Butterfly genome reveals promiscuous exchange of mimicry adaptations among species

Supplementary Information

S1. Shotgun genome sequencing 4
S1.1 454 library construction 4
S1.2 454 sequencing 4
S1.3 Illumina library construction and sequencing 5
Table S1.3.1 Sequencing data for *H. melpomene* assembly 5

S2. Genome assembly 6
S2.1 Additional screening for linker 6
S2.2 Redundancy screening 6
S2.3 CABOG assembly 7
Table S2.3.1 Quantitative statistics from the CABOG assembly 7
Figure S2.3.1 Coverage histograms for contigs of varying sizes 8
S2.4 Haplotype separation 8
S2.5 Local re-assembly 9

S3. Assembly verification 10
S3.1 BAC end mapping. 10
S3.2 BAC sequence comparison. 10
Figure S3.2.1. Twelve scaffolds matching the finished sequence of the *Yb* locus. 11

S4. Chromosomal scaffolding using a RAD linkage map 12
S4.1 The mapping cross 12
S4.2 RAD library preparation 12
S4.3 RAD library sequencing and alignment to reference genome 13
S4.4 Defining markers for linkage mapping 13
S4.5 Linkage mapping 14
Figure S4.5.1 RAD linkage map 15
S4.6 Placing scaffolds on chromosomes and correcting assembly errors 16
Figure S4.6.1 Example of chromosomal scaffolding for linkage group 18 17
Table S4.6.1 Summary of chromosome lengths in scaffolded genome. 18

S5. Novel repeat and transposable element (TE) identification 19
S5.1 Methods 19
S5.2 Results and Discussion 19
Table S5.2.1. TE content of the *H. melpomene* genome 21

S6. Gene prediction 22
S6.1 Methods 22
Table S6.1.1 Summary of RNA-seq data used for automated gene prediction 22
Table S6.1.2 Summary of supporting evidence for automated gene prediction 23
S6.2 Manual verification 23

S7. Characterisation of miRNAs 24
S7.1 Identification of known miRNAs 24

S7.2 Prediction of novel *Heliconius* miRNAs 24

S8. Genome evolution and synteny analyses 25

S8.1 Methods 25

S8.2 Results 26

Table S8.2.1 Summary of chromosome homology between *H. melpomene* and *B. mori* 28

Table S8.2.2 Comparison of syntenic block parameters between *H. melpomene* and *B. mori* 29

Table S8.2.3 Comparison of syntenic block parameters between *H. melpomene* and *D. plexippus* 30

Table S8.2.4 Estimates of rearrangement rates between *H. melpomene* and *B. mori* 31

Table S8.2.5 Estimates of rearrangement rates between *H. melpomene* and *D. plexippus* 31

S9. Olfactory and chemosensory proteins 32

S9.1 Methods 32

S9.2 Odorant-binding proteins (OBPs) 33

Table S9.2.1 CSP and OBP proteins previously reported in silkmoth antennae and female pheromone glands and their homologs in *H. melpomene*. 34

Figure S9.2.1 Maximum likelihood tree of odorant-binding proteins 35

S9.3 Chemosensory proteins (CSPs) 36

S9.4 Olfactory receptors (ORs) 37

Figure S9.4.1 Maximum likelihood tree of olfactory receptors 38

S10. Homeobox genes 39

S10.1 Initial characterization and local reassembly of the *Hox* cluster 39

S10.2 Identification of lepidopteran *Hox* genes 39

S10.3 Phylogenetic analysis of insect *Hox* genes 40

S10.4 Results 40

Figure S10.4.1: Alignment of insect homeobox domains 41

Figure S10.4.2 Phylogenetic tree of *Hox* genes from *D. melanogaster* (Dm), *T. castaneum* (Tc), *A. mellifera* (Ac), *B. mori* (Bm), *D. plexippus* (Dp) and *H. melpomene* (Hm). 42

Figure S10.4.3 Expansion of the lepidopteran *Hox* cluster 43

S11. Immunity Genes 44

S11.1 Methods 44

S11.2 Results 44

Table S11.2.1 Immunity related genes in seven insect species 45

S12. Genomics methods for introgression study 48

S12.1 Sample collection and DNA extraction 48

S12.2 SureSelect targeted sequencing 48

Table S12.2.1 Details of samples used in SureSelect targeted sequencing 49

Table S12.2.2 Oligonucleotides used to prepare libraries for SureSelect sequencing 50

S12.3 RAD genotyping 50

Table S12.3.1 Samples used for RAD sequencing 51

S12.4 Alignment and SNP calling 52

S12.5 F_{ST} analyses 53

S12.6 ABBA-BABA tests of introgression 53

S12.7 Estimating the proportion of genomic introgression 54

S12.8 Estimation of linkage disequilibrium and block jack-knife standard errors 55

S12.9 Phylogenetic analyses 55

S13. Alignment statistics **56**
Table S13.1 Alignment statistics for RAD sequencing 56
Table S13.2 Alignment statistics for SureSelect resequencing 58

S14. Genomic divergence (F_{ST}) among populations in *B/D* and *N/Yb* colour pattern regions, and in non-colour pattern regions **59**
S14.1 The effect of repetitive DNA and repeat masking 59
Figure S14.1.1 Genetic differentiation across colour pattern regions, showing effect of removing repeats 60
S14.2 Results 60
Figure S14.2.1 Genetic differentiation between populations at colour-pattern regions 62
Figure S14.2.2 Genetic differentiation between populations at non-colour pattern regions 65

S15. *D*-statistics of chromosomes **66**
Table S15.1 *D*-statistics among chromosomes 66

S16. Linkage disequilibrium in *Heliconius* **67**
S16.1 Methods 67
S16.2 Results 67
Figure S16.2.1 Decline of linkage disequilibrium with physical distance 68

S17. Evidence for adaptive introgression at the *N/Yb* mimicry locus **69**
Figure S17.1 F_{ST} and ABBA-BABA site patterns along the *N/Yb* colour pattern region 69

S18. Phylogenetic analysis of resequenced individuals **70**
Figure S18.1a Phylogenetic analysis of resequenced individuals based on RAD sequence 70
Figure S18.1b Phylogenetic analysis of resequenced individuals based on SureSelect sequence 71

S19. Phylogenetic analysis across the *B/D* region **72**
Figure S19.1 Phylogenetic analysis across the *B/D* region 72

S20. Phylogenetic analysis across the *N/Yb* region **75**
Figure S20.1 Phylogenetic analysis across the *N/Yb* region (50 kb scale) 75
Figure S20.2 Fine-scale phylogenetic analysis across the *N/Yb* region (10 kb scale) 80

S21. Distribution of ABBA and BABA site patterns in non-colour pattern regions **83**
Figure S21.1 Distribution of ABBA and BABA site patterns in non-colour pattern regions 83

S22. Distribution map **84**
Figure S22.1 Distribution map of *H. melpomene* showing subspecies nomenclature 84

S23. Author contributions **85**

S24. References cited **86**

S1. Shotgun genome sequencing

S1.1 454 library construction

We prepared shotgun genomic, 3 kb paired-end, 8 kb paired-end libraries for sequencing on a XLR/Titanium Genome Sequencer. Shotgun genomic and 3 kb paired-end libraries were constructed from DNA that was isolated from a single *H. melpomene melpomene* pupa using a Qiagen (Valencia, CA) DNA isolation kit. The 8 kb mate-pair library was constructed from DNA from a sibling. Our 3 kb and 8 kb 454 mate pair libraries were prepared according to the manufacturer's protocol with modifications. 5 µg (15 µg for 8 kb) genomic DNA was sheared to 2-4 kb with a Covaris (Covaris, Inc. Woburn, MA) or to 6-9 kb by Hydoshear (Digilab INC, Holliston, MA). 8 kb mate pairs fragments were further size selected on a 0.7% agarose gel. The DNA fragments were end-repaired (NEBNext End-Repair Module; Cat. No. E6050L), and LoxP adaptor ligated (NEBNext Quick Ligation Module Cat. No. E6056L). Nicked DNA was repaired by strand displacement with the *Bst* DNA Polymerase and the DNA fragments quantitated. 100 ng (300 ng for 8 kb) size-selected fragments were circularized by Cre Recombinase (NEB, Cat No. M0298L). Any remaining linear molecules were removed by DNase/Exonuclease digestion.

The circularized DNA fragments were sheared again with a Covaris to a fragment length with an average size of 500 bp. After end repair, fragments containing the biotinylated junction linker from the circularized size-selected fragments were purified using streptavidin-coated magnetic beads (Invitrogen, Carlsbad, CA). These purified fragments were adapter ligated and PCR enriched. The library was size-selected using AMPure size exclusion beads (Beckman Coulter Genomics, Inc.; Cat. No. A63882). These dsDNA amplified molecules were immobilized once more on streptavidin-coated magnetic beads, and single-stranded Paired End DNA library was released by alkaline treatment, then neutralized and cleaned using MinElute PCR purification columns from Qiagen (Valencia, CA). All libraries were quality-checked on an Agilent 2100 Bioanalyzer (Santa Clara, CA) using an RNA Pico 6000 Lab Chip. Library concentrations were determined using a Ribogreen assay and each library diluted to 10^8 molecules prior to sequencing.

S1.2 454 sequencing

Single-stranded library was used as template for single-molecule emulsion PCR on 28 µm diameter beads. The amplified template beads were recovered after emulsion breaking and selective enrichment. The sequencing primer was annealed to the template and the beads were incubated with *Bst* DNA polymerase, apyrase and single-stranded binding protein. A slurry of the template beads, enzyme beads (required for signal transduction) and packing beads (for *Bst* DNA polymerase retention) was loaded into the wells of a picotiter plate. The picotiter plate was inserted in the flow cell and subjected to pyro-sequencing on the

Genome Sequencer XLR Titanium instrument (Roche). The XLR/Titanium Genome Sequencer flows 400 cycles of four solutions containing either dTTP, dATP, dCTP and dGTP reagents, in that order, over the cell. Each dNTP flow was imaged by a charge-coupled device camera on the sequencer, and images were processed in real time to identify template-containing wells and to compute associated signal intensities. The images were further processed for chemical and optical cross-talk, phase errors and read quality before base calling was performed for each template bead.

S1.3 Illumina library construction and sequencing

We also utilized Illumina technology for the reference strain, to correct for homopolymer errors inherent in the 454 technology and improve assembly. High molecular weight double strand genomic DNA samples were constructed into an Illumina paired-end library according to the manufacturer's protocol (Illumina Inc.). The library was sequenced on Illumina's Genome Analyzer IIx system according to the manufacturer's specifications. Briefly, cluster generations were performed on an Illumina cluster station. 36-76 cycles of sequencing were carried out in a separate, single flow cell lane on the Illumina GA II. Sequencing analysis was first done with Illumina analysis pipeline. Sequencing image files were processed to generate base calls and phred-like base quality scores and to remove low-quality reads.

After initial quality control, we generated a total of ~20M 454 and ~43M Illumina high quality reads (Table S1.3.1).

Table S1.3.1 Sequencing data for *H. melpomene* assembly

Only the 454 and Illumina reads that passed the initial quality control (i.e. reads with >63 bases of above zero reported quality) are counted here.

Data type	Number of reads	Average read length (bp)	Library mean	Library standard deviation
454 Shotgun	11,996,548	365	n/a	n/a
454 3 kb	5,369,849	220	2521	607
454 8 kb	3,832,283	212	4998	1111
Illumina paired-end	42,564,386	95	222	35

S2. Genome assembly

We applied the CABOG assembler 6.1 [1] to the combined 454 and Illumina data. As preliminary results using the Newbler assembler on the same data were inferior to those produced by CABOG we devoted our effort to optimizing the CABOG results. The success of any assembly project depends strongly on presenting the highest quality data to the assembly program. Therefore we applied additional pre-processing to the 454 paired-end data. The 454 shotgun and Illumina paired-end data required no extra pre-processing.

S2.1 Additional screening for linker

The 454 paired-end library construction process involves circularization of the DNA fragments with a 42-base linker whose sequence is known. Bioinformatic post-processing was performed to detect the linker and "unwrap" the paired-ends using the built-in sffToCA routine in the CABOG assembler [1]. Although sffToCA detects possible matches to linker and correctly forms mated reads in over 99% of cases, a small proportion of reads remain in the data with linker either at the end or in the middle of the read. Some of these reads led to linker appearing in our initial test assemblies. Because linker sequence will break contigs prematurely and sometimes leads to mis-assembly, it is imperative to ensure that linker sequence is removed from 454 paired-end data before assembly. We used NUCmer, part of the MUMmer package [2], to screen out and then trim off any sequences that mapped fully or partially to linker in the data.

S2.2 Redundancy screening

The circularization and the subsequent amplification of the 454 paired-end library results in redundancy, and the ends of the same DNA fragment may appear many times in the data. Redundancy breaks the assumption of uniform coverage that is in the design of most assemblers, including CABOG. All but one of the redundant mate pairs must be removed from the data before assembly. We therefore created a partial test assembly and looked for mate pairs that start and end at exactly the same base in the contigs, although we allowed the two ends to be in two different contigs. All but one copy of each mate pair were removed. Out of about 4.5 million 454 pairs from both the 3 kb and 8 kb libraries we found 15% to be redundant.

S2.3 CABOG assembly

We applied the CABOG assembler 6.1 [1] to the filtered set of the 454 and Illumina data, using the "mer" overlapper that includes subroutines to adjust for homopolymer errors in the 454 data. For the primary assembly, contig N50 is 51 kb and scaffold N50 is 212 kb (Table 2.3.1), which compares favorably with the recently published *Acropora digitifera* genome [3], also sequenced with a combination of 454 and Illumina reads, which had a contig N50 of 11 kb and a scaffold N50 of 192 kb. The final reported N50 was improved further by means of the linkage map and local scaffolding (Supplementary Information S4).

Table S2.3.1 Quantitative statistics from the CABOG assembly

Half of the total sequence in the assembly is contained in contigs/scaffolds larger than the contig/scaffold N50. Note that final assembly statistics reported in the Article refer to the assembly after linkage map verification and local scaffolding (Supplementary Section 4).

	Initial assembly	Primary non-redundant scaffolds	Redundant haplotype scaffolds
Bases in scaffolds	337 Mb	269 Mb	67 Mb
N50 scaffold size	169 kb	212 kb	n/a
N50 contig size	33 kb	51 kb	n/a

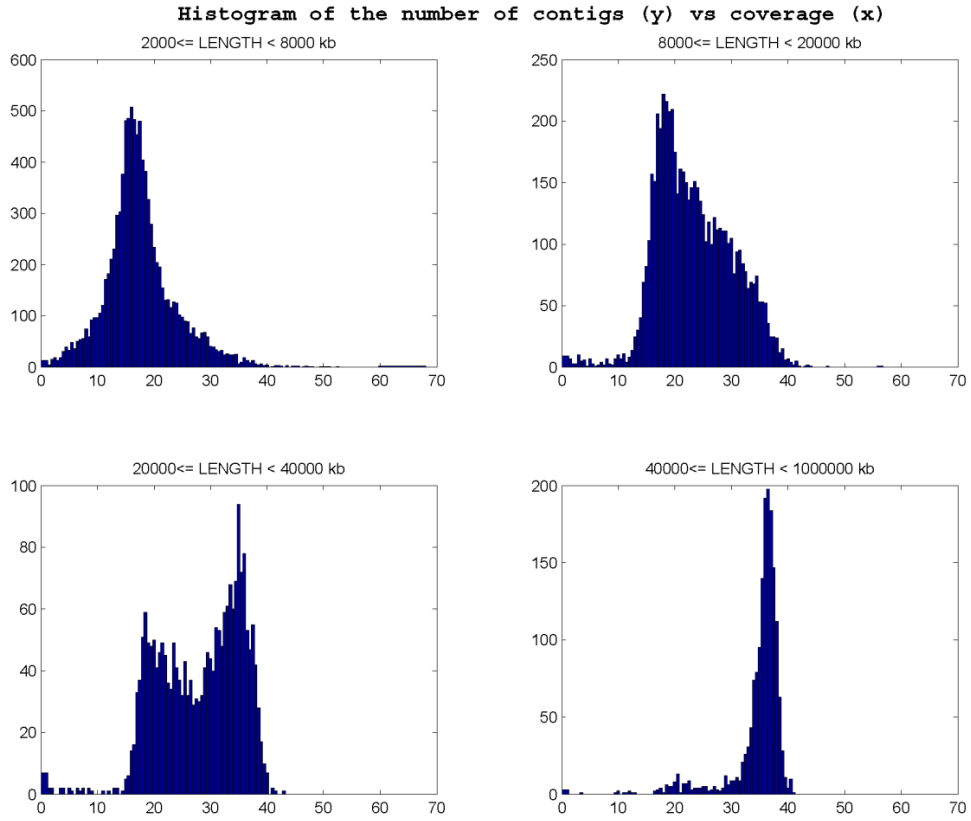


Figure S2.3.1 Coverage histograms for contigs of varying sizes

The bin size is 1. The *x* axis shows the coverage and the height of each bar shows the number of contigs that are covered *x* times.

S2.4 Haplotype separation

The total amount of sequence in scaffolds (336 Mb) exceeded the genome size estimate of 295 Mb. As the sequenced individual is diploid, it is likely that the additional sequence represented divergent haplotype copies of homologous chromosomal regions. To test this hypothesis, we examined the coverage and sizes of the contigs. We assumed that islands of haplotype differences on the chromosomes are relatively small and therefore large (>20 kb) contigs would likely contain reads from both homologous chromosomes. Therefore coverage of the large contigs would be consistent with overall average genome coverage. The smaller contigs are more likely to include divergent haplotype sequences that belong to the same location on the two homologous chromosomes. Figure S2.3.1 shows the histograms of coverage for sets of contigs of varying lengths. It is apparent that coverage of most contigs from 2000 bp to 8000 bp long is 19x, approximately half of the overall expected coverage of

38x. Longer contigs show a bimodal distribution, and contigs above 40 kb are nearly all covered at 38x.

Almost all contigs shorter than 8000 bp show reduced coverage and therefore most likely represent redundant haplotype copies of chromosomal regions already represented in larger scaffolds (Figure S2.3.1). In order to present a non-redundant reference sequence for a single haploid genome sequence, we needed to separate these redundant haplotype contigs. We therefore chose a set of simple rules in order to determine those contigs that were redundant. Briefly, we declared contig A to be a haplotype variant of contig B if all of the following apply:

B is longer than A
The mate pairs between A and B suggest that A is contained in B
There is a sequence alignment of >85% similarity between A and B that is consistent with mate pairs
Average coverage of A and B is less than 19/log(2) – from Poisson distribution

The last condition derives from the assumption that coverage of the contigs is distributed according to Poisson distribution. We estimated that each homologous chromosome is covered at about 19x. The 19/log(2) is the cutoff for distinguishing between the likelihood that the contig represents the region on one of the homologous chromosomes and the likelihood that the contig represents both homologous chromosomes.

Applying these three conditions we are able to report a primary assembly of 269 Mb, with an additional 67 Mb separated into haplotype contigs. Based on the new smaller assembly size of 269 Mb, the N50 contig size of the final assembly is 51 kb and N50 scaffold size is 212 kb.

S2.5 Local re-assembly

In addition, a 1.2 Mb super-scaffold was assembled based on a BAC tile path previously sequenced across the *HmYb* wing patterning locus, and local reassembly and manual super-scaffolding was performed for the *Hox* gene cluster (detailed below in section S10). Mitochondrial scaffolds were separated from the primary assembly and a full mitochondrial genome was assembled manually using Geneious v. 5.5.2.

S3. Assembly verification

S3.1 BAC end mapping.

16,000 BAC clones from a mixed race strain of *H. melpomene* library (insert size ~110 kb) were previously end sequenced at the Wellcome Trust Sanger Centre. These clones did not derive from the genome reference strain and so were not used in genome assembly. However they do provide an independent source of information for assembly verification. BAC end sequences were mapped to scaffolds using BLAST, and clones with mappings that suggested additional linkages noted. BAC end sequences with multiple mappings (suggesting repeat content) or low-quality mapping (suggesting the possibility of paralogy) were discarded. 6,142 BACs were mapped to 1,633 scaffolds, covering 235 Mb of the genome. Of these 6,142 BACs, 2,698 had both ends mapping to the same scaffold, and 3,444 spanned two different scaffolds. In 2698 cases both BAC end sequences from a single clone mapped uniquely in the genome and to the same scaffold. In 2694 of these, the end sequences were oriented with the expected orientation and insert size (109 kb ± 26 kb), providing strong support for the quality of the assembly.

S3.2 BAC sequence comparison.

Comparison with ~2.4 Mb of high quality finished BAC sequence provided further support for the quality of our assembly. This included tile paths across two contiguous regions associated with wing-patterning loci, *Yb* (1,149,501 bp) and *B*/*D* (726,198 bp), as well as another four unlinked individual BACs representing 528,150 bp. We used this sequence to evaluate the consensus quality and correctness of the assembly. The primary scaffolds were mapped to these BAC sequences using NUCmer. For example, Figure S3.2.1 shows twelve scaffolds that match the *Yb* locus. These scaffolds were subsequently linked into a single scaffold (Supplementary Information 2). The average percent identity of matches was 95%, which matches exactly the level of haplotype difference observed between the two haplotypes of the sequenced individual. The layout of matches shows that the scaffolds matching the *Yb* locus were assembled correctly. Results were similar across all of the remaining BAC sequences when aligned to the genome assembly, with mean coverage >90%, average match identity ranging form 92-95%, and no obvious inversions, misassemblies, or redundancies. To further determine the expected level of divergence between haplotypes sampled within *H. melpomene*, we also aligned 24 finished fosmid clones (mean sequence length of 35.6 kb) generated from *H. melpomene* of various wing races [4] to the tiled BACs across the *Yb* region. Average match identity of these fosmid clones ranged from 92% to 99%, similar to that seen between the reference genome and the finished BAC sequences. In summary, alignment to finished BAC sequences provides strong support for the quality of our assembly, with a percentage identity similar to that observed between other haplotypes sampled from the study species.

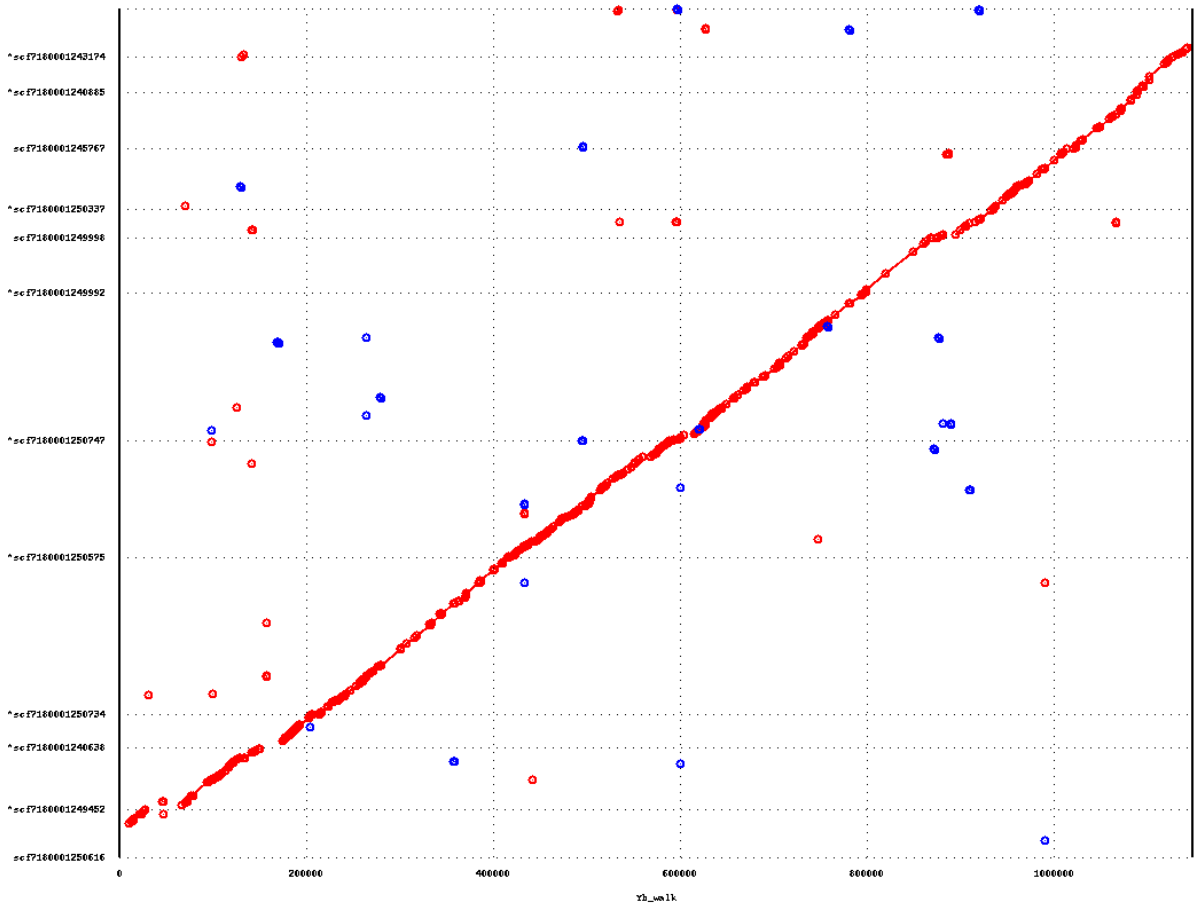


Figure S3.2.1. Twelve scaffolds matching the finished sequence of the *Yb* locus.

The contiguous matches are shown by lines and the beginning and end of each match are shown by circles. The x coordinate of each match is the location of the match on the finished sequence and the y coordinate is the location of the match in the scaffold. Scaffolds were laid out and oriented and horizontal dotted lines separate matches for individual scaffolds. Colour indicates the direction of the match – red means forward and blue means the reverse. The near-absence of blue regions in the matching alignment indicates that there are no inversion misassemblies in the scaffolds.

S4. Chromosomal scaffolding using a RAD linkage map

In order to provide an independent verification of the assembly quality and to assign scaffolds to their chromosomal linkage groups, we generated a linkage map using RAD sequencing method. This led to an improvement of the scaffold N50 from 212 kb in the initial assembly, to 399 kb for mapped superscaffolds. It also permitted identification and correction of a small number of errors in the whole-genome assembly, and construction of a chromosomal assembly including all of the known 21 chromosomes of *Heliconius melpomene*.

S4.1 The mapping cross

After four generations of inbreeding, a male *H. melpomene melpomene* from the same lineage used for the reference sequence, was crossed with a female *H. melpomene rosina* derived from a laboratory strain established from Gamboa, Panama (Figure S4.5.1). An F_1 intercross was performed between two siblings and to produce F2 progeny, many of which were frozen at a larval stage. Sex was determined from wing morphology of individuals that successfully eclosed. DNA was isolated using the DNeasy Blood & Tissue kit (Qiagen) and contaminating RNA was removed by treatment with 4 μl of RNase A (100 mg/ml) per sample.

S4.2 RAD library preparation

RAD library preparation was carried out according to Baxter et al.[5]. The method in brief, including some minor modifications, used 400 ng of genomic DNA from 45 progeny, 800 ng of each parent and 1200 ng of the *H. m. rosina* grandmother. Samples were digested with PstI restriction enzyme (NEB) in 50 μl volumes for one hour at 37°C then inactivated at 80°C for 20 min. P1 adapters containing a unique 5 bp Multiplex IDentifier (MID) were ligated (1 μl of 100 nM stock) to each sample using 1 μl T4 DNA ligase (NEB, 400,000 cohesive end units/mL) for 1 hour in 60 μl reactions at room temperature, then inactivated at 65°C for 20 min. Samples were diluted with 40 μl water, and combined into one of three pools, and sheared using a Bioruptor, set to high for 8 minutes. The sheared DNA was separated using agarose gel electrophoresis (0.5X TBE) then fragments in the 200-400 bp range were excised, purified and blunt-ended (NEB). dATP overhangs were added to fragments with Klenow exo- (NEB) and P2 adapter (1 μl of 10 mM stock) ligated. Products were purified, quantified, then PCR amplified using Phusion High-Fidelity DNA polymerase. A 100 μl master mix was made for each of the three libraries [1X HF buffer, 0.25 μM primers, 1 ng/μl library template, 0.1 mM dNTP]. To minimize the likelihood of PCR error, master mixes were divided into 8 separate 12.5 μl reactions for amplification [98°C 30 s. 14 cycles of 98°C 10 s, 65°C 50 s, 72°C 30 s, then a final extension for 5 min. at 72°C].

S4.3 RAD library sequencing and alignment to reference genome

Three prepared libraries were quantified, pooled and sequenced on a single lane of an Illumina HiSeq flowcell using 100 bp paired-end sequencing. DNA sequencing was carried out in the GenePool Genomics Facility in the University of Edinburgh (http://genepool.bio.ed.ac.uk). 108,307,444 100 bp raw read pairs were sequenced, of which 2,437,682 read pairs were discarded because they did not feature a valid sample MID or restriction site overhang (TGCAG in the case of PstI). Reads were separated by MID using RADtools v1.2.2 (available from http://radseq.info)[5]. Two progeny were discarded due to very low coverage, leaving 43 progeny. The genome scaffolds contained 27,120 PstI sites and so we expected to sequence on the order of 54,000 RAD tags (because PstI is a symmetric cutter). We achieved ~59x coverage of the F_1 mother (3,206,337 reads), ~66x coverage of the F_1 father (3,564,867 reads), ~112x coverage of the F_0 *rosina* grandmother and mean ~39x coverage of each progeny (mean 2,120,774 reads, SD 702,089).

Reads from each butterfly were aligned to the *H. melpomene* genome scaffolds using Stampy v1.0.13[6] with default parameters except for an insert size of mean 500, SD 100 and a substitution rate of 0.01 (to reflect the divergence between *H. m. melpomene* and *H. m. rosina*). BAQ scores were calculated for the alignments [7]. Stampy was parallelised using the processpart option and run using the Edinburgh Compute and Data Facility (ECDF, http://www.ecdf.ed.ac.uk/), partially supported by the eDIKT initiative (http://www.edikt.org.uk). Picard v1.48 tools were used to merge, sort and remove duplicates from the resulting alignments (using MergeSamFiles, SortSam, MarkDuplicates tools). Indels were realigned using the Genome Analysis Tool Kit (GATK) v1.1 RealignerTargetCreator and IndelRealigner tools[8]. Finally, genotypes were called across all individuals using the GATK UnifiedGenotyper[9] using heterozygosity 0.01 and incorporating BAQ scores from Stampy where possible. 11.6 Mb of the genome were covered by at least one individual, with 187,585 bases present in all individuals. These 187 kb bases are distributed across 2,341 scaffolds, covering 262 Mb (97%) of the reference genome. Bases where any individual had Genotype Quality less than 20 and where the base had Mapping Quality less than 67 or Coverage over 3600 were discarded (thresholds chosen empirically and iteratively based on quality of downstream linkage maps), leaving 30,452 bases on 1,797 scaffolds, covering 246 Mb (91%) of the genome.

S4.4 Defining markers for linkage mapping

Meiotic crossing over between sister chromatids does not occur during oogenesis in Lepidoptera, yet does occur during spermatogenesis [10]. Thus markers inherited from the female parent are non-recombining and completely define maternally inherited chromosomes. These 'chromosome prints' [11] are used to associate linkage groups with recombining paternally inherited markers, which are then used to construct linkage maps based on recombination[11]. A custom Perl script was used to extract candidate markers in

JoinMap 3.0 format from the bases present in all individuals. Bases homozygous in one parent and heterozygous in the other were retained as potential parent-specific markers. Bases heterozygous in both parents were retained as intercross markers. Bases homozygous in both parents but where the two parents differed were retained as potential maternal sex chromosome markers. In addition, bases where the mother's genotype was for a single allele *a* and the father's genotype was heterozygous for alleles *a* and *b* were retained as potential paternal sex chromosome markers if all male offspring were genotyped as *aa* or *ab* and all female offspring were genotyped as *a-* or *b-*. Positions with identical segregation patterns across all offspring were collapsed together into candidate markers. In total, 41 maternal segregation patterns derived from 7,463 SNPs, 391 paternal segregation patterns (7,921 SNPs), 807 intercross segregation patterns (14,341 SNPs), 16 maternal sex chromosome segregation patterns (640 SNPs) and 19 paternal sex chromosome segregation patterns (87 SNPs) were identified as candidates.

Candidate segregation patterns were processed by a second custom Perl script. The 41 candidate maternal segregation patterns (father homozygous and mother heterozygous) and 16 candidate maternal sex chromosome segregation patterns were collapsed to 23 candidate chromosome prints by merging mirror segregation patterns and patterns with 3 mismatches or less. The 807 candidate intercross segregation patterns were filtered to 680 after Chi-square tests for 1:2:1 (autosome) and 3:1 (sex chromosome) ratios. To further filter candidate segregation patterns, intercross segregation patterns were linked to maternal segregation patterns and markers retained where matches could be found. Nineteen maternal segregation patterns and 588 intercross segregation patterns remained after this process. These 588 intercross segregation patterns were converted to paternal segregation patterns by comparison with the respective chromosome print. Of the original 391 candidate paternal segregation patterns, 248 were matched to (and so validated by) a converted intercross segregation pattern, indicating that converting intercross patterns added a further 340 validated unique paternal patterns to the set. 149 of the original candidate paternal markers and 92 validated intercross markers were left unmatched.

S4.5 Linkage mapping

Firstly, we used all 397 paternal segregation patterns, 19 paternal sex chromosome segregation patterns and 588 converted intercross segregation patterns linked to 19 maternal patterns as input to JoinMap 3.0 [12]. In total 977 of 1004 markers were grouped into 19 chromosomes (LOD>4.0). Linkage maps of each group were constructed using the Kosambi mapping function and a LOD>1.0. The 92 intercross markers unlinked to maternal markers were used to construct two further linkage maps, containing 45 and 37 markers. These two maps did not match any of our existing 19 linkage maps, and so they completed a set of 21 chromosomes (Figure S4.5.1).

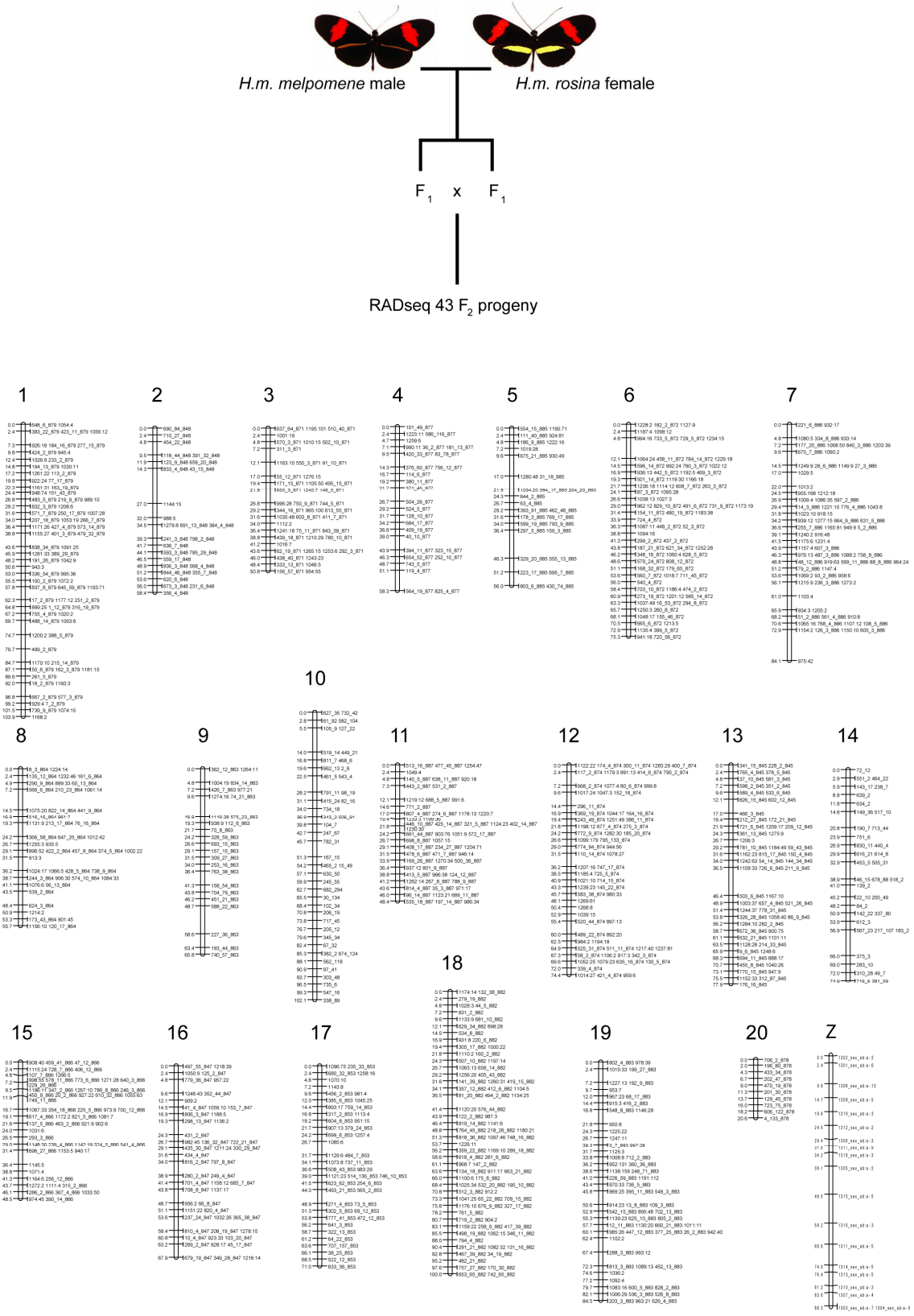


Figure S4.5.1 RAD linkage map

Positions in cM are shown on the left of each linkage group. Numbers on the right indicate Paternal marker ID, number of SNPs supporting a marker pattern and Intercross marker ID (e.g. 648_6_879).

S4.6 Placing scaffolds on chromosomes and correcting assembly errors

Our final map of 21 chromosomes contains 1,055 loci, representing 27,731 bases of the genome, on 1,649 scaffolds, covering 231 Mb (86%) of the genome. Markers could therefore be associated with scaffolds and scaffolds were placed and ordered on chromosomes based on the positions of their markers on the map. Most scaffolds mapped uniquely to one location on one chromosome, providing strong independent support for the quality of our primary assembly. However, in less than 4% of scaffolds (149/4115), we observed markers that mapped to multiple chromosomes, or to distant regions of one chromosome. This is not unexpected with a draft genome, as most draft assemblies contain as many or more errors, most of which go undetected. In a recent study, Salzberg et al. compared eight different assembly programs on multiple genomes, including human chromosome 14, for which the true assembly is known [13]. The data set used for human was very thoroughly cleaned, with no mis-paired reads permitted, and the median number of scaffold mis-joins was 268, even though the target assembly (at 88 Mbp) was only 1/3 the size of *Heliconius*. The quality of our primary assembly therefore compares favourably with this benchmarking study.

However, by using the RAD markers and *Bombyx* synteny data, we were able to correct the misassemblies. These represented 141 scaffolds with markers mapping to multiple chromosomes and 8 scaffolds with markers mapping to distant regions of one chromosome. Thus, the linkage map has significantly improved the quality of the assembly.

Each position in the linkage map corresponds to ~2.4 cM, or on average approximately 370 kb. Thus, in many cases multiple scaffolds were associated with each unique cM position on the linkage map. We improved ordering and orientation of the scaffolds associated with each cM locally using the existing 3 kb and 8 kb 454 mate pair libraries. Mate pair libraries were filtered to retain the 3,838 pairs spanning different scaffolds with less than a 20 kb span. Bambus2 (in the AMOS 3.0.1 distribution) was used to link together scaffolds with mate pair links less than 8 kb long and extend scaffolds further using existing RAD linkage information. Thus, the RAD map allowed us to take advantage of the reduced computational complexity of local assembly, as compared to global assembly, in order to further improve our genome quality. The two rounds of scaffolding that went into the genome assembly are reflected in the AGP file that reports scaffolding information. Thus, the first round of whole-genome scaffolding is encoded in the standard manner in the scaffold AGP as links between the genome contigs. The second round of chromosomal assignment and local scaffolding using Bambus2 is encoded as a second tier, in the chromosomal AGP file.

The final genetic map of the 21 chromosomes is 1455.6 cM long. 223 Mb (83%) of the genome could be mapped to a chromosome in 1,273 scaffolds with an N50 of 399,555 bp (Figure S4.6.1 and Table S4.6.1 for summary of chromosome lengths and scaffolds). The remaining 17% of the genome is assembled into 2,534 scaffolds; the complete genome is contained in 3,807 scaffolds with an N50 of 276,553 bp. Custom Perl scripts and R code written to scaffold the genome will be made available at Dryad (http://datadryad.org).

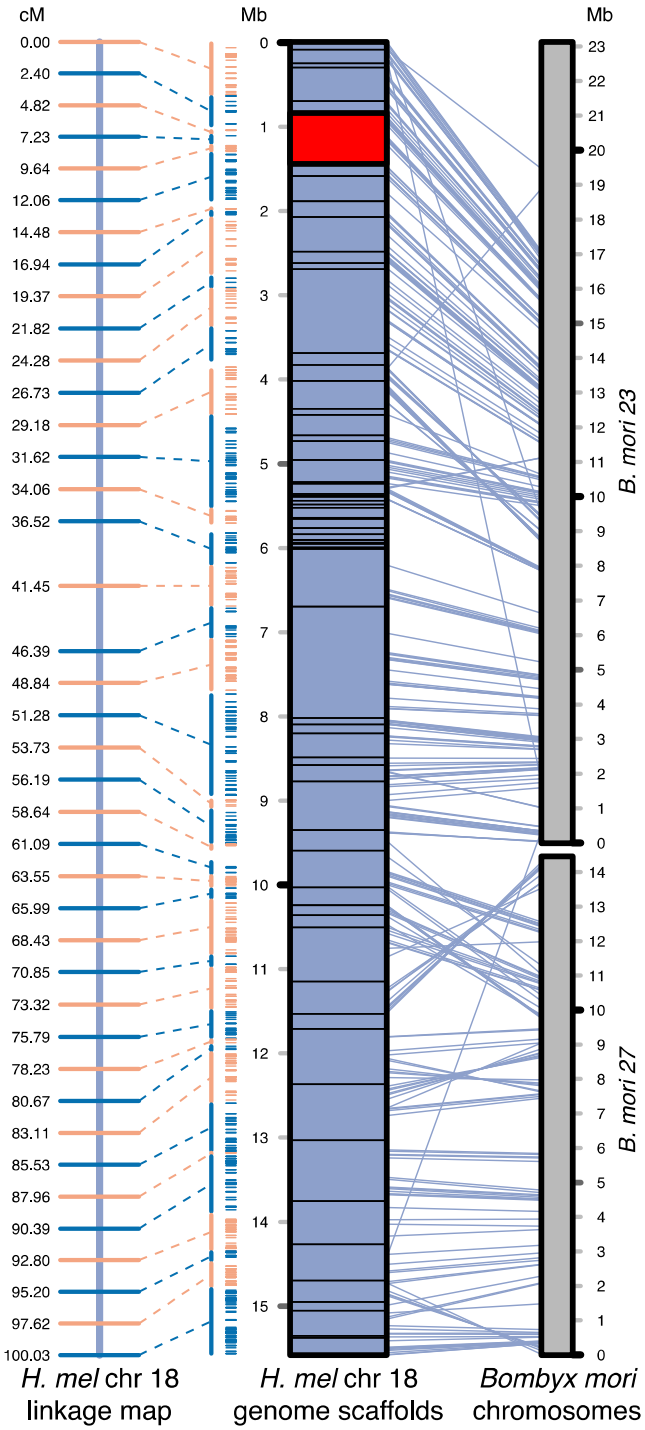


Figure S4.6.1 Example of chromosomal scaffolding for linkage group 18

The RAD linkage map and *Bombyx mori* synteny reveal the structure of *Heliconius melpomene* chromosome 18. Genome scaffolds were placed on chromosomes using a RAD linkage map containing 40 unique loci (orange and blue lines on linkage map) and 1,479 SNPs (orange and blue horizontal dashes left of *H. melpomene* chromosome) on 62 scaffolds (purple, separated by black lines). Gene orthologues in *H. melpomene* and *B. mori* were compared to assess synteny (purple lines). Scaffolds unordered by linkage data were ordered using synteny data where possible. The *B/D* red locus scaffold is highlighted in red.

Table S4.6.1 Summary of chromosome lengths in scaffolded genome.

Placed scaffolds have a chromosomal location based on the linkage map, while assigned scaffolds are associated with a chromosome but do not have a chromosomal location.

Chromosome	cM	Placed Scaffolds	Bases	Assigned Scaffolds	Bases
1	103.9	86	15,755,848	7	390,702
2	20.6	17	3,590,614	1	88,212
3	50.8	109	8,943,778	6	179,747
4	58.3	72	6,670,749	0	0
5	56.0	52	8,000,463	1	127,035
6	75.3	96	13,155,959	3	182,540
7	72.9	115	11,792,147	9	332,915
8	55.7	104	6,828,437	10	298,181
9	65.8	41	8,311,378	0	0
10	102.1	62	17,492,759	0	0
11	48.4	49	11,233,900	3	319,812
12	74.4	68	15,835,627	1	21,850
13	77.9	182	13,526,828	7	318,633
14	74.9	100	6,527,076	0	0
15	48.5	122	7,742,867	8	118,651
16	67.9	61	9,269,283	5	138,287
17	71.0	100	13,935,249	2	47,215
18	100.0	63	15,453,272	5	197,152
19	84.5	118	14,778,520	9	354,749
20	58.4	48	5,748,428	3	874,012
Z	88.3	27	4,088,377	3	478,145
TOTAL	1455.6	1692	218,681,559	83	4,467,838

S5. Novel repeat and transposable element (TE) identification

S5.1 Methods

We analyzed the assembled genome using two algorithms: PILER [14] and RepeatScout [15]. The PILER analysis was performed to search for recently active TEs of all types. Minimum length for discovered repetitive families was set to 100 bp and minimum identity for repeat families was set at 95%. The output from PILER was organized into families (all sequences with 95% and higher similarity) and superfamilies (sequences from two or more families that exhibited more extensive sequence similarity). Each superfamily and family alignment was given a numerical designation and 50% majority-rule consensus sequences were generated. Repeats consisting of low sequence complexity (satellite sequences) were removed and the filtered output was used to query the assembly using BLASTN. The top 40 hits obtained (generally E-value << 10-5) were extracted along with 500 bp of flanking sequence. Extracted sequences were aligned with MUSCLE [16,17] and revised consensus sequences were constructed. Full-length elements were presumed only when single copy DNA was identifiable throughout the element and up to both the 5' and the 3' ends of the alignments. In cases where the full length of the consensus sequence had not been captured the process was repeated until single copy DNA sequence was identifiable at both ends. A similar process was used for the RepeatScout analysis using the initial output as the starting point. The resulting library was then submitted to CENSOR [18] to ascertain their identity with regard to previously classified elements in RepBase [19]. RepBase classifications and likely TEs not present in RepBase were validated through identification of diagnostic characteristics of each TE class or family (A and B boxes, terminal repeats, poly-A tails, etc.). Any elements that could not be identified using these methods were classified as Unknown. After generation of the final TE library, we performed a RepeatMasker (v3.2.3) analysis of final genome assembly.

S5.2 Results and Discussion

Results are summarized in the Table S5.2.1. TEs comprise around 25% of the genome, a value similar to that found in *Drosophila melanogaster* (~22%) [20] but higher than in the African malaria mosquito, *Anopheles gambiae* (~16%), [21] and lower than in the yellow fever mosquito, *Aedes aegypti* (~47%) [22].

To date, only two other lepidopteran genomes have been fully sequenced and characterized with regard to TEs. The moth, *Bombyx mori*, harbors a TE load of ~35.4% consisting mainly of non-LTR retrotransposons [23]. The monarch butterfly, *Danaus plexippus,* apparently has a much lower TE content, ~11%[24]. However, less than half of that was readily classifiable into known TE families. Thus, making firm statements with regard to TE content in the monarch is difficult.

We used BLAST to query the *Bombyx* and *Danaus* genome drafts with our *Heliconius* TE library and retained elements exhibiting ten or more hits with e-values lower than 10^{-50}. Multiple elements from our de novo analysis of *Heliconius* were identifiable in each genome (*Bombyx*, n=9; *Danaus*, n=7). However, only two, the RTE family of retrotransposons and one unidentified family were shared by all three taxa.

A single SINE family and multiple DNA transposons from the Tc/Mariner and Helitron families appear to be recently active on the basis of high sequence similarity and, in the case of some autonomous elements, the identification of intact open reading frames at multiple locations. For example, we identified 109 full-length or nearly full-length autonomous insertions for a novel Tc3 element, Tc3_1_Hm. Of these, 42 harbored intact open reading frames.

LINEs comprise around 4% of the genome, in the same range as the 2.4% identifiable in *Danaus* but much smaller than the 13.8% in *Bombyx.* Like both genomes, the major LINE family is identified as belonging to the RTE family. There was evidence of potential recent activity for a subset of these elements; for example fifty-three full-length copies of RTE3_Hm were identified, five of which harbored intact ORFs.

Finally, as in *Bombyx*, the vast majority (94%) of TE fragments in the genome are ≤500 bp in length and many of the TE families show little divergence among individual insertions. In light of these data, the large numbers of fragmented elements along with the observation that few elements are shared in great numbers by all three taxa, suggests that some mechanism is acting to remove TEs from the genome. Thus, we predict that TE content is probably highly variable in lepidopterans and await additional genome drafts to confirm this prediction.

Table S5.2.1. TE content of the *H. melpomene* genome

	#hits	length (bp)	%genome
NonLTR elements	**196,035**	**32,569,663**	**12.07%**
SINEs	157,258	22,153,884	8.22%
LINEs	38,777	10,415,779	3.85%
RTE	7471	2,407,309	0.89%
Daphne	3870	1,222,570	0.45%
L2	3,336	1,114,806	0.41%
Jockey	1,805	925,997	0.34%
Zenon	3565	854,199	0.32%
Other/Unidentified	18730	3,890,898	1.44%
LTR elements	**4,923**	**1,230,130**	**0.45%**
LTR/Unknown	2,137	642,839	0.24%
Gypsy	2757	575,483	0.21%
Copia	29	11,808	0.004%
DNA transposons	**148,985**	**27,084,606**	**10.05%**
Helitron	52922	14,471,330	5.37%
Mariner	50125	5,744,437	2.13%
Tc3	37068	4,009,553	1.49%
hobo/Activator/Tam	2674	1,037,789	0.38%
PiggyBac	2604	853,461	0.32%
Other/Unidentified	3592	968,036	0.36%
Unclassified	**55,627**	**6,382,494**	**2.37%**
Total interspersed repeats	**405,570**	**67,266,893**	**24.94%**

S6. Gene prediction

S6.1 Methods

The MAKER pipeline was used to generate consensus gene predictions derived from *ab initio* models, transcriptome data, and protein similarity [25]. *Ab initio* predictors Augustus [26] and SNAP[27] were trained specifically for *H. melpomene* using transcriptome data. A broad range of transcriptomic sequences were generated to support gene predictions. These include 1) 6004 Sanger ESTs from developing wing discs 2) 1,879,038 Roche 454 ESTs from developing wing discs and larval midgut and 3) paired-end Illumina RNA-seq data from developing wings, eggs, whole pupal body, as well as abdomen from adult males and females (Table S6.1.1). Sanger and 454 ESTs were assembled together to create a non-redundant set of contigs. RNA-seq data were aligned to the masked reference genome and assembled into transcript models with the tophat/cufflinks pipeline [28]. These assembled ESTs and transcript models were used as transcript evidence in the MAKER pipeline. In addition to the transcript sequences, several protein databases were provided to the MAKER pipeline for homology evidence: the proteomes of *Bombyx mori* (SilkDB), *Apis melifera* (RefSeq [29]), *Drosophila melanogaster* (RefSeq), *Tribolium castaneum* (beetlebase), *Acyrthosiphon pisum* (RefSeq), and Uniref90 [30] (May 26, 2011). The MAKER pipeline incorporates a repeat masking step before gene predictions. We applied a repeat database consisting of 1) All complex hexapoda repeats from RepBase 2) our *de novo* characterisation of repeats from the *H. melpomene* genome 3) 'mobile_element' features from all lepidopteran nucleotide GenBank entries, and 4) TEfam entries for *Anopheles gambiae*. A total of 12,562 protein coding genes were predicted by MAKER before manual curation.

Table S6.1.1 Summary of RNA-seq data used for automated gene prediction

Tissue	No. reads
Body	6,958,901
Brain	14,847,007
Egg	15,300,176
Pupae	4,783,134
Pupal hindwing	16,286,492
Larval hindwing	15,704,677
Adult abdomen male	9,447,660
Adult abdomen female	15,765,253

To quantify the amount of support for these gene predictions we examined how many predicted exons and translatable coding sequence (CDS) coincided with the various evidence

tiers included in the MAKER pipeline. Using scripts from the GFFtools package (biowiki.org/GffTools) we counted exon and CDS features where 90% or more of the sequence overlapped with evidence features arising from 1) EST & Sanger alignments [est2genome], 2) RNAseq alignments [cufflinks], 3) included insect proteomes [protein2genome] and 4) the UniRef90 database [protein2genome] (Table S6.1.2).

Non-coding RNA genes were predicted with RNAmmer and tRNAScan-SE [31,32]. 87 rRNA and 2373 tRNA genes were predicted.

Table S6.1.2 Summary of supporting evidence for automated gene prediction

Evidence tier	No. exons	No. CDS
total features	90,934	84,311
est2genome	29,472	29,620
cufflinks	71,436	69,090
prot2gen_insect	41,445	45,348
prot2gen_uniref	32,127	34,790

S6.2 Manual verification

Gene families of interest were manually curated. We focused on UDP-glucuronosyltransferase genes (UGT), Cytochrome P450s, pigmentation, olfactory and chemosensory genes (SI Section S9), the *Hox* genes (SI Section S10) and immunity-related genes (SI Section S11). Putative P450 gene sequences were sent to the cytochrome P450 nomenclature committee for naming and refinement of gene models [33]. Additional scaffolds were manually curated to confirm the quality of gene predictions, including the *B/D* and *Yb* regions. In total 51 UGT genes, 94 UGT genes, 129 immunity-related genes, 150 olfactory and chemosensory and a total of ~800 gene models were manually curated. This manual effort raised the total count of predicted protein coding genes to 12,657.

S7. Characterisation of miRNAs

S7.1 Identification of known miRNAs

Genome-wide annotation of previously described miRNA families was performed by extracting all known mature metazoan miRNA sequences from miRBase v17.0 [34] and identifying their location in the *Heliconius melpomene* genome using MapMi v1.0.4-b01 [35]. Briefly, this package scans miRNA input sequences against the genome using the Bowtie algorithm [36] allowing for two mismatches but no gaps. Potential precursor sequences were then identified by extending to 110 nucleotides surrounding each match and folding using ViennaRNA [37]. Sequences were filtered to contain mature miRNAs of between 21 and 23 nucleotides and duplicate sequence matches were discarded with the exception of the one containing the fewest number of mismatches. Sequence hairpins were inspected manually and results were grouped according to miRNA family and genome location.

S7.2 Prediction of novel *Heliconius* miRNAs

In a previous study, a library of *Heliconius* small RNAs (sRNAs) was created using total RNA extracted from larval and pupal wing tissue of two races (*H. m. melpomene* and *H. m. rosina*) and enriched for miRNAs [38]. Raw Illumina small RNA reads were processed by converting FASTQ to FASTA format, then removing any adaptor sequences with exact matches to the first eight bases of the 3' adaptor. Any sequences without adaptor matches, with less than three distinct nucleotides or matching known transfer RNA or ribosomal RNA extracted from Rfam [39] were excluded from further analyses. This generated a total of 6,895,260 processed sRNA sequences. These sequences were mapped to the *Heliconius* genome using miRCat [40], which identifies mature miRNAs and their precursors. Briefly, this program identifies clusters of sRNA reads that map to the genome and computes the most likely miRNA from the cluster (based on sequence abundance). The surrounding flanking sequence was then folded using RNAfold [37] and the secondary structure analysed for miRNA hairpins. The precursor miRNA candidate was tested using randfold [41]. Visual representations of novel candidates were generated using the UEA RNA Hairpin Folding and Annotation Tool [40].

95 previously known miRNA loci were detected in the *Heliconius* genome, representing 81 unique miRNAs. These include the conserved miRNAs *bantam* and *let-7*, in addition to 10 miRNAs previously characterised in *Heliconius* using Northern blot (miR-10, -184, -193, -263, -275, -276, -277, -308, -317 and -2788) [38]. We also predicted an additional 14 miRNA loci, corresponding to 12 novel miRNAs that are currently unique to *Heliconius* (i.e. there are no known homologues deposited in the current version of miRBase [34]). Hence, a total of 93 different miRNAs were identified in *Heliconius*. Of the 93 miRNA loci identified in the *Heliconius* genome, 23% were conserved between vertebrates and invertebrates, 16% were conserved among invertebrates, 34% were insect specific, 14% were Lepidoptera specific and 13% were found only in *Heliconius*.

S8. Genome evolution and synteny analyses

The completion of the *H. melpomene* genome enables the first comprehensive assessment of genomic rearrangement rates among Lepidoptera through comparison to the silkmoth *Bombyx mori* and to the Monarch butterfly *Danaus plexippus*, which are estimated to have diverged from *H. melpomene* 100 MYA and 60 MYA respectively [42]. The analyses show that synteny is highly conserved, with 10 chromosomal fusions observed in *H. melpomene* relative to *B. mori*. At a microsynteny level, rates of chromosomal breakage are similar to those previously observed in *Drosophila*.

S8.1 Methods

Bombyx mori sequences and genome annotations (Build 2 Version 3) were obtained from KAIKObase (http://sgp.dna.affrc.go.jp/KAIKO/index.html and *Danaus plexippus* sequences (version 1 and version 2) were obtained from MonarchBase (http://monarchbase.umassmed.edu). Pairwise one-to-one orthologous relationships between the *H. melpomene* core gene set and the *B. mori* GLEAN consensus gene set or the *D. plexippus* official geneset v1.0 were identified by running Inparanoid version 4.0 on the peptide sequences [43]. Only matches with bootstrap support of >95% and a score of >50 were retained for analysis. In total, we identified 7252 orthologous genes on primary scaffolds in comparisons with *B. mori*, of which 6010 were mapped to chromosomes in both species. 8172 orthologous genes were identified in comparisons between *H. melpomene* and *D. plexippus*, of which 8084 could be localised unambiguously to *D. plexippus* version 2 genome scaffolds by mapping with Exonerate est2genome [44].

Synteny blocks and chromosomal rearrangement counts were estimated using OrthoCluster in –rs mode [45]. Perfect synteny blocks were detected by allowing no mismatches between the two genomes. Imperfect synteny blocks, in which a degree of local rearrangement is permitted, provide an indication of higher-order synteny and were identified by relaxing the run parameters to allow up to 25% in-map or out-map mismatches, as described by Vergara and Chen [46]. For analyses of syntenic block parameters, all genes without a 1-to-1 orthologue match were excluded from the analysis.

Estimates of the rate of chromosome breakages per Mb per My were obtained from the rearrangement counts reported by OrthoCluster for perfect synteny (minimum block size = 1). These counts were converted from rearrangements to breakages by applying conversion factors of 2 breaks per inversion and reciprocal translocation, 1.5 per insertion/deletion and 3 breaks per transposition [47,48]. The genomic length was calculated by summing the lengths of the analysed scaffolds, with the exception of the large-scaffolds analysis in *B. mori*, where a genome size correction factor was applied to the *H. melpomene* scaffold length (since *B. mori* scaffolds are generally large and only a fraction of each scaffold was involved in the comparison).

Chromosome-level visualisations were made using CIRCOS [49], with the bundlelinks tool applied to simplify visualisation. For Figure 2b, different scales were applied to the two genomes so that each occupied a semi-circle in the plot. Minimum bundle membership was set to 3 and the maximum gap between adjacent orthologues was limited to 500 kb.

S8.2 Results

For chromosome-level analyses with *B. mori*, we used the subset of 6010 one-to-one orthologues assigned to both the *H. melpomene* RAD linkage map and one of the 28 *B. mori* chromosomes to identify homologous segments. Overall, 11 of 21 *H. melpomene* linkage groups show homology to a single *B. mori* chromosome and ten linkage groups have contributions from two *B. mori* chromosomes (Figure 2 and Table S8.2.1). These findings are consistent with previous studies [42], with the higher resolution of our analysis revealing an additional four major chromosome fusion events. Little inter-chromosomal gene exchange is seen: 5645 (94%) of predicted orthologues are found on homologous chromosomes (Figure S8.2.1 and Table 8.2.1).

To assess microsynteny, we analysed the relative gene order and orientation of all *H. melpomene-B. mori* or *H. melpomene-D. plexippus* orthologues using OrthoCluster [46]. Within multi-gene scaffolds, extensive microsynteny was apparent, with local rearrangements also frequently observed. In comparisons with *B. mori,* perfect synteny blocks containing multiple orthologues spanned a median distance of 45 kb (88 kb when controlling for genome fragmentation into scaffolds) and contained an average of 4.2 (6.2) orthologues (Table S8.2.2). The *D. plexippus* genome assembly is more fragmented than that of *B. mori*, and we obtained estimates of the median syntenic block size of 40.8 kb (129.1 kb in the large scaffolds dataset), and average orthologue content of 4.2 (9.3) genes (Table S8.2.3). In total, 5751 *H. melpomene-B. mori* orthologues (79%) and 5856 *H. melpomene-D. plexippus* orthologues (72%) are found in synteny blocks of 2 or more genes.

To obtain upper and lower bounds for these estimates, we analysed two datasets in parallel: 'All scaffolds' and 'Large scaffolds'. In the 'All scaffolds' dataset all gene-containing scaffolds were included and, in contrast to the analysis of syntenic blocks in Tables S8.2.2 and S8.2.3, all annotated genes were incorporated. This dataset might overestimate the number of chromosome breaks, to an unknown extent, for example due to genome fragmentation into scaffolds or discrepancies in gene prediction and annotation between the two genomes. A lower bound is provided by the analysis of the 'Large' scaffolds dataset. Here, only orthologous genes were considered and, to minimize the impact of genome fragmentation, only scaffolds greater than 500kb in *H. melpomene* and *D. plexippus* were retained for analysis.

We obtained estimates of 0.05-0.13 chromosome breaks per Mb per My in analysis of *H. melpomene-B. mori*, and 0.04-0.29 chromosome breaks per Mb per My between *H.*

melpomene and *D. plexippus.* The holocentric chromosomes of Lepidoptera have been reported to show extensively conserved synteny at the chromosome level [42] but high rates of local rearrangements [47]. d'Alençon *et al* [47] conducted a fine scale analysis of microsynteny in BAC clones derived from the moth species *Helicoverpa armigera* and *Spodoptera frugiperda* relative to 15 homologous regions in the *B. mori* genome. They reported chromosome breakages to occur at a rate of approximately 2 breaks/Mb/My, higher than the 0.5-0.7 breaks/Mb/My seen in the holocentric chromosomes of *Caenorhabditis* [48,50], which is itself four-fold higher than the rate observed in *Drosophila* [51] and in excess of other characterized organisms with non-holocentric chromosomes [46]. In contrast, our analyses give lower estimates, suggesting that rates of chromosome breakages in the Lepidoptera are more comparable to those previously reported in *Drosophila*.

Table S8.2.1 Summary of chromosome homology between *H. melpomene* and *B. mori*

H. melpomene autosomes and their *B. mori* counterparts are shown. Homologies not detected by Pringle *et al.* [42] are highlighted with an asterisk. The partitioning of the 6010 one-to-one orthologues mapped to chromosomes in both genomes is indicated. Where multiple *B. mori* chromosomes contribute to a single *H. melpomene* chromosome, the orthologue count for each contributing chromosome is shown in sequence.

H. melpomene chromosome	*B. mori* chromosome	No. of orthologues on homologous linkage groups	No. of orthologues on non-homologous linkage-groups
1	4, 24*	348, 110	21
2	16	105	0
3	6	183	26
4	21	124	17
5	3	229	11
6	9 , 11*	202, 123	25
7	2, 11	138 231	11
8	25	231	5
9	7	201	12
10	5 , 28	359, 114	21
11	15	378	14
12	8 , 20	277, 145	41
13	14 , 22	108, 214	13
14	19	164	6
15	17	177	12
16	18	186	19
17	13 24	257, 92	40
18	23 , 27	218, 107	10
19	12 , 26*	216, 138	25
20	10 , 23*	121, 43	33
Z	1	106	3

Table S8.2.2 Comparison of syntenic block parameters between *H. melpomene* and *B. mori*

	All Scaffolds		Large Scaffolds	
Number of 1-to-1 orthologues	7252		1869	
Number of scaffolds in *H. melpomene*	1496		81	
Total scaffold length in *H. melpomene* (Mb)	223.9		59.5	
Number of scaffolds in *B. mori*	315		111	
Total scaffold length in *B. mori* (Mb)	432.3		114.8	
	Block size		**Block size**	
PERFECT SYNTENY BLOCKS	≥ 1 gene	≥ 2 genes	≥ 1 gene	≥ 2 genes
No. of synteny blocks	2864	1363	489	266
Mean (genes)	2.5	4.2	3.8	6.2
Interquartile range (genes)	1-3	2-5	1-4	2-8
Mean (kb)	41.1	76.6	83.5	146.1
Median (kb)	14.9	44.7	21.0	87.8
Interquartile range (kb)	5.0-43.7	22.2-94.2	4.8-100.0	33.1-198.4
IMPERFECT SYNTENY BLOCKS				
Number of synteny blocks	2541	1177	350	177
Mean (genes)	3.0	5.2	5.8	10.5
Interquartile range (genes)	1-3	2-6	1-5	2- 15
Mean (kb)	50.8	99.6	134.4	258.6
Median (kb)	14.6	50.5	15.6	142.0
Interquartile range (kb)	4.8-46.9	23.6-114.5	3.9-143.3	35.9-418.4

Table S8.2.3 Comparison of syntenic block parameters between *H. melpomene* and *D. plexippus*

	All Scaffolds		Large Scaffolds	
Number of 1-to-1 orthologues	8084		562	
Number of scaffolds in *H. melpomene*	1545		42	
Total Scaffold length in *H. melpomene* (Mb)	227.2		32.5	
Number of Scaffolds in *D. plexippus*	1358		37	
Total Scaffold length in *D. plexippus* (Mb)	209.7		34.2	
	Block size		**Block size**	
PERFECT SYNTENY BLOCKS	≥ 1 gene	≥ 2 genes	≥ 1 gene	≥ 2 genes
Number of synteny blocks	3301	1498	90	57
Mean (genes)	2.4	4.2	6.2	9.3
Interquartile range (genes)	1-3	2-5	1-5.8	3-9
Mean (kb)	35.0	65.6	121.1	186.3
Median (kb)	14.3	40.8	33.3	129.1
Interquartile range (kb)	5.4-39.0	21.0-84.4	7.3-179.5	42.2-301.1
IMPERFECT SYNTENY BLOCKS				
Number of synteny blocks	3031	1392	75	52
Mean (genes)	2.7	4.8	7.6	10.5
Interquartile range (genes)	1-3	2-6	1-8.5	3-15.3
Mean (kb)	39.7	75.3	153.4	218.5
Median (kb)	14.8	45.1	75.4	157.8
Interquartile range (kb)	5.4-43.0	22.8-94.8	7.8-283.9	65.5-373.9

Table S8.2.4 Estimates of rearrangement rates between *H. melpomene* and *B. mori*

		All Scaffolds	Large Scaffolds
H. melpomene	Number of genes	16199	1869
	Number of scaffolds	2462	81
	Total scaffold length (Mb)	257.9	59.5
B. mori	Number of genes	14622	1869
	Number of scaffolds	517	111
	Total scaffold length (Mb)	438.6	114.8
Orthologues	Number of 1-to-1 orthologues	7252	1869
	Number of synteny blocks	4804	489
Rearrangements	Number of inversions	2956	285
	Number of transpositions	381	118
	Number of insertions/deletions	652	0
	Number of reciprocal translocations	4804	489
Chromosome breakages	Total number of breakages	17641	1902
	Breakages per Mb per MY	0.13	0.05

Table S8.2.5 Estimates of rearrangement rates between *H. melpomene* and *D. plexippus*

		All Scaffolds	Large Scaffolds
H. melpomene	Number of genes	16199	562
	Number of scaffolds	2462	42
	Total scaffold length (Mb)	257.9	32.5
D. plexippus	Number of genes	16355	562
	Number of scaffolds	2520	37
	Total scaffold length (Mb)	238.5	34.2
Orthologues	Number of 1-to-1 orthologues	8084	562
	Number of synteny blocks	4606	90
Rearrangements	Number of inversions	2664	45
	Number of transpositions	218	15
	Number of insertions/deletions	1325	0
	Number of reciprocal translocations	4606	90
Chromosome breakages	Total number of breakages	17182	315
	Breakages per Mb per MY	0.29	0.04

S9. Olfactory and chemosensory proteins

Odorant-binding proteins (OBPs), chemosensory proteins (CSPs) and odorant receptors (ORs) are families of genes known to play an important role in chemosensation [52]. These are large gene families whose diversity is hypothesized to reflect ecological niches occupied by different insects [53,54]. To explore how ecological divergence between moths and butterflies might be reflected in these groups of proteins we comprehensively surveyed OBPs, CSPs, and ORs in *H. melpomene* as well as *B. mori* and *D. plexippus*.

S9.1 Methods

S9.1.1 Gene discovery, curation, and nomenclature

Protein sequences of all known *B. mori* and other selected lepidopteran OBPs, CSPs, and ORs were queried against the *H. melpomene* and the *D. plexippus* (version 1) genome [24] using a variety of search algorithms (tBLASTn, Exonerate, Genewise [44]). These alignments were used to generate novel gene predictions and validate automated gene predictions. In order to avoid over-estimation of gene copy number, we used a conservative criterion by only including genes that differ from one another by at least 7 amino acids (>5% divergence). After initial rounds of phylogenetic analysis, we also searched for previously uncharacterized members of these gene families in *B. mori via* reciprocal tblastn searches.

H. melpomene OBPs, CSPs, and ORs were numbered according to their closest *B. mori* homologues in phylogenetic analysis [55–59]. Where there were conflicting gene names for *B. mori* CSPs, we first used the nomenclature of Foret et al. [58] and, secondarily, those of Gong et al. [57]. When no *B. mori* OR homologue existed, the number of the closest *Spodoptera littoralis or Manduca sexta* OR was assigned. When possible, *H. melpomene* OR, OBP and CSP sequences present on the same scaffold were assigned consecutive numbers.

S9.1.3 Phylogenetic analysis

Curated OBP, CSP and OR protein sequences from *B. mori*, *D. plexippus*, and *H. melpomene* were aligned with ClustalW. These three alignments were visually inspected and manually adjusted as needed. Maximum likelihood trees were estimated in PhyML [60] with 500 bootstrap replicates. Tree images were created using the iTOL web server [61].

S9.2 Odorant-binding proteins (OBPs)

Odorant-binding proteins are small globular proteins which vary in length from ~140-220 amino acids and contain six conserved cysteine residues. We have identified 43 putative OBPs in the *Heliconius melpomene* genome and 35 OBPs in the *Danaus plexippus* genome (this study and ref [24]). Recent antennal transcriptome studies [62,63] reported 17 and 18 putative OBPs in, respectively, *S. littoralis* and *M. sexta*, while we identified one additional OBP in *B. mori* on scaffold nscaf3027*,* an ortholog of *H. melpomene* OBP44, bringing the total known number of OBPs in silkmoth to 46 [55,56]. The OBP family notably includes two Lepidoptera specific sub-families: the pheromone-binding proteins (PBPs), thought to transport pheromone molecules, and the general odorant-binding proteins (GOBPs), thought to transport general odorants such as plant volatiles. Very recently, a female pheromone gland enriched OBP has been identified in *B. mori* [64] (Table S9.2.1). We identified two *Heliconius* GOBPs, HmelOBP1 and HmelOBP2, one in each of the two lepidopteran GOBP lineages. Of the three putative pheromone-binding proteins expressed in silkmoth antennae, we found only two homologs in *H. melpomene*, HmelOBP3 and HmelOBP6 (Figure S9.2.1). Lastly, we identified *Heliconius* specific duplicates (HmelOBP11 and HmelOBP12) of the female pheromone gland specific BmorOBP11. In addition, unlike in *Bombyx* and *Danaus* where OBP gene expansions are rare, there is a striking expansion consisting of eight *H. melpomene* OBPs that appear to have duplicated since *Heliconius* and *Danaus* shared a common ancestor (Figure S9.2.1).

Table S9.2.1 CSP and OBP proteins previously reported in silkmoth antennae and female pheromone glands and their homologs in *H. melpomene*.

Tissue protein expression in *Bombyx mori*	***Bombyx mori* gene name**	***Heliconius melpomene* gene name**
Female pheromone gland	**CSP1[1] (CSP4[2])**	**HmelCSP4, 22, 23, 24,**
	CSP2[1] (CSP13[2])	HmelCSP13
	CSP6[1] (CSP1[2])	HmelCSP1
	CSP8[1] (CSP3[2])	HmelCSP3
	CSP9[1] (CSP5[2])	N/A
	CSP11[1] (CSP8[2])	**HmelCSP8, 19**
	CSP15[1] (CSP11[2])	HmelCSP11
	OBP11[3]	**HmelOBP11, 12**
Male and female antennae	OBP2[3,4]	HmelOBP2
	OBP3[3,5] (PBP)	HmelOBP3*
	OBP27[3]	N/A
	CSP1[1] (CSP4[2])	HmelCSP4, 22, 23, 24,
	CSP2[1] (CSP13[2])	HmelCSP13
Male antennae	OBP20[3]	HmelOBP20
Female antennae	OBP1[3.4]	HmelOBP1
	OBP25[3]	N/A
	CSP8[1] (CSP3[2])	HmelCSP3
	CSP9[1] (CSP5[2])	N/A
Male antennae	OBP20[3]	HmelOBP20

Note: Expansions in the *Heliconius* lineage of homologs with known pheromone gland expression in *Bombyx mori* are shown in bold. Gene names in parentheses are shown in Fig. 3a and Fig. S9.3.1.

[1] Gene names from Gong et al.[57]

[2] Gene names from Gong et al., Foret et al., Dani et al. and Viera et al. [55,56,58,64]

[3] Gene names from Gong et al. [55]

[4] General odorant-binding protein (GOBP)

[5] Pheromone binding protein (PBP)

**B. mori* OBP3 and OBP4 are homologous to HmelOBP3

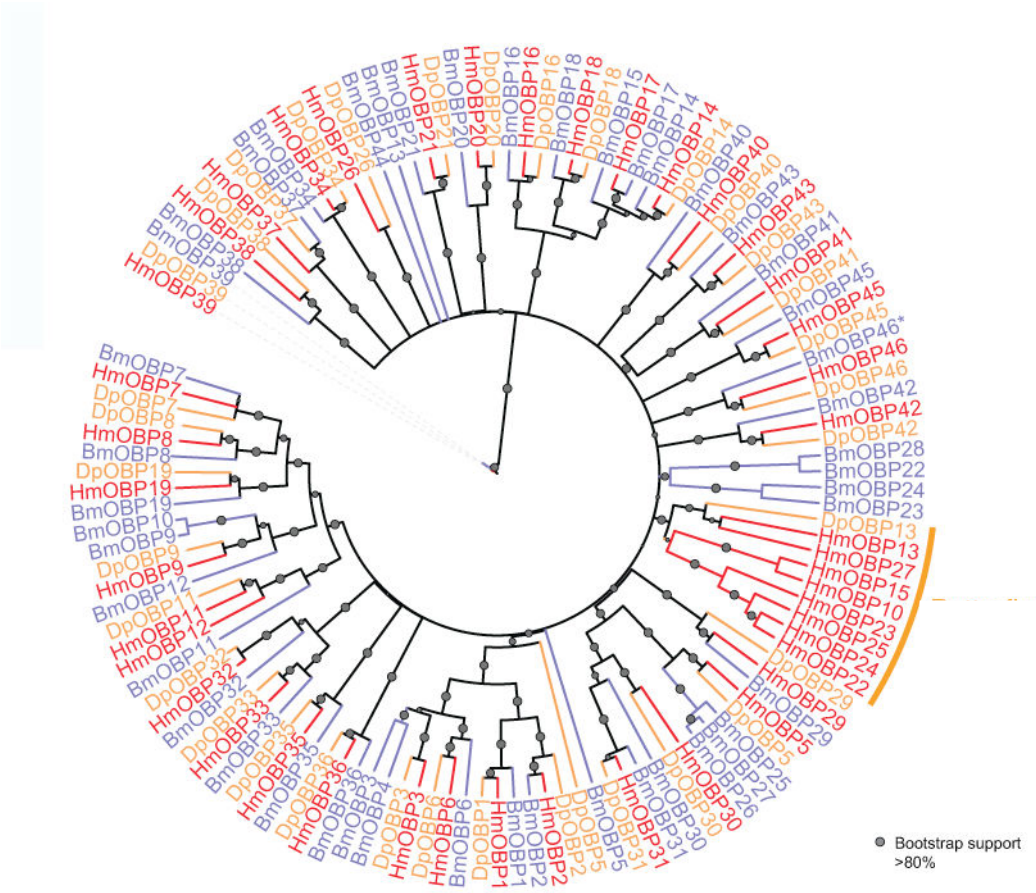


Figure S9.2.1 Maximum likelihood tree of odorant-binding proteins

Bm, *Bombyx mori*; Hm, *Heliconius melpomene;* Dp, *Danaus plexippus*. Genes from *D. plexippus, H. melpomene* and *B. mori* are shown in orange, red, and blue, respectively. Grey circles on branches indicate bootstrap values >80% from 500 bootstrap replicates. Branches highlighted by an orange arc indicate *Heliconius* specific OBP expansions.

S9.3 Chemosensory proteins (CSPs)

We have identified 33 putative chemosensory protein (CSP) genes in the *Heliconius melpomene* genome and 34 CSPs in the *D. plexippus* genome (Fig. 3a). By contrast, 24 genes have been annotated in the *B. mori* genome[56–58] and 9 have been identified in the *S. littoralis* antennal transcriptome[62]. Our data confirm the occurrence of a large number of CSPs in phytophagous insects (Lepidoptera, Orthoptera) compared to the small number identified in genomes from other orders[65]. Notably, the number of predicted CSPs in *Heliconius* and *Danaus* is considerably larger than that found in any other insect whose genome has been sequenced so far (e.g., *Drosophila* spp., 3-4; *Anopheles gambiae*, 8; *Tribolium castaneum,* 19; *Apis mellifera*, 6; *Acyrthosiphon pisum*, 19; *Pediculus humanis,* 7)[56]. Unlike other insects, where few lineage-specific gene duplications have been observed, at least two large independent butterfly-specific expansions have occurred since butterflies and *Bombyx* shared a common ancestor (Figure 3a). Since CSPs are widely distributed in insect tissues, a function as general hydrophobic ligand carriers has been proposed for this protein family[52]. Recent proteomic work on *B. mori* has revealed an enrichment of seven CSPs in female pheromone gland and five CSPs in male and female silkmoth antennae (Table S9.2.1). For one abundant female pheromone gland CSP in silkmoth, which is also enriched in male and female antennae, the *H. melpomene* genome contains six homologs (HmelCSP4, 22, 23, 24, 27, 28) (Table S9.2.1). For one other female pheromone gland CSP in silkmoth, the *H. melpomene* genome contains independent duplicates (HmelCSP8 and HmelCSP19, respectively). The largest CSP expansion in *D. plexippus* consists of 12 CSPs while the largest expansion in *H. melpomene* consists of 6 CSPs. The CSP expansions in *D. plexippus* and *H. melpomene* compared to other Lepidoptera may possibly be related to sequestration/storing of hydrophobic toxins that caterpillars ingest from their toxic host-plants and concentrate in their tissues as a protection from predation[66].

S9.4 Olfactory receptors (ORs)

We identified 70 putative OR genes in the *Heliconius melpomene* genome, including the insect obligatory co-receptor ORco (HmelOR2)[67], one homologue of the *cis*-jasmonate specific larval receptor BmorOR56[59], one homologue of a monoterpene citral receptor (HmelOR49)[68], and one homolog of a putative pseudogene in *B. mori* BmorOR52 (Figure S9.4.1). A total of 64 ORs were reported in the *D. plexippus* genome [24], although nine of these were too short to be included in our alignment and phylogenetic analysis. We further identified two additional ORs in *D. plexippus* with clear homology to *H. melpomene* ORs 67 and 71 as well as five new ORs in *B. mori* (BmorOR69-73, extending the nomenclature of Tanaka et al. [59]). For comparison, 47 OR mRNAs have been reported in an antennal transcriptome of *M. sexta*[63] and from our analysis and previous work, 72 ORs were found in the *B. mori* genome, three being pseudogenes[59 64]

Several ORs associated with male- or female-enriched pheromone detection in *B. mori* have duplicated in *Heliconius*. We found two homologues of the female-enriched receptor BmorOR30[69] (HmelOR30, 31) and four members of the male-enriched pheromone receptor subgroup, three of which are *Heliconius* specific (HmelOR9, 10, 11). No homologues of the female-enriched receptor subgroup containing BmorOR45, 46, and 47 were found. The largest butterfly expansions of ORs appears to have occurred in the same clade both in *Danaus* and *Heliconius* independently.

Phylogenetic analysis revealed several HmelORs arranged on the same scaffold, often clustered with only one BmorOR (e.g., HmelOR5, 6, scf7180001250321; HmelOR62, 63, scf7180001250807). Another cluster of *Heliconius* specific ORs (HmelOR35, 36, 38 to 40) occurred on scf7180001250804. Interestingly, the male-enriched pheromone receptor subgroup (HmelOR9, 10, 11) had 3 genes located on the same scaffold (scf7180001250419). While the total number of ORs (11) in *B. mori* associated with male- or female-enriched pheromone detection is higher than the total number in *H. melpomene* (6), the relatively high proportion of *Heliconius* and *Danaus* specific ORs may be related to the complex pheromonal communication observed in butterflies [70–72].

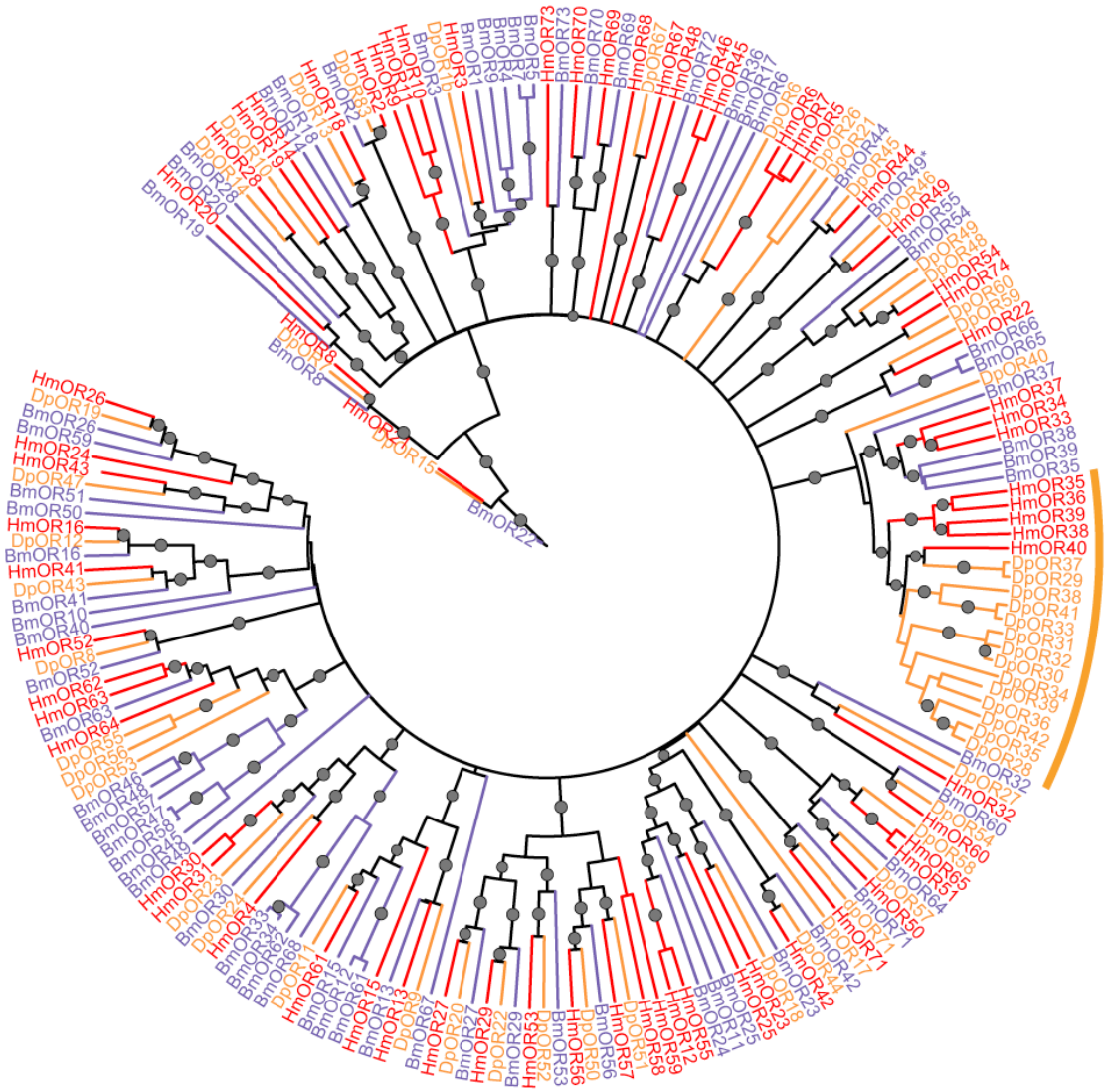


Figure S9.4.1 Maximum likelihood tree of olfactory receptors

Bm, *Bombyx mori*; Hm, *Heliconius melpomene;* Dp, *Danaus plexippus*. Genes from *D. plexippus, H. melpomene* and *B. mori* are shown in orange, red, and blue, respectively. Grey circles on branches indicate bootstrap values >80% from 500 bootstrap replicates. Branches highlighted by an orange arc indicate butterfly specific OR expansions.

S10. Homeobox genes

S10.1 Initial characterization and local reassembly of the *Hox* cluster

Putative homeobox (*Hox*) genes were identified in *H. melpomene* by BLAST searches of both scaffolds and predicted genes using the homeobox domains of *D. melanogaster* and *A. mellifera Hox* genes, and the *B. mori Shx* homeodomains [73]. In addition, scaffolds were searched for the highly conserved motif within the canonical homeobox domain 'WFQNRR'. These initial efforts revealed the *Hox* genes were spread across at least 10 scaffolds. An *in silico* tile-path (i.e. an ordering of scaffolds into a super-scaffold for the relevant region) was constructed across the *Hox* cluster using BLAST searches of scaffolds from the main assembly, the haplotype scaffolds, and an independent, preliminary genome assembly (Helmel5). Ultimately 26 scaffolds were required to span the complete *Hox* cluster region. Given the importance of this genomic region, we decided to create an improved assembly of the *Hox* region by local reassembly.

For the reassembly we initially collected all reads in original scaffolds that mapped to the *Hox* region to produce a set of read sequences R. We then used all mates of reads in R to the set to obtain set M. The recently developed MSR-CA 1.3 assembler (http://www.genome.umd.edu/SR_CA_MANUAL.htm) was then used to assemble the combined set {R U M}. This reassembly produced a substantial increase in scaffold sizes for the desired genomic region. Aligning the reassembled scaffolds to the original scaffolds yielded a tile-path between assemblies where all original scaffolds were covered by a set of seven reassembled scaffolds. This tiling also provided the order and orientation of the reassembled scaffolds.

S10.2 Identification of lepidopteran *Hox* genes

H. melpomene gene models were generated for the seven reassembled scaffolds using MAKER similarly to the original scaffolds. Putative *Hox* genes were re-identified via BLAST searches as previously described. These gene models were manually edited based on evidence from sequence similarity with other insect *Hox* genes and alignment of *H. melpomene* transcripts. The *B. mori* and *D. plexippus* genomes were searched for putative *Hox* genes using published sequences from *D. melanogaster* and *B. mori* where available, and the *H. melpomene* gene models. BLAST searches were carried out against both the predicted gene sets and genome scaffolds.

S10.3 Phylogenetic analysis of insect *Hox* genes

The full-length *Hox* genes for *H. melpomene* and *D. plexippus* were aligned with homeobox sequences from *D. melanogaster*, *A. mellifera*, *T. castaneum* and *B. mori* using ClustalW, implemented in Geneious (V. 5.4.6). Homeobox sequences were downloaded from Homeodb [74], and for *B. mori* manually entered based on published information[73]. The alignment was then cropped to the canonical homeobox domain. Prottest was used to determine the model of sequence evolution (LG+G+I) which best fitted the data[75], and maximum likelihood phylogenetic analysis was carried out using Phyml with 100 bootstrap replicates (Figure S10.4.2).

S10.4 Results

Phylogenetic analysis indicates that tandem duplication of *PG3/zen* gave rise to three distinct groups of *shx* genes in the common ancestor of *Bombyx* and the nymphalid butterflies: *A, B* and *C* (Figure S10.4.1 and Figure S10.4.2). The *ShxA* gene has undergone additional rounds of duplication to give eight genes in *B. mori* (*Bm/Shx1-8*), some of which have multiple homeodomains. *B. mori* has also undergone a further duplication of *ShxC* to yield *Bm/Shx10* and *11* (Figure S10.4.1). No *Bombyx* orthologue was recovered for the nymphalid butterfly *ShxD* gene. The evolutionary origin and orthology of lepidopteran *Shx* genes was further confirmed by their position and orientation within the *Hox* cluster, with only *B. mori shx7* showing an altered direction of transcription relative to other *ShxA*-derived genes (Figure 10.4.3).

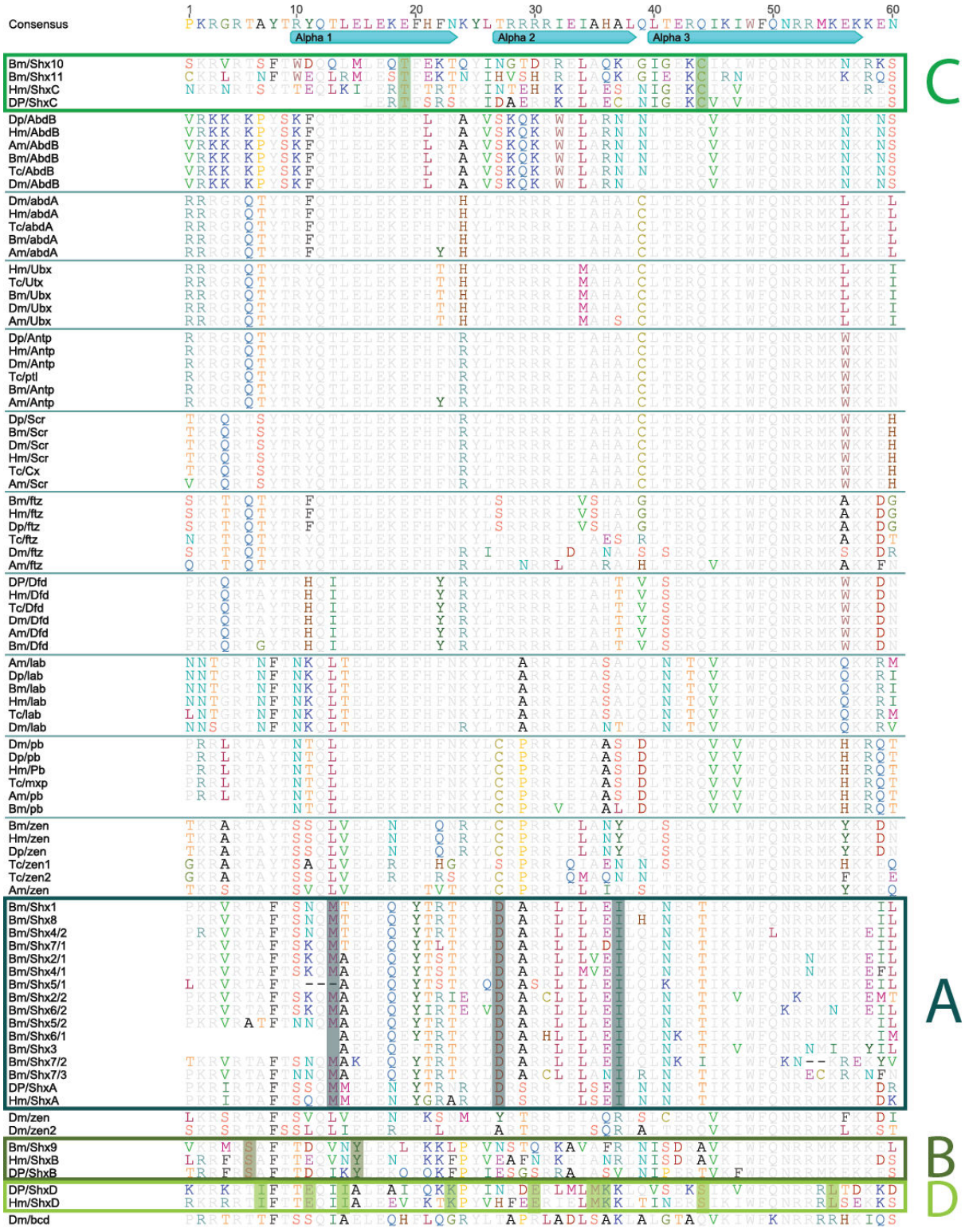


Figure S10.4.1: Alignment of insect homeobox domains

Homeodomains were extracted from full length genes, aligned using Clustalw and annotated in Geneious. Where *B. mori* genes have more than one homeodomain they are numbered in order of the direction of transcription. Amino acids differing from the consensus are highlighted, and three highly conserved alpha helices are illustrated. Numbering refers to position within the canonical homeodomain. The lepidopteran *Shx* clades from phylogenetic

analysis (Fig. S10.4.2) are boxed, and diagnostic residues (those which occur in the homeodomains of an *Shx* but in no other *Hox* genes) are shaded. Positions 5, 12, 16, 17, 25, 38, 48 and 49 were invariant across all *Hox* homeodomains analysed, including the *Shx* genes. The *H. melpomene ftz* homeodomain was taken from a previous genome assembly version, as it was not present following the local assembly of the region.

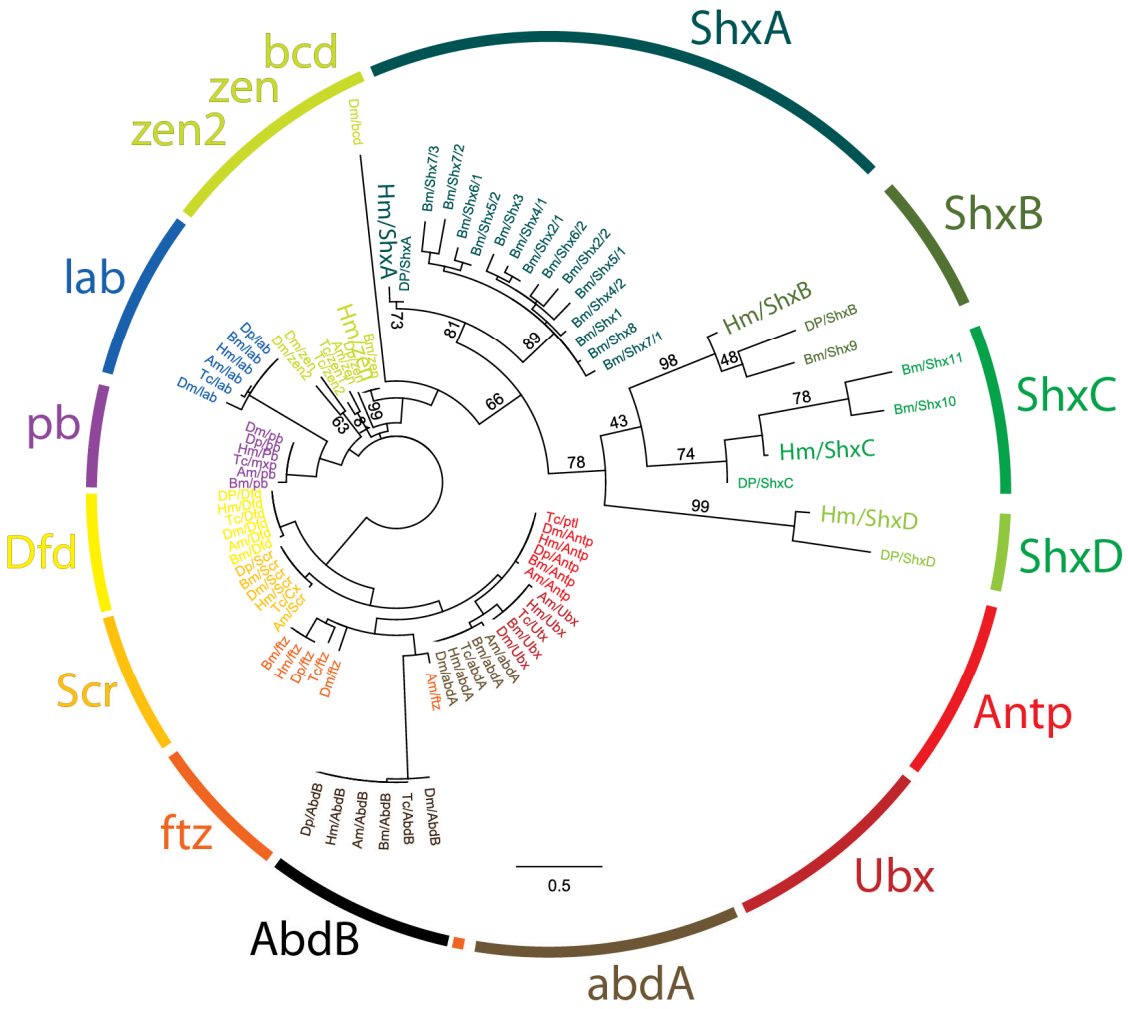


Figure S10.4.2 Phylogenetic tree of *Hox* genes from *D. melanogaster* (Dm), *T. castaneum* (Tc), *A. mellifera* (Ac), *B. mori* (Bm), *D. plexippus* (Dp) and *H. melpomene* (Hm).

Maximum likelihood tree generated with 100 bootstrap replicates using PhyML. Only those support values relating to *Shx* genes, and over 40, are shown. There is good support for three lepidopteran *Shx* clades – *A*, *B* and *C*; and for a nymphalid butterfly *ShxD* clade. *H. melpomene Shx* genes are highlighted.

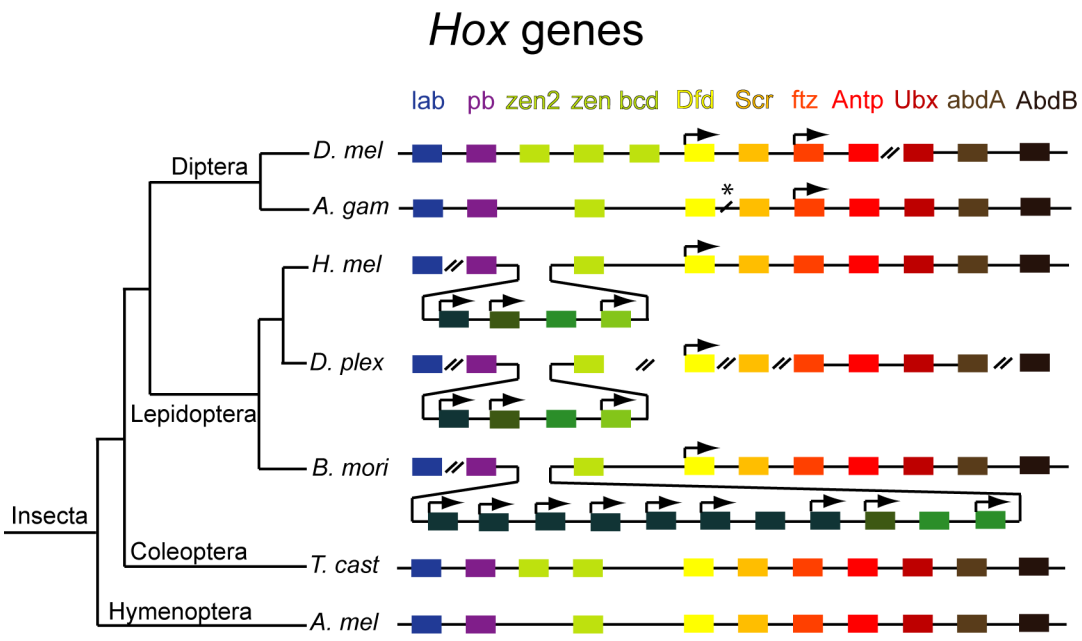


Figure S10.4.3 Expansion of the lepidopteran *Hox* cluster

Colouring indicates *Hox* gene orthology. Phylogenetic analysis suggests that *zen* has independently duplicated in *D. melanogaster* (*zen*, *zen-2*, *bcd*), and *T. castaneum* (*zen*, *zen-2*). Known physical linkage is indicated by a solid line. // indicates a break in the *Hox* cluster, whilst the *A. gambiae* */ indicates that the order of *lab-Dfd* is inverted in the current genome assembly. Gene orientation within the cluster is from right to left, unless indicated by an arrow.

S11. Immunity Genes

Immune systems are under strong selection from pathogens [76,77], and it is thought that this has driven the expansion and contraction of many families of immunity genes in different insect taxa [77,78]. The ecology of butterflies makes it likely that their immune systems experience different selection pressures to well-studied groups like mosquitoes and *Drosophila*, and it is therefore of interest to see how this has shaped their immune systems. Here we survey the genomes of *Heliconius melpomene* and *Danaus plexippus* for homologues of genes that are known or suspected of having immune functions in *Drosophila melanogaster* and *Bombyx mori*.

S11.1 Methods

A consolidated core of *H. melpomene* Maker predictions from both primary and haplotype scaffolds as well as the *D. plexippus* Official Gene Set (OGS1.0) [24] were searched against known immunity genes from both *B. mori* [79] and *D. melanogaster* [78] using Blastp. To confirm that there were no immunity genes that had not been predicted by Maker, *H.melpomene* and *D.plexippus* scaffolds were also searched with tBlastn. All gene predictions were also searched for conserved domains known to function in immunity [78,79] with the NCBI Batch CD-search [80]. Putative orthologues between *D.melanogaster, B. mori* and each of *D. plexippus* and *H. melpomene* were also identified as best reciprocal hits with InParanoid 4.1 [81]. All putative homologues were additionally queried against the NCBI Protein Database with apparently spurious hits filtered from the final gene counts.

For *H. melpomene* only, gene models of predicted orthologues were manually examined and curated using Apollo [82]. Updated gene models were then re-run through the same process, and low quality hits were removed. Where relevant, results from both *H. melpomene* and *D. plexippus* were grouped into gene families by sequence homology to known members in other taxa [78,79,83,84]. For each family longest-isoform peptides were retreived for six further taxa: *A. aegypti* (AaegL1.2,[78]), *A. gambiae* (AgamP3.6, [78]), *A. mellifera* (PreRelease2. [83]), *D. melanogaster* (r5.41), *B. mori* (Release2.0,[79]) and *T. castaneum* (Tcas3.0,[84]).

The peptide sequences were then aligned using ClustalX2.1 [85]. In cases where it was difficult to align sequences, only conserved sections or domains were used as identified by Gblocks [86]. Gene trees were reconstructed using maximum-likelihood by PhyML [87] with 1,000 bootstrap replicates and the best ranking model fit of sequence evolution for each protein set as identified by Modelgenerator [88].

S11.2 Results

The four main immunity signalling pathways — IMD, TOLL, JAK/STAT, and JNK — were present in both *Heliconius* and *Danaus*, with most pathway members in 1:1 relationships with homologues in *Bombyx* and other insects (Table S11.2.1).

Gene families involved in the recognition of parasites are particularly prone to expansions and contractions. However, most putative recognition molecules in both *Heliconius* and *Danaus* (including FREPs, TEPs, Draper, Eater, nimrods and DSCAM) are 1:1 orthologues with *Bombyx* (Table S11.2.1). We found fewer peptidoglycan recognition proteins (PGRPs) in both *Heliconius* and *Danaus* than in *Bombyx*, but a greater number of beta-1,3 glucan recognition proteins (βGRPs) - probably due to a recent lineage-specific expansion in butterflies (tree not shown). The overall similarity of recognition molecules between *Bombyx, Danaus, and Heliconius* contrasts with the Diptera, where families of proteins such as FREPs, TEPs and nimrods are extensively duplicated and lost [77,78] (Table S11.2.1).

Homologues were also found for all *Bombyx* effector gene classes including all anti-microbial peptides (AMPs) as well as RNAi, PPO and NOS genes (Table S11.2.1). Canonical *Drosophila* RNAi genes were found in both *Heliconius* and *Danaus* in 1:1 relationships with *Drosophila.* The size of other AMP families can differ considerably between closely related insects [77,78] and this is also the case when comparing *Bombyx*, *Danaus*, and *Heliconius*. Across AMP families, there tends to be a lower gene copy number in *Heliconius* than *Bombyx* or *Danaus. Danaus* also possesses an additional copy of the attacin encoding gene. A greater number of PPO genes were found in both *Danaus* and *Heliconius* (Table S11.2.1).

Overall it is striking how similar immune systems within the Lepidoptera are, with most genes being 1:1 orthologues (data not shown). In particular there is no evidence for the wide copy number variation observed in some gene families in the Diptera, or the loss of canonical signalling components seen in other insect taxa [89].

Table S11.2.1 Immunity related genes in seven insect species

Counts of *H. melpomene* and *D. plexippus* genes were generated by this analysis. Those from six other insect species: *Anopheles gambiae* (Ag), *Aedes aegypti* (Aa), *Drosophila melanogaster* (Dm)*, Apis mellifera* (Am), *Bombyx mori* (Bm) and *Tribolium castaneum* (Tc) are from † Tanaka et al. 2008, †† Evans et al. 2006, * Waterhouse et al. 2007, ** Zou et al. 2007 respectively. Counts of RNAi genes are from Zhan et al. 2011 (Dp, Bm, Tc), Tomoyasu et al. 2007 (Dm, Tc) and the NCBI Protein Database. '-' Denotes putative absence or unknown.

Recognition and Related	**Hm**	**Dp**	**Bm†**	**Am††**	**Aa***	**Ag***	**Dm***	**Tc****
Peptidoglycan Recognition Proteins (PGRPs)	7	8	11	4	7	7	13	6
ß-Glucan recognition proteins (ßGRPs)	7	6	4	2	7	7	3	3
Fibrinogen-related Proteins (FREPS)	3	3	3	2	34	46	13	7
Scavenger receptor (SCR)	15	17	18	14	19	19	22	21
Thioester Containing Protein (TEPs)	3	1	3	3	6	13	6	4
Nimrods	2	4	3	4	2	-	10	4
Modulation								

Recognition and Related	Hm	Dp	Bm†	Am††	Aa*	Ag*	Dm*	Tc**
CLIP Serine Proteases	22	33	15	16	66	53	46	46
Signalling								
Toll Pathway								
Späetzle-like Proteins (SPZ)	2	2	3	2	7	9	6	6
Toll-like Receptor	15	13	13	4	11	10	9	9
TOLLIP	1	1	2	1	1	2	1	1
MYD88	1	1	1	1	-	1	1	-
TUBE	1	1	1	1	1	1	1	1
PELLINO	1	1	1	1	-	1	1	1
PELLE	1	1	1	1	1	1	1	1
TRAF2	1	1	1	1	-	1	1	1
ECSIT	1	1	1	1	1	1	1	-
CACTUS	1	1	1	3	1	1	1	1
DIF/DORSAL	1	1	1	2	-	1	1	2
IMD Pathway								
IMD	1	1	1	1	1	1	1	1
TAK1	1	1	1	1	1	1	1	1
IKKG	1	1	1	1	1	1	1	1
IKKB	1	1	1	1	1	1	1	1
FADD	1	1	1	1	1	1	1	1
DREDD	1	1	1	1	1	1	1	1
TAB2	1	1	1	1	1	1	1	1
IAP2	1	1	1	1	1	1	1	1
UBC13	1	1	1	1	-	1	1	-
RELISH	1	1	1	2	1	1	1	1
JNK Pathway								
HEM	1	1	1	1	-	1	1	-
JNK	1	1	1	1	1	1	1	1
FOS	1	1	1	1	-	1	1	-
JUN	1	1	1	1	1	1	1	1
JAK/STAT Pathway								
PIAS	1	1	1	1	-	1	1	-
SOCS	1	1	1	1	1	1	1	-
DOMELESS	1	1	1	1	1	1	1	1

Recognition and Related	Hm	Dp	Bm†	Am††	Aa*	Ag*	Dm*	Tc**
STAT	1	1	1	1	1	2	1	1
Effectors								
Antimicrobial Peptides								
Attacins	2	3	2	-	1	-	4	3
Cecropins	3	6	12	-	9	4	5	2
Gloverins	1	3	4	-	-	-	-	-
Moricin	1	2	1	-	-	-	-	-
RNAi								
Dicer-1	1	1	1	1	1	1	1	1
Dicer-2	1	1	1	-	1	-	1	1
Drosha	1	1	1	1	-	-	1	1
R2D2	1	1	1	-	-	-	1	2
Loquacious	1	1	1	-	-	-	1	1
Pasha	1	1	1	-	-	-	1	1
Argonaute-1	1	1	1	1	-	-	1	1
Argonaute-2	1	1	1	1	-	-	1	2
Argonaute-3	1	1	1	-	-	-	1	1
Piwi	1	1	1	1	1	1	1	1
Aubergine	1	1	1	-	-	-	1	1
Other								
Nitric Oxide Synthetases (NOS)	5	3	2	1	1	2	1	1
Prophenoloxidases (PPOs)	5	4	2	2	10	9	2	3

S12. Genomics methods for introgression study

S12.1 Sample collection and DNA extraction

Samples were collected from wild populations (Tables S12.2.1 and S12.3.1). These include *H. melpomene amaryllis* (postman pattern), *H. melpomene aglaope* (rayed pattern), *H. timareta* ssp. nov. (postman pattern), *H. timareta florencia* (rayed pattern), five outgroup species with divergent wing patterning from the silvaniform clade (Fig. 1a, main paper) including the ray-patterned *H. elevatus*, as well as several additional races of *H. melpomene*, *H, timareta* and *H. cydno* that were used only in a RAD phylogeny. Wings were removed from specimens and kept separately for reference and identification. Tissues were preserved in a NaCl-saturated DMSO solution at -20°C. Whole genomic DNA was extracted from one-third of the thorax of each specimen using the DNeasy Blood and Tissue Kit (QIAGEN), and DNA extracts stored at -20°C.

S12.2 SureSelect targeted sequencing

SureSelect probes (Agilent Technologies) were used to enrich genomic DNA for specific targeted regions prior to sequencing. Genomic regions targeted for sequencing included those known to contain the *HmYb* (~1.1 Mb BAC walk) and *HmB/D* (~0.7 Mb BAC walk) loci, controlling respectively yellow and red wing pattern element differences between races of *H. melpomene* [4], as well as a further ~1.8 Mb of sequence located in 55 separate genome scaffolds outside the two colour pattern regions. These non-colour pattern regions included three sequenced BACs (13H8, 27N4 and 7E22) as well as several contigs from a preliminary draft *H. melpomene* genome assembly.

Paired-end libraries were prepared for 22 individuals (Supplementary Table S12.2.1) as recommended in SureSelect protocol (Agilent, G3362-90001_SureSelect_Exome_2.0.1, May 2010) with some modifications. Purified DNA (1-2 ug per sample) was sheared using a Covaris S2 or E210 to approximately 150-200 bp (DNA1000 Assay, Agilent Bioanalyzer). Post-shearing and subsequent purifications were performed with SPRI beads (AMPure XP, Beckman Coulter) as recommended, with the exception that bead drying was performed at room temperature. End-repair, A-tailing, and ligation were performed using NEBNext DNA Sample Prep Reagent Set 1 (E6000L, NEB) and custom Illumina adapters modified to include a 5 bp barcode (MID) at the 3' end (Supplementary Table S12.2.2). Pre-capture PCR was performed for 6 or 10 cycles as described in the SureSelect manual using half of the purified ligation template, the Illumina PCR primers PE1.0 and PE2.0 and Herculase II polymerase (Agilent). After purification and quantification, pools of four libraries containing 500-700 ng of template library were dried and resuspended in 3.4 ul water. Capture, wash, elution, and PCR was performed as described in the SureSelect protocol using 8 cycles or 12 cycles of

amplification. After purification, libraries were determined to have a peak fragment length approximately 300-350bp (DNA1000 Assay, Agilent Bioanalyzer).

Post-capture libraries were prepared for Illumina sequencing as recommended by the manufacturer, run on a single Illumina HiSeq2000 lane and 100 base paired-end sequence data collected to deliver 30-150 fold coverage of targeted sequences (GenePool, University of Edinburgh).

Table S12.2.1 Details of samples used in SureSelect targeted sequencing

Sample ID	Taxon	Sample locality	Latitude	Longitude
09-246	*H. melpomene aglaope*	Km-103.1 Tarapoto-Yurimaguas, Peru	05° 58' 18'' S	76° 13' 55'' W
09-247	”	”	”	”
09-357	”	”	”	”
09-268	”	Idea Religiosa, Munichis, Peru	05° 54' 37'' S	76° 13' 33'' W
09-75	*H. melpomene amaryllis*	Urahuasha, Km-8 Tarapoto-Yurimaguas, Peru	06° 27' 49'' S	76° 20' 07'' W
09-79	”	”	”	”
09-332	”	Tarapoto - Urahuasha trail, Peru	06° 28' 40'' S	76° 21' 06'' W
09-333	”	”	”	”
09-312	*H. timareta* ssp. nov.	El Tunel trail, Km-18 Tarapoto-Yurimaguas, Peru	06° 27' 11'' S	76° 17' 19'' W
8624	”	La Antena, Km-15 Tarapoto-Yurimaguas, Peru	06° 27' 24'' S	76° 17' 54'' W
8628	”	”	”	”
8631	”	”	”	”
2403	*H. timareta florencia*	Quebrada Doraditas, Suaza, Caqueta, Colombia	1° 43' 04''N	75° 42' 35''W
2406	”	”	”	”
2407	”	”	”	”
2410	”	”	”	”
09-343	*H. elevatus*	Km-103.1 Tarapoto-Yurimaguas, Peru	05° 58' 18'' S	76° 13' 55'' W
09-63	*H. ethilla aerotome*	Urahuasha Km-8 Tarapoto-Yurimaguas, Peru	06° 27' 49'' S	76° 20' 07'' W
09-345	*H. hecale felix*	Km-103.1 Tarapoto-Yurimaguas, Peru	05° 58' 18'' S	76° 13' 55'' W
09-387	*H. pardalinus ssp. nov.*	Caño Tushmo, Lago Yarinacocha, Peru	08° 20' 33'' S	74° 35' 32'' W
09-326	*H. pardalinus sergestus*	Tarapoto - Urahuasha trail, Peru	06° 28' 40'' S	76° 21' 06'' W
09-364	*H. numata silvana*	El Tunel trail, Km-18 Tarapoto-Yurimaguas, Peru	06° 27' 11'' S	76° 17' 19'' W

Table S12.2.2 Oligonucleotides used to prepare libraries for SureSelect sequencing

Barcode		Sequence
ACTGC	oligo 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**ACTGC***T
	oligo 2	P-**GCAGT**AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
AGAGT	oligo 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGAGT***T
	oligo 2	P-**ACTCT**AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
GAGAT	oligo 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGAGT***T
	oligo 2	P-**ATCTC**AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
GTACA	oligo 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GTACA***T
	oligo 2	P-**TGTAC**AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
TCGAG	oligo 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**TCGAG***T
	oligo 2	P-**CTCGA**AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

* (phosphotioate bond)
P (phosphate)

S12.3 RAD genotyping

Separate paired-end RAD libraries were created for 84 samples (Supplementary Table S12.3.1) using standard RAD library preparation protocols [5,90]. The PstI restriction enzyme was used to achieve a high density of markers. Paired-end 100 base sequencing was performed on the libraries using Illumina Genome Analyzer IIx and HiSeq2000 sequencers. The *H. melpomene* genome sequence contains ~27K PstI cut sites; therefore, on average, RAD markers are expected every 10 kb. We multiplexed three to six individuals per sequencing lane by barcoding each individual RAD sample with five-base MIDs.

Table S12.3.1 Samples used for RAD sequencing

Sample ID	Taxon	Sample locality	Latitude	Longitude
09-102	*H. melpomene aglaope*	Parque Ecologico Munichis, Peru	05° 53' 56'' S	76° 13' 58'' W
09-107	”	Idea Religiosa, Munichis, Peru	05° 54' 37'' S	76° 13' 33'' W
09-119	”	”	”	”
09-128	”	Km-103.1 Tarapoto-Yurimaguas, Peru	05° 58' 18'' S	76° 13' 55'' W
09-247	”	”	”	”
09-1	*H. melpomene amaryllis*	Río Shilcayo, Tarapoto, Peru	06° 26' 58'' S	76° 20' 51'' W
09-7	”	”	”	”
09-8	”	”	”	”
09-76	”	Urahuasha trail, Km-8 Tarapoto-Yurimaguas, Peru	06° 27' 49'' S	76° 20' 07'' W
09-334	”	Tarapoto - Urahuasha trail, Peru	06° 28' 40'' S	76° 21' 06'' W
09-231	*H. timareta* ssp. nov.	El Afluente-Nuevo Eden, Peru	05° 39' 46'' S	77° 41' 46'' W
09-232	”	”	”	”
09-233	”	”	”	”
09-234	”	”	”	”
09-88	”	La Antena, Km-14.7 Tarapoto – Yurimaguas, Peru	06° 27' 18'' S	76° 17' 54'' W
09-118	*H. elevatus*	Idea Religiosa, Munichis, Peru	05° 54' 37'' S	76° 13' 33'' W
09-291	”	Km-103.1 Tarapoto-Yurimaguas, Peru	05° 58' 18'' S	76° 13' 55'' W
09-342	”	”	”	”
09-302	”	Biodiversidad, Km-17.2 Tarapoto-Yurimaguas, Peru	06° 27' 43'' S	76° 17' 31'' W
09-315	”	”	”	”
09-22	*H. ethilla aerotome*	Río Shilcayo, Tarapoto, Peru	06° 26' 58'' S	76° 20' 51'' W
09-58	”	La Antena, Km-14.7 Tarapoto – Yurimaguas, Peru	06° 27' 18'' S	76° 17' 54'' W
09-188	”	Tarapoto - Urahuasha trail, Peru	06° 28' 40'' S	76° 21' 06'' W
09-200	”	”	”	”
09-98	*H. hecale felix*	Km-28 Tarapoto-Yurimaguas, Peru	06° 24' 35'' S	76° 18' 11'' W
09-103	”	Idea Religiosa, Munichis, Peru	05° 54' 37'' S	76° 13' 33'' W
09-271	”	Km-103.1 Tarapoto-Yurimaguas, Peru	05° 58' 18'' S	76° 13' 55'' W
09-341	”	”	”	”
09-310	”	El Tunel trail, Km-18 Tarapoto-Yurimaguas, Peru	06° 27' 11'' S	76° 17' 19'' W
09-105	*H. pardalinus butleri*	Idea Religiosa, Munichis, Peru	05° 54' 37'' S	76° 13' 33'' W
09-106	”	”	”	”
09-269	”	”	”	”
09-346	”	Km-103.1 Tarapoto-Yurimaguas, Peru	05° 58' 18'' S	76° 13' 55'' W
09-396	*H. pardalinus ssp. nov.*	Caño Tushmo, Lago Yarinacocha, Peru	08° 20' 33'' S	74° 35' 32'' W
09-397	”	”	”	”
09-211	*H. pardalinus sergestus*	Tarapoto - Urahuasha trail, Peru	06° 28' 40'' S	76° 21' 06'' W
09-212	”	”	”	”
09-213	”	”	”	”
09-214	”	”	”	”
09-325	”	”	”	”
09-286	*H. numata aurora*	Km-103.1 Tarapoto-Yurimaguas, Peru	05° 58' 18'' S	76° 13' 55'' W
09-344	”	”	”	”
09-367	*H. numata bicoloratus*	El Tunel trail, Km-18 Tarapoto-Yurimaguas, Peru	06° 27' 11'' S	76° 17' 19'' W
09-11	”	Río Shilcayo, Tarapoto, Peru	06° 26' 58'' S	76° 20' 51'' W
09-3	*H. numata tarapotensis*	”	”	”
09-10	”	”	”	”
09-4	*H. numata timaeus*	”	”	”
09-143	*H. numata seraphion*	Idea Religiosa, Munichis, Peru	05° 54' 37'' S	76° 13' 33'' W
09-358	*H. numata illustris*	Km-103.1 Tarapoto-Yurimaguas, Peru	05° 58' 18'' S	76° 13' 55'' W
09-359	*H. numata elegans*	”	”	”
9179	*H. timareta timareta*	El Topo, Tungurahau, Ecuador	01° 24' 43" S	78° 11' 08" W
9180	”	“	“	“
CH7	*H. heurippa*	Buenavista, Meta, Colombia	04° 10' 30'' N	73° 40' 41'' W

CH8	”	”	”	”
CH9	”	”	”	”
CH11	”	”	”	”
CH12	”	”	”	”
CH14	”	”	”	”
M2158	*H. cydno cordula*	San Cristobal (UNET-Paramillo), Venezuela	07° 47' 56" N	72° 11' 56'' W
M2166	”	”	”	”
M2186	”	”	”	”
M2253	”	”	”	”
M2255	”	”	”	”
M2259	”	”	”	”
2440	*H. cydno chioneus*	Pipeline road, Colon, Panama	09° 07' 43" N	79° 42' 55" W
8265	”	Cerro Campana, Panama, Panama	08° 38' 25" N	79° 59' 35" W
204	*H. cydno weymeri*	El Saladito, near Cali, Valle del Cauca, Colombia	03° 22' 10" N	76° 39' 04" W
221	”	“	“	“
216	*H. cydno cydnides*	Yotoco, Valle del Cauca, Colombia	03° 53' 20" N	76° 25' 59" W
217	”	“	“	“
14671	*H. melpomene melpomene*	Puerto Lara, Darien, Panama	08° 38' 34" N	78° 07' 13" W
114671	”	“	“	“
CM1	”	Virgen de Chirajara, Meta, Colombia	4°12'48'' N	73°47'70'' W
CM2	”	”	”	”
CM3	”	”	”	”
CM6	”	Morcote, Casanare, Colombia	5°37'00.52''N	72°18'00.00''W
CM7	”	”	”	”
CM8	”	”	”	”
8228	”	Pointe Macouria, French Guyana	4°53'47"N	52°21'36"W
8229	”	”	”	”
2071	*H. melpomene rosina*	Gamboa, Panama, Panama	09° 07' 09" N	79° 41' 51" W
2097	”	“	“	“
9111	*H. melpomene ecuadorensis*	Old Zamora Road, Zamora-Chinchipe, Ecuador	“	“
9112	”	“	“	“

S12.4 Alignment and SNP calling

Image analysis and base calling was performed using the Illumina Pipeline v1.7. Reads were separated by sample and MIDs removed with RADtools. RAD sequences were aligned to the *H. melpomene* reference genome, while the SureSelect targeted sequences were aligned to modified *H. melpomene* reference genome in which the *HmYb* and *HmB*/*D* BAC walks and the three non-colour pattern BAC sequences, from which SureSelect baits were designed, replaced the relevant genome scaffolds [4]. Reads from each individual were aligned to the *H. melpomene* genome scaffolds and BAC sequences using Stampy v1.0.13 with default parameters except for a substitution rate of 0.01 and, for RAD data, an insert size of mean 500, SD 100. BAQ scores were calculated for the alignments [7]. Picard v1.48 tools (http://picard.sourceforge.net) were used to merge, sort and remove duplicates from the resulting alignments. Indels were realigned using the Genome Analysis Tool Kit (GATK) v1.1 RealignerTargetCreator and IndelRealigner tools [8]. Finally, genotypes were called across all individuals using the GATK UnifiedGenotyper [9] with a heterozygosity of 0.01 and incorporating BAQ scores from Stampy where possible. Only high quality (Q ≥ 30) genotypes were used in subsequent analyses.

S12.5 F_{ST} analyses

F_{ST} was calculated between races of *H. melpomene*, between races of *H. timareta*, and between *H. melpomene* and *H. timareta* using the SureSelect sequence dataset. F_{ST} was calculated in 10 kb sliding windows, moving in 100 base intervals, across the *B*/*D* and *N*/*Yb* colour pattern regions and the non-colour pattern regions, using the equation:

$$F_{ST} = \frac{H_T - H_S}{H_T},$$

where H_T is the expected heterozygosity in the total paired populations and H_S is the mean expected heterozygosity within each of the two races [91]. Expected heterozygosity within each 10 kb window was calculated as the mean heterozygosity across all biallelic SNPs within the window, where heterozygosity at each biallelic SNP was calculated as *2pq*, the allele frequencies across the individuals we sampled, based on the Hardy-Weinberg principle. This formula does not correct for bias when the true $F_{ST} = 0$. However, since we are uninterested here in testing the null hypothesis, bias correction was not necessary [4]. Windows with more than 90% missing data were excluded.

S12.6 ABBA-BABA tests of introgression

To test for differential gene flow between populations of *H. timareta* and of *H. melpomene* we used the four taxon "ABBA-BABA" testing procedure (Fig. 4a, main paper) [92,93].

1) SureSelect colour-pattern enriched data set. The SureSelect dataset was used to test for introgression at the *B*/*D* and *N*/*Yb* colour pattern regions. We focussed on positions along the genome at which biallelic SNPs have either ABBA or BABA site patterns in two comparisons: (*H. melpomene aglaope*, *H. melpomene amaryllis*, *H. timareta* ssp. nov., silvaniform) and (*H. melpomene aglaope*, *H. melpomene amaryllis*, *H. timareta florencia*, silvaniform), where 'silvaniform' consists of a pool of the ithomiine-mimicking species from the same region (*H. hecale*, *H. numata*, *H. ethilla*, and *H. pardalinus*, Supplementary Figure S18.1b). At these sites, the two *H. melpomene* races carry different alleles (**A** and **B**), and *H. timareta* carries the derived allele (**B**) compared to the silvaniform outgroup species (**A**) (Fig. 3a). As the *B*/*D* and *N*/*Yb* colour pattern regions are known to be divergent between our races of *H. melpomene*, we restricted this analysis to nucleotide sites fixed within each of *aglaope*, *amaryllis* and *timareta*, as well as among the silvaniform species. Using bases fixed among all four silvaniform outgroup species ensures a high likelihood that the derived allele "B" evolved in the ancestor of *H. timareta* and *H. melpomene*. In these analyses the ray-patterned silvaniform species, *H. elevatus*, was not included among the outgroup species due to the possibility of introgressed colour pattern genes between *H. elevatus* and the rayed *H. melpomene aglaope*. We examined the frequency and distribution of ABBA and

BABA sites in 10 kb sliding windows, moving in 1 kb increments along the *B*/*D* and *N*/*Yb* colour pattern regions, and also, as a control, in the non-colour pattern regions.

2) RAD whole genome sample dataset. The RAD dataset was used to test for evidence of genome-wide introgression between *H. melpomene* and *H. timareta* using the comparison: *H. melpomene aglaope*, *H. melpomene amaryllis*, *H. timareta* ssp. nov., silvaniform. Due to an almost complete absence of fixed nucleotide differences over most of the genome between the two *H. melpomene* races [4], instead of using fixed nucleotide sites we estimated Patterson's *D*-statistic based on ABBA and BABA SNP frequency differences using the expression:

$$D(P_1,P_2,P_3,P_4)=\frac{\sum_{i=1}^{n} p_{i3}\left(1-p_{i4}\right)\left[\left(1-p_{i1}\right)p_{i2}-p_{i1}\left(1-p_{i2}\right)\right]}{\sum_{i=1}^{n} p_{i3}\left(1-p_{i4}\right)\left[\left(1-p_{i1}\right)p_{i2}+p_{i1}\left(1-p_{i2}\right)\right]}$$ (Equation 2 from ref. [93])

where P_1, P_2, P_3 and P_4 are the four taxa in the comparison, p_{ij} is the observed frequency of the derived "B" SNP *i* in taxon *j*, and *n* is the total number of SNPs.

This analysis was restricted to sites fixed for an ancestral allele in the silvaniform outgroup taxa because "SNP frequency" is meaningless when applied across multiple species. In addition to calculating *D* for the entire genome, to examine variation in *D* across the genome, separate *D*-statistics were evaluated for each of the 21 chromosomes.

S12.7 Estimating the proportion of genomic introgression

We estimated the genomic "admixture fraction" from *H. melpomene* into *H. timareta ssp. nov*., as well as in the opposite direction, from *H. timareta ssp. nov.* into *H. melpomene amaryllis* using the RAD dataset (Equation S18.5 from ref. [92]).

Genomic admixture from H. timareta *ssp. nov. into* H. melpomene amaryllis*:* This calculation compares the numerator of Patterson's *D*-statistic from the grouping (*H. melpomene aglaope*, *H. melpomene amaryllis*, *H. timareta* ssp. nov., silvaniform) with the maximum expected if there was complete introgression, i.e. the numerator of the *D*-statistic from the grouping (*H. melpomene aglaope*, *H. timareta* ssp. nov.$_1$, *H. timareta* ssp. nov.$_2$, silvaniform) where *H. timareta* ssp. nov.$_1$ and *H. timareta* ssp. nov.$_2$ are two subsets of the five *H. timareta* ssp. nov. samples. Using this method, we estimate 2.2 ± 0.3% of the genome introgressed from *H. timareta* ssp. nov. into *H. melpomene amaryllis* to the exclusion of *H. melpomene aglaope*. Standard errors were obtained by using a block jack-knife (S12.8).

Genomic admixture from H. melpomene *into* H. timareta *ssp. nov.:* To estimate the amount of introgression in the opposite direction we took advantage of the fact that we have RAD sequence from an additional race of *H. timareta*, *H. timareta timareta* from Ecuador (Table S12.3.1). This taxon has a peculiar partially non-mimetic, rayed colour pattern, but is also closely related to Peruvian *H. timareta* ssp. nov. (Fig. S18.1a). The numerator of Patterson's *D*-statistic from the grouping (*H. timareta timareta*, *H. timareta* ssp. nov., *H. melpomene amaryllis*, silvaniform) was compared with the maximum expected if there was complete introgression, i.e. the numerator of the *D*-statistic from the grouping (*H. timareta timareta*, *H. melpomene amaryllis* $_1$, *H. melpomene amaryllis* $_2$, silvaniform) where *H. melpomene amaryllis* $_1$ and *H. melpomene amaryllis* $_2$ are two subsets of the five *H. melpomene amaryllis* samples. Using this method, we estimate 4.6 ± 0.7% of the genome introgressed from *H. melpomene amaryllis* into *H. timareta* ssp. nov., to the exclusion of *H. timareta timareta*.

S12.8 Estimation of linkage disequilibrium and block jack-knife standard errors

Standard errors on *D*-statistics and other measures were estimated using a block jack-knife to overcome the problem of autocorrelation among sites due to linkage disequilibrium (LD) [92]. Linkage disequilibrium estimation methods and results are given in section S16 (below). For the block jack-knife procedure the block size was selected to be greater than the extent of LD. Beyond 10-100 kb, LD declines to the empirically found asymptote for unlinked comparisons (r^2 = 0.23) when using only four individuals (Supplementary Figure S16.2.1). Therefore, in the calculation of standard errors on *D*-statistics for the whole genome we used a block size of 500 kb. A block size of 100 kb was used across individual chromosomes and the colour pattern regions.

S12.9 Phylogenetic analyses

Phylogenetic relationships between the taxa included in this study (Supplementary Tables S12.2.1 and S12.3.1) were established by neighbour-joining phylogenetic analysis of two independent sequence datasets: genome-wide RADs and SureSelect non-colour pattern regions. In both cases all available sequence was concatenated into a single alignment prior to analysis. To explore genealogical changes along the colour pattern regions, maximum likelihood phylogenetic hypotheses were generated from 50 kb non-overlapping windows along the *B*/*D* and *N*/*Yb* regions. Neighbour-joining (using K2P distances) and maximum likelihood (GTR+Γ substitution model) phylogenies were constructed with PHYML [60], and node support was established from 1000 and 100 bootstrap replicates respectively.

S13. Alignment statistics

Table S13.1 Alignment statistics for RAD sequencing

SNPs are heterozygous and homozygous variants relative to the *H. melpomene* reference. Values for bases mapped and SNPs are for calls with genotype qualities ≥ 30.

[a] the proportion of mapped bases containing SNPs relative to the *H. melpomene* reference

Sample ID	Taxon	Bases mapped	SNPs ([a])
09-102	*H. melpomene aglaope*	14961419	527799 (0.0353)
09-107	”	15691443	554931 (0.0354)
09-119	”	14162692	494505 (0.0349)
09-128	”	14582413	512074 (0.0351)
09-247	”	14542734	530070 (0.0364)
09-1	*H. melpomene amaryllis*	13769506	523756 (0.0380)
09-7	”	8031714	296734 (0.0369)
09-8	”	13163214	477523 (0.0363)
09-76	”	12880800	464502 (0.0361)
09-334	”	12673480	466779 (0.0368)
09-231	*H. timareta* ssp. nov.	10547427	377308 (0.0358)
09-232	”	9849582	351058 (0.0356)
09-233	”	13599719	478641 (0.0352)
09-234	”	14060773	500024 (0.0356)
09-88	”	13069819	450693 (0.0345)
09-118	*H. elevatus*	9647197	462720 (0.0480)
09-291	”	12698256	613503 (0.0483)
09-342	”	11618224	576452 (0.0496)
09-302	”	10416536	501494 (0.0481)
09-315	”	12671787	632252 (0.0499)
09-22	*H. ethilla aerotome*	11484515	504689 (0.0439)
09-58	”	11912469	530661 (0.0445)
09-188	”	11930599	529136 (0.0444)
09-200	”	7668234	325341 (0.0424)
09-98	*H. hecale felix*	12386058	549230 (0.0443)
09-103	”	13053149	590334 (0.0452)
09-271	”	12815692	576551 (0.0450)
09-341	”	12741384	575758 (0.0452)
09-310	”	13544206	615993 (0.0455)
09-105	*H. pardalinus butleri*	11106825	535321 (0.0482)
09-106	”	12167521	588111 (0.0483)
09-269	”	12269751	595078 (0.0485)
09-346	”	9649192	488471 (0.0506)
09-396	*H. pardalinus ssp. nov.*	12987203	620361 (0.0478)
09-397	”	12219756	584246 (0.0478)
09-211	*H. pardalinus sergestus*	11722387	492665 (0.0420)
09-212	”	10470216	435543 (0.0416)
09-213	”	11357892	476960 (0.0420)
09-214	”	11890938	502914 (0.0423)

09-325	”	13064503	564952 (0.0432)
09-286	*H. numata aurora*	15829213	808434 (0.0511)
09-344	”	13481678	659498 (0.0489)
09-367	*H. numata bicoloratus*	13441842	712295 (0.0530)
09-11	”	8699835	429965 (0.0494)
09-3	*H. numata tarapotensis*	11253297	557572 (0.0495)
09-10	”	9617065	465640 (0.0484)
09-4	*H. numata timaeus*	9272777	450190 (0.0485)
09-143	*H. numata seraphion*	14456038	685939 (0.0474)
09-358	*H. numata illustris*	13488021	653073 (0.0484)
09-359	*H. numata elegans*	12268832	588950 (0.0480)
9179	*H. timareta timareta*	10638340	376992 (0.0354)
9180	”	9106908	324462 (0.0356)
CH7	*H. heurippa*	10617178	384694 (0.0362)
CH8	”	10051585	404872 (0.0370)
CH9	”	10051585	361394 (0.0360)
CH11	”	11322154	420104 (0.0371)
CH12	”	10638693	392880 (0.0370)
CH14	”	10595115	387247 (0.0366)
M2158	*H. cydno cordula*	9436419	376997 (0.0400)
M2166	”	10844106	437195 (0.0403)
M2186	”	10417857	417890 (0.0401)
M2253	”	11032390	442662 (0.0401)
M2255	”	12059846	486816 (0.0404)
M2259	”	10951202	443180 (0.0405)
2440	*H. cydno chioneus*	11675159	472565 (0.0405)
8265	”	9985663	402320 (0.0403)
204	*H. cydno weymeri*	14404694	472565 (0.0405)
221	”	15128962	402320 (0.0403)
216	*H. cydno cydnides*	14451204	573164 (0.0340)
217	”	15846795	618555 (0.0390)
14671	*H. melpomene melpomene*	12057712	368596 (0.0306)
114671	”	15684501	477164 (0.0304)
CM1	”	9295405	296453 (0.0319)
CM2	”	8189498	261553 (0.0319)
CM3	”	12221302	401149 (0.0328)
CM6	”	11000623	348615 (0.0317)
CM7	”	11082409	345506 (0.0312)
CM8	”	11068384	345616 (0.0312)
8228	”	11619029	436985 (0.0376)
8229	”	14791682	552509 (0.0374)
2071	*H. melpomene rosina*	14716312	436372 (0.0297)
2097	”	16167546	469090 (0.0290)
9111	*H. melpomene ecuadorensis*	10797171	421420 (0.0390)
9112	”	10431028	405598 (0.0389)

Table S13.2 Alignment statistics for SureSelect resequencing

SNPs are heterozygous and homozygous variants relative to the *H. melpomene* reference. Values for bases mapped and SNPs are for calls with genotype qualities ≥ 30. The lengths of the *N*/*Yb* walk, *B*/*D* walk and the non-colour pattern regions targeted for sequencing are 1,149,452 bp, 716,635bp and 1,827,245 bp respectively.

[a] the proportion of mapped bases containing SNPs relative to the *H. melpomene* reference.

Sample ID	Taxon	*B*/*D* walk colour pattern region		*N*/*Yb* walk colour pattern region		Non-colour pattern regions	
		Bases mapped	SNPs ([a])	Bases mapped	SNPs ([a])	Bases mapped	SNPs ([a])
09-246	*H. melpomene aglaope*	484800	12895 (0.0266)	637763	21372 (0.0335)	1194076	47719 (0.0400)
09-247	”	549746	15233 (0.0277)	765795	27716 (0.0362)	1394710	66077 (0.0474)
09-357	”	523063	13968 (0.0267)	705490	24655 (0.0349)	1281157	55267 (0.0431)
09-268	”	547867	14993 (0.0274)	763852	27608 (0.0361)	1384688	64905 (0.0469)
09-75	*H. melpomene amaryllis*	529991	13146 (0.0248)	701321	21665 (0.0309)	1293469	55206 (0.0427)
09-79	”	569481	14794 (0.0260)	794210	27285 (0.0344)	1413143	68524 (0.0485)
09-332	”	556882	14573 (0.0262)	757949	24071 (0.0318)	1376730	63435 (0.0461)
09-333	”	559174	14172 (0.0253)	765569	24583 (0.0321)	1371398	63922 (0.0466)
09-312	*H. timareta* ssp. nov.	530124	14509 (0.0274)	731991	21391 (0.0292)	1292988	54853 (0.0424)
8624	”	543397	14659 (0.0270)	769753	24854 (0.0323)	1335295	56691 (0.0425)
8628	”	561907	14261 (0.0254)	813471	26381 (0.0324)	1406613	64776 (0.0461)
8631	”	556326	15423 (0.0277)	797256	25425 (0.0319)	1373435	60383 (0.0440)
2403	*H. timareta florencia*	547291	14845 (0.0271)	782181	24458 (0.0313)	1352617	57952 (0.0428)
2406	”	549011	15471 (0.0282)	803850	26213 (0.0326)	1369142	61167 (0.0447)
2407	”	563116	15935 (0.0283)	833128	27707 (0.0333)	1411603	65858 (0.0467)
2410	”	546470	15121 (0.0277)	789920	25512 (0.0323)	1367889	61294 (0.0448)
09-343	*H. elevatus*	490010	17707 (0.0361)	670174	29695 (0.0443)	1181247	63198 (0.0535)
09-63	*H. ethilla aerotome*	462412	16837 (0.0364)	608187	24163 (0.0397)	1163979	55701 (0.0479)
09-345	*H. hecale felix*	423223	15689 (0.0371)	561914	21811 (0.0388)	1060182	48543 (0.0458)
09-387	*H. pardalinus ssp. nov.*	373890	14420 (0.0386)	445590	20658 (0.0464)	916419	43885 (0.0479)
09-326	*H. pardalinus sergestus*	452318	15623 (0.0345)	581792	22688 (0.0390)	1136395	50808 (0.0447)
09-364	*H. numata silvana*	427450	18434 (0.0431)	542484	26934 (0.0496)	1036999	54178 (0.0522)

S14. Genomic divergence (F_{ST}) among populations in *B/D* and *N/Yb* colour pattern regions, and in non-colour pattern regions

S14.1 The effect of repetitive DNA and repeat masking

In order to test the effect of repetitive DNA on our results, repetitive regions were identified in the *B*/*D* and *N*/*Yb* regions using Repeatmasker (http://www.repeatmasker.org) with a repeat database generated from repeats found in *Heliconius* genomic resources as well as other known lepidopteran repetitive elements (Supplementary Information S5). These repeat regions were removed from the alignments and F_{ST} calculated from 10 kb sliding windows as before. The same process was then carried out using only the regions identified as repetitive, comprising approximately 20-30% of the data (Supplementary Figure S14.1.1)

F_{ST} results based on unmasked and repeat masked datasets are virtually indistinguishable apart from a region located at ~400 kb in *B/D* (Supplementary Figure S14.1.1). Indeed, F_{ST} analyses based only on the regions removed by repeat masking produce almost identical results (Supplementary Figure S14.1.1). This indicates that our mapping and filtering process has largely removed reads that may have incorrectly mapped to these regions from elsewhere in the genome and that the data remaining in these regions are giving a signal consistent with flanking non-repetitive regions. Therefore our strategy of using a dataset that is not repeat-masked for our primary analysis appears justified. This assertion is further supported by the fact that only a few narrow sections comprising ~2% of the sequenced regions have the elevated levels of sequence coverage caused by sequences being incorrectly aligned to a repeat. The one exception is the repetitive region at approximately 400 kb in the *HmB/D* region. This produces a large trough of F_{ST} at the centre of the highest peak of F_{ST} between colour pattern races, most likely due to non-unique mapping of similar reads in both species to this region.

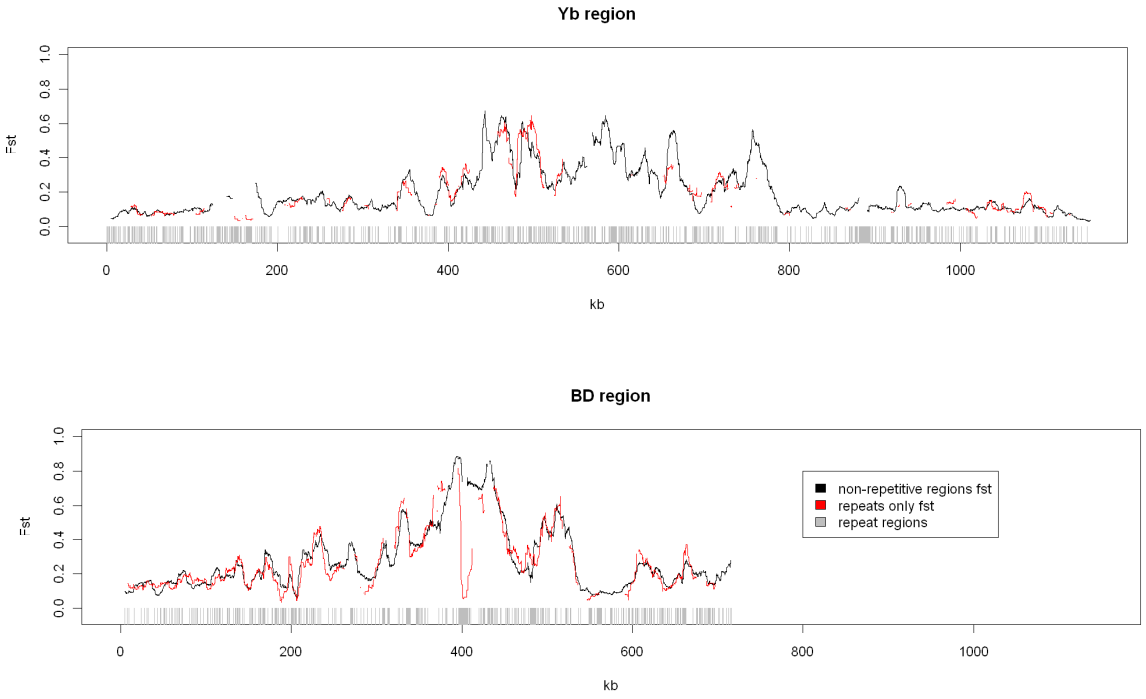


Figure S14.1.1 Genetic differentiation across colour pattern regions, showing effect of removing repeats

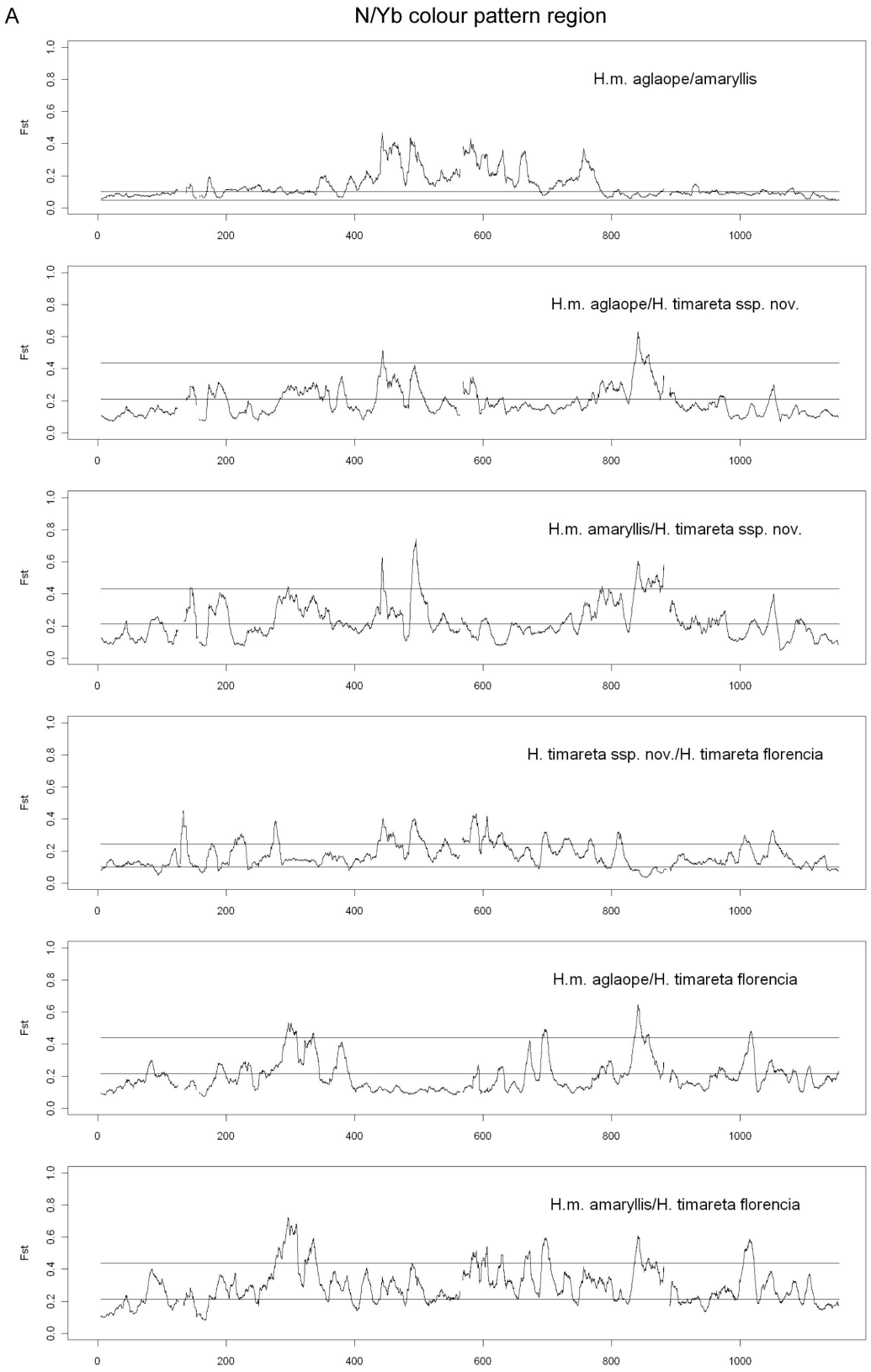
Genetic differentiation (F_{ST}) between *H. melpomene aglaope* and *H. melpomene amaryllis* across the colour pattern regions with repeats removed (black) and for only repeats (red). The positions of the repeats are show in grey at the bottom of the plots.

S14.2 Results

As a measure of divergence we calculated F_{ST} (Supplementary Information S12.5) between all pairs of populations (Supplementary Figure S14.2.1). As found previously [4], peaks of divergence were present in the centre of the colour pattern regions between races/species that differ in colour pattern. The SureSelect-captured non-colour pattern regions were used to assess the level of background divergence elsewhere in the genome (Supplementary Figure S14.2.2) and 95% confidence intervals from bootstrapping this data serve as thresholds for the background levels. The within-species *H. melpomene aglaope/amaryllis* comparison shows more distinct peaks of divergence compared with background levels than do the other comparisons. Contrary to the results reported previously [4], peaks of divergence do not appear to be broader in the *H. m. aglaope* to *H. timareta* ssp. nov. comparison. This is due to the wider genomic sampling of unlinked regions, which has raised the 95% CI for background divergence. In general the between-species comparisons show both peaks above and troughs below the background divergence levels. This indicates that the colour pattern regions have experienced both higher divergence and also more introgression, with these effects localised to fairly small regions. Overall, regions showing highest divergence

between populations with different colour patterns are also the regions with evidence for introgression among species with the same colour patterns, supporting our finding that the colour pattern convergence among *melpomene-cydno* group species has arisen by introgression of these regions. This evidence for gene flow is most marked in the *H. m. aglaope*/ *H. timareta florencia* comparison with fairly wide and low troughs of $F_{ST.}$ This may suggest that introgression between these populations has been more recent than that between *H. m. amaryllis* and *H. timareta* ssp. nov.

Figure S14.2.1 Genetic differentiation between populations at colour-pattern regions



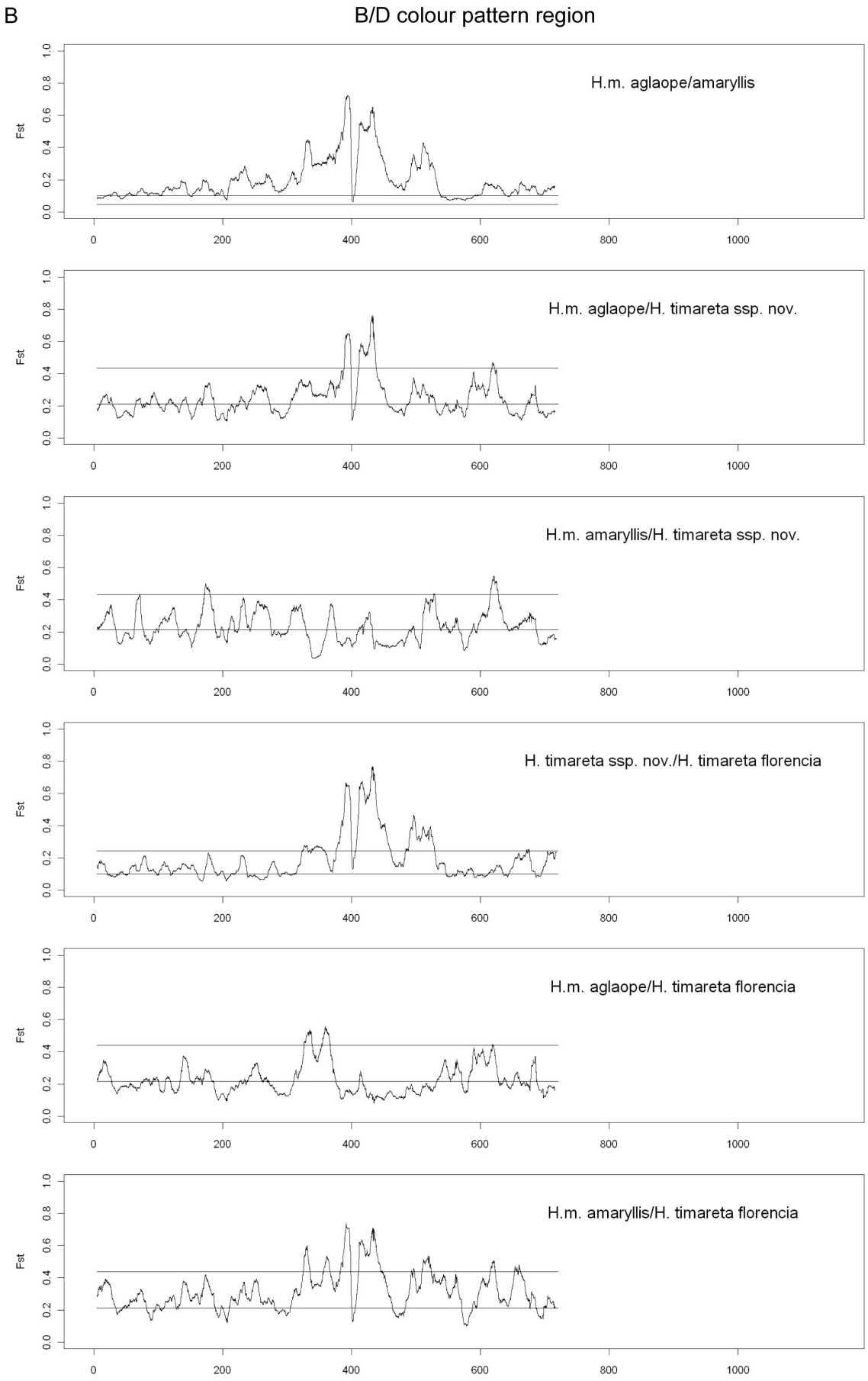


Figure S14.2.1 contd.

Figure S14.2.1, legend. Genetic differentiation between populations at colour-pattern regions

Genetic differentiation (F_{ST}) between all pairs of populations across the colour pattern regions: A. *N*/*Yb* and B. *B/D*. F_{ST} is uncorrected for sample size (n = 4 for each population) and calculated as a 10 kb sliding window moving at 100 bp increments. Thresholds indicate the upper and lower 95% CIs from 10,000 bootstrap resampling replicates of 1,000 bps (the minimum number of sites with data in each 10 kb window) from the non-colour-pattern regions. Peaks above the upper threshold indicate regions of high divergence compared to background levels; troughs below the lower threshold indicate regions of high gene-flow compared to background levels. The x-axis shows distance in kilobases.

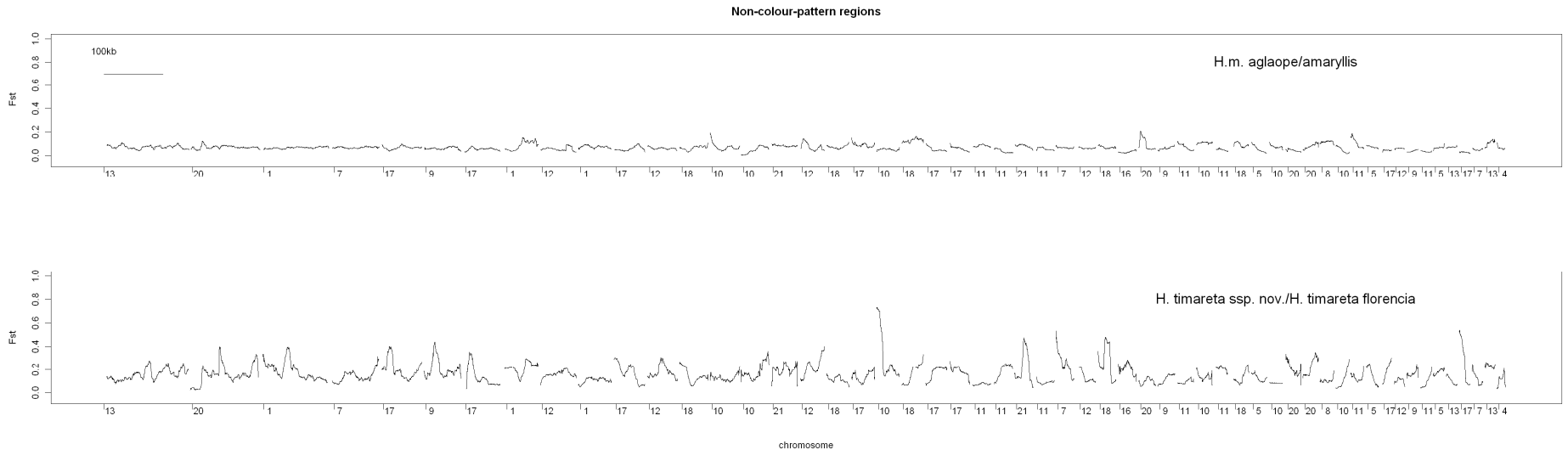


Figure S14.2.2 Genetic differentiation between populations at non-colour pattern regions

Genetic differentiation (F_{ST}) between *H. melpomene aglaope* and *H. melpomene amaryllis* (top) and between *H. timareta* ssp. nov. and *H. timareta florencia* (bottom) across the non-colour-pattern regions. The chromosome that each region is found on is indicated on the x-axis.

S15. *D*-statistics of chromosomes

Table S15.1 *D*-statistics among chromosomes

Patterson's *D*-statistics significantly different from zero are shown in bold (two-tailed Z-test for $D = 0$). Chromosomes 15 and 18 contain the *N*/*Yb* and *B*/*D* colour pattern regions respectively.

* *D*-statistics calculated for chromosomes 15 and 18 with the colour pattern regions removed.

Chromosome	*D*-statistic ± se	Z	*P* (two-tailed)
1	**0.043 ± 0.017**	**2.04**	**0.04**
2	0.038 ± 0.031	1.06	0.29
3	0.028 ± 0.020	0.44	0.66
4	-0.004 ± 0.018	-0.11	0.91
5	0.027 ± 0.023	1.14	0.26
6	0.007 ± 0.016	0.42	0.67
7	**0.051 ± 0.016**	**4.81**	**2×10^{-6}**
8	0.013 ± 0.013	0.64	0.52
9	**0.057 ± 0.016**	**2.30**	**0.02**
10	-0.007 ± 0.019	0.17	0.86
11	**0.039 ± 0.017**	**2.18**	**0.03**
12	**0.055 ± 0.019**	**2.11**	**0.03**
13	**0.029 ± 0.013**	**2.10**	**0.04**
14	0.002 ± 0.010	1.30	0.19
15	**0.082 ± 0.020**	**5.15**	**3×10^{-7}**
15 *	**0.054 ± 0.008**	**7.34**	**3×10^{-13}**
16	**0.044 ± 0.014**	**4.28**	**2×10^{-5}**
17	**0.060 ± 0.015**	**4.90**	**1×10^{-6}**
18	**0.081 ± 0.021**	**4.73**	**3×10^{-6}**
18 *	**0.072 ± 0.021**	**4.25**	**3×10^{-5}**
19	0.042 ± 0.025	0.94	0.35
20	**0.041 ± 0.018**	**2.52**	**0.01**
Z	0.029 ± 0.031	0.96	0.34
Whole genome	**0.037 ± 0.003**	**13.5**	**1×10^{-40}**

S16. Linkage disequilibrium in *Heliconius*

S16.1 Methods

Standard errors on *D*-statistics and other measures were calculated using a block jack-knife procedure. For the block jack-knife procedure the block size needs to be greater than the extent of linkage disequilibrium (LD). To establish what block size we should use we investigated the rate at which LD declines with distance across the *B*/*D* and *N*/*Yb* regions for *H. melpomene amaryllis*, *H. melpomene aglaope*, *H. timareta* ssp. nov. and *H. timareta florencia* using all pairwise comparisons of biallelic SNPs across these regions. Biallelic SNPs between the three non-colour pattern BAC regions were used to determine the expected value of r^2 corresponding to linkage equilibrium between SNPs. We calculated the pairwise LD across the *B*/*D* and *N*/*Yb* regions within each of *H. melpomene amaryllis*, *H. melpomene aglaope*, *H. timareta* ssp. n. and *H. timareta florencia* using all pairwise comparisons of biallelic SNPs across these regions. The program EMLD [94], which utilises an expectation maximization algorithm [95], was used to estimate the correlation coefficient of pairwise LD (r^2).

S16.2 Results

As we have genotypes from only four individuals within each taxon the LD estimate for each pair of SNPs is very noisy. We therefore averaged LD estimates from pairs of SNPs binned in 100 bp windows across the *Yb* and *B/D* regions to produce very accurate average LD estimates between 100 bp windows at different distances apart (100 bp, 200 bp, etc.) (Supplementary Figure S16.2.1). LD is found to decline smoothly, approaching a horizontal asymptote representing absence of any LD, which coincides with our estimate of LD calculated between SNPs in unlinked comparisons. This horizontal asymptote will tend towards zero as the number of individuals used in the LD estimation increases. The variance about this curve increases at longer genetic distances as the averages in each window at larger distances are based on smaller numbers of SNP comparisons.

In all cases LD declines to background levels beyond 100 kb, and usually this decline is reached within 50 kb (Supplementary Figure S16.2.1). Therefore, for the calculation of standard errors on *D*-statistics across chromosomes and the colour pattern regions we used a block size of 100 kb, and for analyses across the whole genome the block size was set at 500 kb.

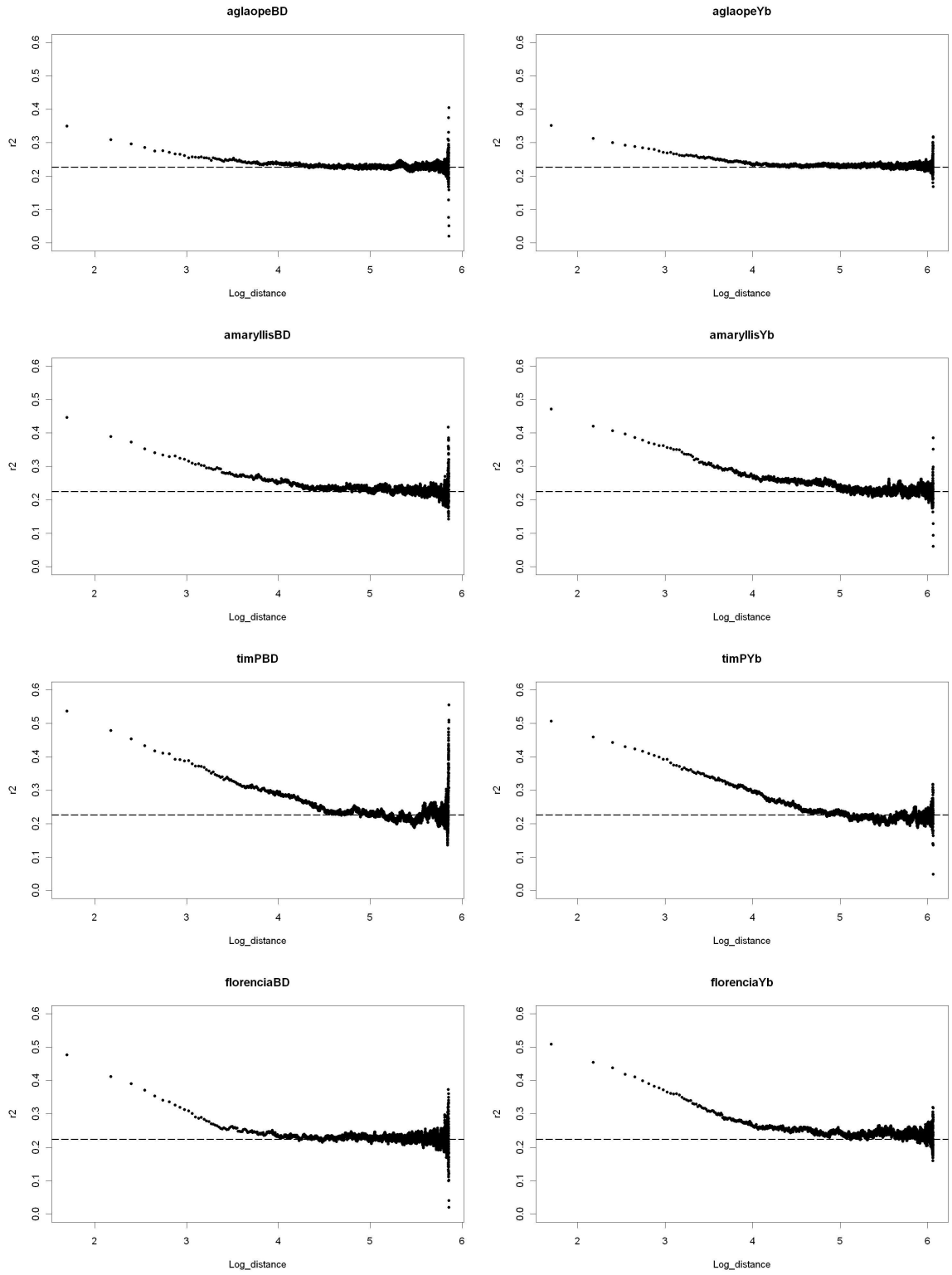


Figure S16.2.1 Decline of linkage disequilibrium with physical distance

Linkage disequilibrium variation with $\log_{10}$ physical distance along *B/D* and *N/Yb* regions for *H. melpomene amaryllis*, *H. melpomene aglaope*, *H. timareta* ssp. nov. and *H. timareta florencia*. Linkage disequilibrium asymptotes to a value of $r^2 = 0.23$ (which is also the empirically derived value found between unlinked comparisons) at a physical distance of 10^4-10^5 bp.

S17. Evidence for adaptive introgression at the *N/Yb* mimicry locus

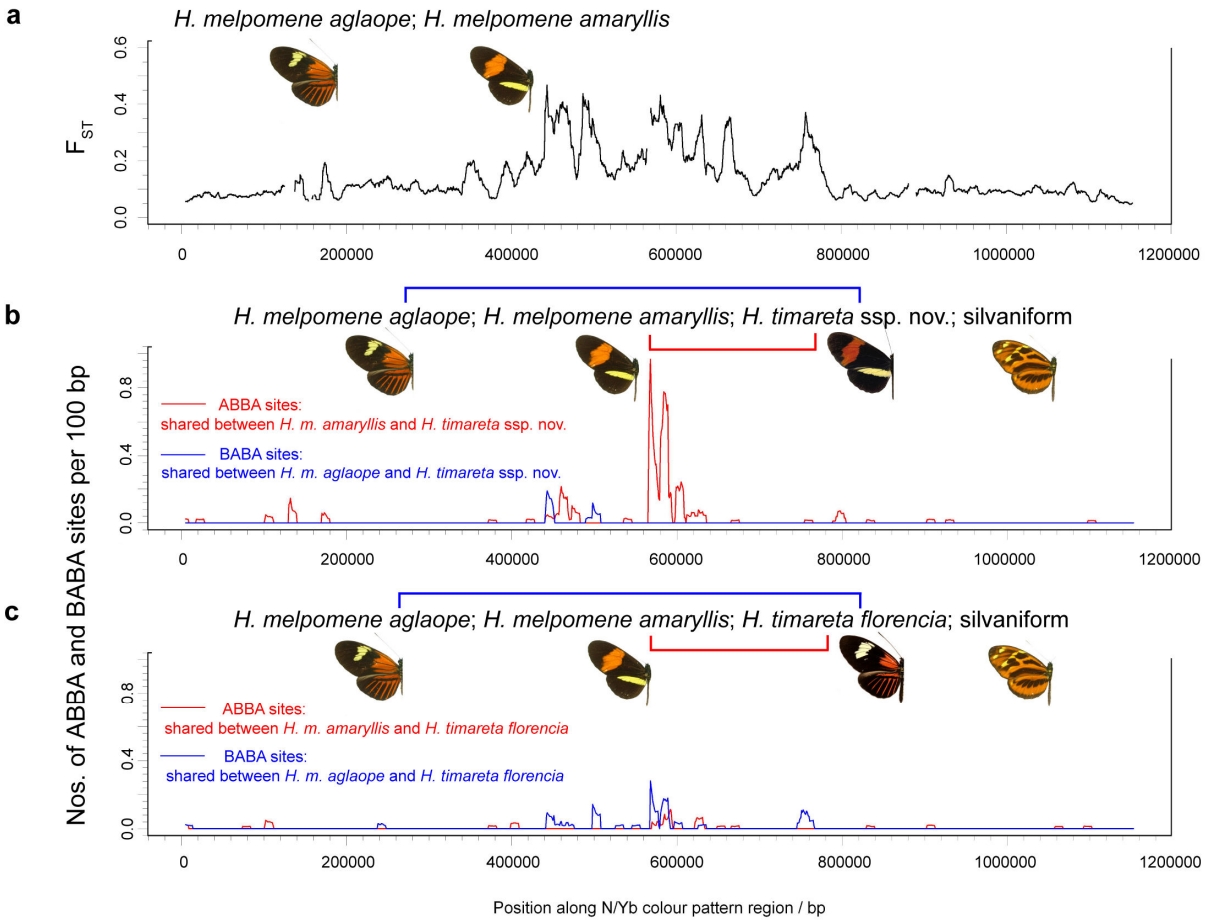


Figure S17.1 F_{ST} and ABBA-BABA site patterns along the *N/Yb* colour pattern region

a, Genetic divergence, F_{ST}, between *H. melpomene aglaope* (rayed) and *H. melpomene amaryllis* (postman) along the *N/Yb* colour pattern region containing loci controlling differences in yellow wing pattern elements. F_{ST} is calculated in 10 kb sliding windows at 100 bp increments. Distribution of fixed ABBA and BABA single nucleotide sites along *N/Yb* in 10 kb sliding windows at 1 kb increments for the two comparisons shown in **b** (*H. melpomene aglaope*, *H. melpomene amaryllis*, *H. timareta* ssp. nov., pooled silvaniforms) and **c** (*H. melpomene aglaope*, *H. melpomene amaryllis*, *H. timareta florencia*, pooled silvaniforms). There are significantly more ABBA compared to BABA sites in this region for the *H. melpomene aglaope*; *H. melpomene amaryllis*; *H. timareta* ssp. nov. silvaniform comparison (*D*-statistic = 0.85; two-tailed Z-test for $D = 0$, $Z = 9.8$, $P = 6 \times 10^{-23}$). For the second comparison (*H. melpomene aglaope*; *H. melpomene amaryllis*; *H. timareta florencia*; silvaniform) the *D*-statistic is not significantly different from zero (*D*-statistic = -0.27; two-tailed Z-test for $D = 0$, $Z = -1.5$, $P = 0.13$). However, the two comparisons have significantly different ABBA:BABA ratios ($X^2 = 42.7$, df = 1, $P = 7 \times 10^{-11}$).

S18. Phylogenetic analysis of resequenced individuals

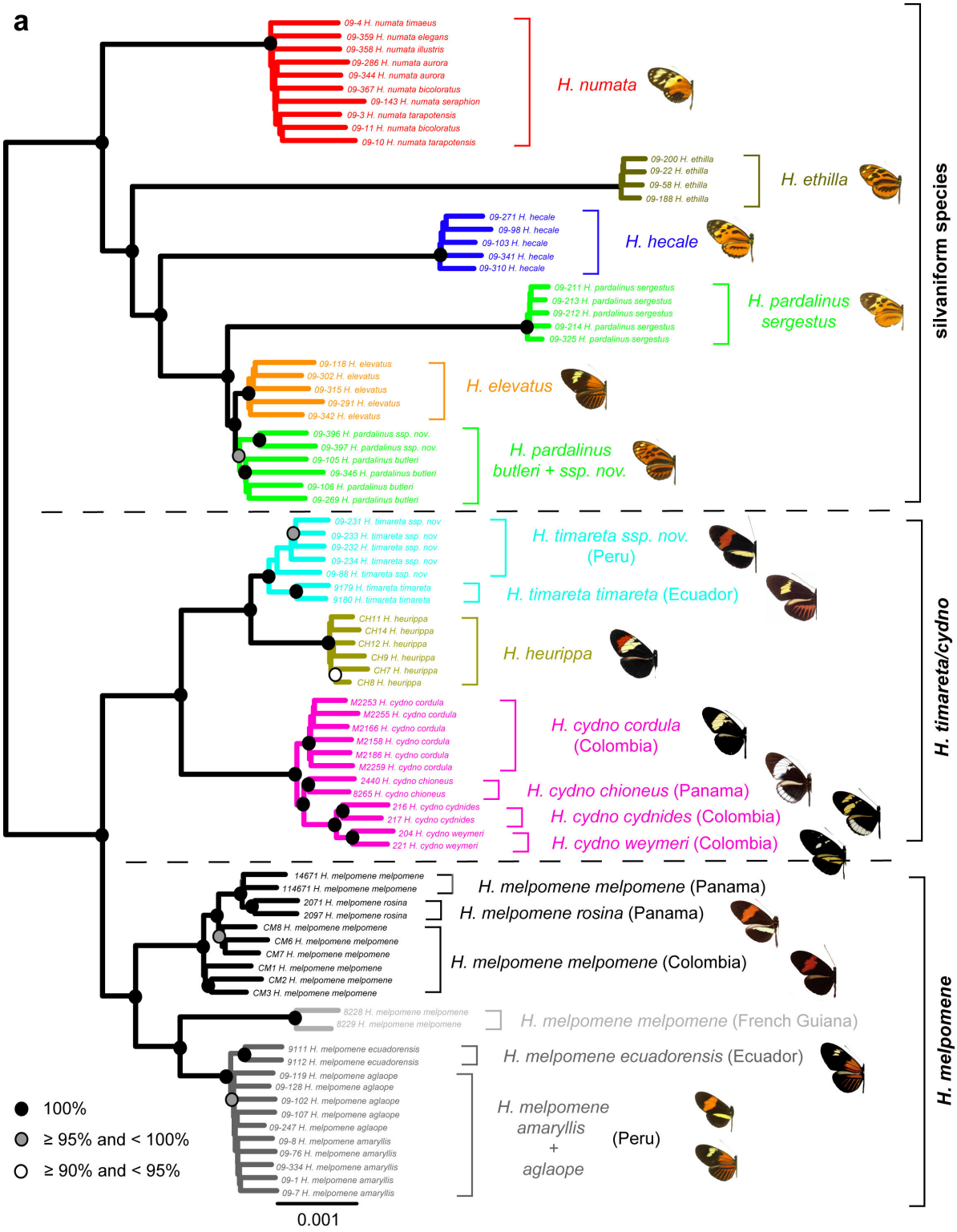


Figure S18.1a Phylogenetic analysis of resequenced individuals based on RAD sequence

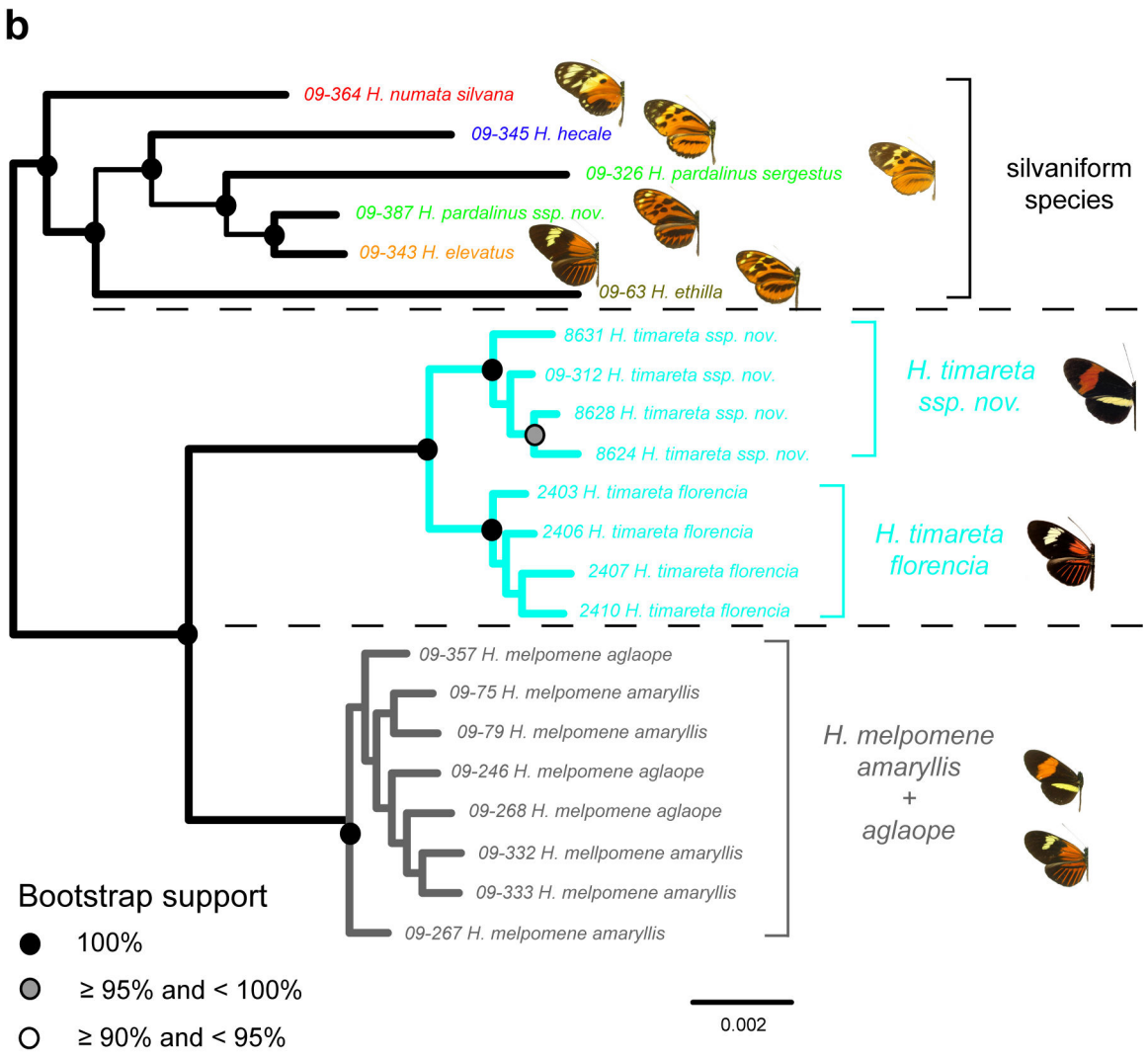


Figure S18.1b Phylogenetic analysis of resequenced individuals based on SureSelect sequence

Neighbour-joining phylogenetic inferences based on K2P distances for sequence data from non-colour pattern regions: Figure S18.1a) 530,898 bp of RAD sequence, Figure S18.1b) 657,745 bp of SureSelect sequence. Analyses were carried out on complete datasets, i.e. there is no missing data. Node support was derived from 1000 bootstrap replicates.

S19. Phylogenetic analysis across the *B*/*D* region

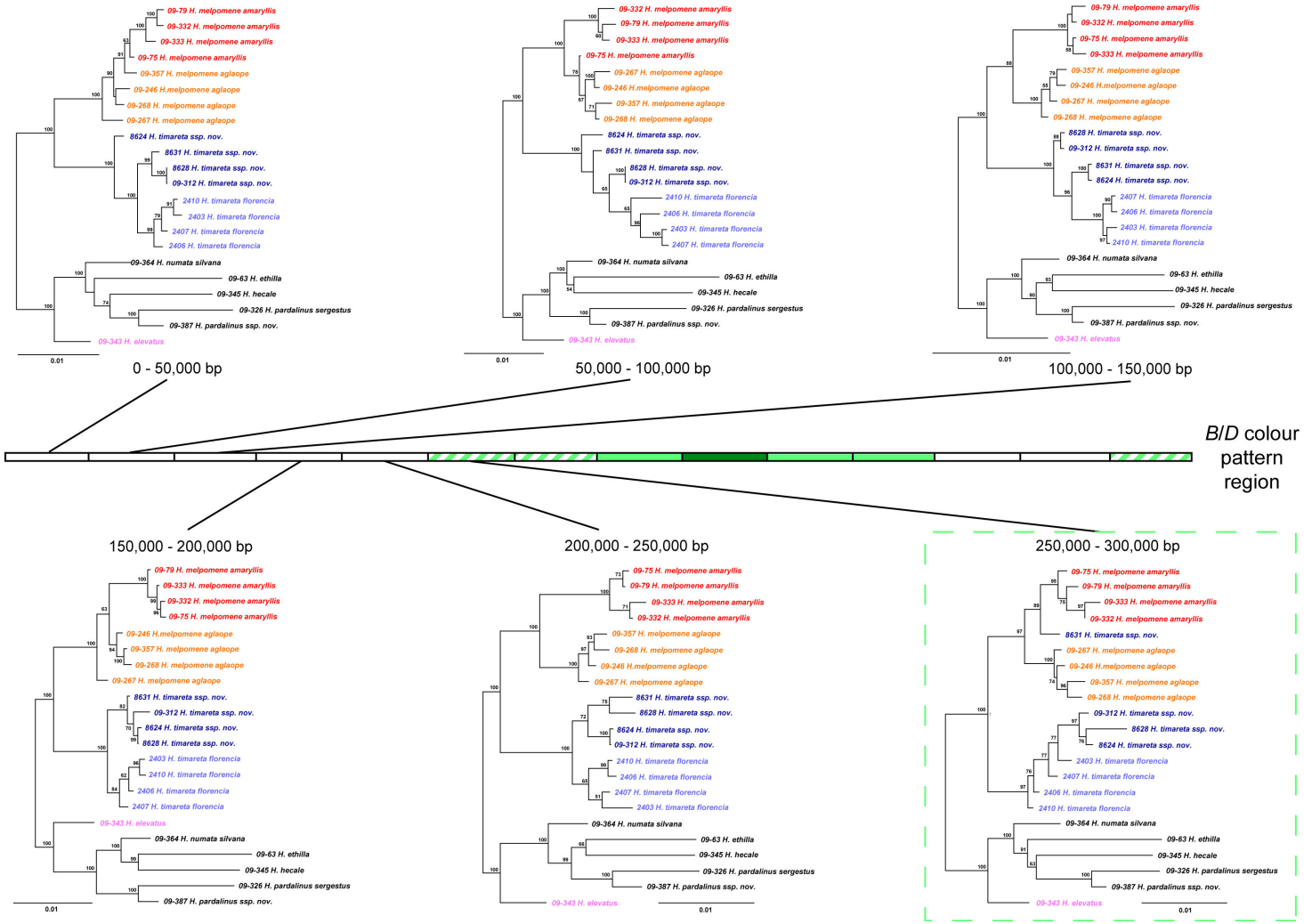


Figure S19.1 Phylogenetic analysis across the *B*/*D* region

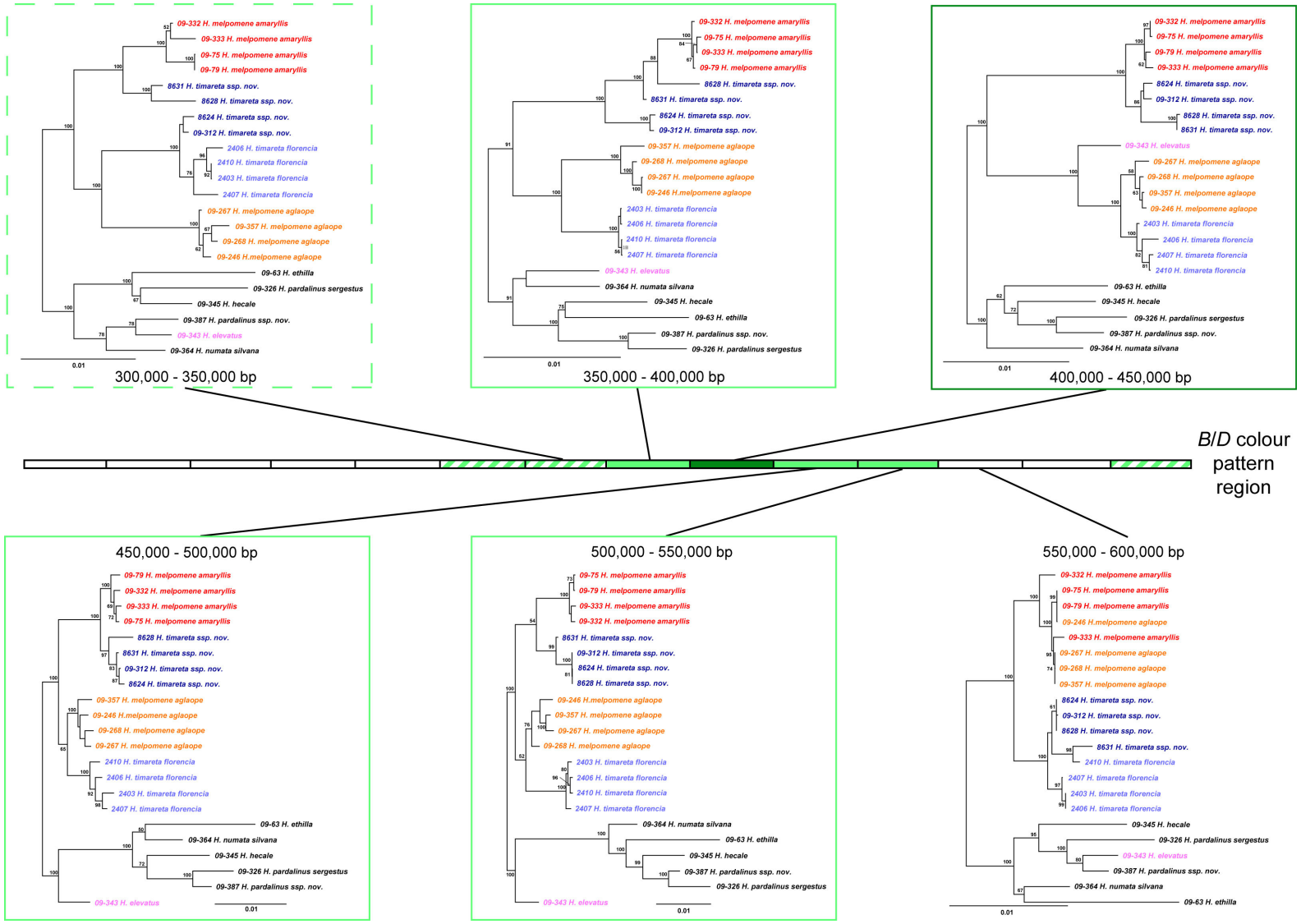


Figure S19.1, contd.

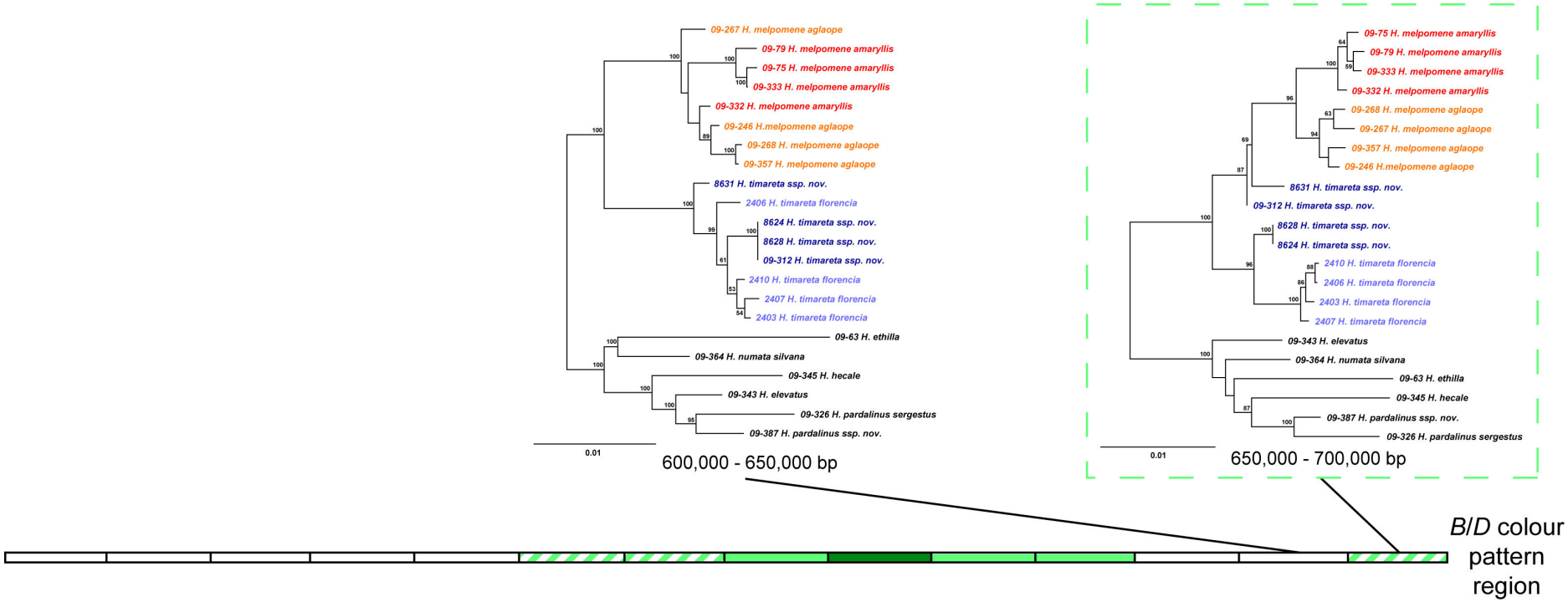


Figure S19.1 Phylogenetic analysis across the *B/D* region

Maximum likelihood phylogenies based on non-overlapping 50 kb windows across the *B/D* colour pattern region. Bootstrap supports for nodes are based on 100 bootstrap replicates and are shown when support is greater than 50. In the phylogenies from the white windows, the taxa are grouped by species. In the phylogenies from pale and dark green windows the taxa are grouped by colour pattern. In phylogenies from the striped boxes, either *H. melpomene* or *H. timareta* are paraphyletic but the taxa do not group cleanly by colour pattern.

S20. Phylogenetic analysis across the *N/Yb* region

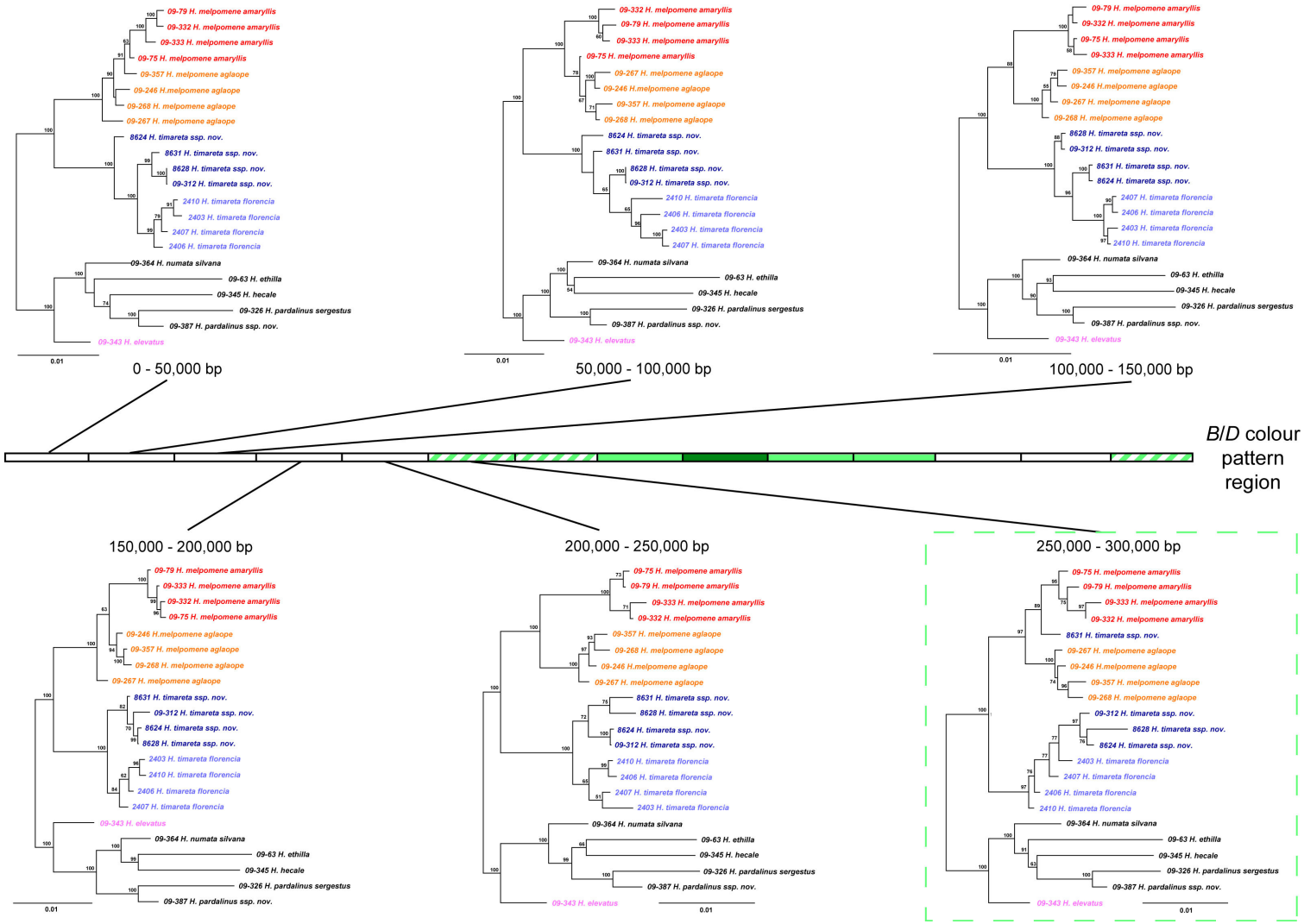


Figure S20.1 Phylogenetic analysis across the *N/Yb* region (50 kb scale)

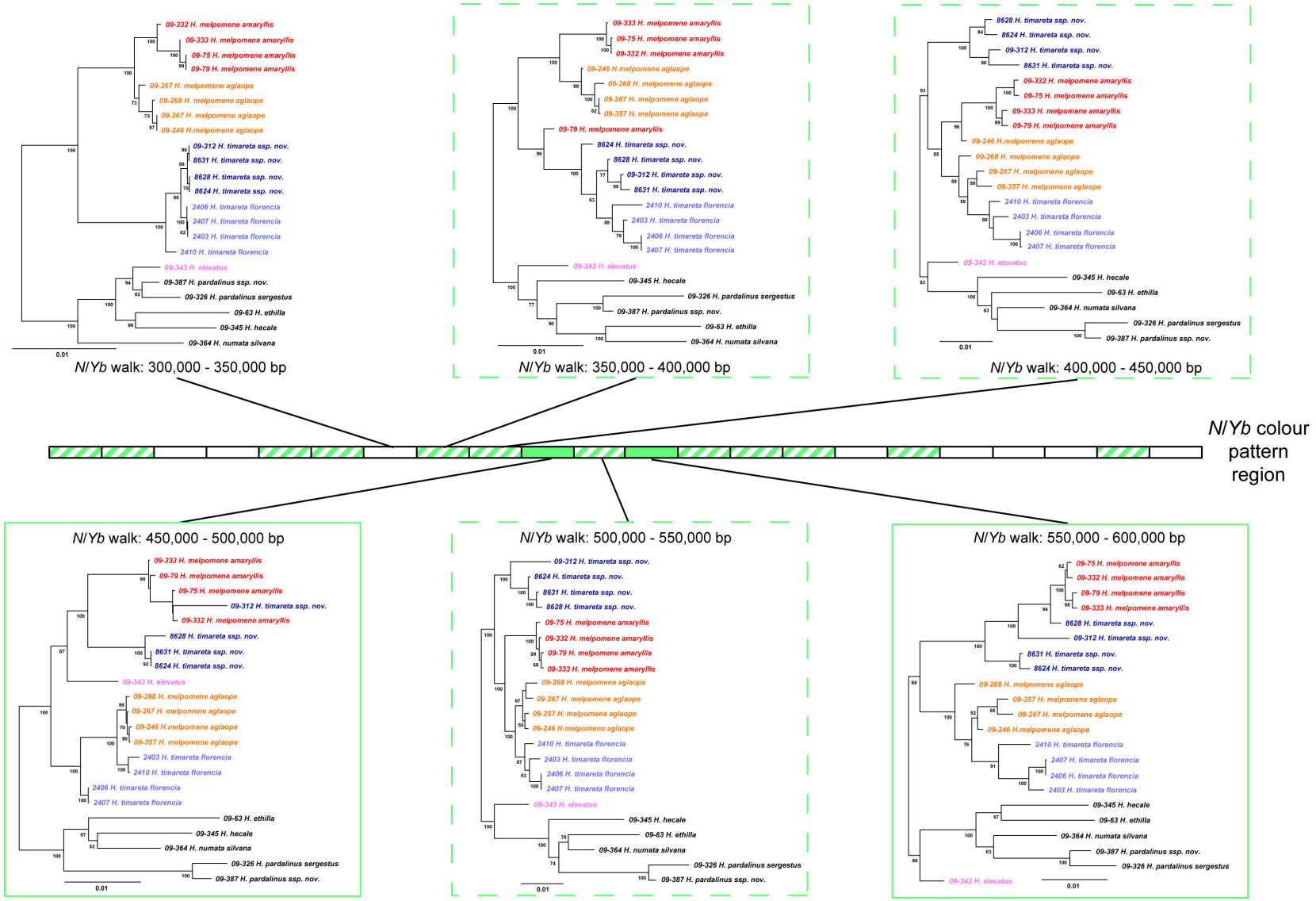


Figure S20.1, contd.

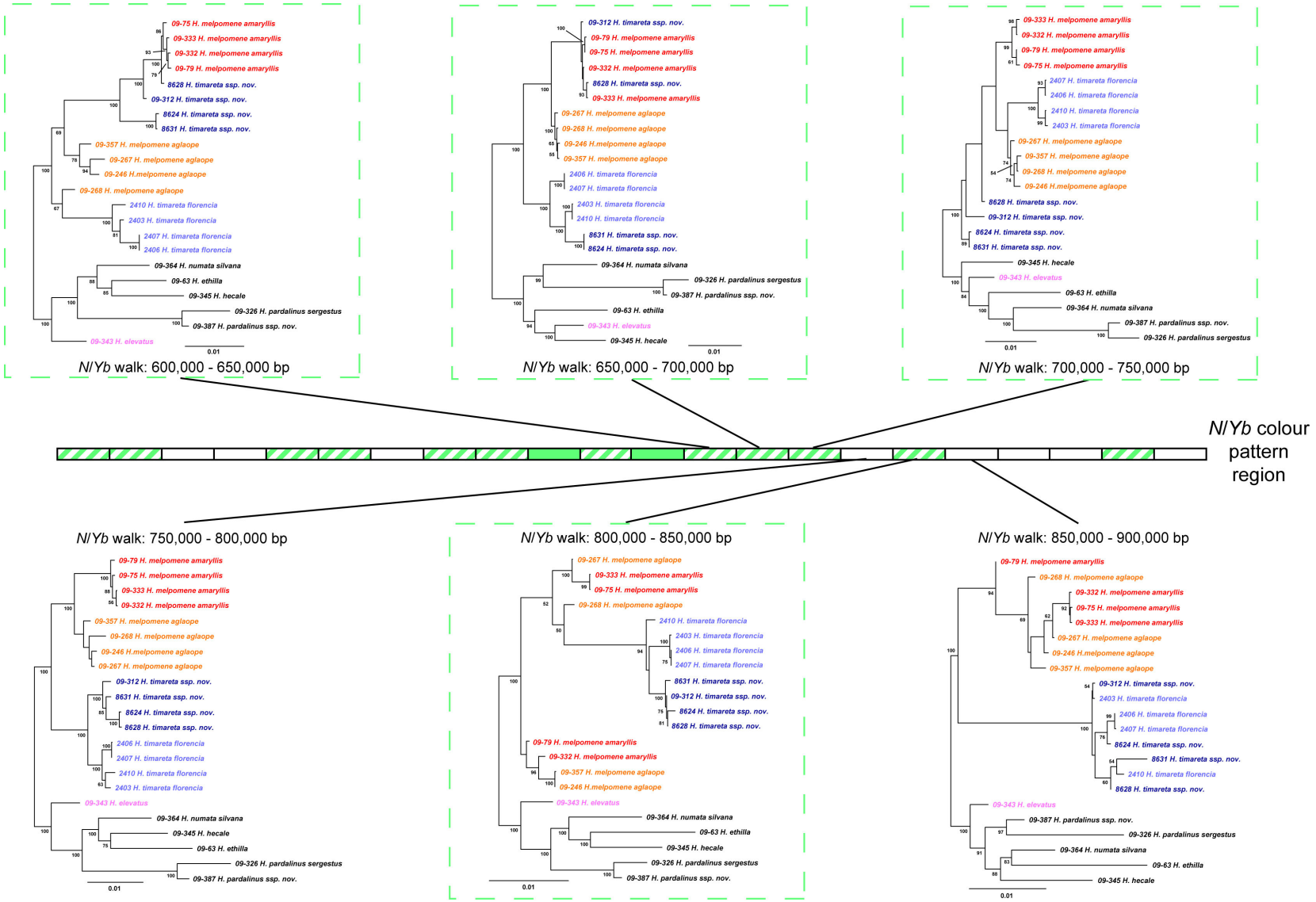


Figure S20.1, contd.

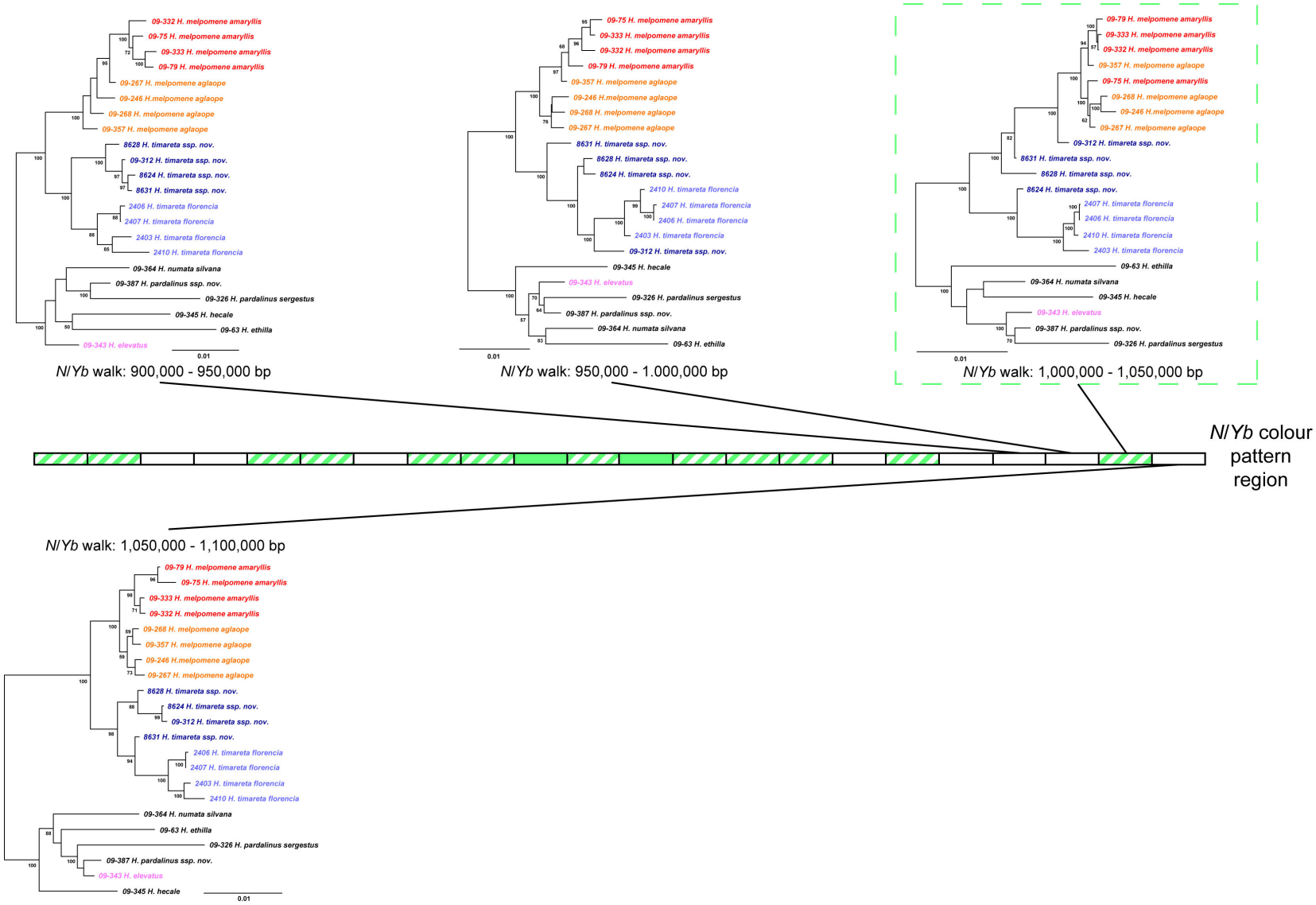


Figure S20.1, contd.

Figure S20.1 Phylogenetic analysis across the N/Yb region (50 kb scale)

Maximum likelihood phylogenies based on non-overlapping 50 kb windows across the *N*/*Yb* colour pattern region. Bootstrap supports for nodes are based on 100 bootstrap replicates and are shown when support is greater than 50. In the phylogenies from the white windows, the taxa are grouped by species. In the phylogenies from pale and dark green windows the taxa are grouped by colour pattern. In phylogenies from the striped boxes, either *H. melpomene* or *H. timareta* are paraphyletic but the taxa do not group cleanly by colour pattern.

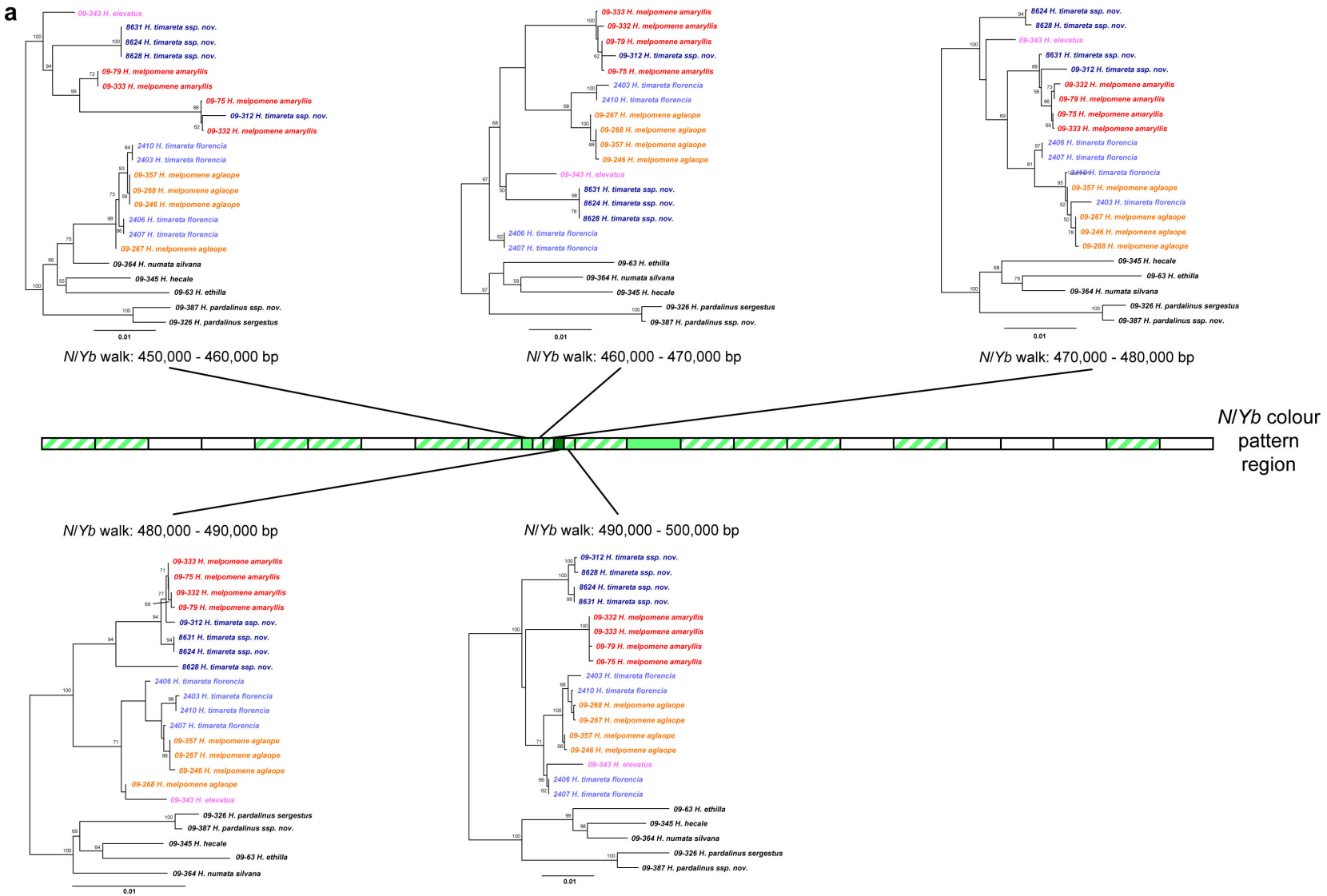


Figure S20.2 Fine-scale phylogenetic analysis across the *N/Yb* region (10 kb scale)

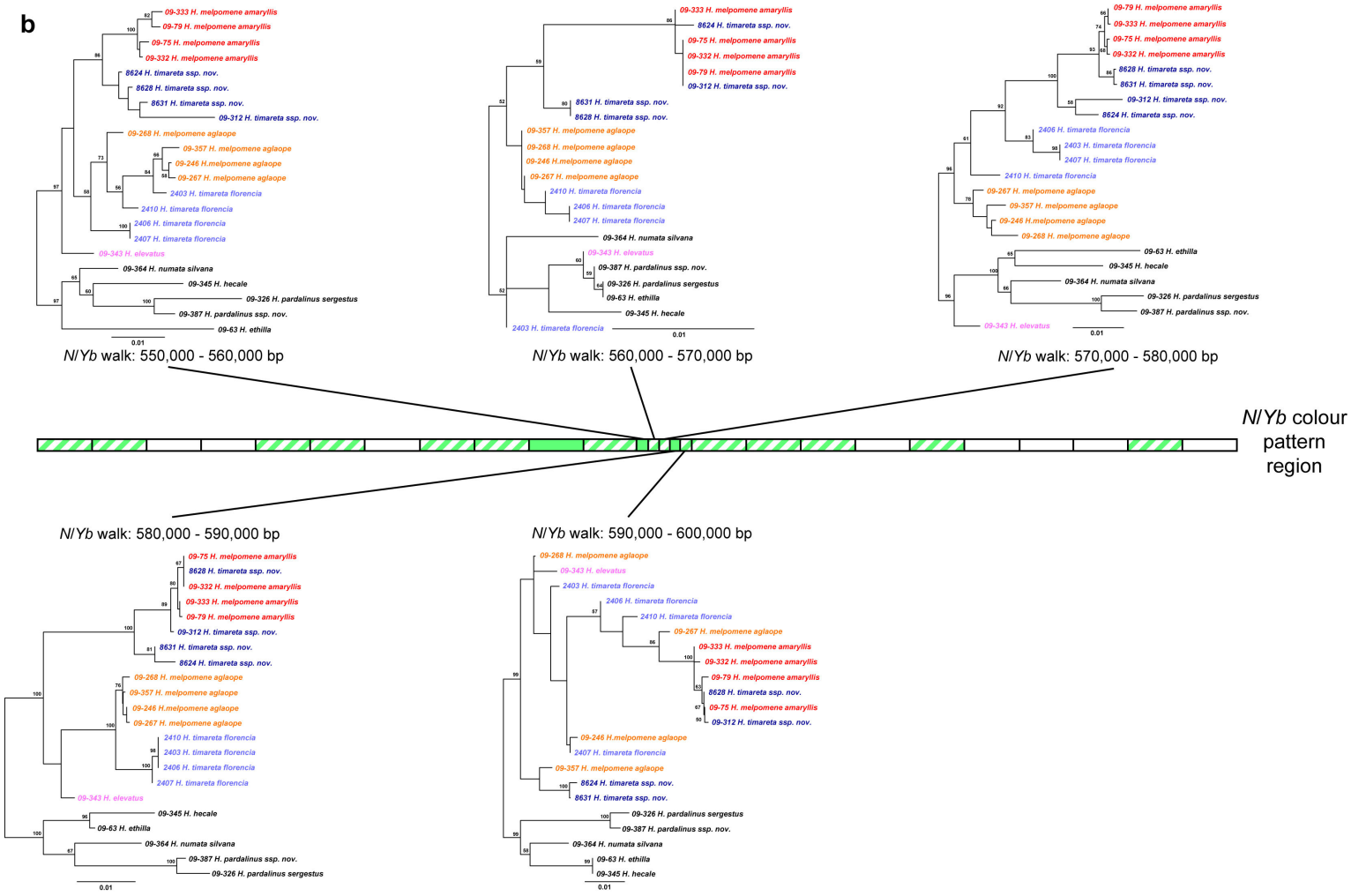


Figure S20.2, contd.

Figure S20.2 Fine-scale phylogenetic analysis across the N/Yb region (10 kb scale)

Maximum likelihood phylogenies based on non-overlapping 10 kb windows across two sections of the *N*/*Yb* colour pattern region. a) 450,000-500000 bp and b) 550,000-600000 bp. Bootstrap supports for nodes are based on 100 bootstrap replicates and are shown when support is greater than 50. In the phylogenies from the white windows, the taxa are grouped by species. In the phylogenies from pale and dark green windows the taxa are grouped by colour pattern. In phylogenies from the striped boxes, either *H. melpomene* or *H. timareta* are paraphyletic but the taxa do not group cleanly by colour pattern.

S21. Distribution of ABBA and BABA site patterns in non-colour pattern regions

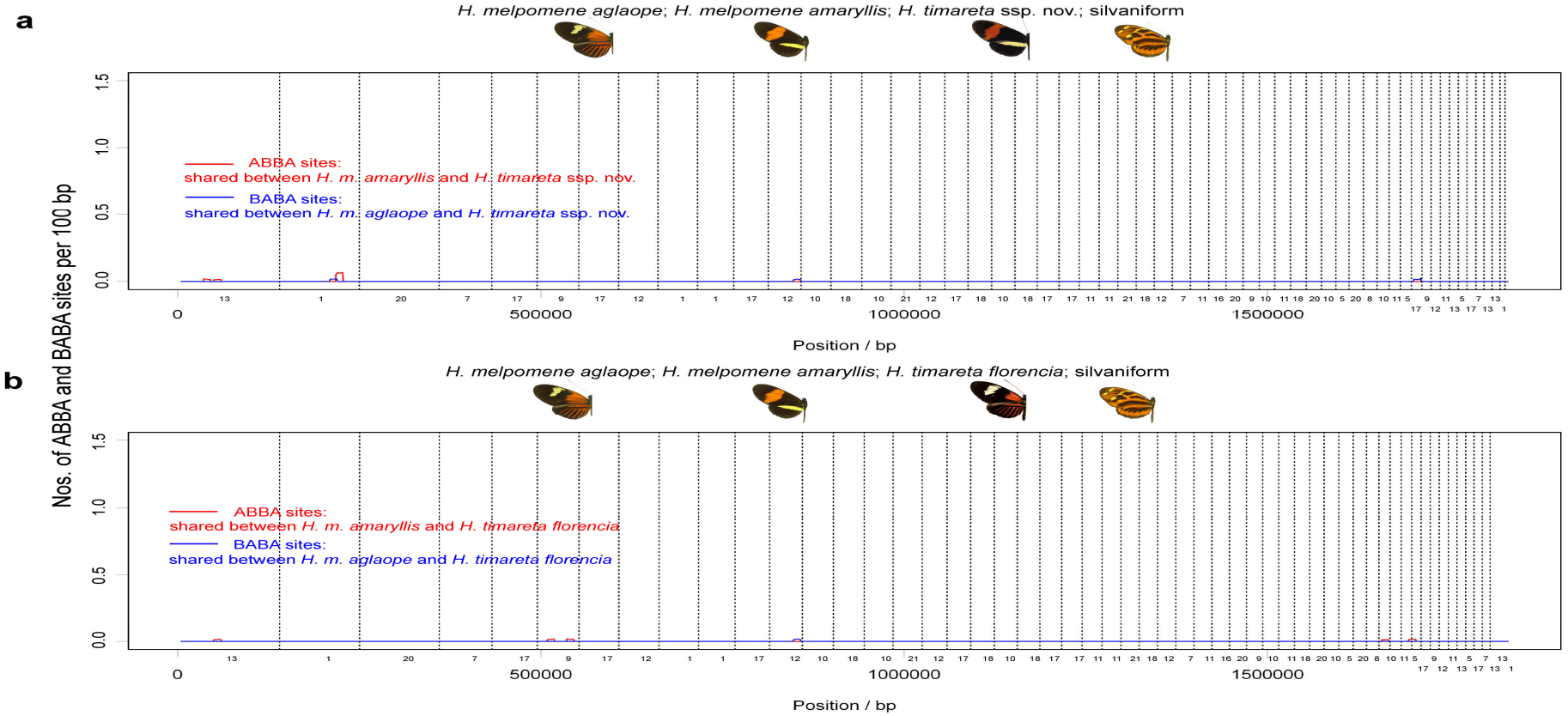


Figure S21.1 Distribution of ABBA and BABA site patterns in non-colour pattern regions

Distribution of ABBA and BABA single nucleotide sites within 10 kb sliding windows at 1 kb increments along non-colour pattern regions comprising ~1.8 Mb over 55 genome scaffolds (SureSelect dataset). The two comparisons shown are: **a,** (*H. melpomene aglaope*; *H. melpomene amaryllis*; *H. timareta* ssp. nov.; silvaniform) and **b,** (*H. melpomene aglaope*; *H. melpomene amaryllis*; *H. timareta florencia*; silvaniform). Vertical dotted lines represent boundaries between separate genomic regions that were targeted for sequencing. These plots only show positions that are fixed with each of the four taxa used in each comparison. The chromosome that each region is found on is indicated on the x-axis.

S22. Distribution map

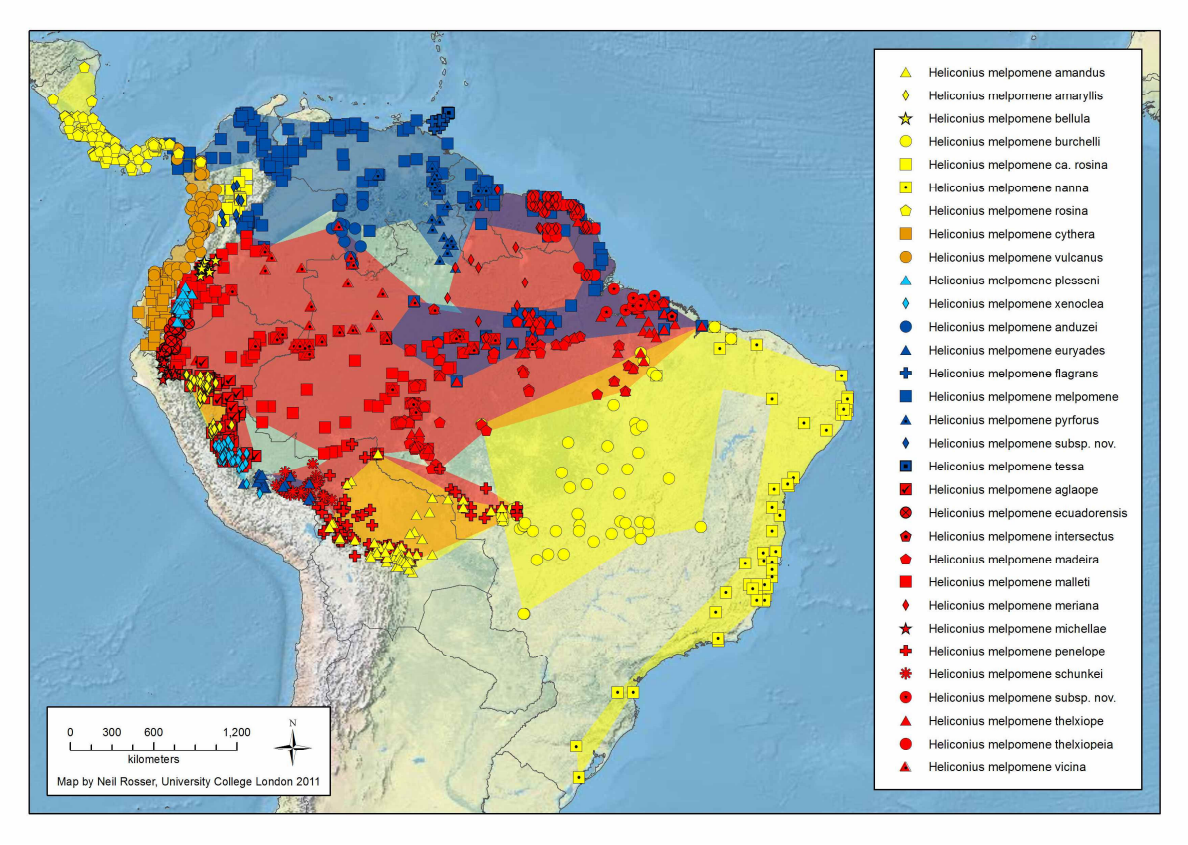


Figure S22.1 Distribution map of *H. melpomene* showing subspecies nomenclature

This distribution map is based on 4891 geographical records for *Heliconius melpomene* subspecies with information for 1576 point localities [96]. The principal sources of data were museum collections (primarily those in the Natural History Museum in London and the Florida Museum of Natural History in Gainesville), private research databases and the scientific literature.

S24. References cited

1. Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008).
2. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478–2483 (2002).
3. Shinzato, C. *et al.* Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* **476**, 320–323 (2011).
4. Nadeau, N. J. *et al.* Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philos Trans R Soc Lond B Biol Sci* **367**, 343–353 (2012).
5. Baxter, S. W. *et al.* Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* **6**, e19315 (2011).
6. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
7. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**, 1157–1158 (2011).
8. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
9. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
10. Turner, J. R. & Sheppard, P. M. Absence of crossing-over in female butterflies (*Heliconius*). *Heredity* **34 Part 2**, 265–269 (1975).
11. Jiggins, C. D. *et al.* A genetic linkage map of the mimetic butterfly *Heliconius melpomene*. *Genetics* **171**, 557–70 (2005).
12. Stam, R. & Van Ooijen, J. *JoinMap (TM) version 3.0: Software for the calculation of genetic linkage maps*. (2001).
13. Salzberg, S. L. *et al.* GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res* **22**, 557–67 (2012).
14. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21 Suppl 1**, i152–158 (2005).
15. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351–358 (2005).
16. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113–113
17. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).
18. Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**, 474–474
19. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
20. Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc Natl Acad Sci U S A* **100**, 6569–6574 (2003).
21. Holt, R. A. *et al.* The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149 (2002).
22. Nene, V. *et al.* Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* **316**, 1718–1723 (2007).
23. Osanai-Futahashi, M., Suetsugu, Y., Mita, K. & Fujiwara, H. Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* **38**, 1046–1057 (2008).
24. Zhan, S., Merlin, C., Boore, J. L. & Reppert, S. M. The monarch butterfly genome yields insights into long-distance migration. *Cell* **147**, 1171–1185 (2011).
25. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–96 (2008).
26. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215–225 (2003).

27. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
28. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
29. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61–D65 (2007).
30. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).
31. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100–3108 (2007).
32. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**, W686–W689 (2005).
33. Nelson, D. R. The cytochrome p450 homepage. *Hum. Genomics* **4**, 59–65 (2009).
34. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152–D157 (2011).
35. Guerra-Assunção, J. A. & Enright, A. J. MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* **11**, 133–133
36. 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–R25 (2009).
37. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie Chemical Monthly* **125**, 167–188 (1994).
38. Surridge, A. *et al.* Characterisation and expression of microRNAs in developing wings of the neotropical butterfly *Heliconius melpomene*. *BMC Genomics* **12**, 62 (2011).
39. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121–D124 (2005).
40. Moxon, S. *et al.* A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics* **24**, 2252–2253 (2008).
41. Bonnet, E., Wuyts, J., Rouzé, P. & Van de Peer, Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20**, 2911–2917 (2004).
42. Pringle, E. G. *et al.* Synteny and chromosome evolution in the lepidoptera: evidence from mapping in *Heliconius melpomene*. *Genetics* **177**, 417–26 (2007).
43. O'Brien, K. P., Remm, M. & Sonnhammer, E. L. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476–480 (2005).
44. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
45. Ng, M.-P. *et al.* OrthoClusterDB: an online platform for synteny blocks. *BMC Bioinformatics* **10**, 192–192
46. Vergara, I. A. & Chen, N. Large synteny blocks revealed between *Caenorhabditis elegans* and *Caenorhabditis briggsae* genomes using OrthoCluster. *BMC Genomics* **11**, 516–516
47. d' Alençon, E. *et al.* Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. *Proc Natl Acad Sci U S A* **107**, 7680–7685 (2010).
48. Coghlan, A. & Wolfe, K. H. Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genome Res.* **12**, 857–867 (2002).
49. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
50. Stein, L. D. *et al.* The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1**, E45 (2003).
51. Ranz, J. M., Casals, F. & Ruiz, A. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* **11**, 230–239 (2001).
52. Pelosi, P., Zhou, J.-J., Ban, L. P. & Calvello, M. Soluble proteins in insect chemical communication. *Cell. Mol. Life Sci.* **63**, 1658–1676 (2006).
53. Touhara, K. & Vosshall, L. B. Sensing odorants and pheromones with chemosensory receptors. *Annu. Rev. Physiol.* **71**, 307–332 (2009).

54. Hekmat-Scafe, D. S., Scafe, C. R., McKinney, A. J. & Tanouye, M. A. Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. *Genome Res* **12**, 1357–1369 (2002).
55. Gong, D.-P., Zhang, H.-J., Zhao, P., Xia, Q.-Y. & Xiang, Z.-H. The odorant binding protein gene family from the genome of silkworm, *Bombyx mori*. *BMC Genomics* **10**, 332–332
56. Vieira, F. G. & Rozas, J. Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biol Evol* **3**, 476–490 (2011).
57. Gong, D.-P. *et al.* Identification and expression pattern of the chemosensory protein gene family in the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* **37**, 266–277 (2007).
58. Forêt, S., Wanner, K. W. & Maleszka, R. Chemosensory proteins in the honey bee: Insights from the annotated genome, comparative analyses and expressional profiling. *Insect Biochem. Mol. Biol.* **37**, 19–28 (2007).
59. Tanaka, K. *et al.* Highly selective tuning of a silkworm olfactory receptor to a key mulberry leaf volatile. *Curr. Biol.* **19**, 881–890 (2009).
60. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
61. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
62. Legeai, F. *et al.* An Expressed Sequence Tag collection from the male antennae of the Noctuid moth *Spodoptera littoralis*: a resource for olfactory and pheromone detection research. *BMC Genomics* **12**, 86 (2011).
63. Grosse-Wilde, E. *et al.* Antennal transcriptome of *Manduca sexta*. *Proc Natl Acad Sci U S A* **108**, 7449–7454 (2011).
64. Dani, F. R. *et al.* Odorant-binding proteins and chemosensory proteins in pheromone detection and release in the silkmoth *Bombyx mori*. *Chem. Senses* **36**, 335–344 (2011).
65. Ozaki, K., Utoguchi, A., Yamada, A. & Yoshikawa, H. Identification and genomic structure of chemosensory proteins (CSP) and odorant binding proteins (OBP) genes expressed in foreleg tarsi of the swallowtail butterfly *Papilio xuthus*. *Insect Biochem. Mol. Biol.* **38**, 969–976 (2008).
66. Engler-Chaouat, H. S. & Gilbert, L. E. De novo synthesis vs. sequestration: negatively correlated metabolic traits and the evolution of host plant specialization in cyanogenic butterflies. *J. Chem. Ecol.* **33**, 25–42 (2007).
67. Vosshall, L. B. & Hansson, B. S. A unified nomenclature system for the insect olfactory coreceptor. *Chem. Senses* **36**, 497–498 (2011).
68. Jordan, M. D. *et al.* Odorant receptors from the light brown apple moth (*Epiphyas postvittana*) recognize important volatile compounds produced by plants. *Chem. Senses* **34**, 383–394 (2009).
69. Wanner, K. W. *et al.* Female-biased expression of odourant receptor genes in the adult antennae of the silkworm, *Bombyx mori*. *Insect Mol. Biol.* **16**, 107–119 (2007).
70. Schulz, S., Estrada, C., Yildizhan, S., Boppré, M. & Gilbert, L. E. An antiaphrodisiac in *Heliconius melpomene* butterflies. *J. Chem. Ecol.* **34**, 82–93 (2008).
71. Estrada, C., Yildizhan, S., Schulz, S. & Gilbert, L. E. Sex-specific chemical cues from immatures facilitate the evolution of mate guarding in *Heliconius* butterflies. *Proc Biol Sci* **277**, 407–413 (2010).
72. Estrada, C., Schulz, S., Yildizhan, S. & Gilbert, L. E. Sexual selection drives the evolution of antiaphrodisiac pheromones in butterflies. *Evolution* **65**, 2843–2854 (2011).
73. Chai, C.-L. *et al.* A genomewide survey of homeobox genes and identification of novel structure of the Hox cluster in the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* **38**, 1111–1120 (2008).
74. Zhong, Y. & Holland, P. W. H. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution & Development* **13**, 567–568 (2011).
75. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).
76. Obbard, D. J., Welch, J. J., Kim, K.-W. & Jiggins, F. M. Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genet* **5**, e1000698 (2009).

77. Sackton, T. B. *et al.* Dynamic evolution of the innate immune system in *Drosophila*. *Nat. Genet.* **39**, 1461–1468 (2007).
78. Waterhouse, R. M. *et al.* Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* **316**, 1738–1743 (2007).
79. Tanaka, H. *et al.* A genome-wide analysis of genes and gene families involved in innate immunity of *Bombyx mori*. *Insect Biochem. Mol. Biol.* **38**, 1087–1110 (2008).
80. Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–229 (2011).
81. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
82. Lewis, S. E. *et al.* Apollo: a sequence annotation editor. *Genome Biol* **3**, 1–14 (2002).
83. Evans, J. D. *et al.* Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Mol. Biol.* **15**, 645–656 (2006).
84. Zou, Z. *et al.* Comparative genomic analysis of the *Tribolium* immune system. *Genome Biol.* **8**, R177 (2007).
85. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
86. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
87. Guindon, S., Delsuc, F., Dufayard, J.-F. & Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* **537**, 113–137 (2009).
88. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & Mclnerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6**, 29 (2006).
89. Gerardo, N. M. *et al.* Immunity and other defenses in pea aphids, Acyrthosiphon pisum. *Genome Biol* **11**, R21 (2010).
90. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376 (2008).
91. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics*. (Sinauer Associates: Mass., 1997).
92. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
93. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
94. Huang, Q., Shete, S., Swartz, M. & Amos, C. I. Examining the effect of linkage disequilibrium on multipoint linkage analysis. *BMC Genet* **6**, S83 (2005).
95. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
96. Rosser, N., Phillimore, A. B., Huertas, B., Willmott, K. R. & Mallet, J. Testing historical explanations for gradients in species richness in heliconiine butterflies of tropical America. *Biological Journal of the Linnean Society* **105**, 479–497 (2012).