A major gene controls mimicry and crypsis in butterflies and moths

Nicola J. Nadeau[1,2], Carolina Pardo-Diaz[3], Annabel Whibley[4,5], Megan 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[2,6], Heather Hines[11], Camilo Salazar[3], Andrea Dowling[12], Richard ffrench-Constant[12], Violaine Llaurens[4], Mathieu Joron[4], W. Owen McMillan[2], Chris D. Jiggins[6,2]

[1]Department of Animal and Plant Sciences, University of Sheffield, UK; [2]Smithsonian Tropical Research Institute, Panama; [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, UK, NR4 7UH , [6]The Australian National University, ACT, Australia; [7]Department of Zoology, University of Cambridge, UK; [8]Energy and Resources Group, University of California at Berkeley, CA, USA; [9]Department of Biology, Williams College, MA, USA; [10]Department of Zoology, University of Oxford, UK; [11] Penn State University, 517 Mueller, University Park, PA 16802; [12]School of Biosciences, University of Exeter in Cornwall, Penryn, UK TR10 9EZ

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 if there is any commonality between divergent species. Here, we identify a gene, *cortex*, through fine-scale mapping using population genomics and gene expression analyses, which 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 most likely regulates pigmentation patterning through regulation of 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* (*Hm*) and *H. erato* (*He*), but also whole wing variation in black, yellow, white, and orange/red elements in *H. numata* (*Hn*)[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 localised 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 *He, Hm* and *Hn* identified a genomic interval of ~1Mb[11–13] (Extended Data Table 1), which also overlaps with the 1.4Mb region containing the *carbonaria* locus in *B. betularia*[9] and a 100bp, apparently gene-free region containing the *Ws* mutation in *B. mori*[10] (Fig 1). We took a population genomics approach to identify single nucleotide polymorphisms (SNPs) most strongly associated with phenotypic variation within the ~1Mb *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[14,15]. 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 *He* the strongest associations with the presence of a yellow hindwing bar were centred around the genomic region containing *cortex* (Fig 2A). We identified 108 SNPs that were fixed for one allele in *He favorinus*, and fixed for the alternative allele in all individuals lacking the yellow bar, the majority of which were in introns of *cortex* (Extended Data Table 2). 15 SNPs showed a similar fixed pattern for *He petiverana*, which also has a yellow bar. These were non-overlapping with those in *He favorinus*, consistent with the hypothesis that this phenotype evolved independently in the two disjunct populations[16].

Previous work has suggested that alleles at the *Yb* locus are shared between *Hm* and the closely related species *H. timareta*, and also the more distantly related species *H. elevatus*, resulting in mimicry between these species[17]. Across these species, the strongest associations with the yellow hindwing bar phenotype were again found at *cortex* (Fig 2D, Extended Data Fig 1A and Table 3). Similarly, the strongest associations with the yellow forewing band were found around the 5’ UTRs of *cortex* and gene *HM00036*, an orthologue of *D.*

melanogaster washout gene. A single SNP ~17kb upstream of *cortex* (the closest gene) was perfectly associated with the yellow forewing band across all *Hm, H. timareta* and *H. elevatus* individuals (Extended Data Fig 1A, Fig 2 and Table 3). We found no fixed coding sequence variants at *cortex* in a larger sample (43-61 individuals) of *Hm aglaope* and *Hm amaryllis* (Extended Data Figure 3, Supplementary Information), which differ in *Yb* controlled phenotypes[18], suggesting that functional variants are likely to be regulatory rather than coding. We found extensive transposable element variation around *cortex* but it is unclear if any of these associate with phenotype (Extended Data Figure 3 and Table 4; Supplementary Information).

Finally, in *Hn* large inversions at the *P* supergene locus (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 *He* and *Hm*, this analysis showed a narrow region of associated SNPs corresponding exactly to the *cortex* gene (Fig 2E), again with the majority of SNPs 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[17], as well as probes tiled across the central portion of the *Yb* locus, focussing on two naturally hybridising *Hm* races (*plesseni* and *malleti*) that differ in *Yb* controlled phenotypes[7]. *cortex* was the only gene across the entire interval to show significant expression differences both between races with different wing patterns and between wing sections with different pattern elements (Fig 3).

This finding was reinforced in the tiled probe set, where we observed strong differences in expression of *cortex* exons and introns but few differences outside this region (Extended Data Table 2). *cortex* expression was higher in *Hm malleti* than *Hm plesseni* in all three wing sections used (but not eyes) (Fig 3C; Extended Data Fig 4C). When different wing sections were compared within each race, *cortex* expression in *Hm malleti* was higher in the distal section that contains the *Yb* controlled yellow forewing band, consistent with *cortex* producing this band. In contrast, *Hm plesseni*, which lacks the yellow band, had higher *cortex* expression in the proximal forewing section (Fig 3F; Extended Data Fig 4J). Expression differences were found only in day 1 and day 3 pupal wings rather than day 5 or day 7 (Extended Data Fig 4), similar to the pattern observed previously for the transcription factor *optix*[19].

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 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. RT-PCR revealed a diversity of splice variants (Extended Data Fig 5), and sequenced products revealed 8 non-constitutive exons and 6 variable donor/acceptor sites, but this was not exhaustive (Supplementary Information). We cannot rule out the possibility that some of the differentially expressed intronic regions could be distinct non-coding RNAs. However, qRT-PCR in other hybridising races with divergent *Yb* alleles (*aglaope/amaryllis* and *rosina/melpomene)* also identified expression differences at *cortex* and allele-specific splicing differences between both pairs of races (Extended Data Figs 1 and 5, Supplementary Information).

Finally, *in situ* hybridisation 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 *Hn* 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 time-points 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), this strongly implicates *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 100bp 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 implicate *cortex* as controlling melanism in the peppered moth[4].

It seems likely that *cortex* controls pigmentation patterning through control of scale cell development. The *cortex* gene falls in an insect specific lineage within the fizzy/CDC20 family of cell cycle regulators (Extended Data Fig 6A). The phylogenetic tree of the 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)[3,20]. The third group, cortex, is evolving rapidly, with low amino acid identity between *D. melanogaster* and *Hm* cortex (14.1%), contrasting with much higher identities for orthologues between these species in the other two groups (fzy, 47.8% and rap/fzr,47.2%, Extended Data Fig 6A). *Drosophila melanogaster* cortex acts through a

similar mechanism to fzy in order to control meiosis in the female germ line[21–23]. *Hm* cortex also has some conservation of the fizzy family C-box and IR elements (Supplementary Information) that mediate binding to the APC/C[22], suggesting that it may have retained a cell cycle function, although we found that expressing *Hm cortex* in *D. melanogaster* wings produced no detectable effect (Extended Data Fig 6, Supplementary Information).

Previously identified butterfly wing patterning genes have been transcription factors or signalling molecules[19,24]. Developmental rate has long been thought to play a role in lepidopteran patterning[25,26], but *cortex* was not a likely *a priori* candidate, because its *Drosophila* orthologue has a highly specific function in meiosis[22]. 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[1,7,9,27] to generate both cryptic[4] and aposematic patterns.

References

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* **This issue,**
5. Joron, M. *et al.* A Conserved Supergene Locus Controls Colour Pattern Diversity in Heliconius Butterflies. *PLoS Biol.* **4,** (2006).

6. Sheppard, P. M., Turner, J. R. G., Brown, K. S., Benson, W. W. & Singer, M. C. Genetics and the Evolution of Muellerian Mimicry in Heliconius Butterflies. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **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* (2015). doi:10.1038/hdy.2015.69
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.* **108,** 19666–19671 (2011).
15. Pardo-Diaz, C. *et al.* Adaptive Introgression across Species Boundaries in Heliconius Butterflies. *PLoS Genet* **8,** e1002752 (2012).
16. 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).

17. Consortium, T. H. G. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487,** 94–98 (2012).
18. Mallet, J. The Genetics of Warning Colour in Peruvian Hybrid Zones of Heliconius erato and H. melpomene. *Proc. R. Soc. Lond. B Biol. Sci.* **236,** 163–185 (1989).
19. Reed, R. D. *et al.* optix Drives the Repeated Convergent Evolution of Butterfly Wing Pattern Mimicry. *Science* **333,** 1137–1141 (2011).
20. Barford, D. Structural insights into anaphase-promoting complex function and mechanism. *Philos. Trans. R. Soc. B Biol. Sci.* **366,** 3605–3624 (2011).
21. 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).
22. 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).
23. 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. *Dev. Camb. Engl.* **134,** 891–899 (2007).
24. Martin, A. *et al.* Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. *Proc. Natl. Acad. Sci.* **109,** 12632–12637 (2012).
25. 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).
26. 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. Lepidoptera* **26,** 141–160 (1988).
27. Mallet, J. & Barton, N. H. Strong Natural Selection in a Warning-Color Hybrid Zone. *Evolution* **43,** 421–431 (1989).

Figure 1. A homologous genomic region controls a diversity of phenotypes across the Lepidoptera. Left: phylogenetic relationships[28]. Right: chromosome maps with colour pattern intervals in grey, coloured bars represent markers used to assign homology[5,8–10], the first and last genes from Fig 2 shown in red. In *He* the *HeCr* locus controls the yellow hind-wing bar phenotype (grey boxed races). In *Hm* it controls both the yellow hind-wing bar (*HmYb*, pink box) and the yellow forewing band (*HmN*, blue box). In *Hn* 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.

Figure 2. Association analyses across the genomic region known to contain major colour pattern loci in *Heliconius.* A) Association in *He* with the yellow hind-wing bar (n=45). Coloured SNPs are fixed for a unique state in *He petiverana* (orange) or *He favorinus* (purple). B) Genes in *He* with direct homologs in *Hm*. Gene are in different colours with exons (coding and UTRs) connected by a line. Grey bars are transposable elements. C) *Hm* genes and transposable elements: colours correspond to homologous *He* genes; MicroRNAs[29] in black. D) Association in the *Hm/timareta*/silvaniform group with the yellow hind-wing bar (red) and yellow forewing band (blue) (n=49). E) Association in *Hn* with the *bicoloratus* morph (n=26); inversion positions[13] shown below. In all cases black/dark coloured points are above the strongest associations found outside the colour pattern scaffolds (*He* p=1.63e-05; *Hm* p=2.03e-05 and p=2.58e-05; *Hn* p=6.81e-06)

Figure 3. Differential gene expression across the genomic region known to contain major colour pattern loci in *Heliconius melpomene.* Day 3 pupae expression differences for all

241 genes in the *Yb* interval (A,D) and tiling probes spanning the central portion of the interval
242 (B,C,E,F). Expression is compared between races for each wing region (A,B,C) and between
243 proximal and distal forewing sections for each race (D,E,F). C and F: magnitude and
244 direction of expression difference (log_2 fold-change) for tiling probes showing significant
245 differences ($p \leq 0.05$); probes in known *cortex* exons shown in dark colours. See Extended
246 Data Fig 4 for other stages. Gene *HM00052* was differentially expressed between other races
247 in RNA sequence data (Supplementary Information) but is not differentially expressed here.

248

249 Figure 4. *In situ* hybridisations of *cortex* in hind-wings from final instar larvae. B) *Hn*
250 *tarapotensis*; adult wing shown in A, coloured points indicate landmarks based on wing
251 venation and yellow arrows highlight adult pattern elements corresponding to the *cortex*
252 staining observed here. D) *Hm rosina*; adult wing shown in C, staining patterns in other *Hm*
253 races (*meriana* and *aglaope*) appeared similar. The probe used was complementary to the
254 *cortex* isoform with the longest ORF (which we also detected as being the most common,
255 Supplementary Information)

256

257 **Methods**

258 ***He Cr reference***

259 *Cr* is the homologue of *Yb* in *He* (Fig 1). An existing reference for this region was available
260 in 3 pieces (467,734bp, 114,741bp and 161,149bp, GenBank: KC469893.1)[30]. We screened
261 the same BAC library used previously[11,30] using described procedures[11] with probes designed
262 to the ends of the existing BAC sequences and the *HmYb* BAC reference sequence. Two
263 BACs (04B01 and 10B14) were identified as spanning one of the gaps and sequenced using

Illumina 2x250 bp paired-end reads collected on the Illumina MiSeq. The raw reads were screened to remove vector and *E.* coli bases. The first 50k read pairs were taken for each BAC and assembled individually with the Phrap[31] software and manually edited with consed[32]. 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 where the terminal 200bp 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.

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

The final contig was 1,009,595bp in length of which 2,281bp were unknown (N's). 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[32], with some minor modifications. Briefly this involved first generating a reference based transcriptome assembly with existing *H. erato* RNA-seq wing tissue (GenBank accession SRA060220). We used Trimmomatic[33] (v0.22), and FLASh[34] (v1.2.2) to prepare the raw sequencing reads, checking the quality with FastQC[35] (v0.10.0). We then used the Bowtie/TopHat/Cufflinks[36–38] pipeline to generate transcripts for the unmasked reference sequence. We generated gene predictions with the MAKER pipeline[39] (v2.31). Homology and synteny in gene content with the *Hm Yb*

reference were identified by aligning the *Hm* coding sequences to the *He* reference with BLAST. Homologous genes were present in the same order and orientation in *He* and *Hm* (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 *He cortex* was manually curated from the RNA-seq data and using *Exonerate*[40] 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 the Supplementary Information. We used shotgun Illumina sequence reads from 45 *He* individuals from 7 races that were generated as part of a previous study[30] (Supplementary Information). Reads were aligned to an *He* reference containing the *Cr* contig and other sequenced *He* BACs[11,30] with BWA[41] , which has previously been found to work better than Stampy[42] (which was used for the alignments in the other species) with an incomplete reference sequence[30]. 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[43] 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 *Hm / Hn* clade we used previously published sequence data from 19 individuals from enrichment sequencing targeting of the *Yb* region, the unlinked *HmB/D* region that controls the presence/absence of red colour pattern elements, and ~1.8Mb of non-colour pattern genomic regions[44], as well as 9 whole genome shotgun sequenced individuals[17,45]. We added

313 targeted sequencing and shotgun whole genome sequencing of an additional 47 individuals
314 (Supplementary Information). Alignments were performed using Stampy[42] with default
315 parameters except for substitution rate which was set to 0.01. We again removed duplicates
316 and used GATK to re-align indels and call SNPs with expected heterozygosity set to 0.1.

317 The analysis of the *Hm/timareta*/silvaniform included 49 individuals, which were aligned to
318 v1.1 of the *Hm* reference genome with the scaffolds containing *Yb* and *HmB/D* swapped with
319 reference BAC sequences[17], which contained fewer gaps of unknown sequence than the
320 genome scaffolds. 232,631 SNPs were present in the *Yb* region and a further 370,079 SNPs
321 were used to establish background association levels.

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

327 We measured associations between genotype and phenotype using a score test (qtscore) in the
328 GenABEL package in R[46]. This was corrected for background population structure using a
329 test specific inflation factor, λ, calculated from the SNPs unlinked to the major colour pattern
330 controlling loci (described above), as the colour pattern loci are known to have different
331 population structure to the rest of the genome[14,15,17]. We used a custom perl script to convert
332 GATK vcf files to Illumina SNP format for input to genABEL[46]. genABEL does not accept
333 multiallelic sites, so the script also converted the genotype of any individuals for which a
334 third (or fourth) allele was present to a missing genotype (with these defined as the lowest
335 frequency alleles). Custom R scripts were used to identify sites showing perfect associations
336 with calls for >75% of individuals.

We generated two long-range PCR products covering 88.8% of the 1,344bp coding region of *cortex* (excluding 67bp at the 5' end and 83bp at the 3' end, further details in Supplementary Information). A product spanning coding exons 5 to 9 (the final exon) was obtained from 29 *Hm amaryllis* individuals and 29 *Hm aglaope* individuals; a product spanning coding exons 2 to 5 was obtained from 32 *Hm amaryllis* individuals and 14 *Hm aglaope*. In addition, a product spanning exons 4 to 6 was obtained from 6 *Hm amaryllis* and 5 *Hm aglaope* that failed to amplify one or both of the larger products. 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 (*Hm aglaope* and *amaryllis)* were then prepared as two separate libraries with molecular identifiers and sequenced on a single lane of an Illumina GAIIx.

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[44] with sequence from the *Yb* BAC walk[12] used to fill any gaps. Alignments were performed with BWA v0.5.6[41] and converted to pileup format using Samtools v0.1.12 before being filtered based on quality (≥20) and coverage (≥10). We then calculated coverage and minor allele frequencies for each race and the difference between these using custom scripts in R[47].

Gene Expression Analyses

We designed a Roche NimbleGen microarray (12x135K format) with probes for all annotated *Hm* genes[17] and tiling the central portion of the *Yb* BAC sequence contig that was previously identified as showing the strongest differentiation between *Hm* races[44]. This 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 *Hm plesseni* and *malleti*. Pupae were from captive stocks maintained in insectary facilities in Gamboa, Panama. Tissue was stored in RNA later at -80°C prior to 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 randomised and each hybridised to a separate array. The *HmYb* probe array contained 9,979 probes distanced on average at 10bp. The whole-genome expression array contained on average 9 probes per annotated gene in the genome (v1.1[17]) 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 (version 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-normalised, $\log_2$-transformed, quality controlled and analysed for differences in expression between individuals and wing regions. P-values were adjusted for multiple hypotheses testing using the False Discovery Rate (FDR) method [49].

We detected isoform-specific expression differences between *Hm aglaope*/*amaryllis* and *Hm rosina/melpomene* using RT-PCR and qRT-PCR on RNA extracted from developing hind-wing tissue (further details in Supplementary Information). Previously published RNAseq data was also used to assess gene expression differences between *Hm aglaope* and *amaryllis*[17] (further details in Supplementary Information).

In situ hybridisations

Hn and *Hm* 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[24] with a *cortex* riboprobe synthesized from a 831-bp cDNA amplicon from *Hn*. 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 1:3000 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.

References

28. Wahlberg, N., Wheat, C. W. & Peña, C. Timing and Patterns in the Taxonomic Diversification of Lepidoptera (Butterflies and Moths). *PLoS ONE* **8,** e80875 (2013).
29. Surridge, A. *et al.* Characterisation and expression of microRNAs in developing wings of the neotropical butterfly Heliconius melpomene. *BMC Genomics* **12,** 62 (2011).
30. Supple, M. A. *et al.* Genomic architecture of adaptive color pattern divergence and convergence in Heliconius butterflies. *Genome Res.* **23,** 1248–1257 (2013).
31. de la Bastide, M. & McCombie, W. R. Assembling genomic DNA sequences with PHRAP. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* **Chapter 11,** Unit11.4 (2007).
32. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8,** 195–202 (1998).
33. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* btu170 (2014). doi:10.1093/bioinformatics/btu170
34. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27,** 2957–2963 (2011).
35. Andrews, S. *FastQC*. (2011).

410 36. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of
411 short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
412 37. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq.
413 *Bioinformatics* **25,** 1105–1111 (2009).
414 38. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated
415 transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515
416 (2010).
417 39. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool
418 for second-generation genome projects. *BMC Bioinformatics* **12,** 491 (2011).
419 40. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison.
420 *BMC Bioinformatics* **6,** 31 (2005).
421 41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
422 *Bioinforma. Oxf. Engl.* **25,** 1754–1760 (2009).
423 42. Lunter, G. & Goodson, M. Stampy: A statistical algorithm for sensitive and fast mapping of
424 Illumina sequence reads. *Genome Res.* **21,** 936–939 (2011).
425 43. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation
426 DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).
427 44. Nadeau, N. J. *et al.* Genomic islands of divergence in hybridizing Heliconius butterflies identified
428 by large-scale targeted sequencing. *Philos. Trans. R. Soc. B Biol. Sci.* **367,** 343–353 (2012).
429 45. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in Heliconius
430 butterflies. *Genome Res.* **23,** 1817–1828 (2013).
431 46. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide
432 association analysis. *Bioinforma. Oxf. Engl.* **23,** 1294–1296 (2007).
433 47. R Development Core Team. *R: A language and environment for statistical computing*. (R
434 Foundation for Statistical Computing, 2011).

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 New York, 2005).

49. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57,** 289–300 (1995).

Extended Data Figure Legends

Extended Data Figure 1. A) Exons and splice variants of *cortex* in *Hm*. Orientation is reversed with respect to figures 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 *Hm amaryllis* and *Hm aglaope* whole hindwings (N=11 and N=10 respectively). C) Expression of a *cortex* isoform lacking exon 3 is found in *Hm aglaope* but not *Hm amaryllis* hindwings. D) Expression of an isoform lacking exon 5 is found in *Hm rosina* but not *Hm 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 position of the primers used in the assay.

Extended Data Figure 2. Alignments of *de novo* assembled fragments containing the top associated SNPs from *Hm* 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 box = individuals with a hindwing yellow bar; blue box = 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 2 *Hm* races. B) Minor allele frequency difference from these reads between *Hm aglaope* and *Hm amaryllis*. Exons of *cortex* are indicated by boxes, numbered as in Extended Data Figure 2. C) Alignments of sequenced fosmids overlapping *cortex* from 3 *Hm* 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*). *Hm amaryllis* and *rosina* have the same phenotype, but do not share any TEs that are not present in the other races. Hm_BAC = BAC reference sequence, Hm_mel = *melpomene* from new unpublished assembly of *Hm* genome[50], Hm_ros = *rosina* (2 different alleles were sequenced from this individual), Hm_ama = *amaryllis* (2 non-overlapping clones were sequenced in this individual), Hm_agla = *aglaope* (4 clones were sequenced in this individual 2 of which represent alternative alleles). Alignments were performed with Mauve: coloured bars represent homologous genomic regions. *cortex* is annotated in black above each clone. Variable TEs 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, related to Figure 4. A-G: comparisons between races (*H. m. plesseni* and *H. m. malleti*) for 3 wing regions. H-N: comparisons between proximal and distal forewing regions for each race. Significance values (-log10(p-value)) 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.

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. m. aglaope* and *H. m. amaryllis*. Products amplified with a primer spanning the exon 2/4 junction at 3 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 Figure 2C) C) Lack of consistent differences between *H. m. melpomene* and *H. m. rosina* in splicing of exon 3. Top panel shows products amplified with a primer spanning the exon 2/4 junction, lower panel is the same samples amplified between exons 2 and 4. D) Differences in splicing of exon 5 between *H. m. melpomene* and *H. m. rosina*. Products amplified with a primer spanning the exon 4/6 junction at 3 developmental stages. E) Subset of samples from D amplified with primers between exons 4 and 6 for verification (middle, 24hr pupae samples are replicated in Extended Data Figure 2D). F) Lack of consistent differences between *H. m. aglaope* and *H. m. amaryllis* in splicing of exon 5. Products amplified with a primer spanning the exon 4/6 junction. G) *H. m. cythera* also expresses the isoform lacking exon 5, while a pool of 6 *H. m. malleti* individuals do not. H) Expression of the isoform lacking exon 5 from an F2 *H. m. melpomene* x *H. m. rosina* cross. Individuals homozygous or heterozygous for the *H. m. rosina HmYb* allele express the isoform while those homozygous for the *H. m. 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. m. 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 characterised proteins (in bold) from *Saccharomyces cerevisiae, Homo sapiens* and *Drosophila 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 mellifica, Nasonia vitripennis, Anopheles gambiae, Tribolium castaneum*) including the lepidoptera *H. melpomene* (in blue), *Danaus plexippus* and *Bombyx mori*. Branch colours: dark blue, CDC20/fzy; light blue, CDH1/fzr/rap; red, lepidoptran cortex. B-E) Ectopic expression of *cortex* in *Drosophila 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 IHC in the last instar *Drosophila* larva wing imaginal disc (D, red), with DAPI staining in blue.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Christopher Saski, Clemson University, for assembly of the *He* BACs. Richard Merrill, Moises Abanto and Adriana Tapia assisted with raising butterflies. Thanks to Mathieu Chouteau, Jake Morris and Kanchon Dasmahapatra for providing larvae for *in situ* hybridisations. Anna Morrison, Robert Tetley, Sarah Carl and Hanna Wegener assisted with lab work at the University of Cambridge. Simon Baxter made the *Hm* fosmid libraries. We thank the governments of Colombia, Ecuador, Panama and Peru for permission to collect butterflies. This work was funded by a Leverhulme Trust award and

530 BBSRC grant (H01439X/1) to CDJ, NSF grants (DEB 1257689, IOS 1052541) to WOM, an
531 ERC starting grant to MJ and a French National Agency for Research (ANR) grant to VL
532 (ANR-13-JSV7-0003-01). NJN is funded by a NERC fellowship (NE/K008498/1).

533

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

544 **Author Information** Short read sequence data generated for this study are available from
545 ENA (http://www.ebi.ac.uk/ena) under study accession PRJEB8011 and PRJEB12740 (see
546 Supplementary table S1 for previously published data accessions). The updated Cr contig is
547 deposited in Genbank with accession KC469893. The assembled *Hm* fosmid sequences are
548 deposited in Genbank with accessions KU514430-KU514438. The microarray data are
549 deposited in GEO with accessions GSM1563402- GSM1563497. Reprints and permissions
550 information is available at www.nature.com/reprints. Correspondence and requests for
551 materials should be addressed to n.nadeau@sheffield.ac.uk or c.jiggins@zoo.cam.ac.uk

552

Extended Data Tables

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

			Heliconius melpomene									*H. erato*			*H. numata*	
Hm gene ID	*He* gene ID	Putative gene name	Yb^{I}	Sb^{I}	A^{Yb}	A^{N}	E^{1}	E^{gw}	E^{gr}	E^{tw}	E^{tr}	Cr^{I}	A^{pet}	A^{fav}	P^{I}	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	

555 Yb^{I}, within the previously mapped *Yb* interval[12].
556 Sb^{I}, within the previously mapped *Sb* interval[12]. *Sb* controls a white/yellow hindwing margin and is not
557 investigated in this study. The *N* locus has not been fine-mapped previously.
558 A^{Yb}, number of above background SNPs associated with the hindwing yellow bar in this study.
559 A^{N}, number of above background SNPs associated with the forewing yellow band in this study.
560 E^{1}, detected as differentially expressed between *Hm aglaope* and *amaryllis* from RNAseq data in this study
561 (Supplementary Information S2.5).
562 E^{gw}, detected as differentially expressed between forewing regions in the gene array in this study.
563 E^{gr}, detected as differentially expressed between *Hm plesseni* and *malleti* in in the gene array in this study.
564 E^{tw}, numbers of probes showing differential expression between forewing regions in the tilling array in this
565 study.
566 E^{tr}, numbers of probes showing differential expression between *Hm plesseni* and *malleti* in in the tiling array in
567 this study.
568 Cr^{I}, within the previously mapped *HeCr* interval[11].
569 A^{pet}, number of SNPs fixed for the alternative allele in *He petiverana*.
570 A^{fav}, number of SNPs fixed for the alternative allele in *He favorinus*.
571 P^{I}, within the previously mapped P interval[13].
572 A^{bic}, number of above background SNPs associated with the *Hn bicoloratus* phenotype in this study.
573
574

575 Extended Data Table 2. Locations of fixed/above background SNPs and differentially
576 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 petiverana 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

577

578

579

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 5844655§ (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

582 †between exons 3 and 4 of *cortex*, *downstream of *cortex*, ‡upstream of *cortex*, §between
583 exons U4 and U3 of *cortex*. None of these SNPs are within known TEs.

584

585

586 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

587

588