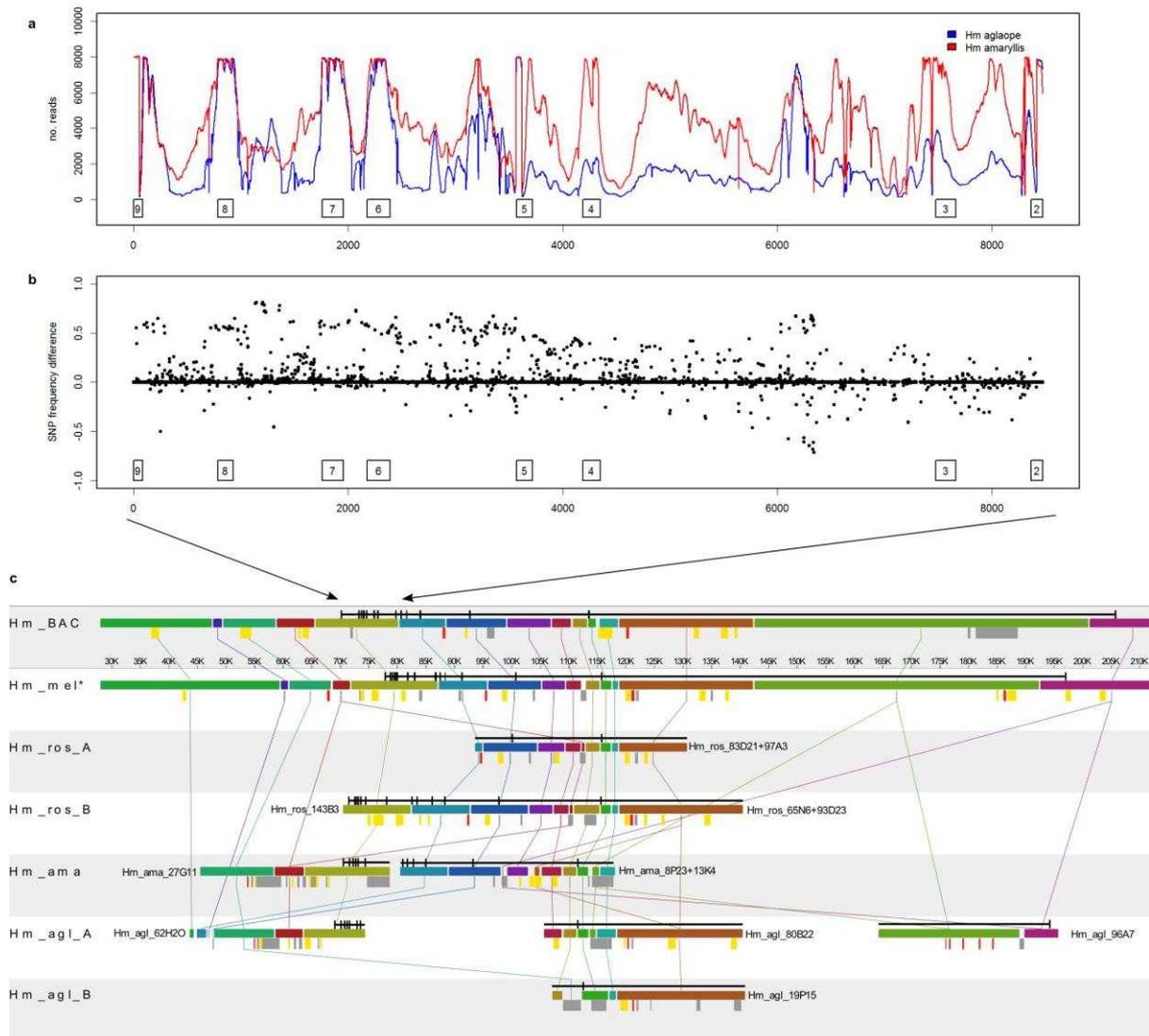
688 cortex isoform lacking exon 3 is found in Hm aglaope but not Hm amaryllis hindwings. D)
689 Expression of an isoform lacking exon 5 is found in Hm rosina but not Hm melpomene
690 hindwings. Green triangles indicate predicted start codons and red triangles predicted stop
691 codons, with usage dependent on which exons are present in the isoform. Schematics of the
692 targeted exons are shown for each (q)RT-PCR product, black triangles indicate the position
693 of the primers used in the assay.



694

695 Extended Data Figure 2. Alignments of de novo assembled fragments containing the top
 696 associated SNPs from Hm and related taxa short-read data. Identified indels do not show
 697 stronger associations with phenotype that those seen at SNPs (as shown in Extended Data
 698 Table 2), although some near-perfect associations are seen in fragment C. Black regions =

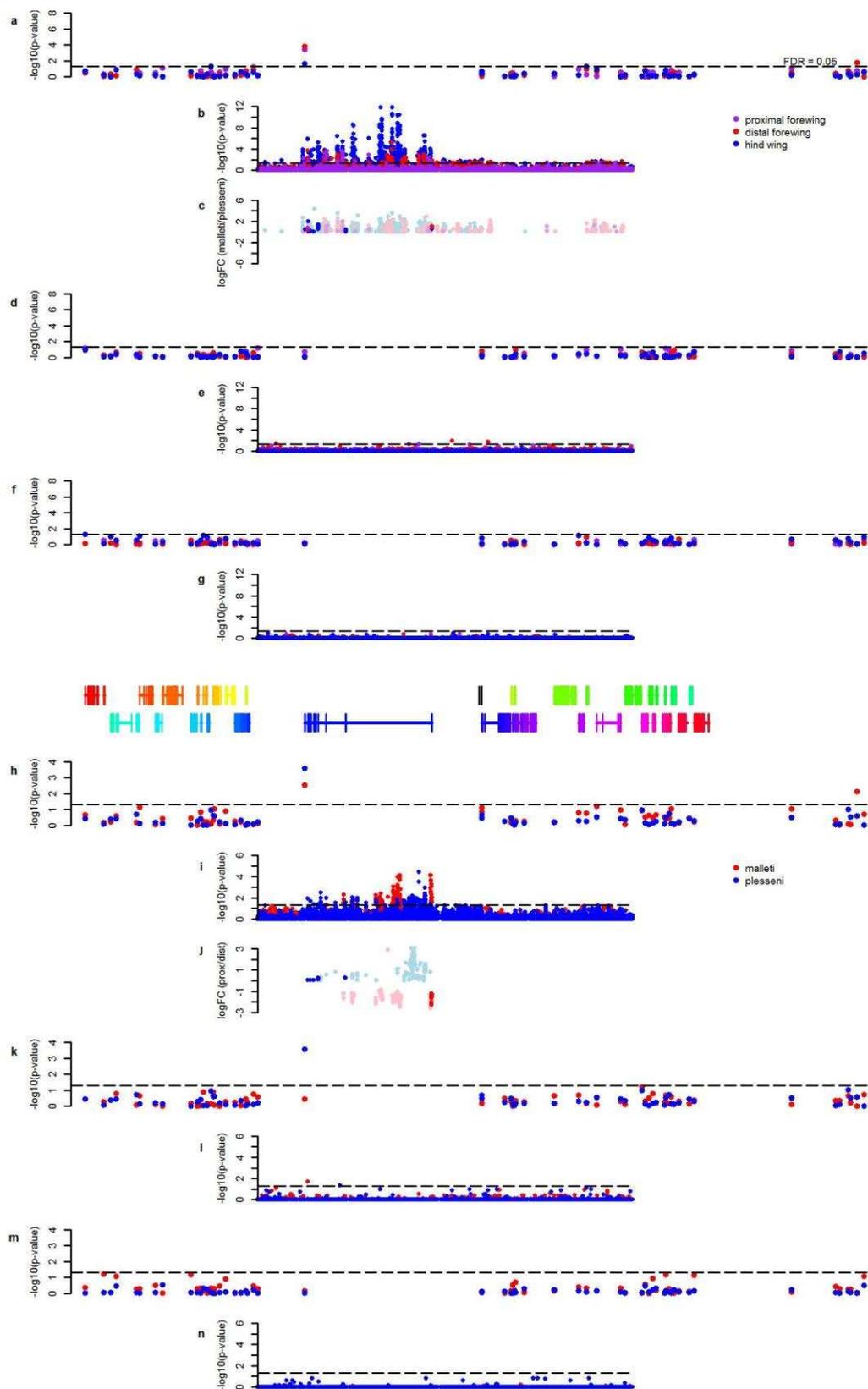
699 missing data; yellow box = individuals with a hindwing yellow bar; blue box = individuals
 700 with a yellow forewing band.



701

702 Extended Data Figure 3. Sequencing of long-range PCR products and fosmids spanning
 703 cortex. A) Sequence read coverage from long-range PCR products across the cortex coding
 704 region from 2 Hm races. B) Minor allele frequency difference from these reads between Hm
 705 aglaope and Hm amaryllis. Exons of cortex are indicated by boxes, numbered as in Extended
 706 Data Figure 2. C) Alignments of sequenced fosmids overlapping cortex from 3 Hm
 707 individuals of difference races. No major rearrangements are observed, nor any major
 708 differences in transposable element (TE) content between closely related races with different

709 colour patterns (melpomene/rosina or amaryllis/aglaope). Hm amaryllis and rosina have the
710 same phenotype, but do not share any TEs that are not present in the other races. Hm_BAC =
711 BAC reference sequence, Hm_mel = melpomene from new unpublished assembly of Hm
712 genome⁵¹, Hm_ros = rosina (2 different alleles were sequenced from this individual),
713 Hm_ama = amaryllis (2 non-overlapping clones were sequenced in this individual), Hm_agla
714 = aglaope (4 clones were sequenced in this individual 2 of which represent alternative
715 alleles). Alignments were performed with Mauve: coloured bars represent homologous
716 genomic regions. cortex is annotated in black above each clone. Variable TEs are shown as
717 coloured bars below each clone: red = Metulj-like non-LTR, yellow = Helitron-like DNA,
718 grey = other.

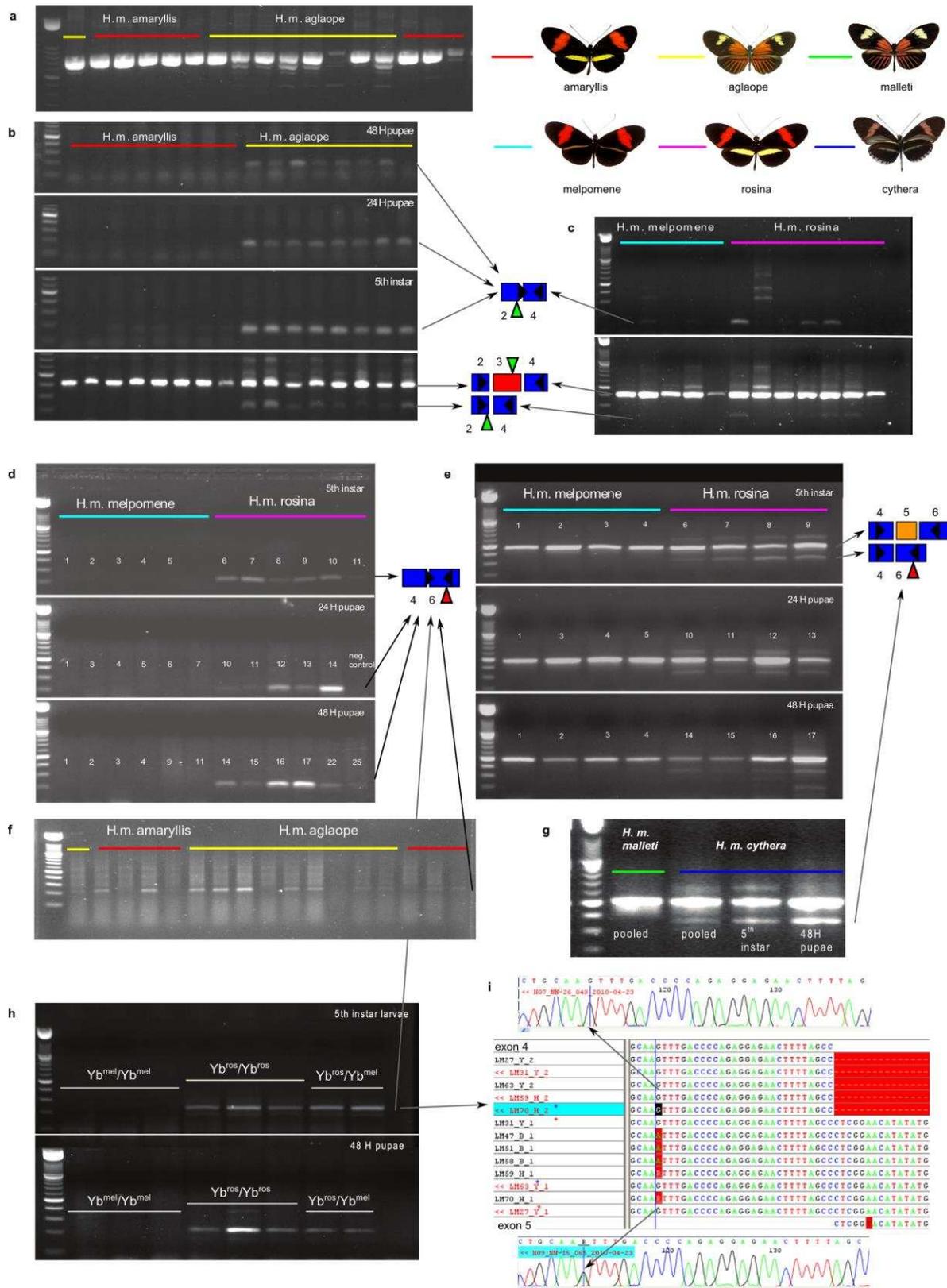


719

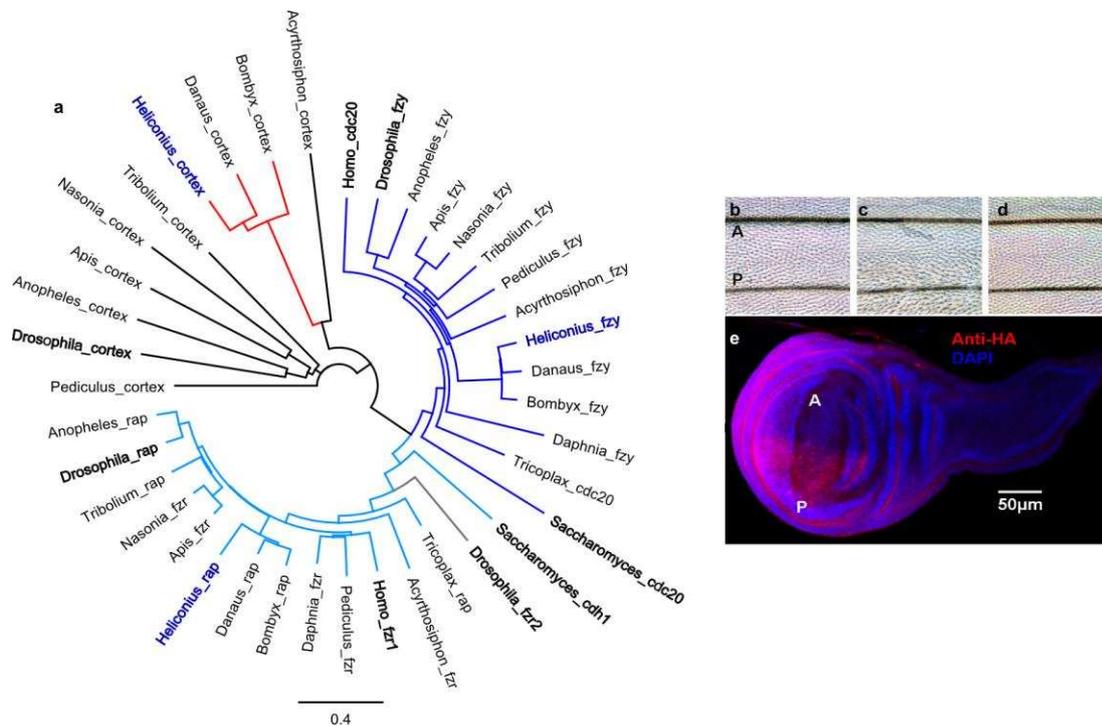
720 Extended Data Figure 4. Expression array results for additional stages, related to Figure 4. A-

721 G: comparisons between races (*H. m. plesseni* and *H. m. malleti*) for 3 wing regions. H-N:

722 comparisons between proximal and distal forewing regions for each race. Significance values
723 ($-\log_{10}(\text{p-value})$) are shown separately for genes in the HmYb region from the gene array
724 (A,D,F,H,K,M) and for the HmYb tiling array (B,E,G,I,L,N) for day 1 (A,B,H,I), day 5
725 (D,E,K,L) and day 7 (F,G,M,N) after pupation. The level of expression difference (log fold
726 change) for tiling probes showing significant differences ($p \leq 0.05$) is shown for day 1 (C and
727 J) with probes in known cortex exons shown in dark colours and probes elsewhere shown as
728 pale colours.



732 Differences in splicing of exon 3 between *H. m. aglaope* and *H. m. amaryllis*. Products
733 amplified with a primer spanning the exon 2/4 junction at 3 developmental stages. The lower
734 panel shows verification of this assay by amplification between exons 2 and 4 for the same
735 final instar larval samples (replicated in Extended Data Figure 2C) C) Lack of consistent
736 differences between *H. m. melpomene* and *H. m. rosina* in splicing of exon 3. Top panel
737 shows products amplified with a primer spanning the exon 2/4 junction, lower panel is the
738 same samples amplified between exons 2 and 4. D) Differences in splicing of exon 5 between
739 *H. m. melpomene* and *H. m. rosina*. Products amplified with a primer spanning the exon 4/6
740 junction at 3 developmental stages. E) Subset of samples from D amplified with primers
741 between exons 4 and 6 for verification (middle, 24hr pupae samples are replicated in
742 Extended Data Figure 2D). F) Lack of consistent differences between *H. m. aglaope* and *H.*
743 *m. amaryllis* in splicing of exon 5. Products amplified with a primer spanning the exon 4/6
744 junction. G) *H. m. cythera* also expresses the isoform lacking exon 5, while a pool of 6 *H. m.*
745 *malleti* individuals do not. H) Expression of the isoform lacking exon 5 from an F2 *H. m.*
746 *melpomene* x *H. m. rosina* cross. Individuals homozygous or heterozygous for the *H. m.*
747 *rosina* HmYb allele express the isoform while those homozygous for the *H. m. melpomene*
748 HmYb allele do not. I) Allele specific expression of isoforms with and without exon 5.
749 Heterozygous individuals (indicated with blue and red stars) express only the *H. m. rosina*
750 allele in the isoform lacking exon 5 (G at highlighted position), while they express both
751 alleles in the isoform containing exon 5 (G/A at this position).



752

753 Extended Data Figure 6. Phylogeny of fizzy family proteins and effects of expressing cortex

754 in the *Drosophila* wing. A) Neighbour joining phylogeny of Fizzy family proteins including755 functionally characterised proteins (in bold) from *Saccharomyces cerevisiae*, *Homo sapiens*756 and *Drosophila melanogaster* as well as copies from the basal metazoan *Trichoplax*757 *adhaerens* and a range of annotated arthropod genomes (*Daphnia pulex*, *Acyrthosiphon*758 *pisum*, *Pediculus humanus*, *Apis mellifica*, *Nasonia vitripennis*, *Anopheles gambiae*,759 *Tribolium castaneum*) including the lepidoptera *H. melpomene* (in blue), *Danaus plexippus*760 and *Bombyx mori*. Branch colours: dark blue, CDC20/fzy; light blue, CDH1/fzr/rap; red,761 lepidopteran cortex. B-E) Ectopic expression of cortex in *Drosophila melanogaster*.762 *Drosophila cortex* produces an irregular microchaete phenotype when expressed in the763 posterior compartment of the fly wing (C) whereas *Heliconius cortex* does not (D), when764 compared to no expression (B). A, anterior; P, posterior. Successful *Heliconius cortex*765 expression was confirmed by anti-HA IHC in the last instar *Drosophila* larva wing imaginal

766 disc (D, red), with DAPI staining in blue.

767 Extended Data Table 1. Genes in the Yb region and evidence for wing patterning control in
 768 *Heliconius*

<i>Hm</i> gene ID	<i>He</i> gene ID	Putative gene name	<i>Heliconius melpomene</i>										<i>H. erato</i>			<i>Hn</i>	
			Yb ^l	Sb ^l	A ^{Yb}	A ^N	E ^l	E ^{9w}	E ^{9r}	E ^{hw}	E ^{lr}	Cr ^l	A ^{not}	A ^{fav}	P ^l	A ^{bic}	
HM00002	HERA000036	Acylpeptide hydrolase			2								x				
HM00003	HERA000037	HM00003											x				
HM00004	HERA000038	Trehalase-1B	x										x				
HM00006	HERA000038.1	Trehalase-1A	x										x				
HM00007	HERA000039	B9 protein	x										x				
HM00008	HERA000040	HM00008	x		2								x				
HM00010	HERA000041	WD40 repeat domain 85	x										x				
HM00012	HERA000042	CG2519	x					x					x				
HM00013	HERA000045	Unkempt	x										x				
HM00014	HERA000046	Histone H3	x										x				
HM00015	HERA000047	HM00015	x										x				
HM00016	HERA000048	HM00016	x										x				
HM00017	HERA000049	RecQ Helicase	x										x				
HM00018	HERA000051	HM00018	x										x				
HM00019	HERA000052	BmSuc2	x					x					x				
HM00020	HERA000053	CG5796	x										x				
HM00021	HERA000054	HM00021	x										x				
HM00022	HERA000055	Enoyl-CoA hydratase	x										x				
HM00023	HERA000056	ATP binding protein	x										x				
HM00024	HERA000057	HM00024	x										x				
HM00025	HERA000059	cortex	x	x	56	74	x	x	x	603	1796	x	2	99	x	51	
HM00026	HERA000077	Poly(A)-specific ribonuclease (parr)		x	10						1	34	x				x
HM00027	HERA000079	CG31320		x									x				x
HM00028	HERA000080	ARP-like		x									x				x
HM00029	HERA000081	CG4692		x									x				x
HM00030	HERA000082	Proteasome 26S non ATPase subunit 4		x									x				x
HM00031	HERA000083	HM00031		x									x				x
HM00032	HERA000084	Zinc phosphodiesterase		x								1	x				x
HM00033	HERA000085	Serine/threonine-protein kinase (LMTK1)		x								8	x				x
HM00034	HERA000086	WD repeat domain 13 (Wdr13)			1	4						5	x				x
HM00035	HERA000087	Domeless			1	2							x				x
HM00036	HERA000061	WAS protein family homologue 1			5	36						37	x				x
HM00038	HERA000062	Lethal (2) k05819 CG3054											x	2			x
HM00039	HERA000064	Mitogen-activated protein kinase (MAPKK)											x				x
HM00040	HERA000064.1	DNA excision repair protein ERCC-6											x				x
HM00041	HERA000065	Penguin											x				x
HM00042	HERA000066	Thymidylate kinase											x				x
HM00043	HERA000067	Caspase-activated DNase											x				x
HM00044	HERA000068	Regulator of ribosome biosynthesis											x				x
HM00045	HERA000069	CG12659											x				x
HM00046	HERA000070	CG33505											x				x
HM00047	HERA000071	Sr protein											x				x
HM00048	HERA000073	HM00048											x				x
HM00049	HERA000073.1	HM00049											x				x
HM00050	HERA000074	Shuttle craft											x				x
HM00051	HERA000075	HM00051											x				x
769 HM00052	HERA000076	HM00052					x						x				x

770 Yb^l, within the previously mapped Yb interval¹². Sb^l, within the previously mapped Sb
 771 interval¹². Sb controls a white/yellow hindwing margin and is not investigated in this study.
 772 The N locus has not been fine-mapped previously. A^{Yb}, number of above background SNPs

773 associated with the hindwing yellow bar in this study. A^N , number of above background
774 SNPs associated with the forewing yellow band in this study. E^1 , detected as differentially
775 expressed between Hm aglaope and amaryllis from RNAseq data in this study
776 (Supplementary Information). E^{gw} , detected as differentially expressed between forewing
777 regions in the gene array in this study. E^{gr} , detected as differentially expressed between Hm
778 plesseni and malleti in in the gene array in this study. E^{tw} , numbers of probes showing
779 differential expression between forewing regions in the tiling array in this study. E^{tr} ,
780 numbers of probes showing differential expression between Hm plesseni and malleti in in the
781 tiling array in this study. Cr^I , within the previously mapped HeCr interval¹¹. A^{pet} , number of
782 SNPs fixed for the alternative allele in He demophoon. A^{fav} , number of SNPs fixed for the
783 alternative allele in He favorinus. P^I , within the previously mapped P interval¹³. A^{bic} , number
784 of above background SNPs associated with the Hn bicoloratus phenotype in this study.
785

786 Extended Data Table 2. Locations of fixed/above background SNPs and differentially
 787 expressed (DE) tiling array probes

		Positions of SNPs in the <i>He</i> and <i>Hn</i> association analyses								
		<i>cortex</i> coding exons	<i>cortex</i> UTR exons	<i>cortex</i> introns (nonTE)	<i>cortex</i> flanking intergenic (nonTE)	TEs	Other genes (exons or introns)	Other intergenic	Total	
<i>erato favorinus</i> fixed		2	0	96	8	2	0	0	108	
<i>erato demophoon</i> fixed		0	0	1	5	1	2	6	15	
<i>numata bicoloratus</i> above background		1	3	47	16	0	2	0	69	
		Positions of DE tiling array probes								
		Known <i>cortex</i> coding exons	<i>cortex</i> UTR exons	<i>cortex</i> introns (nonTE)	miRNAs	TEs	Other gene exons	Other introns/intergenic	Total	
Day3	malleti vs plesseni	Forewing proximal	8	7	323	0	13	1	7	359
		Forewing distal	12	2	327	0	8	0	8	357
		Hindwing	5	14	378	0	9	1	6	413
Proximal vs distal		malleti	0	1	68	0	0	0	12	81
		plesseni	2	4	222	0	10	0	4	242
Day1	malleti vs plesseni	Forewing proximal	1	0	22	0	3	0	7	33
		Forewing distal	2	3	116	1	9	5	112	248
		Hindwing	9	10	500	1	20	2	80	622
Proximal vs distal		malleti	0	12	95	0	1	0	0	108
		plesseni	3	3	81	0	99	0	0	186

788

789

790 Extended Data Table 3. SNPs showing the strongest phenotypic associations in the H.
 791 melpomene/timareta/silvaniform comparison.

Species	Race	Sample Code	SNP pos		SNP pos		SNP pos		SNP pos		SNP pos		SNP pos	
			HW 457083† bar (p=6.07E-10)	439063* (p=1.72E-09)	602131‡ (p=2.42E-09)	457056† (p=2.42E-09)	FW band	584465§ (p=1.37E-07)	584418§ (p=1.41E-07)	584633§ (p=2.10E-07)	603344‡ (p=2.19E-07)			
<i>H. melpomene</i>	<i>aglaope</i>	09-246	0	A/A	A/G	A/A	C/C	1	T/T	A/A	NA	T/T		
<i>H. melpomene</i>	<i>aglaope</i>	09-267	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T		
<i>H. melpomene</i>	<i>aglaope</i>	09-268	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T		
<i>H. melpomene</i>	<i>aglaope</i>	09-357	0	A/A	G/G	G/A	C/C	1	T/T	NA	C/C	T/T		
<i>H. melpomene</i>	<i>aglaope</i>	aglaope.1	0	A/A	G/G	NA	C/C	1	C/T	T/A	T/C	T/T		
<i>H. melpomene</i>	<i>amandus</i>	2221	1	A/A	NA	G/G	C/C	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>amandus</i>	2228	1	A/A	NA	G/G	C/C	0	C/T	T/A	T/C	A/A		
<i>H. melpomene</i>	<i>amaryllis</i>	09-332	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>amaryllis</i>	09-333	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>amaryllis</i>	09-075	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>amaryllis</i>	09-079	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>amaryllis</i>	amaryllis.1	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>bellula</i>	228	1	T/T	NA	G/G	T/T	0	C/C	T/T	T/T	NA		
<i>H. melpomene</i>	<i>bellula</i>	231	1	T/T	NA	G/A	T/T	0	C/T	T/A	T/C	NA		
<i>H. melpomene</i>	<i>cythera</i>	2856	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>cythera</i>	2857	1	NA	NA	NA	NA	0	NA	NA	NA	NA		
<i>H. melpomene</i>	<i>malleti</i>	17162	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T		
<i>H. melpomene</i>	<i>melpomene</i>	melpomene18038	0	A/A	G/G	G/G	C/C	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>melpomene</i>	melpomene18097	0	NA	G/G	NA	C/C	0	C/C	T/T	T/T	NA		
<i>H. melpomene</i>	<i>melpomene</i>	melpomenem0.06	0	A/A	G/G	G/G	C/C	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>melpomene</i>	melpomenegen_ref	0	A/A	G/G	NA	C/C	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>melpomene</i>	melpomene13435	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>melpomene</i>	melpomene9315	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>melpomene</i>	melpomene9316	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>melpomene</i>	melpomene9317	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>plesseni</i>	9156	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	NA		
<i>H. melpomene</i>	<i>plesseni</i>	16293	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	NA		
<i>H. melpomene</i>	<i>rosina</i>	rosina.1	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>rosina</i>	2071	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>rosina</i>	531	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>rosina</i>	533	1	T/T	NA	G/G	T/T	0	C/C	T/T	T/T	NA		
<i>H. melpomene</i>	<i>rosina</i>	546	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. melpomene</i>	<i>thelxiopoeia</i>	13566	0	A/A	G/G	A/A	C/C	1	C/T	T/A	T/C	T/T		
<i>H. melpomene</i>	<i>vulcanus</i>	14632	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	NA		
<i>H. melpomene</i>	<i>vulcanus</i>	519	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. timareta</i>	<i>florencia</i>	2403	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T		
<i>H. timareta</i>	<i>florencia</i>	2406	0	A/A	A/G	A/A	C/C	1	T/T	A/A	C/C	T/T		
<i>H. timareta</i>	<i>florencia</i>	2407	0	A/A	A/G	A/A	C/C	1	T/T	A/A	C/C	T/T		
<i>H. timareta</i>	<i>florencia</i>	2410	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T		
<i>H. timareta</i>	<i>timareta</i>	8533	0	A/A	G/G	A/A	C/C	1	C/T	T/A	T/C	T/T		
<i>H. timareta</i>	<i>timareta</i>	9184	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T		
<i>H. timareta</i>	<i>timareta</i>	8520	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T		
<i>H. timareta</i>	<i>timareta</i>	8523	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T		
<i>H. timareta</i>	<i>thelxinoe</i>	09-312	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. timareta</i>	<i>thelxinoe</i>	8624	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. timareta</i>	<i>thelxinoe</i>	8628	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. timareta</i>	<i>thelxinoe</i>	8631	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A		
<i>H. elevatus</i>		09-343	0	A/T	G/G	A/A	T/T	1	C/T	NA	C/C	T/T		
<i>H. pardalinus</i>	<i>sergestus</i>	09-326	0	A/A	A/A	A/A	NA	0	C/C	T/T	T/T	NA		

792

793 *downstream of cortex, †between exons 3 and 4 of cortex, ‡upstream of cortex, §between

794 exons U4 and U3 of cortex. None of these SNPs are within known TEs. Colours show

795 phenotypic associations: yellow = yellow hindwing bar; pink = no yellow hindwing bar;

796 green = yellow forewing band; blue = no yellow forewing band; grey = allele does not match
 797 expected pattern.

798

799 Extended Data Table 4. Transposable Elements (TEs) found within the Yb region.

Unique Occurrences					No.	TE name	Superfamily	Type
BAC	mel	ros	ama	agl				
1					1	BEL-1	BEL	LTR retrotransposon
					1	CR1-2	Jockey	Non-LTR retrotransposon
	1				1	Daphne-1	Jockey	Non-LTR retrotransposon
1					1	Daphne-6	Jockey	Non-LTR retrotransposon
1					1	DNA-like-8		DNA transposon
	1	2			1	Helitron-like-14	Helitron_A	DNA transposon
	2				4	Helitron-like-12	Helitron_A	DNA transposon
1					5	Helitron-like-12b	Helitron_A	DNA transposon
	1	1	1	1	7	Helitron-like-4a	Helitron_A	DNA transposon
						Helitron-like-4b	Helitron_A	DNA transposon
						Helitron-N2	Helitron_A	DNA transposon
					3	Helitron-like-7	Helitron_A	DNA transposon
5	3	3	1	2	16	Helitron-like-6a	Helitron_B	DNA transposon
						Helitron-like-6b	Helitron_B	DNA transposon
						Helitron-like-11	Helitron_B	DNA transposon
2	2	1		1	11	Helitron-like-15	Helitron_B	DNA transposon
6	5	3	1		18	Helitron-like-5	Helitron_B	DNA transposon
		1			2	Hmel_Unknown_50		
	1		1		2	Hmel_Unknown_174a/b		
	1				1	Hmel_Unknown_187b		
			1	1	2	Hmel_Unknown_230		
					1	Hmel_Unknown_234a		
					1	Hmel_Unknown_236a		
	1				1	Jockey-4	Jockey	Non-LTR retrotransposon
	1				1	LTR-3_gypsy	Gypsy	LTR retrotransposon
				1	1	Mariner-4	Mariner/Tc1	DNA transposon
1				3	29	Metulj-0	Metulj	SINE Non-LTR retrotransposon
						Metulj-1	Metulj	SINE Non-LTR retrotransposon
						Metulj-2	Metulj	SINE Non-LTR retrotransposon
						Metulj-3	Metulj	SINE Non-LTR retrotransposon
						Metulj-4	Metulj	SINE Non-LTR retrotransposon
						Metulj-5	Metulj	SINE Non-LTR retrotransposon
						Metulj-6	Metulj	SINE Non-LTR retrotransposon
						Metulj-7	Metulj	SINE Non-LTR retrotransposon
						nTc3-4	Mariner/Tc1	DNA transposon
						SINE-1	SINE	Non-LTR retrotransposon
1	1				2	nMar-3	Mariner/Tc1	DNA transposon
1					1	nMar-16	Mariner/Tc1	DNA transposon
			1		1	nMar-12/20	Mariner/Tc1	DNA transposon
				1	1	nPIF-3	PIF/Harbinger	DNA transposon
1					1	nTc3-2	Mariner/Tc1	DNA transposon
1					2	nTc3-3	Mariner/Tc1	DNA transposon
	1				2	R4-1	R2	Non-LTR retrotransposon
			1	1	6	Rep-1	REP	Non-LTR retrotransposon
				1	4	RTE-3	RTE	Non-LTR retrotransposon
2		1			2	RTE-11	RTE	Non-LTR retrotransposon
				1	3	Zenon-1	Jockey	Non-LTR retrotransposon
			1		1	Zenon-3	Jockey	Non-LTR retrotransposon

800