doi:10.1038/nature17961

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

Nicola J. Nadeau[1,2], Carolina Pardo-Diaz[3], Annabel Whibley[4,5], Megan A. Supple[2,6], Suzanne V. Saenko[4], Richard W. R. Wallbank[2,7], Grace C. Wu[8], Luana Maroja[9], Laura Ferguson[10], Joseph J. Hanly[2,7], Heather Hines[11], Camilo Salazar[3], Richard M. Merrill[2,7], Andrea J. Dowling[12], Richard H. ffrench-Constant[12], Violaine Llaurens[4], Mathieu Joron[4,13], W. Owen McMillan[2] & Chris D. Jiggins[2,7]

The wing patterns of butterflies and moths (Lepidoptera) are diverse and striking examples of evolutionary diversification by natural selection[1,2]. Lepidopteran wing colour patterns are a key innovation, consisting of arrays of coloured scales. We still lack a general understanding of how these patterns are controlled and whether this control shows any commonality across the 160,000 moth and 17,000 butterfly species. Here, we use fine-scale mapping with population genomics and gene expression analyses to identify a gene, *cortex*, that regulates pattern switches in multiple species across the mimetic radiation in *Heliconius* butterflies. *cortex* belongs to a fast-evolving subfamily of the otherwise highly conserved fizzy family of cell-cycle regulators[3], suggesting that it probably regulates pigmentation patterning by regulating scale cell development. In parallel with findings in the peppered moth (*Biston betularia*)[4], our results suggest that this mechanism is common within Lepidoptera and that *cortex* has become a major target for natural selection acting on colour and pattern variation in this group of insects.

In *Heliconius*, there is a major effect locus, *Yb*, that controls a diversity of colour pattern elements across the genus. It is the only locus in *Heliconius* that regulates all scale types and colours, including the diversity of white and yellow pattern elements in the two co-mimics *H. melpomene* and *H. erato*, and whole-wing variation in black, yellow, white, and orange/red elements in *H. numata*[5–7]. In addition, genetic variation underlying the *Bigeye* wing pattern mutation in *Bicyclus anynana*, melanism in the peppered moth, *Biston betularia*, and melanism and patterning differences in the silkmoth, *Bombyx mori*, have all been localized to homologous genomic regions[8–10] (Fig. 1). Therefore, this genomic region appears to contain one or more genes that act as major regulators of wing pigmentation and patterning across the Lepidoptera.

Previous mapping of this locus in *H. erato*, *H. melpomene* and *H. numata* identified a genomic interval of about 1 Mb (refs 11–13) (Extended Data Table 1), which also overlaps with the 1.4-Mb region containing the *carbonaria* locus in *B. betularia*[9] and a 100-bp non-coding region containing the *Ws* mutation in *B. mori*[10] (Fig. 1). We used a population genomics approach to identify the single nucleotide polymorphisms (SNPs) that were most strongly associated with phenotypic variation within the approximately 1-Mb *Heliconius* interval. The diversity of wing patterning in *Heliconius* arises from divergence at wing pattern loci[7], while convergent patterns generally involve the same loci and sometimes even the same alleles[14–16]. We used this pattern of divergence and sharing to identify SNPs associated with colour pattern elements across many individuals from a wide diversity of colour pattern phenotypes (Fig. 2).

In three separate *Heliconius* species, our analysis consistently implicated the gene *cortex* as being involved in adaptive differences in wing colour pattern. In *H. erato* the strongest associations with the presence of a yellow hindwing bar were centred around the genomic region

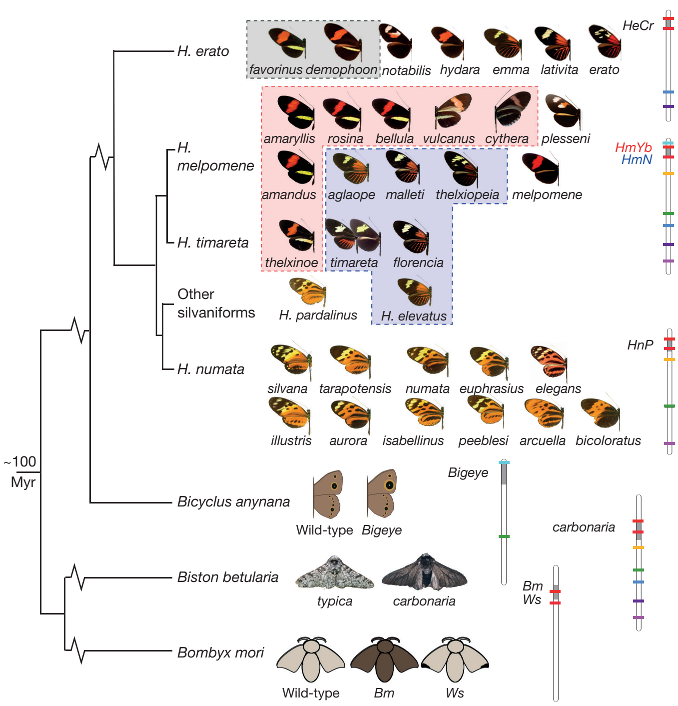


Figure 1 | A homologous genomic region controls a diversity of phenotypes across the Lepidoptera. Left, phylogenetic relationships[29]. Right, chromosome maps with colour pattern intervals in grey; coloured bars represent markers used to assign homology[5,8–10] and the first and last genes from Fig. 2 are shown in red. In *H. erato* the *HeCr* locus controls the yellow hindwing bar phenotype (grey boxed races). In *H. melpomene* it controls both the yellow hindwing bar (*HmYb*, pink box) and the yellow forewing band (*HmN*, blue box). In *H. numata* it modulates black, yellow and orange elements on both wings (*HnP*), producing phenotypes that mimic butterflies in the genus *Melinaea*. Morphs/races of *Heliconius* species included in this study are shown with names. All images are by the authors or are in the public domain.

[1]Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield, S10 2TN UK. [2]Smithsonian Tropical Research Institute, Apartado Postal 0843-00153, Panamá, República de Panamá. [3]Biology Program, Faculty of Natural Sciences and Mathematics, Universidad del Rosario, Cra. 24 No 63C-69, Bogotá D.C., 111221, Colombia. [4]Institut de Systématique, Evolution et Biodiversité (UMR 7205 CNRS, MNHN, UPMC, EPHE, Sorbonne Université), Museum National d'Histoire Naturelle, CP50, 57 rue Cuvier, 75005 Paris, France. [5]Cell and Developmental Biology, John Innes Centre, Norwich, Norfolk NR4 7UH, UK. [6]Research School of Biology, The Australian National University, 134 Linnaeus Way, Acton, ACT, 2601, Australia. [7]Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, UK. [8]Energy and Resources Group, University of California at Berkeley, California, 94720, USA. [9]Department of Biology, Williams College, Williamstown, Massachusetts 01267, USA. [10]Department of Zoology, University of Oxford, South Parks Rd, Oxford OX1 3PS, UK. [11]Penn State University, 517 Mueller, University Park, Pennsylvania 16802, USA. [12]School of Biosciences, University of Exeter in Cornwall, Penryn, Cornwall TR10 9FE, UK. [13]Centre d'Ecologie Fonctionnelle et Evolutive (CEFE, UMR 5175 CNRS, Université de Montpellier, Université Paul-Valéry Montpellier, EPHE), 1919 route de Mende, 34293 Montpellier, France.

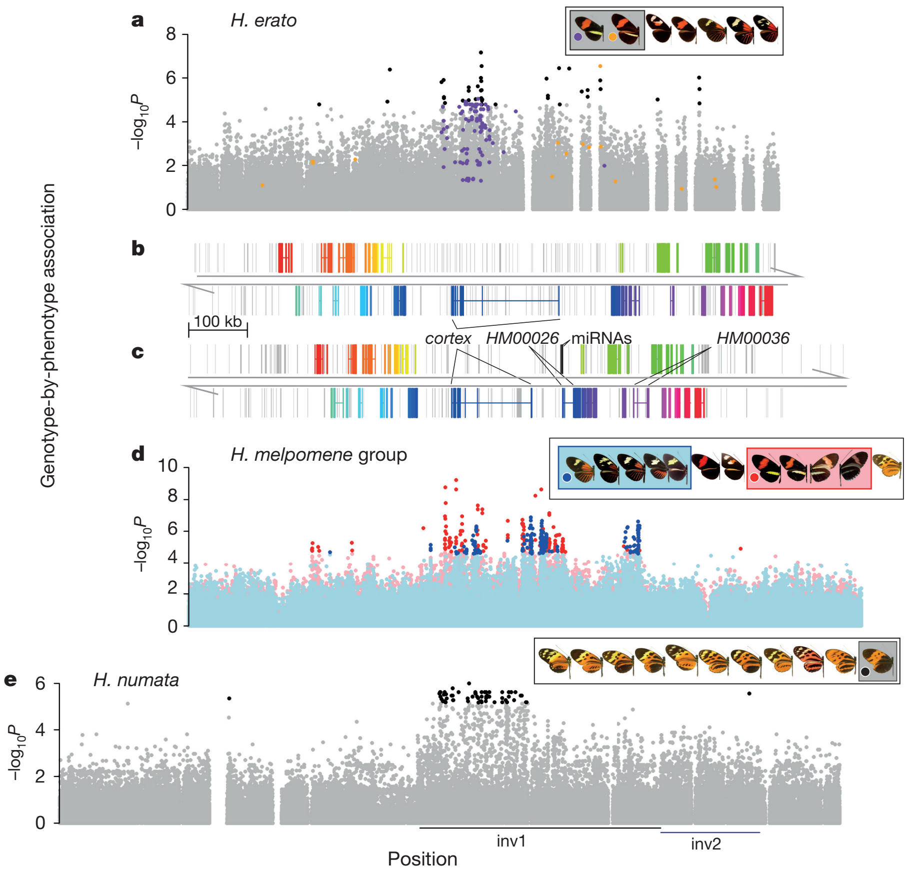


Figure 2 | Association analyses across the genomic region known to contain major colour pattern loci in *Heliconius*. **a**, Association in *H. erato* with the yellow hindwing bar ($n = 45$). Coloured SNPs are fixed for a unique state in *H. erato demophoon* (orange) or *H. erato favorinus* (purple). **b**, Genes in *H. erato* with direct homologues in *H. melpomene*. Genes are in different colours with exons (coding and UTRs) connected by lines. Grey bars are transposable elements. **c**, *H. melpomene* genes and transposable elements. Colours correspond to homologous *H. erato* genes and microRNAs[30] are black. **d**, Association in the *H. melpomene*/*timareta*/*silvaniform* group with the yellow hindwing bar (red) and yellow forewing band (blue) ($n = 49$). **e**, Association in *H. numata* with the *bicoloratus* morph ($n = 26$); inversion positions[13] shown below. In all cases black or dark coloured points are above the strongest associations found outside the colour pattern scaffolds (*H. erato* $P = 1.63 \times 10^{-5}$; *H. melpomene* $P = 2.03 \times 10^{-5}$ and $P = 2.58 \times 10^{-5}$ for hindwing bar and forewing band, respectively; *H. numata* $P = 6.81 \times 10^{-6}$). P values are from score tests for association.

containing *cortex* (Fig. 2a). We identified 108 SNPs that were fixed for one allele in *H. erato favorinus*, and fixed for the alternative allele in all individuals lacking the yellow bar; the majority of these SNPs were in introns of *cortex* (Extended Data Table 2). Fifteen SNPs showed a similar fixed pattern for *H. erato demophoon*, which also has a yellow bar. These SNPs did not overlap with those in *H. erato favorinus*, consistent with the hypothesis that this phenotype evolved independently in the two disjunct populations[17].

Previous work has suggested that alleles at the *Yb* locus are shared between *H. melpomene*, the closely related species *H. timareta*, and the more distantly related species *H. elevatus*, resulting in mimicry among these species[18]. Across these species, the strongest associations with the yellow hindwing bar phenotype were again found at *cortex* (Fig. 2d, Extended Data Fig. 1a and Extended Data Table 3). Similarly, the strongest associations with the yellow forewing band were found around the 5′ untranslated regions (UTRs) of *cortex* and *HM00036*, an orthologue of the *wash* gene in *Drosophila melanogaster*. A single SNP about 17 kb upstream of *cortex* (the closest gene) was perfectly associated with the yellow forewing band across all *H. melpomene*, *H. timareta* and *H. elevatus* individuals (Extended Data Figs 1a, 2 and Extended Data Table 3). We found no fixed coding sequence variants at *cortex* in larger samples (14–38 individuals) of *H. melpomene aglaope* and *H. melpomene amaryllis* (Extended Data Fig. 3 and Supplementary Information), which differ in *Yb*-controlled phenotypes[19], suggesting that functional variants are likely to be regulatory rather than coding. We found extensive transposable element variation around *cortex* but it is unclear whether any of these are associated with phenotypic differences (Extended Data Fig. 3, Extended Data Table 4 and Supplementary Information).

Finally, large inversions at the *P* supergene locus in *H. numata* (Fig. 1) are associated with different morphs[13]. There is a steep increase in genotype-by-phenotype association at the breakpoint of inversion 1, consistent with the role of these inversions in reducing recombination (Fig. 2e). However, the *bicoloratus* morph can recombine with all other morphs across one or the other inversion, permitting finer-scale association mapping of this region. As in *H. erato* and *H. melpomene*, this analysis showed a narrow region of associated SNPs corresponding exactly to the *cortex* gene (Fig. 2e), again with the majority of SNPs being found in introns (Extended Data Table 2). This associated region does not correspond to any other known genomic feature, such as an inversion or inversion breakpoint.

To determine whether sequence variants around *cortex* were regulating its expression, we investigated gene expression across the *Yb* locus. We used a custom designed microarray including probes from all predicted genes in the *H. melpomene* genome[18] as well as probes tiled across the central portion of the *Yb* locus, focusing on two naturally hybridizing *H. melpomene* races (*plesseni* and *malleti*) that differ in *Yb*-controlled phenotypes[7]. *cortex* was the only gene across the entire interval to show significant differences in expression both between races with different wing patterns (false discovery rate (FDR)

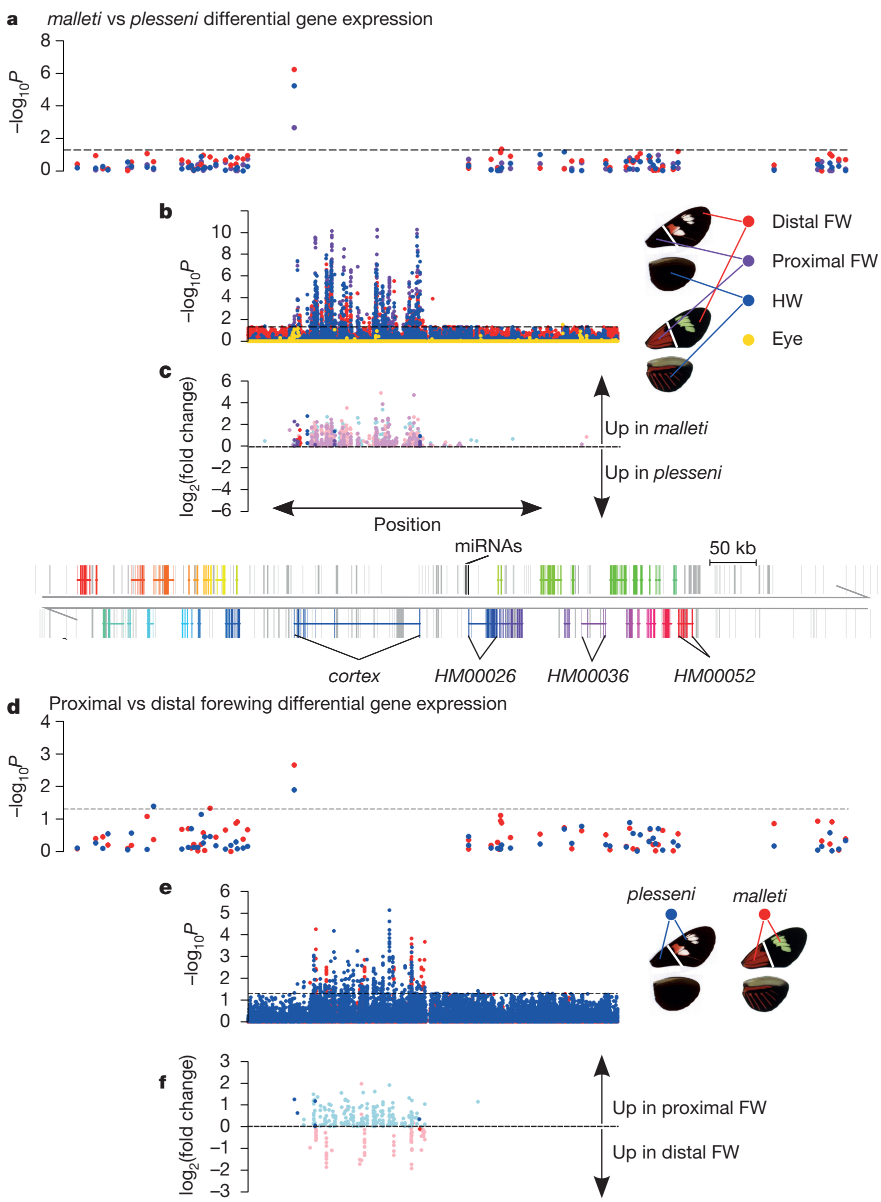


Figure 3 | Differential gene expression across the genomic region known to contain major colour pattern loci in *H. melpomene*. **a–f**, Expression differences in day 3 pupae, for all genes in the *Yb* interval (**a**, **d**) and tiling probes spanning the central portion of the interval (**b**, **c**, **e**, **f**). Expression is compared between races for each wing region (**a–c**) and between proximal and distal forewing sections for each race (**d–f**). **c**, **f**, Magnitude and direction of expression difference ($\log_2$ fold change) for tiling probes showing significant differences ($P \leq 0.05$); probes in known *cortex* exons shown in dark colours. Gene *HM00052* was differentially expressed between other races in RNA-seq data (Supplementary Information) but is not differentially expressed here. *P* values are based on FDR-adjusted *t*-statistics.

adjusted *t*-test $P = 6.09 \times 10^{-7}$) and between wing sections with different pattern elements (FDR adjusted *t*-test $P = 0.00224$; Fig. 3). This finding was reinforced in the tiled probe set, where we observed strong differences in the expression of *cortex* exons and introns but few differences outside this region (Extended Data Table 2). *cortex* expression was higher in *H. melpomene malleti* than in *H. melpomene plesseni* in all three wing sections used (but not eyes) (Fig. 3c and Extended Data Fig. 4c). When different wing sections were compared within each race, *cortex* expression in *H. melpomene malleti* was higher in the distal section that contains the *Yb*-controlled yellow forewing band than in the proximal section, consistent with *cortex* producing this band. In contrast, *H. melpomene plesseni*, which lacks the yellow band, had higher *cortex* expression in the proximal forewing section than in the distal section (Fig. 3f and Extended Data Fig. 4j). Differences in expression were found in pupal wings only on days 1 and 3 but not on days 5 or 7 (Extended Data Fig. 4), similar to the pattern observed previously for the transcription factor *optix*[20].

Differential expression was not confined to the exons of *cortex*; the majority of differentially expressed probes in the tiling array corresponded to *cortex* introns (Fig. 3). This differential expression of introns does not appear to be due to transposable element variation (Extended Data Table 2), but may be due to elevated background transcription and unidentified splice variants. PCR with reverse transcription (RT–PCR) revealed a diversity of splice variants (Extended Data Fig. 5), and their sequenced products included eight non-constitutive exons and six variable donor/acceptor sites, but we did not exhaustively sequence all transcripts (Supplementary Information). We cannot rule out the possibility that some of the differentially expressed intronic regions could be distinct non-coding RNAs. However, quantitative RT–PCR (qRT–PCR) in other hybridizing races with

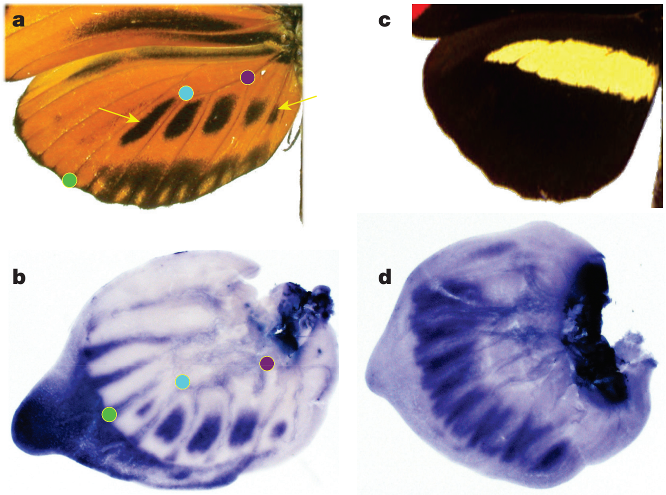


Figure 4 | *In situ* hybridizations of *cortex* in hindwings of final instar larvae. a, **b**, *H. numata tarapotensis* (replicated three times in the lab). Adult wing shown in **a**; coloured points indicate landmarks, yellow arrows highlight adult pattern elements corresponding to *cortex* staining. **c**, **d**, *H. melpomene rosina* (replicated twice in the lab). Adult wing shown in **c**; staining patterns in other *H. melpomene* races (*meriana*, $n = 11$, and *aglaope*, $n = 6$) appeared similar. The probe used was complementary to the *cortex* isoform with the longest open reading frame (also the most common; see Extended Data Fig. 5).

divergent *Yb* alleles (*aglaope*/*amaryllis* and *rosina*/*melpomene*) also identified differences in *cortex* expression and allele-specific splicing differences between both pairs of races (Extended Data Figs 1, 5 and Supplementary Information).

Finally, *in situ* hybridization of *cortex* in final instar larval hindwing discs showed expression in wing regions fated to become black in the adult wing, most strikingly in their correspondence to the black patterns on adult *H. numata* wings (Fig. 4). In contrast, the array results from pupal wings were suggestive of higher expression in non-melanic regions. This may suggest that *cortex* is upregulated at different timepoints in wing regions fated to become different colours.

Overall, *cortex* shows significant differential expression and is the only gene in the candidate region to be consistently differentially expressed in multiple race comparisons and between differently patterned wing regions. Coupled with the strong genotype-by-phenotype associations across multiple independent lineages (Extended Data Table 1), these findings strongly implicate *cortex* as a major regulator of colour and pattern. However, we have not excluded the possibility that other genes in this region also influence pigmentation patterning. A prominent role for *cortex* is also supported by studies in other taxa; our identification of distant 5′ untranslated exons of *cortex* (Supplementary Information) suggests that the 100-bp interval containing the *Ws* mutation in *B. mori* is likely to be within an intron of *cortex* and not in intergenic space, as previously thought[10]. In addition, fine mapping and gene expression also suggest that *cortex* controls melanism in the peppered moth[4].

It seems likely that *cortex* controls pigmentation patterning by controlling scale cell development. The *cortex* gene falls in an insect-specific lineage within the fzy (also known as Cdc20/fizzy) family of cell-cycle regulators (Extended Data Fig. 6a). The phylogenetic tree of this gene family highlighted three major orthologous groups, two of which have highly conserved functions in cell-cycle regulation, mediated through interaction with the anaphase-promoting complex/cyclosome (APC/C)[3,21]. The third group, containing cortex proteins, is evolving rapidly, with low amino acid identity between *D. melanogaster* and *H. melpomene* cortex (14.1%), contrasting with much higher identities for orthologues between these species in the other two groups (fzy, 47.8% and rap (also known as fzr, cdh1, rap/Fzr), 47.2%; Extended Data Fig. 6a). *D. melanogaster* cortex acts through a similar mechanism to fzy to control meiosis in the female germ line[22–24]. *H. melpomene* cortex also has some conservation of the fizzy family C-box and IR (isoleucine–arginine) tail elements (Supplementary Information) that mediate binding to the APC/C[23], suggesting that it may have retained a cell-cycle function, although we found that expressing *H. melpomene cortex* in *D. melanogaster* wings produced no detectable effect (Extended Data Fig. 6 and Supplementary Information).

Previously identified butterfly wing patterning genes have been transcription factors or signalling molecules[20,25]. Developmental rate has long been thought to play a role in lepidopteran patterning[26,27], but *cortex* was not a likely a priori candidate, because its *Drosophila* orthologue has a highly specific function in meiosis[23]. The recruitment of *cortex* to wing patterning appears to have occurred before the major diversification of the Lepidoptera and this gene has repeatedly been targeted by natural selection[1,7,9,28] to generate both cryptic[4] and aposematic patterns.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 July 2015; accepted 29 March 2016.

1. Cook, L. M., Grant, B. S., Saccheri, I. J. & Mallet, J. Selective bird predation on the peppered moth: the last experiment of Michael Majerus. *Biol. Lett.* **8,** 609–612 (2012).
2. Jiggins, C. D. Ecological speciation in mimetic butterflies. *Bioscience* **58,** 541–548 (2008).
3. Dawson, I. A., Roth, S. & Artavanis-Tsakonas, S. The *Drosophila* cell cycle gene fizzy is required for normal degradation of cyclins A and B during mitosis and has homology to the CDC20 gene of *Saccharomyces cerevisiae*. *J. Cell Biol.* **129,** 725–737 (1995).
4. van't Hof, A. E. *et al.* The industrial melanism mutation in British peppered moths is a transposable element. *Nature* http://dx.doi.org/10.1038/nature17951 (this issue).
5. Joron, M. *et al.* A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biol.* **4,** e303 (2006).
6. Sheppard, P. M., Turner, J. R. G., Brown, K. S., Benson, W. W. & Singer, M. C. Genetics and the evolution of Müllerian mimicry in *Heliconius* butterflies. *Phil. Trans. R. Soc. Lond. B* **308,** 433–610 (1985).
7. Nadeau, N. J. *et al.* Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. *Genome Res.* **24,** 1316–1333 (2014).
8. Beldade, P., Saenko, S. V., Pul, N. & Long, A. D. A. Gene-based linkage map for *Bicyclus anynana* butterflies allows for a comprehensive analysis of synteny with the lepidopteran reference genome. *PLoS Genet.* **5,** e1000366 (2009).
9. van't Hof, A. E., Edmonds, N., Dalíková, M., Marec, F. & Saccheri, I. J. Industrial melanism in British peppered moths has a singular and recent mutational origin. *Science* **332,** 958–960 (2011).
10. Ito, K. *et al.* Mapping and recombination analysis of two moth colour mutations, Black moth and Wild wing spot, in the silkworm *Bombyx mori*. *Heredity* **116,** 52–59 (2016).
11. Counterman, B. A. *et al.* Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in *Heliconius erato*. *PLoS Genet.* **6,** e1000796 (2010).
12. Ferguson, L. *et al.* Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the *HmYb/Sb* locus. *Mol. Ecol.* **19,** 240–254 (2010).
13. Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477,** 203–206 (2011).
14. Hines, H. M. *et al.* Wing patterning gene redefines the mimetic history of *Heliconius* butterflies. *Proc. Natl Acad. Sci. USA* **108,** 19666–19671 (2011).
15. Pardo-Diaz, C. *et al.* Adaptive introgression across species boundaries in *Heliconius* butterflies. *PLoS Genet.* **8,** e1002752 (2012).
16. Wallbank, R. W. R. *et al.* Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLoS Biol.* **14,** e1002353 (2016).
17. Maroja, L. S., Alschuler, R., McMillan, W. O. & Jiggins, C. D. Partial complementarity of the mimetic yellow bar phenotype in *Heliconius* butterflies. *PLoS ONE* **7,** e48627 (2012).
18. The *Heliconius* Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487,** 94–98 (2012).
19. Mallet, J. The genetics of warning colour in peruvian hybrid zones of *Heliconius erato* and *H. melpomene*. *Proc. R. Soc. Lond. B* **236,** 163–185 (1989).
20. Reed, R. D. *et al.* optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* **333,** 1137–1141 (2011).
21. Barford, D. Structural insights into anaphase-promoting complex function and mechanism. *Philos. Trans. R. Soc. B* **366,** 3605–3624 (2011).

22. Chu, T., Henrion, G., Haegeli, V. & Strickland, S. *Cortex*, a Drosophila gene required to complete oocyte meiosis, is a member of the Cdc20/fizzy protein family. *Genesis* **29,** 141–152 (2001).
23. Pesin, J. A. & Orr-Weaver, T. L. Developmental role and regulation of cortex, a meiosis-specific anaphase-promoting complex/cyclosome activator. *PLoS Genet.* **3,** e202 (2007).
24. Swan, A. & Schüpbach, T. The Cdc20/Cdh1-related protein, Cort, cooperates with Cdc20/Fzy in cyclin destruction and anaphase progression in meiosis I and II in *Drosophila*. *Development* **134,** 891–899 (2007).
25. Martin, A. *et al.* Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. *Proc. Natl Acad. Sci. USA* **109,** 12632–12637 (2012).
26. Koch, P. B., Lorenz, U., Brakefield, P. M. & ffrench-Constant, R. H. Butterfly wing pattern mutants: developmental heterochrony and co-ordinately regulated phenotypes. *Dev. Genes Evol.* **210,** 536–544 (2000).
27. Gilbert, L. E., Forrest, H. S., Schultz, T. D. & Harvey, D. J. Correlations of ultrastructure and pigmentation suggest how genes control development of wing scales of *Heliconius* butterflies. *J. Res. Lepid.* **26,** 141–160 (1988).
28. Mallet, J. & Barton, N. H. Strong natural selection in a warning-color hybrid zone. *Evolution* **43,** 421–431 (1989).
29. Wahlberg, N., Wheat, C. W. & Peña, C. Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). *PLoS ONE* **8,** e80875 (2013).
30. Surridge, A. K. *et al.* Characterisation and expression of microRNAs in developing wings of the neotropical butterfly *Heliconius melpomene*. *BMC Genomics* **12,** 62 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Saski for assembly of the *H. erato* BACs; M. Abanto and A. Tapia for assistance with raising butterflies; M. Chouteau, J. Morris and K. Dasmahapatra for providing larvae for *in situ* hybridizations; A. Morrison, R. Tetley, S. Carl and H. Wegener for assistance with laboratory work; S. Baxter for the *H. melpomene* fosmid libraries; and the governments of Colombia, Ecuador, Panama and Peru for permission to collect butterflies. This work was funded by a Leverhulme Trust award (RPG-2014-167), BBSRC (H01439X/1), ERC (SpeciationGenetics 339873), and NERC small project (MGF 280) grants to C.D.J., NSF grants (DEB 1257689, IOS 1052541) to W.O.M., an ERC starting grant (StG-243179) to M.J. and French National Agency for Research (ANR) grants to M.J. (ANR-12-JSV7-0005) and V.L. (ANR-13-JSV7-0003-01). N.J.N. is funded by a NERC fellowship (NE/K008498/1).

Author Contributions N.J.N. performed the association analyses, 5′ RACE, RT–PCR and qRT–PCR and prepared the manuscript. N.J.N. and C.D.J. co-ordinated the research. C.P.-D. performed and analysed the microarray and RNA-seq experiments. A.W. performed the *H. numata* association analysis. M.A.S. assembled and annotated the *HeCr* BAC reference and the *H. erato* alignments. S.V.S. performed *in situ* hybridizations. R.W.R.W. performed the transgenic experiments and analysis of *de novo* assembled sequences and fosmids together with J.J.H. G.C.W. and L.F. initially identified splicing variants of *cortex*. L.M. performed crosses between *H. melpomene* races. H.H. screened the *HeCr* BAC library. C.S. and R.M.M. provided samples. A.J.D. contributed to the *H. melpomene* BAC sequencing and annotation. R.H.f.-C., M.J., V.L., W.O.M. and C.D.J. are PIs who obtained funding and led the project elements. All authors commented on the manuscript.

Author Information Short read sequence data generated for this study are available from ENA (http://www.ebi.ac.uk/ena) under study accession PRJEB8011 and PRJEB12740 (see Supplementary Table 1 for previously published data accessions). The updated *Cr* contig is deposited in Genbank with accession KC469893.2. The assembled *H. melpomene* fosmid sequences are deposited in Genbank with accessions KU514430–KU514438. The microarray data are deposited in GEO with accessions GSM1563402–GSM1563497. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.J.N. (n.nadeau@sheffield.ac.uk) or C.D.J. (c.jiggins@zoo.cam.ac.uk).

METHODS

No statistical methods were used to predetermine sample size.

***H. erato Cr* reference.** *Cr* is the homologue of *Yb* in *H. erato* (Fig. 1). An existing reference for this region was available in three pieces[31] (467,734 bp, 114,741 bp and 161,149 bp; GenBank KC469893.1). We screened the same bacterial artificial chromosome (BAC) library used previously[11,31] using described procedures[11] with probes designed to the ends of the existing BAC sequences and the *HmYb* BAC reference sequence. Two BACs (04B01 and 10B14) were identified as spanning one of the gaps and sequenced using Illumina 2 × 250-bp paired-end reads collected on the Illumina MiSeq. The raw reads were screened to remove vector and *Escherichia coli* bases. The first 50,000 read pairs were taken for each BAC and assembled individually with the Phrap[32] software and manually edited with consed[33]. Contigs with discordant read pairs were manually broken and properly merged using concordant read data. Gaps between contig ends were filled using an in-house finishing technique in which the terminal 200 bp of the contig ends were extracted and queried against the unused read data for spanning pairs, which were added using the addSolexaReads.perl script in the consed package. Finally, a single reference contig was generated by identifying and merging overlapping regions of the two consensus BAC sequences.

To fill the remaining gap (between positions 800,387 and 848,446) we used the overhanging ends to search the scaffolds from a preliminary *H. erato* genome assembly of five Illumina paired-end libraries with different insert sizes (250, 500, 800, 4,300 and 6,500 bp) from two related *H. erato demophoon* individuals. We identified two scaffolds (scf1869 and scf1510) that overlapped and spanned the gap (using 12,257 bp of the first scaffold and 35,803 bp of the second).

The final contig was 1,009,595 bp in length, of which 2,281 bp were unknown (N). The *HeCr* assembly was verified by aligning to the *HmYb* genome scaffold (HE667780) with mummer and blast. The *HeCr* contig was annotated as described previously[31] with some minor modifications. Briefly this involved first generating a reference-based transcriptome assembly with existing *H. erato* RNA-sequenced (RNA-seq) wing tissue (GenBank accession SRA060220). We used Trimmomatic[34] (v0.22), and FLASh[35] (v1.2.2) to prepare the raw sequencing reads, checking the quality with FastQC[36] (v0.10.0). We then used the Bowtie/TopHat/Cufflinks[37–39] pipeline to generate transcripts for the unmasked reference sequence. We generated gene predictions with the MAKER pipeline[40] (v2.31). Homology and synteny in gene content with the *H. melpomene Yb* reference were identified by aligning the *H. melpomene* coding sequences to the *H. erato* reference with BLAST. Homologous genes were present in the same order and orientation in *H. erato* and *H. melpomene* (Fig. 2b, c). Annotations were manually adjusted if genes had clearly been merged or split in comparison to *H. melpomene* (which has been extensively manually curated[12]). In addition, *H. erato cortex* was manually curated from the RNA-seq data and using *Exonerate*[41] alignments of the *H. melpomene* protein and mRNA transcripts, including the 5′ UTRs.

Genotype-by-phenotype association analyses. Information on the individuals used and ENA accessions for sequence data are given in Supplementary Table 1. We used shotgun Illumina sequence reads from 45 *H. erato* individuals from 7 races that were generated as part of a previous study[31] (Supplementary Information). Reads were aligned to an *H. erato* reference containing the *Cr* contig and other sequenced *H. erato* BACs[11,31] with BWA[42], which has previously been found to work better than Stampy[43] (which was used for the alignments in the other species) with an incomplete reference sequence[31]. The parameters used were as follows: maximum edit distance (n), 8; maximum number of gap opens (o), 2; maximum number of gap extensions (e), 3; seed (l), 35; maximum edit distance in seed (k), 2. We then used Picard tools to remove PCR and optical duplicate sequence reads and GATK[44] to re-align indels and call SNPs using all individuals as a single population. Expected heterozygosity was set to 0.2 in GATK. 132,397 SNPs were present across *Cr*. A further 52,698 SNPs not linked to colour pattern loci were used to establish background association levels.

For the *H. melpomene*/*H. numata* clade we used previously published sequence data from 19 individuals from enrichment sequencing targeting the *Yb* region, the unlinked *HmB*/*D* region that controls the presence or absence of red colour pattern elements, and ~1.8 Mb of non-colour pattern genomic regions[45], as well as 9 whole-genome shotgun-sequenced individuals[18,46]. We added targeted sequencing and shotgun whole-genome sequencing of an additional 47 individuals (Supplementary Information). Alignments were performed using Stampy[43] with default parameters except for substitution rate which was set to 0.01. We again removed duplicates and used GATK to re-align indels and call SNPs with expected heterozygosity set to 0.1.

The analysis of *H. melpomene*/*timareta*/silvaniform included 49 individuals, which were aligned to v1.1 of the *H. melpomene* reference genome with the scaffolds containing *Yb* and *HmB*/*D* swapped with reference BAC sequences[18], which contained fewer gaps of unknown sequence than the genome scaffolds. The *Yb* region contained 232,631 SNPs and a further 370,079 SNPs were used to establish background association levels.

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

We measured associations between genotype and phenotype using a score test (qtscore) in the GenABEL package in R (ref. 47). This was corrected for background population structure using a test specific inflation factor (λ) calculated from the SNPs unlinked to the major colour pattern controlling loci (described above), as the colour pattern loci are known to have a different population structure from the rest of the genome[14,15,18]. We used a custom perl script to convert GATK vcf files to Illumina SNP format for input to GenABEL[47]. GenABEL does not accept multiallelic sites, so the script also converted the genotype of any individuals for which a third (or fourth) allele was present to a missing genotype (with these defined as the lowest frequency alleles). Custom R scripts were used to identify sites showing perfect associations with calls for >75% of individuals.

Microarray gene expression analyses. We designed a Roche NimbleGen microarray (12 × 135K format) with probes for all annotated *H. melpomene* genes[18] and tiling of the central portion of the *Yb* BAC sequence contig that was previously identified as showing the strongest differentiation between *H. melpomene* races[45]. In addition to the *HmYb* tiling array probes there were 6,560 probes tiling *HmAc* (a third unlinked colour pattern locus) and 10,716 probes tiling *HmB*/*D*, again distanced on average at 10-bp intervals. The whole-genome gene expression array contained 107,898 probes in total.

This array was interrogated with Cy3-labelled double-stranded cDNA generated from total RNA (with a SuperScript double-stranded cDNA synthesis kit (Invitrogen) and a one-colour DNA labelling kit (Niblegen)) from four pupal developmental stages of *H. melpomene plesseni* and *malleti*. Pupae were from captive stocks maintained in insectary facilities in Gamboa, Panama. Tissue was stored in RNA later (Ambion) at −80 °C before RNA extraction. RNA was extracted using TRIzol (Invitrogen) followed by purification with RNeasy (Qiagen) and DNase treated with DNA-free (Ambion). Quantification was performed using a Qubit 2.0 fluorometer (Invitrogen) and purity and integrity assessed using a Bioanalyzer 2100 (Agilent). Samples were randomized and each hybridized to a separate array. The *HmYb* probe array contained 9,979 probes distanced on average at 10 bp. The whole-genome expression array contained on average 9 probes per annotated gene in the genome (v.1.1 (ref. 18)) as well as any transcripts not annotated but predicted from RNA-seq evidence.

Background corrected expression values for each probe were extracted using NimbleScan software (v.2.3). Analyses were performed with the LIMMA package implemented in R/Bioconductor[48]. The tiling array and whole-genome data sets were analysed separately. Expression values were extracted and quantile-normalized, $\log_2$-transformed, quality controlled and analysed for differences in expression between individuals and wing regions. *P* values were adjusted for multiple hypothesis testing using the false discovery rate (FDR) method[49].

***In situ* hybridization.** *H. numata* and *H. melpomene* larvae were reared in a greenhouse at 25–30 °C and sampled at the last instar. *In situ* hybridizations were performed according to previously described methods[25] with a *cortex* riboprobe synthesized from a 831-bp cDNA amplicon from *H. numata*. Wing discs were incubated in a standard hybridization buffer containing the probe for 20–24 h at 60 °C. For secondary detection of the probe, wing discs were incubated in a 1:3,000 dilution of anti-digoxigenin alkaline phosphatase Fab fragments and stained with BM Purple for 3–6 h at room temperature. Stained wing discs were photographed with a Leica DFC420 digital camera mounted on a Leica Z6 APO stereomicroscope.

***De novo* assembly of short read data in *H. melpomene* and related taxa.** To better characterize indel variation from the short-read sequence data used for the genotype-by-phenotype association analysis, we performed *de novo* assemblies of a subset of *H. melpomene* individuals and related taxa with a diversity of phenotypes (Extended Data Fig. 2). Assemblies were performed using the *de novo* assembly function of CLCGenomics Workbench v.6.0 under default parameters. The assembled contigs were then BLASTed against the *Yb* region of the *H. melpomene melpomene* genome[18], using Geneious v.8.0. The contigs identified by BLAST were then concatenated to generate an allele sequence for each individual. Occasionally two unphased alleles were generated when two contigs were matched to a given region. If more than two contigs of equal length matched then this was considered an unresolvable repeat region and replaced with Ns. The assembled alleles were then aligned using the MAFFT alignment plugin in Geneious v.8.0.

Long-range PCR targeted sequencing of *cortex* in *H. melpomene aglaope* and *H. melpomene amaryllis*. We generated two long-range PCR products covering

88.8% of the 1,344-bp coding region of *cortex* (excluding 67 bp at the 5′ end and 83bp at the 3′ end; see Supplementary Information). A product spanning coding exons 5–9 (the final exon) was obtained from 29 *H. melpomene amaryllis* individuals and 29 *H. melpomene aglaope* individuals; a product spanning coding exons 2–5 was obtained from 32 *H. melpomene amaryllis* individuals and 14 *H. melpomene aglaope* individuals. In addition, a product spanning exons 4–6 was obtained from six *H. melpomene amaryllis* and five *H. melpomene aglaope* individuals that failed to amplify one or both of the larger products. Long-range PCR was performed using Extensor long-range PCR mastermix (Thermo Scientific) following the manufacturer's guidelines with a 60 °C annealing temperature in a 10–20-μl volume. The product spanning coding exons 5–9 was obtained with primers HM25_long_F1 and HM25_long_R4 (see Supplementary Table 2 for primer sequences); the product spanning coding exons 2–5 was obtained with primers HM25_long_F4 and HM25_long_R2; the product spanning exons 4–6 was obtained with primers 25_ex5-ex7_r1 and 25_ex5-ex7_f1. Products were pooled for each individual, including five additional products from the *Yb* locus and seven products in the region of the *HmB*/*D* locus. They were then cleaned using QIAquick PCR purification kit (QIAgen) before being quantified with a Qubit Fluorometer (Life Technologies) and pooled in equimolar amounts for each individual, taking into account variation in the length and number of PCR products included for each individual (because of some PCR failures, that is, proportionally less DNA was included if some PCR products were absent for a given individual).

Products were pooled within individuals (including additional products for other genes not analysed here) and then quantified and pooled in equimolar amounts for each individual within each race. The pooled products for each race (*H. melpomene aglaope* and *amaryllis*) were then prepared as two separate libraries with molecular identifiers and sequenced on a single lane of an Illumina GAIIx. Analysis was performed using Galaxy and the history is available at https://usegalaxy.org/u/njnadeau/h/long-pcr-final. Reads were quality filtered with a minimum quality of 20 required over 90% of the read, which resulted in 5% of reads being discarded. Reads were then quality trimmed to remove bases with quality less than 20 from the ends. They were then aligned to the target regions using the fosmid sequences from known races[45] with sequence from the *Yb* BAC walk[12] used to fill any gaps. Alignments were performed with BWA v.0.5.6 (ref. 42) and converted to pileup format using Samtools v.0.1.12 before being filtered on the basis of quality (≥20) and coverage (≥10). BWA alignment parameters were as follows: fraction of missing alignments given 2% uniform base error rate (aln -n) 0.01; maximum number of gap opens (aln -o) 2; maximum number of gap extensions (aln -e) 12; disallow long deletion within 12 bp towards the 3′-end (aln -d); number of first subsequences to take as seed (aln -l) 100. We then calculated coverage and minor allele frequencies for each race and the difference between these using custom scripts in R[50].

Sequencing and analysis of *H. melpomene* fosmid clones. Fosmid libraries had previously been made from single individuals of three *H. melpomene* races (*rosina*, *amaryllis* and *aglaope*) and several clones overlapping the *Yb* interval had been sequenced[45]. We extended the sequencing of this region, particularly the region overlapping *cortex*, by sequencing an additional four clones from *H. melpomene rosina* (1051_83D21, accession KU514430; 1051_97A3, accession KU514431; 1051_65N6, accession KU514432; 1051_93D23, accession KU514433), two clones from *H. melpomene amaryllis* (1051_13K4, accession KU514434; 1049_8P23, accession KU514435) and three clones from *H. melpomene aglaope* (1048_80B22, accession KU514437; 1049_19P15, accession KU514436; 1048_96A7, accession KU514438). These were sequenced on a MiSeq 2000, and assembled using the *de novo* assembly function of CLCGenomcs Workbench v.6.0. The individual clones (including existing clones 1051-143B3, accession FP578990; 1049-27G11, accession FP700055; and 1048-62H20, accession FP565804) were then aligned to the BAC and genome scaffold[18] references using the MAFFT alignment plugin of Geneious v.8.0. Regions of general sequence similarity were identified and visualized using MAUVE[51]. We merged overlapping clones from the same individual if they showed no sequence differences, indicating that they came from the same allele. We identified transposable elements using nBLAST with an insect transposable element list downloaded from Repbase Update[52], including known *Heliconius*-specific transposable elements[53].

5′ RACE, RT–PCR and qRT–PCR. All tissues used for gene expression analyses were dissected from individuals from captive stocks derived from wild-caught individuals of various races of *H. melpomene* (*aglaope*, *amaryllis*, *melpomene*, *rosina*, *plesseni* and *malleti*) and F_2 individuals from a *H. melpomene rosina* (female) × *H. melpomene melpomene* (male) cross. Experimental individuals were reared at 28–31 °C. Developing wings were dissected and stored in RNAlater (Ambion Life Technologies). RNA was extracted using a QIAgen RNeasy Mini kit following the manufacturer's guidelines and treated with TURBO DNA-free DNase kit (Ambion Life Technologies) to remove remaining genomic DNA. RNA quantification was performed with a Nanodrop spectrophotometer, and the RNA integrity was assessed using the Bioanalyzer 2100 system (Agilent).

Total RNA was thoroughly checked for DNA contamination by performing PCR for EF1α (using primers ef1-a_RT_for and ef1-a_RT_rev, Supplementary Table 2) with 0.5 μl of RNA extract (50 ng–1 μg of RNA) in a 20-μl reaction using a polymerase enzyme that is not functional with RNA template (BioScript, Bioline Reagents Ltd). If a product amplified within 45 cycles then the RNA sample was re-treated with DNase.

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

5′ RACE (rapid amplification of cDNA ends) was performed using RNA from hindwing discs from one *H. melpomene aglaope* and one *H. melpomene amaryllis* final instar larvae with a SMARTer RACE kit from Clonetech. The gene-specific primer used for the first round of amplification was anchored in exon 4 (fzl_raceex5_R1; Supplementary Table 2). Secondary PCR of these products was then performed using a primer in exon 2 (HM25_long_F2; Supplementary Table 2) and the nested universal primer A. Other isoforms were detected by RT–PCR using primers within exons 2 and 9 (gene25_for_full1 and gene25_rev_ex3). We identified isoforms from 5′ RACE and RT–PCR products by cutting individual bands from agarose gels and if necessary by cloning products before Sanger sequencing. Cloning of products was performed using TOPO TA (Invitrogen) or pGEM-T (Promega) cloning kits. Sanger sequencing was performed using BigDye terminator v3.1 (Applied Biosystems) run on an ABI13730 capillary sequencer. Primers fzl_ex1a_F1 and fzl_ex4_R1 were used to confirm expression of the furthest 5′ UTR. For isoforms that appeared to show some degree of race specificity, we designed isoform-specific PCR primers spanning specific exon junctions (Extended Data Figs 2, 4 and Supplementary Table 2) and used these to either qualitatively (RT–PCR) or quantitatively (qRT–PCR) assess differences in expression between races.

We performed qRT–PCR using SensiMix SYBR green (Bioline Reagents Ltd) with 0.2–0.25 μM of each primer and 1 μl of the diluted product from the cDNA reactions. Reactions were performed in an Opticon 2 DNA engine (MJ Research) with the following cycling parameters: 95 °C for 10 min; 35–50 × (95 °C for 15 s, 55–60 °C for 30 s, 72 °C for 30 s); 72 °C for 5 min. Melting curves were generated between 55 °C and 90 °C with readings taken every 0.2 °C for each of the products to check that a single product was generated. At least one product from each set of primers was also run on a 1% agarose gel to check that a single product of the expected size was produced and the identity of the product was confirmed by direct sequencing (see Supplementary Table 2 for details of primers for each gene). We used two housekeeping genes (*EF1A* and *RPS3A*) for normalization and all results were taken as averages of triplicate PCR reactions for each sample.

C_t values were defined as the point at which fluorescence crossed a threshold (R_{Ct}) adjusted manually to be the point at which fluorescence rose above the background level. Amplification efficiencies (E) were calculated using a dilution series of clean PCR products. Starting fluorescence, which is proportional to the starting template quantity, was calculated as $R_0 = R_{Ct}(1+E)^{-Ct}$. Normalized values were then obtained by dividing R_0 values for the target loci by R_0 values for *EF1α* and *RPS3A*. Results from both of these controls were always very similar, so the results presented are normalized to the mean of *EF1A* and *RPS3A*. All results were taken as averages of triplicate PCR reactions. If one of the triplicate values was more than one cycle away from the mean then this replicate was excluded. Similarly any individuals that were more than two s.d. away from the mean of all individuals for the target or normalization genes were excluded (these are not included in the numbers of individuals reported). Statistical significance was assessed by Wilcoxon rank sum tests performed in R (ref. 50).

RNA-seq analysis of *H. melpomene amaryllis*/*aglaope*. RNA-seq data for hindwings from three developmental stages had previously been obtained for two individuals of each race at each stage (12 individuals in total) and used in the annotation of the *H. melpomene* genome[18] (deposited in ENA under study accessions ERP000993 and PRJEB7951). Four samples were multiplexed on each sequencing lane with the fifth instar larval and day 2 pupal samples sequenced on a GAIIx sequencer and the day 3 pupal wings sequenced on a Hiseq 2000 sequencer.

Two methods were used for alignment of reads to the reference genome and inferring read counts: Stampy[43] and RSEM (RNaseq by Expectation Maximisation)[54]. In addition we used two different R/Bioconductor packages for estimation of differential gene expression: DESeq[55] and BaySeq[56]. Read bases with quality scores <20 were trimmed with FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Stampy was run with default parameters except for mean insert size, which was set to 500, s.d. 100, and substitution rate, which was set to

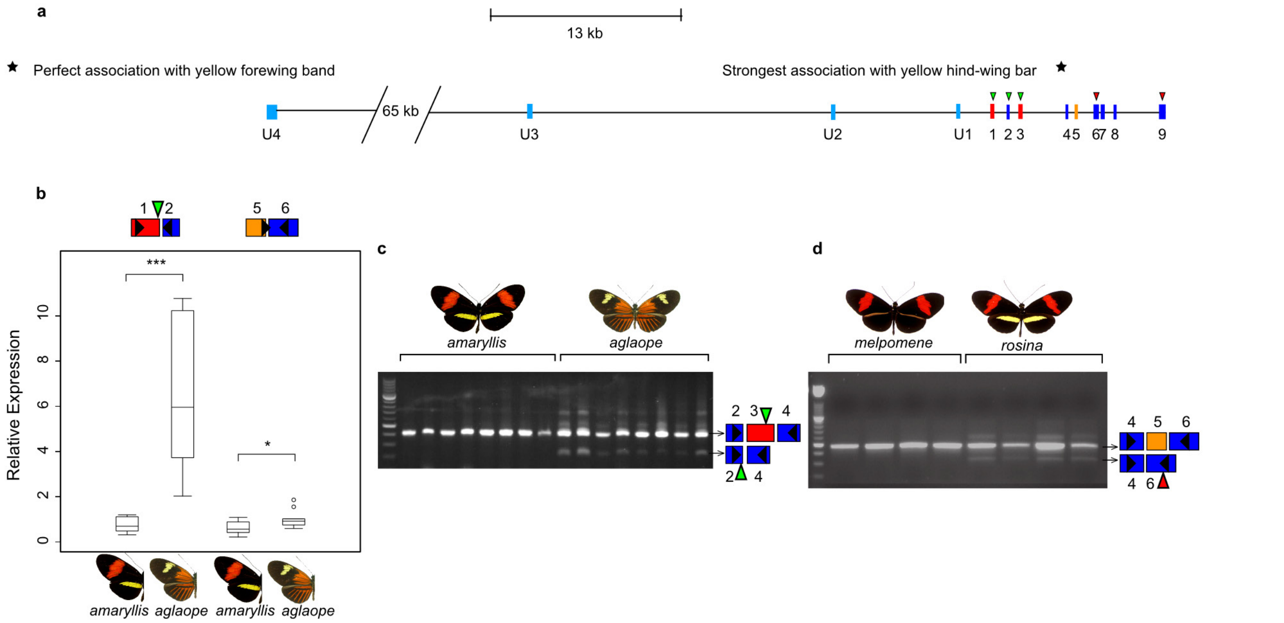
0.01. Alignments were filtered to exclude reads with mapping quality <30 and sorted using Samtools[57]. We used the HT seq-count script with HTseq[58] to infer counts per gene from the BAM files.

RSEM[54] was run with default parameters to infer a transcriptome and then map RNA-seq reads against this using Bowtie[37] as an aligner. This was run with default parameters except for the maximum number of mismatches, which was set to 3.

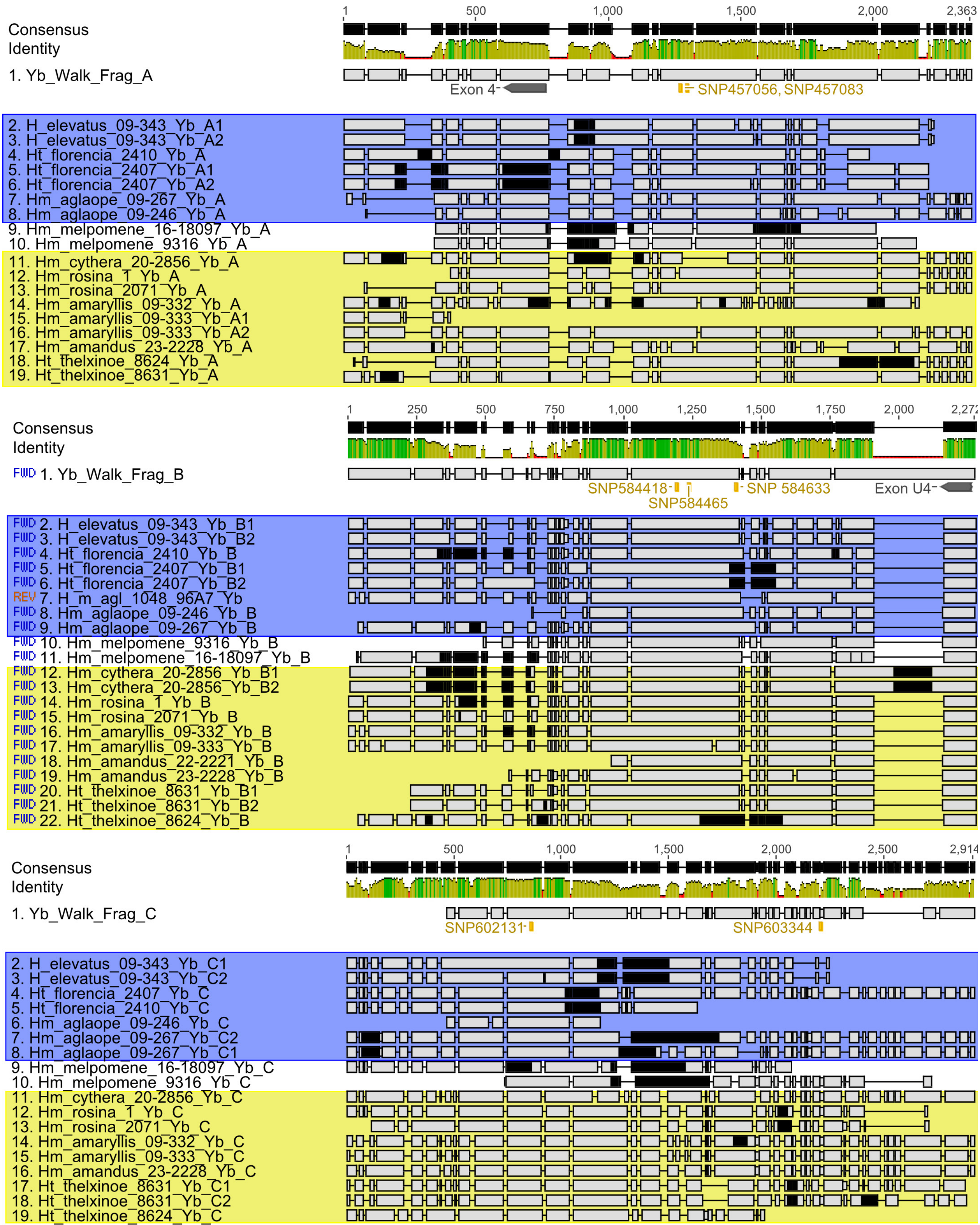
Annotation and alignment of fizzy family proteins. In the arthropod genomes, some fizzy family proteins were found to be poorly annotated based on alignments to other family members. In these cases annotations were improved using well-annotated proteins from other species as references in the program Exonerate[41] and the outputs were manually curated. Specifically, the annotation of *B. mori rap* (also known as *fzr*) was extended based on alignment of *Danaus plexippus rap*; the annotation of *B. mori fzy* was altered based on alignment of *D. melanogaster* and *D. plexippus fzy*; *H. melpomene fzy* was identified as part of the annotated gene *HMEL017486* on scaffold HE671623 (Hmel v.1.1) based on alignment of *D. plexippus fzy;* the *Apis mellifera rap* annotation was altered based on alignment of *D. melanogaster rap*; the annotation of *Acyrthosiphon pisum rap* was altered based on alignment of *D. melanogaster rap.* No one-to-one orthologues of *D. melanogaster fzr2* were found in any of the other arthropod genera, suggesting that this gene is *Drosophila*-specific. Multiple sequence alignment of all the fizzy family proteins was then performed using the Expresso server[59] within T-coffee[60], and this alignment was used to generate a neighbour joining tree in Geneious v.8.1.7.

Expression of *H. melpomene cortex* in *D. melanogaster* wings. *D. melanogaster cortex* is known to generate an irregular microchaete phenotype when ectopically expressed in the posterior compartment of the adult fly wing[24]. We performed the same assay using *H. melpomene cortex* to test whether this functionality was conserved. Following the methods of Swan and Schüpbach[24], we created an upstream activating sequence (UAS)–GAL4 construct using the coding region for the long isoform of *H. melpomene cortex*, plus a *Drosophila cortex* version to act as positive control. The haemagglutinin (HA)-tagged *H. melpomene* UAS-cortex expression construct was generated using cDNA reverse transcribed (Revert-Aid, Thermo-Scientific) from RNA extracted (Qiagen RNeasy) from pre-ommochrome pupal wing material. An HA-tagged *D. melanogaster* UAS–cortex version was also constructed[24]. Expression was driven by the *hsp70* promoter. Constructs were injected into ϕC31-attP40 flies (25709, Bloomington Stock Centre; Cambridge University Genetics Department, UK, fly injection service) by site-directed insertion into CII via an attB site in the construct. Homozygous transgenic flies were crossed with *w*,*y*′;*en–GAL4*;UAS–*GFP* flies (gift from M. Landgraf laboratory, Cambridge University Zoology Department) to drive expression in the engrailed posterior domain of the wing, and adult offspring wings were photographed (Extended Data Fig. 6b–d). Expression of the construct was confirmed by immunohistochemistry (using the standard *Drosophila* protocol) against an HA tag inserted at the N terminus of the protein, using final instar larval wing discs with mouse anti-HA and goat anti-mouse alexa-fluor 568 secondary antibodies (Abcam), imaged by Leica SP5 confocal (Extended Data Fig. 6e).

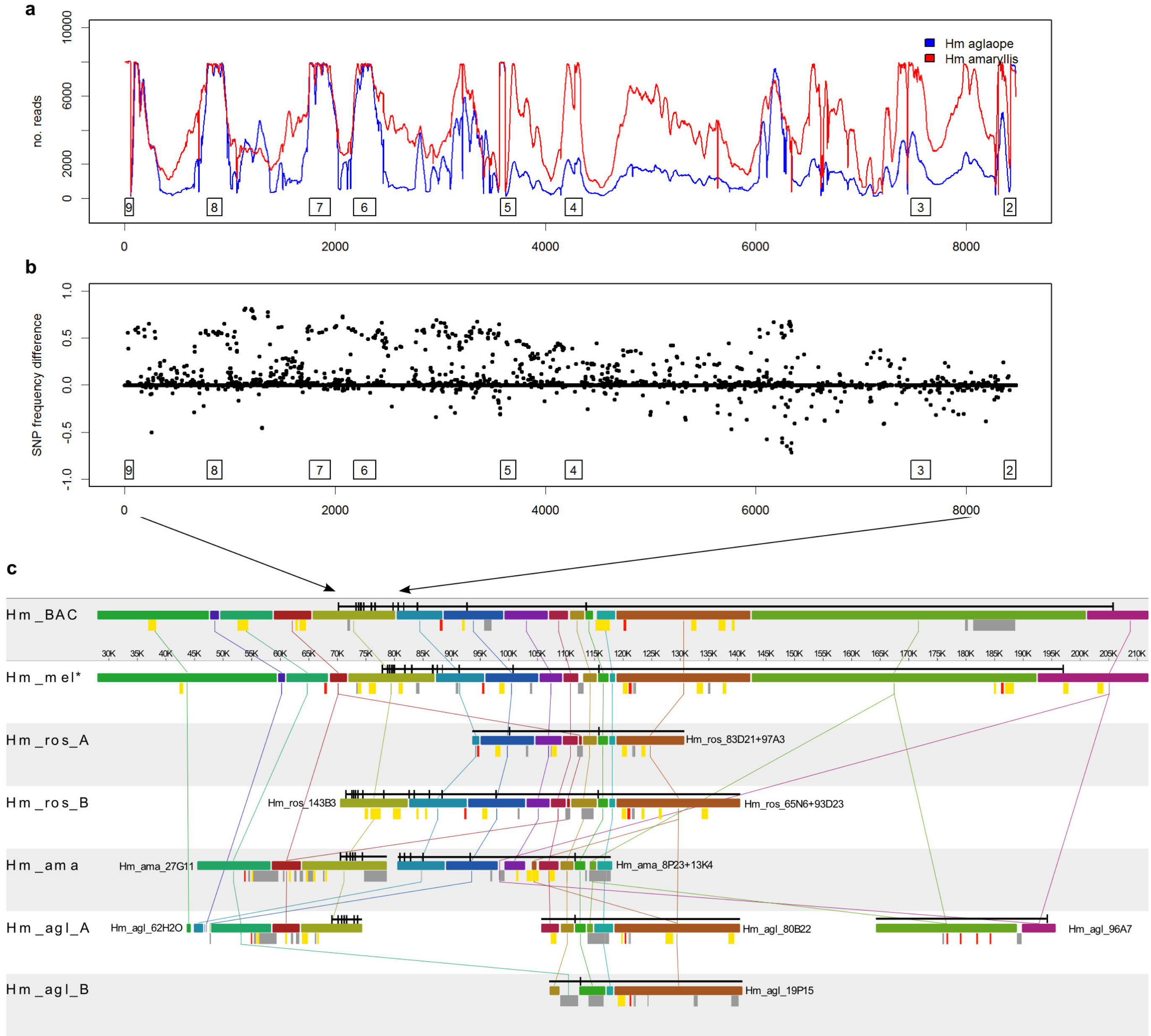
31. Supple, M. A. *et al.* Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. *Genome Res.* **23,** 1248–1257 (2013).
32. de la Bastide, M. & McCombie, W. R. Assembling genomic DNA sequences with PHRAP. *Curr. Protoc. Bioinformatics* **11,** 11.4 (2007).
33. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8,** 195–202 (1998).
34. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
35. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27,** 2957–2963 (2011).
36. Andrews, S. *FastQC* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2011).
37. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
38. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
39. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).
40. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12,** 491 (2011).
41. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
43. Lunter, G. & Goodson, M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21,** 936–939 (2011).
44. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
45. Nadeau, N. J. *et al.* Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Phil. Trans. R. Soc. B* **367,** 343–353 (2012).
46. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* **23,** 1817–1828 (2013).
47. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23,** 1294–1296 (2007).
48. Smyth, G. K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397–420 (Springer, 2005).
49. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).
50. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2011).
51. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14,** 1394–1403 (2004).
52. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–467 (2005).
53. Lavoie, C. A., Platt, R. N., Novick, P. A., Counterman, B. A. & Ray, D. A. Transposable element evolution in *Heliconius* suggests genome diversity within Lepidoptera. *Mob. DNA* **4,** 21 (2013).
54. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
55. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
56. Hardcastle, T. J. & Kelly, K. A. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11,** 422 (2010).
57. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
58. Anders, S., Pyl, P. T. & Huber, W. HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
59. Armougom, F. *et al.* Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* **34,** W604–W608 (2006).
60. Di Tommaso, P. *et al.* T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* **39,** W13–17 (2011).



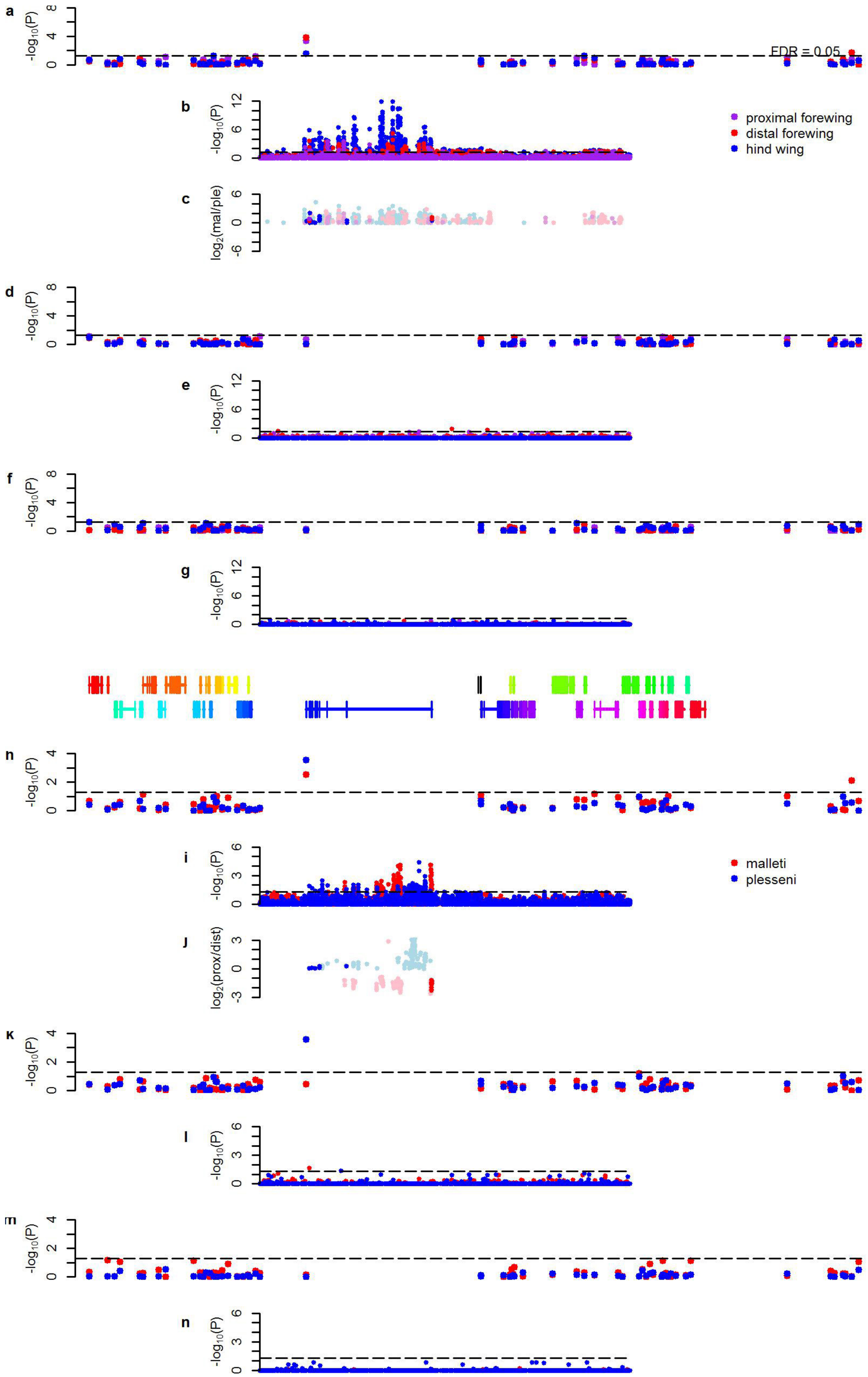
Extended Data Figure 1 | *H. melpomene* race-associated *cortex* splicing variation. **a**, Exons and splice variants of *cortex* in *H. melpomene*. Orientation is reversed with respect to Figs 2 and 4, with transcription going from left to right. SNPs showing the strongest associations with phenotype are shown with stars. **b**, Differential expression of two regions of *cortex* between whole hindwings of *H. melpomene amaryllis* and *H. melpomene aglaope* ($n = 11$ and $n = 10$, respectively). Box plots are standard (median; seventy-fifth and twenty-fifth percentiles; maximum and minimum excluding outliers (shown as discrete points)). ***$P < 0.0001$, *$P < 0.05$, Wilcoxon rank sum test. **c**, Expression of a *cortex* isoform lacking exon 3 is found in *H. melpomene aglaope* but not *H. melpomene amaryllis* hindwings. **d**, Expression of an isoform lacking exon 5 is found in *H. melpomene rosina* but not *H. melpomene melpomene* hindwings. Green triangles indicate predicted start codons and red triangles predicted stop codons, with usage dependent on which exons are present in the isoform. Schematics of the targeted exons are shown for each (q)RT–PCR product; black triangles indicate the positions of the primers used in the assay.



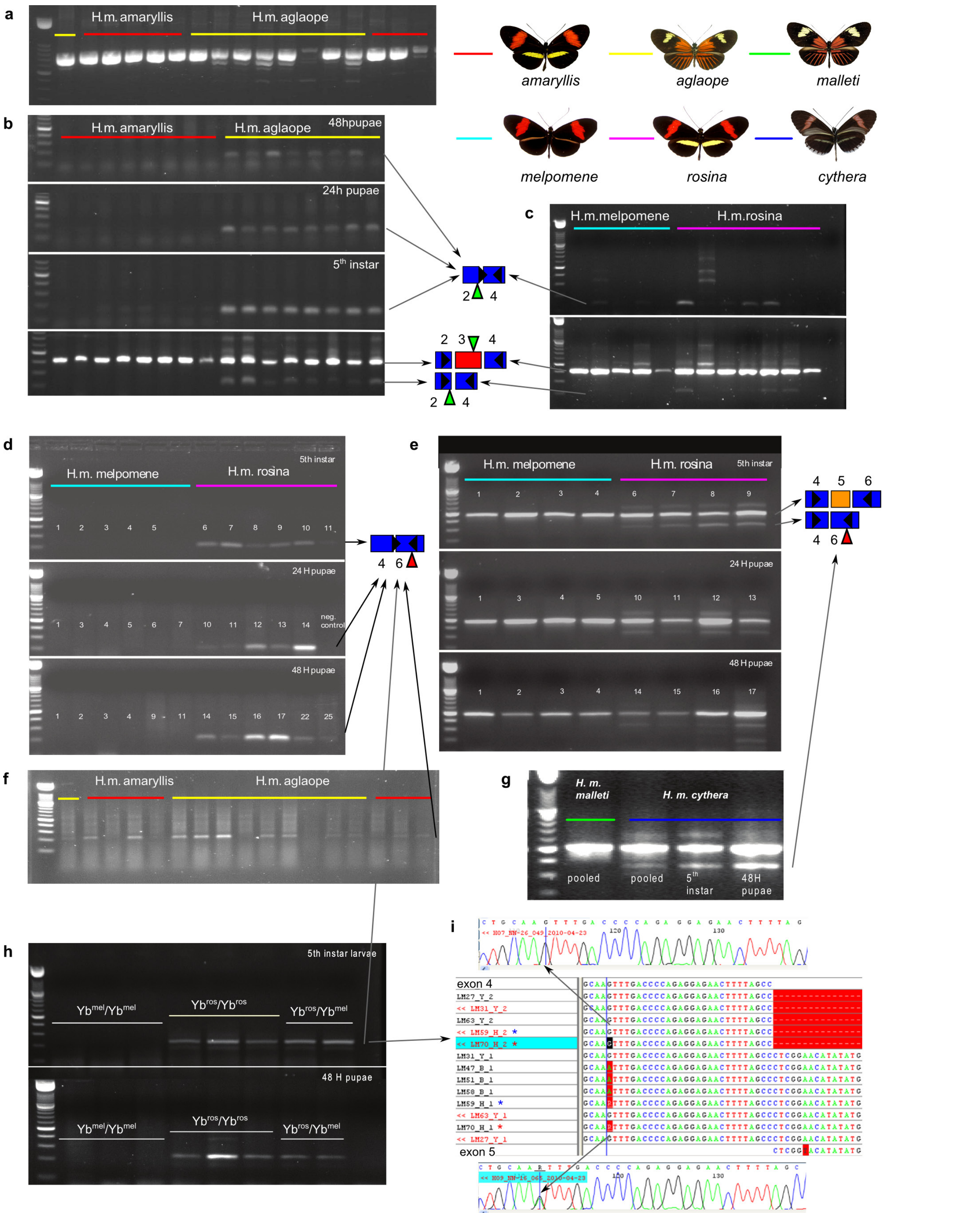
Extended Data Figure 2 | Alignments of *de novo* assembled fragments containing the top associated SNPs from *H. melpomene* and related taxa short-read data. Identified indels do not show stronger associations with phenotype that those seen at SNPs (as shown in Extended Data Table 2), although some near-perfect associations are seen in fragment C. Black regions, missing data; yellow boxes, individuals with a yellow hindwing bar; blue boxes, individuals with a yellow forewing band.



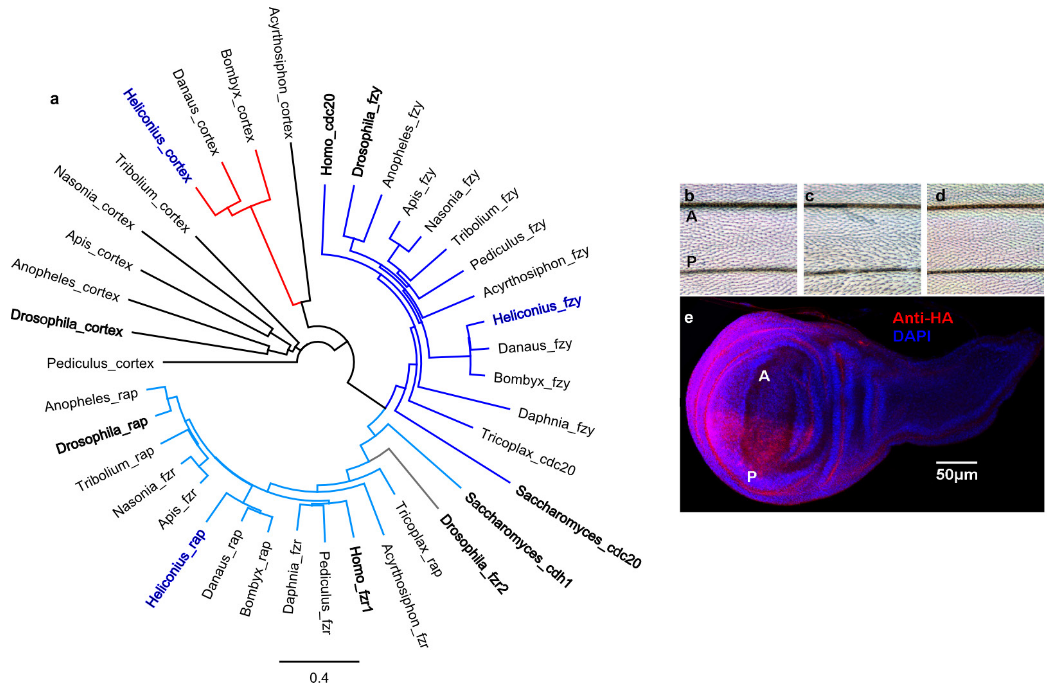
Extended Data Figure 3 | Sequencing of long-range PCR products and fosmids spanning *cortex*. **a**, Sequence read coverage from long-range PCR products across the *cortex* coding region from two *H. melpomene* races. **b**, Minor allele frequency difference from these reads between *H. melpomene aglaope* and *H. melpomene amaryllis*. Exons of *cortex* are indicated by boxes, numbered as in Extended Data Fig. 2. **c**, Alignments of sequenced fosmids overlapping *cortex* from three *H. melpomene* (*H. m.*) individuals of difference races. No major rearrangements are observed, nor any major differences in transposable element (TE) content between closely related races with different colour patterns (*melpomene*/*rosina* or *amaryllis*/*aglaope*). *H. melpomene amaryllis* and *rosina* have the same phenotype, but do not share any transposable elements that are not present in the other races. Hm_BAC, BAC reference sequence; Hm_mel, *melpomene* from new unpublished assembly of *H. melpomene* genome[51]; Hm_ros, *rosina* (two different alleles were sequenced from this individual); Hm_ama, *amaryllis* (two non-overlapping clones were sequenced from this individual); Hm_agla, *aglaope* (four clones were sequenced from this individual, of which two represent alternative alleles). Alignments were performed with Mauve;coloured bars represent homologous genomic regions. *cortex* is annotated in black above each clone. Variable transposable elements are shown as coloured bars below each clone: red, Metulj-like non-LTR; yellow, Helitron-like DNA; grey, other.



Extended Data Figure 4 | Expression array results for additional stages. Array results are related to Fig. 4. **a**–**g**, Comparisons between races (*H. melpomene plesseni* and *H. melpomene malleti*) for three wing regions. **h**–**n**, Comparisons between proximal and distal forewing regions for each race. Significance values ($-\log_{10}P$) are shown separately for genes in the *HmYb* region from the gene array (**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 (**d**, **e**, **k**, **l**) and day 7 (**f**, **g**, **m**, **n**) after pupation. The level of expression difference (log fold change) for tiling probes showing significant differences ($P \leq 0.05$) is shown for day 1 (**c** and **j**) with probes in known *cortex* exons shown in dark colours and probes elsewhere shown as pale colours. *P* values are based on FDR-adjusted *t*-statistics.



Extended Data Figure 5 | Alternative splicing of *cortex*. **a**, Amplification of the whole *cortex* coding region, showing the diversity of isoforms and variation between individuals. **b**, Differences in splicing of exon 3 between *H. melpomene aglaope* and *H. melpomene amaryllis*. Products amplified with a primer spanning the exon 2–4 junction at three developmental stages. The lower panel shows verification of this assay by amplification between exons 2 and 4 for the same final instar larval samples (replicated in Extended Data Fig. 2c). **c**, Lack of consistent differences between *H. melpomene melpomene* and *H. melpomene rosina* in splicing of exon 3. Top panel shows products amplified with a primer spanning the exon 2–4 junction; lower panel shows the same samples amplified between exons 2 and 4. **d**, Differences in splicing of exon 5 between *H. melpomene melpomene* and *H. melpomene rosina*. Products amplified with a primer spanning the exon 4–6 junction at three developmental stages. **e**, Subset of samples from **d** amplified with primers between exons 4 and 6 for verification (middle, 24-h pupae samples are replicated in Extended Data Fig. 2d). **f**, Lack of consistent differences between *H. melpomene aglaope* and *H. melpomene amaryllis* in splicing of exon 5. Products amplified with a primer spanning the exon 4–6 junction. **g**, *H. melpomene cythera* also expresses the isoform lacking exon 5, while a pool of six *H. melpomene malleti* individuals do not. **h**, Expression of the isoform lacking exon 5 from an F_2 *H. melpomene melpomene* × *H. melpomene rosina* cross. Individuals homozygous or heterozygous for the *H. melpomene rosina HmYb* allele express the isoform while those homozygous for the *H. melpomene melpomene HmYb* allele do not. **i**, Allele-specific expression of isoforms with and without exon 5. Heterozygous individuals (indicated with blue and red stars) express only the *H. melpomene rosina* allele in the isoform lacking exon 5 (G at highlighted position), while they express both alleles in the isoform containing exon 5 (G/A at this position).



Extended Data Figure 6 | Phylogeny of fizzy family proteins and effects of expressing *cortex* in the *Drosophila* wing. a, Neighbour joining phylogeny of fizzy family proteins including functionally characterized proteins (in bold) from *Saccharomyces cerevisiae*, *Homo sapiens* and *D. melanogaster* as well as copies from the basal metazoan *Trichoplax adhaerens* and a range of annotated arthropod genomes (*Daphnia pulex*, *Acyrthosiphon pisum*, *Pediculus humanus*, *Apis mellifera*, *Nasonia vitripennis*, *Anopheles gambiae* and *Tribolium castaneum*) including the lepidoptera *H. melpomene* (in blue), *D. plexippus* and *B. mori*. Branch colours: dark blue, cdc20/fzy; light blue, rap; red, lepidopteran cortex. **b**–**e**, Ectopic expression of *cortex* in *D. melanogaster*. *Drosophila cortex* produces an irregular microchaete phenotype when expressed in the posterior compartment of the fly wing (**c**) whereas *Heliconius cortex* does not (**d**), when compared to no expression (**b**). A, anterior; P, posterior. Successful *Heliconius cortex* expression was confirmed by anti-HA immunohistochemistry in the last instar *Drosophila* larva wing imaginal disc (**e**, red), with DAPI staining in blue.

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

			Heliconius melpomene									*H. erato*			*Hn*	
Hm gene ID	*He* gene ID	Putative gene name	Yb^{l}	Sb^{l}	A^{Yb}	A^{N}	E^{1}	E^{gw}	E^{gr}	E^{tw}	E^{tr}	Cr^{l}	A^{pet}	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 (parn)		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			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	
HM00052	HERA000076	HM00052					x					x			x	

A^{bic}, number of above background SNPs associated with the *H. numata* (*Hn*) *bicoloratus* phenotype in this study. A^{fav}, number of SNPs fixed for the alternative allele in *H. erato favorinus*. A^{N}, number of above background SNPs associated with the forewing yellow band in this study. A^{pet}, number of SNPs fixed for the alternative allele in *H. erato demophoon*. A^{Yb}, number of above background SNPs associated with the hindwing yellow bar in this study. Cr^{l}, within the previously mapped *HeCr* interval[11]. P^{l}, within the previously mapped P interval[13]. E^{1}, detected as differentially expressed between *H. melpomene aglaope* and *amaryllis* from RNA-seq data in this study (Supplementary Information). E^{gr}, detected as differentially expressed between *H. melpomene plesseni* and *malleti* in in the gene array in this study. E^{gw}, detected as differentially expressed between forewing regions in the gene array in this study. E^{tr}, numbers of probes showing differential expression between *H. melpomene plesseni* and *malleti* in in the tiling array in this study. E^{tw}, numbers of probes showing differential expression between forewing regions in the tilling array in this study. Sb^{l}, within the previously mapped *Sb* interval[12]. Yb^{l}, within the previously mapped *Yb* interval[12]. *Sb* controls a white–yellow hindwing margin and is not investigated in this study. The *N* locus has not been fine-mapped previously.

Extended Data Table 2 | Locations of fixed or above-background SNPs and differentially expressed (DE) tiling array probes

Positions of SNPs in the *He* and *Hn* association analyses	*cortex* coding exons	*cortex* UTR exons	*cortex* introns (nonTE)	*cortex* flanking intergenic (nonTE)	TEs	Other genes (exons or introns)	Other intergenic	Total
erato favorinus fixed	2	0	96	8	2	0	0	108
erato demophoon fixed	0	0	1	5	1	2	6	15
numata bicoloratus above background	1	3	47	16	0	2	0	69

Positions of DE tiling array probes			Known *cortex* coding exons	*cortex* UTR exons	*cortex* 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

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

Species	Race	Sample Code	HW bar	SNP pos 457083† (p=6.07E-10)	SNP pos 439063* (p=1.72E-09)	SNP pos 602131‡ (p=2.42E-09)	SNP pos 457056† (p=2.42E-09)	FW band	SNP pos 584465§ (p=1.37E-07)	SNP pos 584418§ (p=1.41E-07)	SNP pos 584633§ (p=2.10E-07)	SNP pos 603344‡ (p=2.19E-07)
H. melpomene	*aglaope*	09-246	0	A/A	A/G	A/A	C/C	1	T/T	A/A	NA	T/T
H. melpomene	*aglaope*	09-267	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
H. melpomene	*aglaope*	09-268	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
H. melpomene	*aglaope*	09-357	0	A/A	G/G	G/A	C/C	1	T/T	NA	C/C	T/T
H. melpomene	*aglaope*	aglaope.1	0	A/A	G/G	NA	C/C	1	C/T	T/A	T/C	T/T
H. melpomene	*amandus*	2221	1	A/A	NA	G/G	C/C	0	C/C	T/T	T/T	A/A
H. melpomene	*amandus*	2228	1	A/A	NA	G/G	C/C	0	C/T	T/A	T/C	A/A
H. melpomene	*amaryllis*	09-332	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. melpomene	*amaryllis*	09-333	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. melpomene	*amaryllis*	09-075	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. melpomene	*amaryllis*	09-079	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. melpomene	*amaryllis*	amaryllis.1	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. melpomene	*bellula*	228	1	T/T	NA	G/G	T/T	0	C/C	T/T	T/T	NA
H. melpomene	*bellula*	231	1	T/T	NA	G/A	T/T	0	C/T	T/A	T/C	NA
H. melpomene	*cythera*	2856	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. melpomene	*cythera*	2857	1	NA	NA	NA	NA	0	NA	NA	NA	NA
H. melpomene	*malleti*	17162	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
H. melpomene	*melpomene*	18038	0	A/A	G/G	G/G	C/C	0	C/C	T/T	T/T	A/A
H. melpomene	*melpomene*	18097	0	NA	G/G	NA	C/C	0	C/C	T/T	T/T	NA
H. melpomene	*melpomene*	m0.06	0	A/A	G/G	G/G	C/C	0	C/C	T/T	T/T	A/A
H. melpomene	*melpomene*	gen_ref	0	A/A	G/G	NA	C/C	0	C/C	T/T	T/T	A/A
H. melpomene	*melpomene*	13435	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A
H. melpomene	*melpomene*	9315	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A
H. melpomene	*melpomene*	9316	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A
H. melpomene	*melpomene*	9317	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	A/A
H. melpomene	*plesseni*	9156	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	NA
H. melpomene	*plesseni*	16293	0	A/A	G/G	A/A	C/C	0	C/C	T/T	T/T	NA
H. melpomene	*rosina*	rosina.1	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. melpomene	*rosina*	2071	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. melpomene	*rosina*	531	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. melpomene	*rosina*	533	1	T/T	NA	G/G	T/T	0	C/C	T/T	T/T	NA
H. melpomene	*rosina*	546	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. melpomene	*thelxiopeia*	13566	0	A/A	G/G	A/A	C/C	1	C/T	T/A	T/C	T/T
H. melpomene	*vulcanus*	14632	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	NA
H. melpomene	*vulcanus*	519	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. timareta	*florencia*	2403	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
H. timareta	*florencia*	2406	0	A/A	A/G	A/A	C/C	1	T/T	A/A	C/C	T/T
H. timareta	*florencia*	2407	0	A/A	A/G	A/A	C/C	1	T/T	A/A	C/C	T/T
H. timareta	*florencia*	2410	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
H. timareta	*timareta*	8533	0	A/A	G/G	A/A	C/C	1	C/T	T/A	T/C	T/T
H. timareta	*timareta*	9184	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
H. timareta	*timareta*	8520	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
H. timareta	*timareta*	8523	0	A/A	G/G	A/A	C/C	1	T/T	A/A	C/C	T/T
H. timareta	*thelxinoe*	09-312	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. timareta	*thelxinoe*	8624	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. timareta	*thelxinoe*	8628	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. timareta	*thelxinoe*	8631	1	T/T	A/A	G/G	T/T	0	C/C	T/T	T/T	A/A
H. elevatus		09-343	0	A/T	G/G	A/A	T/T	1	C/T	NA	C/C	T/T
H. pardalinus	*sergestus*	09-326	0	A/A	A/A	A/A	NA	0	C/C	T/T	T/T	NA

*Downstream of *cortex*. †Between exons 3 and 4 of *cortex*. ‡Upstream of *cortex*. §Between exons U4 and U3 of *cortex*. None of these SNPs are within known transposable elements. Colours show phenotypic associations: yellow, yellow hindwing bar; pink, no yellow hindwing bar; green, yellow forewing band; blue, no yellow forewing band; grey, allele does not match expected pattern.

Extended Data Table 4 | Transposable elements (TEs) found within the *Yb* region

Unique Occurrences									
BAC	mel	ros	ama	agl	No.	TE name	Superfamily		Type
1					1	BEL-1	BEL		LTR retrotransposon
					1	CR1-2	Jockey	LINE	Non-LTR retrotransposon
	1				1	Daphne-1	Jockey	LINE	Non-LTR retrotransposon
1					1	Daphne-6	Jockey	LINE	Non-LTR retrotransposon
1					1	DNA-like-8			DNA transposon
					1	Helitron-like-14	Helitron_A		DNA transposon
	1	2			4	Helitron-like-12	Helitron_A		DNA transposon
1	2				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	LINE	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	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	LINE	Non-LTR retrotransposon
			1	1	6	Rep-1	REP	LINE	Non-LTR retrotransposon
2		1		1	4	RTE-3	RTE	LINE	Non-LTR retrotransposon
				1	2	RTE-11	RTE	LINE	Non-LTR retrotransposon
	1				3	Zenon-1	Jockey	LINE	Non-LTR retrotransposon
			1		1	Zenon-3	Jockey	LINE	Non-LTR retrotransposon