MOLECULAR ECOLOGY

Molecular Ecology (2010), **19** (Suppl. 1), 240–254 doi: 10.1111/j.1365-294X.2009.04475.x

Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the *HmYb/Sb* locus

LAURA FERGUSON,* SIU FAI LEE,¶ NICOLA CHAMBERLAIN,† NICOLA NADEAU,* MATHIEU JORON,‡ SIMON BAXTER,* PAUL WILKINSON,** ALEXIE PAPANICOLAOU,** SUJAI KUMAR,§ THUAN-JIN KEE,¶ RICHARD CLARK,†† CLAIRE DAVIDSON,†† REBECCA GLITHERO,†† HELEN BEASLEY,†† HEIKO VOGEL,‡‡ RICHARD FFRENCH-CONSTANT** and CHRIS JIGGINS*

**Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK, †Harvard FAS Center for Systems Biology, Northwest Lab Building, 52 Oxford Street, Cambridge, MA 02138, USA, ‡CNRS UMR5202, Case postale 29, 16, rue Buffon, 75005 Paris, France, §Ashworth Laboratories, University of Edinburgh, West Mains Road, EH9 3JT Scotland, ¶Department of Genetics, Bio21 Institute, University of Melbourne, 30 Flemington Road, Parkville, Melbourne, Australia, **School of Biosciences, University of Exeter in Cornwall, Penryn, Cornwall TR10 9EZ, UK, ††The Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK, ‡‡Max Planck Institute for Chemical Ecology, Hans-Knöll-Straße 8, D-07745 Jena, Germany*

Abstract

The mimetic wing patterns of *Heliconius* butterflies are an excellent example of both adaptive radiation and convergent evolution. Alleles at the *HmYb* and *HmSb* loci control the presence/absence of hindwing bar and hindwing margin phenotypes respectively between divergent races of *Heliconius melpomene*, and also between sister species. Here, we used fine-scale linkage mapping to identify and sequence a BAC tilepath across the *HmYb/Sb* loci. We also generated transcriptome sequence data for two wing pattern forms of *H. melpomene* that differed in *HmYb/Sb* alleles using 454 sequencing technology. Custom scripts were used to process the sequence traces and generate transcriptome assemblies. Genomic sequence for the *HmYb/Sb* candidate region was annotated both using the MAKER pipeline and manually using transcriptome sequence reads. In total, 28 genes were identified in the *HmYb/Sb* candidate region, six of which have alternative splice forms. None of these are orthologues of genes previously identified as being expressed in butterfly wing pattern development, implying previously undescribed molecular mechanisms of pattern determination on *Heliconius* wings. The use of next-generation sequencing has therefore facilitated DNA annotation of a poorly characterized genome, and generated hypotheses regarding the identity of wing pattern at the *HmYb/Sb* loci.

Keywords: mimicry, Heliconius, Lepidoptera, adaptation, alternative splicing

Received 2 June 2009; revision received 9 July 2009; accepted 13 July 2009

Introduction

The genetic basis of adaptive novelty is a major concern of contemporary evolutionary biology. Whilst it was traditionally believed that a large number of genes would contribute to the generation of adaptive phenotypes and morphological change, it is increasingly being shown that morphological evolution both within and between species can be accounted for by a surprisingly small number of loci. Quantitative Trait Locus analysis of phenotypic variation has often shown that loci of major effect control much of the morphological differentiation between related species and populations (Colosimo *et al.* 2004; Shapiro *et al.* 2004; Steiner *et al.* 2007). Even more strikingly, where the genetic basis of such differences has been analysed functionally, repeated involvement of the same genes has been found. Thus,

Correspondence: Chris Jiggins, Fax: 01223 336676; E-mail: cj107@cam.ac.uk

adaptive pigmentation differences can result from alterations at the same loci, as in the case of *MC1R* mutations in a range of vertebrates (Mundy *et al.* 2004; Hoekstra 2006; Nadeau *et al.* 2007; Steiner *et al.* 2007), and *yellow* gene expression within the genus *Drosophila* (Wittkopp *et al.* 2002; Prud'homme *et al.* 2006).

The great phenotypic diversity of butterfly wing patterns offers an opportunity to study adaptive morphological evolution. Previous work on the developmental basis of wing patterning in *Bicyclus anynana* has shown that highly conserved developmental pathways such as wingless, transforming growth factor-β (TGF-β), hedgehog and *Notch*, and transcription factors such as *engrailed* (Caroll *et al.* 1994; Monteiro *et al.* 2006; Keys *et al.* 1999; Brunetti *et al.* 2001; Reed & Gilbert 2004) are re-deployed in generating eyespot patterns. There is also some evidence that population level variation in eyespot size might be controlled by differences in expression of the *Distalless* homeobox gene (Beldade *et al.* 2002). These results lend support to the view that co-option of conserved signalling pathways through *cis*-regulatory evolution has played a key role in morphological innovation during evolution (Carroll 2008).

In contrast to the serially repeated eyespot patterns of species such as *B. anynana*, the broad bands of colour seen in *Heliconius* wing patterns suggest a different developmental basis to wing patterning (Reed & Gilbert 2004; Joron *et al.* 2006a; but see Nijhout 1991). The presence or absence of *Heliconius* wing pattern elements is under simple Mendelian control (Sheppard *et al.* 1985; Mallet 1989; Joron *et al.* 2006a), but thus far a candidate gene approach has not yielded information on the identity of these loci (Joron *et al*., 2006c). Here, we develop previous work in which we have described markers closely linked to the locus controlling presence of a hindwing bar in *Heliconius melpomene*, and shown homology in the genomic location of similar genes for polymorphism in *Heliconius erato* and *Heliconius numata* (Jiggins *et al.* 2005; Joron *et al.* 2006a; Joron *et al*., 2006c).

Heliconius show a continuum of genetic divergence from freely hybridizing colour pattern races through incipient species with partial reproductive barriers to completely reproductively isolated species, offering the opportunity to study multiple stages of speciation (Mallet *et al.* 2007). *Heliconius melpomene*, for example, includes 29 distinct geographic races with bright, simple wing patterns of red, orange, yellow, white and black. Most of the phenotypic variation can be explained by gene complexes located on two linkage groups: *HmYb/HmSb/HmN* (linkage group 15) which controls white and yellow pattern elements, and *HmB/HmD* (linkage group 18) which controls red pattern elements (Gilbert 2003; Naisbit *et al.* 2003; Joron *et al*., 2006c; Kronforst *et al.* 2006). More recently, molecular markers have demonstrated that these patterning loci are also shared with more distantly related species. Most notably, the *Yb* locus of *H. melpomene* is orthologous to *Cr* in the distantly related and mimetic *H. erato*, and in both species controls the yellow hindwing bar phenotype, but also maps to the same genomic location as the *P* locus of *H. numata* which controls phenotypically divergent whole-wing polymorphism of mottled 'tiger' type wing patterns (Joron *et al*., 2006c).

One of the major challenges in characterizing functionally important variation in taxa with poorly characterized genomes is sequence annotation. The protein sequences in public databases are highly taxonomically biased towards the traditional model organisms, meaning that many genes in organisms such as butterflies show no similarity to published sequence databases. Furthermore, automated gene-finding methods are not optimized for poorly characterized genomes (Korf 2004). To overcome these difficulties, next-generation sequencing technologies offer a method for rapidly characterizing the transcriptomes of novel genomes at relatively low cost (Hudson 2008; Vera *et al.* 2008). The high sequence coverage that can potentially be achieved allows for error correction and the detection of rare transcripts or alternative splicing (Weber *et al.* 2007). Here, we identify and fully sequence a genomic region controlling two major phenotypic changes in the mimetic butterfly, *H. melpomene*. We further develop a transcriptomic sequence library from developing wing tissue and use this to annotate and characterize the genomic sequence. We therefore provide a detailed description of genomic regions at two patterning loci, and demonstrate how high-throughput sequencing technologies can be used for annotation of genomic sequence in an evolutionary model system.

Methods

Bacterial artificial chromosome (BAC) tilepath construction

A single clone (AEHM-41c10) was previously identified from an *Heliconius melpomene* BAC library as tightly linked to the *HmYb* locus (Joron *et al*., 2006c). The BAC library provides approximately 7.9× coverage of the *H. melpomene* genome (Joron *et al.* 2006b) and has been fingerprinted using *Hind*III restriction digest. A total of 17 494 successfully fingerprinted clones were assembled into 2086 contigs (with average 6.2 clones) plus 4484 singletons (Baxter *et al*., 2008b). Fingerprint contigs provide a physical genomic map, and, within each contig, a BAC tilepath showing the extent of clone overlap and approximate length of each clone. We also end-sequenced 18 816 clones from the BAC library by Sanger di-deoxy sequencing, generating 32 528 sequence reads (NCBI trace archive, Accession nos 1433293924–1433326451).

To extend the *HmYb* genomic walk from the AEHM-41c10 clone (i) primers were designed from the BAC sequence/end sequence of the current clone, (ii) a product was amplified from *H. melpomene* genomic DNA, (iii) the product was used to probe the BAC library as described previously (Baxter *et al.* 2008a). Next (iv) primers were designed from BAC-end sequence of positive clones, (v) their genomic location was confirmed using linkage mapping and (vi) polymerase chain reaction (PCR) amplification between clones and *in silico* alignment of end sequences with the existing tilepath were used to confirm relative clone positions. Finally, the best clones for sequencing were chosen based on the clone length and extent of overlap as predicted from the fingerprint database.

Target amplicon size for probe design was 500 bp, and where possible primers were designed in coding sequence. It was not possible to identify a clone following 21B20 from either probing the BAC library or the fingerprint database. Therefore, using synteny with the *Bombyx mori* genome, the *Bombyx parn* gene was identified and used to search the 454 transcriptome library (see below) for *H. melpomene* sequence. Primers were designed and a product was amplified for probing the library (Table S2, Supporting information). This library screen identified a single large fingerprint contig (Contig 105, Table S1, Fig. S1, Supporting information), from which four clones were chosen for sequencing. Therefore, the BAC fingerprint, clone end sequences, *Bombyx-Heliconius* synteny and wing transcriptome sequence were all essential tools in constructing the tilepath.

Clones were sequenced, assembled and finished by the Wellcome Trust Sanger Institute to HTG phase 3 (Joron *et al*., 2006c). Part of this BAC walk was used to develop markers for probing the *Heliconius erato* genome and has been published in a previous comparative analysis (Papa *et al.* 2008). Here, we present the first complete description of the candidate genomic region for *HmYb*, and extend the walk significantly to additionally cover the *HmSb* locus.

Linkage mapping

Linkage mapping was carried out using molecular markers developed from genomic sequences, including noncoding regions, putative genes and microsatellites which were identified using the automated 'Microsatellite repeat finder' available on ButterflyBase (Papanicolaou *et al.* 2005). In total, several hundred pairs of primers were designed from which those in Table S2 (Supporting information) were chosen based on reliability, amplification across the four broods of the mapping family, unambiguous genotypes and informative variation. Markers in coding sequence generally performed better than microsatellites or other noncoding markers. Primers were designed using Primer3 (Rozen & Skaletsky 2000) to span an intron where possible and with a target amplicon size of 600 bp.

There is no meiotic crossing-over in female Lepidoptera (Turner & Sheppard 1975), so paternal alleles were used to generate a linkage map of the candidate region. In total, 371 individuals from four mapping families were used for linkage mapping (Fig. 1). Grandparents were wild-caught *H. melpomene melpomene* (individuals 10 064 and 8233 from French Guiana) and *H. melpomene cythera* (8237 and 8292 from Mindo, Ecuador). *HmYb* genotypes were scored as homozygous dominant (no bar, Yb/Yb), homozygous recessive (bar, Yb^cYb^c) and

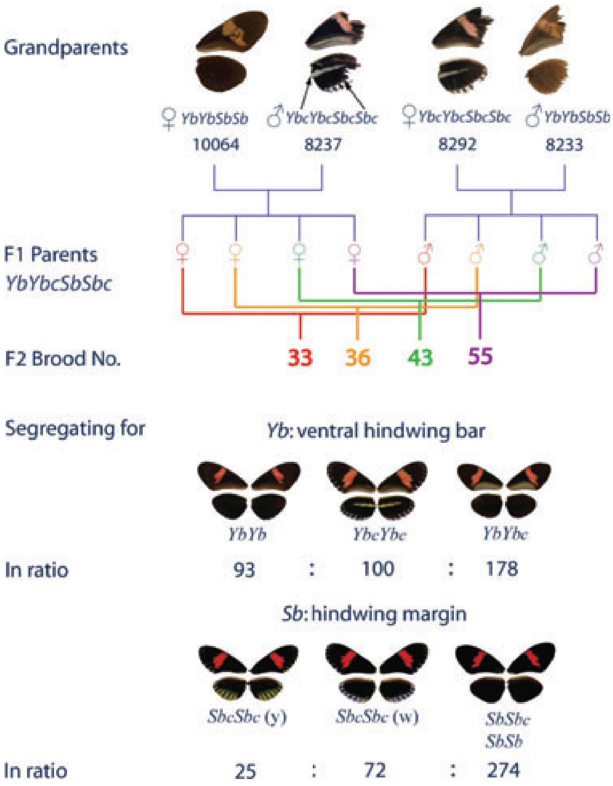

Fig. 1 Mapping families segregating for *HmYb* and *HmSb*. All four mapping broods (33, 36, 43 and 55) originated from two single pair matings between *Heliconius melpomene melpomene* and *Heliconius melpomene cythera*. The F2 progeny segregated for *Yb* and *Sb* phenotypes under expected Mendelian ratios. An additional unlinked locus, *K*, is also segregating in this cross and controls the colour of the hindwing margin, with yellow (y) recessive to white (w). Data from the four broods were combined for mapping of *HmYb* and *HmSb*.

heterozygous (Yb/Yb^c). Heterozygous individuals have a 'shadow' bar as described previously (Joron *et al.*, 2006c). The *HmSb*^{c} allele is recessive and was scored as margin present (Sb^{c}Sb^c), or absent (SbSb, SbSb^c). The complete genotype could therefore be inferred only for those individuals carrying a maternal chromosome with the recessive allele, which gave a total of 175 *Sb* informative individuals. The pigmentation (yellow/white) of the hindwing margin is controlled by an unlinked locus, *K* that is not considered further here (Jiggins *et al.* 2005).

DNA was extracted from the mapping broods using a DNeasy kit (Qiagen). For the parents of each brood, each marker was amplified using PCR and products analysed for visible length variation on a 1.5% agarose gel stained with ethidium bromide. Where variation was apparent, DNA from the brood progeny was amplified and genotyped directly by scoring size variation on agarose. When length variation was not visible, the parental PCR products were cleaned using ethanol precipitation and sequenced directly. Sequence traces were searched using CodonCode Aligner software (CodonCode Corporation) for informative restriction digest sites that would yield different sized bands on agarose. Where such sites were identified, all progeny from each brood were then amplified for the marker, digested and the genotypes scored on agarose as above. All progeny were scored for markers at both ends of the BAC tilepath so that all recombinant individuals were identified. If a variable site in parental sequence could not be cleaved by restriction digest, only those progeny known to be recombinants were amplified and sequenced, and the informative site scored from sequence traces. Although indels were common in noncoding sequence, the markers chosen had sufficient 'clean' sequence to carry out genotyping. The type of assay used for each marker is given in Table S2 (Supporting information). In most cases, sequence identity was confirmed using BLAST, and fine-scale linkage mapping of markers to the predicted location in the tilepath additionally confirmed identity of amplicons. Microsatellite length variation was either scored directly on an agarose gel or by amplifying the loci using fluorescently labelled primers, in which case products were run on an ABI 3730 capillary sequencer (Applied Biosystems) and fragments analysed using Genemapper v. 3.7 (Applied Biosystems).

Generation of H. melpomene *transcriptome sequence*

Heliconius melpomene colonies were established from pupae ordered from Stratford Butterfly Farm, UK. Two pools of normalized wing cDNA were generated from mixed RNA extracted from wing tissue of the races *H. melpomene cythera* and *H. melpomene malleti* and sequenced using 454 FLX technology. Whole fore- and hindwings were dissected from larval and pupal stages of both *H. m. cythera* (17 individuals of unknown sex) and *H. m. malleti* (six individuals of unknown sex) and total RNA was extracted using the Trizol Reagent (Invitrogen) followed by messenger RNA isolation using the NucleoTrap mRNA kit (Nigel Machery). mRNA was pooled by race according to the following proportions: 40% mid-fifth instar larvae, 20% late fifth instar (crawler), 20% early pupae and 20% pre-ommochrome pupae (See Ferguson & Jiggins 2009 for staging). First strand cDNA was synthesized using the Creator SMART cDNA Library Construction Kit (Clontech). Double stranded cDNA synthesized from the first strand cDNA (product of reverse transcription) was normalized using the Trimmer-Direct Kit (Evrogen, Cat# NK002). Briefly, DNA template was denatured and allowed to re-anneal at 68 °C in the presence of duplex-specific nuclease (DSN) solution, which preferentially degrades double stranded nucleic acids (made up mostly of highly abundant transcripts). The reaction was terminated after 20 min of DSN treatment and the resulting equalized single stranded cDNA was reamplified using adapter (introduced during previous reverse transcription) specific primers. The optimal cycle number for each sample was empirically determined by subjecting the same template to a range of PCR cycles and the products checked on a 1.5% agarose gel. Samples with signs of over cycling, characterized by an intense, uniform, high-molecular weight smear, were discarded. To evaluate the quality of the normalization, we included no-DSN controls in the above procedures, whose products were run along side with the tested samples. The no-DSN controls typically displayed a number of strong bands against a cDNA smear whereas the normalized samples showed a smooth cDNA smear (most intense between 1–2 kb) with no obvious bands. Our second validation was to prepare a plasmid library for each sample. An aliquot of normalized cDNA was digested with *Sfi-1* and the purified products were ligated to the pDNR-lib vector (Clontech). Plasmids were transformed into electrocompetent DH10B cells by electroporation. A total of 101 colonies (58 cythera and 43 aglaope) were sequenced and we obtained 99 unique gene objects (2 contigs and 97 singletons). The low redundancy of these expressed sequence tags (ESTs) indicated that the cDNA was successfully normalized. Ds-cDNA was then purified and concentrated using the DNA Clean and Concentrator kit (Zymogen). Approximately 10 μg of the normalized ds-cDNA per strain was sequenced using the 454 FLX system (454 Life Sciences).

Raw 454 Roche pyrosequencing data were obtained in the form of Standard Flowgram Format files (.sff). The sequences in these files were preprocessed using a

Perl script to remove the poly-T primer and the Smart-IV linkers before assembly using the Newbler assembler. Preprocessing involved trimming the sequence reads to remove the adapter or linker if present and removing all bases downstream of a poly-A tail or upstream of a poly-T sequence. The trimming steps, in order, were: (i) if a poly-A sequence was found that matched the pattern 'five or more As, followed by up to two other nucleotides, followed by five or more As', then the first such poly-A sequence and all bases downstream of it were removed from the read. (ii) Conversely, if a poly-T sequence was found that matched the pattern 'five or more Ts, followed by up to two other nucleotides, followed by five or more Ts', then the last such poly-T sequence and all the bases upstream of it were removed from the read. (iii) If any part of the read matched the two known adapters (poly_T_primer: AAGCAGTGGTATCAACGCAGAGT-GGCCGAGGCGGCCTGTTTTGTTTTTTTTTCTTTTTTTT-TTVN and smart-IV: AAGCAGTGGTATCAACGCA GAGTGGCCATTACGGCCGGG), then the matching parts were discarded. The match was established by using BLASTn with e-value cutoff of 0.1) and '-F f' (do not filter query sequence). (v) If the sequence CGGCCGGG was seen within 10 bases of the start in the remaining sequence, it and all bases upstream were removed. Conversely, if the sequence CCCGGCCG was seen within 10 bases of the end, it and all bases downstream were removed. This pattern, although present in step 3, was sometimes not found by the BLASTn algorithm and had to be looked for explicitly. (v) If, in the remaining bases, any Ns were present, the whole read was discarded because the N indicates a low-quality read (Huse *et al.* 2007). The trimmed reads were resaved as a .sff file and the Newbler assembler (Roche, v 1.1.03.24) was used with default settings to assemble the reads into as few contigs as possible. For comparison, the reads were also assembled using MIRA (v2.9.26x3) (Chevreux *et al.* 2004).

Transcriptome annotation

The resulting assembly was annotated using known proteins (uniref100) from the UniProt Consortium (http://www.uniprot.org). For comparison with *B. mori*, we first concatenated the Genome Consortium version 2.0 protein predictions with UniProt and RefSeq (Pruitt *et al.* 2007) manually curated protein sets and then made this database nonredundant at the 100% level using the cd-hit software (Li & Godzik 2006). To compare with existing butterfly transcriptomes, we generated MIRA assemblies from the public EST data available for *Bicyclus anynana* and *Melitaea cinxia*. Using the tBLASTx algorithm and custom Perl scripts, we calculated the number of contigs matching a butterfly or known eukaryotic protein. To identify the proportion of coding vs. noncoding, we parsed the BLAST report to calculate the number of individual bases of our assembly matching a protein. All databases were downloaded on 15 May 2009. BLAST similarity searches against these databases were performed using e-value and bit-score cut-offs of 1e-5 and 80 bits respectively. For insect repeats, we concatenated *B. mori* predicted repeats provided by Dr Chris Smith (University of San Franscisco) with the Insecta class from Repbase (Jurka *et al.* 2005) and made nonredundant at the 95% level. To characterize *H. melpomene* genomic sequence repeats, BAC-end sequences were screened with the ReRep pipeline, which is optimized for identification of repetitive structures in a Genome Survey Sequence data set (Otto *et al.* 2008). Repeat databases for Insecta, *Bombyx* and *Heliconius* were then concatenated for BAC sequence annotation.

Annotation of the HmYb/Sb *region*

We used both automated and manual annotation to generate gene models for the candidate region. The motivation was to develop automated gene prediction methods for *H. melpomene*, but also investigate the extent to which such methods could be improved by manual inspection of gene models and comparison with raw sequence reads.

The *Heliconius* BAC sequences for the *HmYb* locus were individually annotated with the repeat database and the clustered EST contigs (see above) using the annotation pipeline MAKER (Cantarel *et al.* 2008). This identifies repeats, aligns ESTs and proteins to a DNA sequence, produces *ab initio* gene predictions, and automatically synthesizes the data into gene annotations with evidence-based quality indices. BLAST searches were performed against our database of *B. mori* proteins. Gene predictions were generated by the Semi-HMM-based Nucleic Acid Parser (Korf 2004) using a Hidden Markov Model optimized to the *B. mori* genome. The resulting gff files (those describing genes and other features associated with DNA) from the MAKER pipeline were analyzed and viewed with the Apollo Genome Annotation Curation Tool version 1.9.6 (Lewis *et al.* 2002).

The starting point for manual annotation was gene predictions as generated above and from the automated *Bombyx* gene prediction tool 'Kaikogaas' (http://kaikogaas.dna.affrc.go.jp/). Kaikogaas is an online tool for genome annotation based on parameters determined from the *B. mori* genome. Predicted genes from the two methods were concatenated into a single Artemis file (Rutherford *et al.* 2000; Carver *et al.* 2008), and the new models were used to search NCBI nucleotide, protein and EST sequence databases using tBLASTx or tBLASTn (cutoff e^{-5}) as well as local databases of all the unassembled 454 sequence reads from *H. m. cythera* and *H. m. malleti* (BLASTn, cutoff e = 0.001). All significant *Heliconius* hits were assembled against BAC sequence with an alignment cutoff of 90% (*H. melpomene* 454/EST sequence) or 75% (other *Heliconius* EST sequence) nucleotide identity. All alignments were manually checked, and Artemis models were modified where necessary. Significant hits for other species were translated, and the most closely related species aligned with the translated *H. melpomene* gene models using ClustalW (Larkin *et al.* 2007). Where *H. melpomene* models diverged from other species, translated sequence from distantly related insects (e.g. *Apis*, *Tribolium*, *Drosophila*) was aligned, highly conserved regions were identified, and *H. melpomene* BAC sequence was searched for these regions as a starting point for gene model modification in Artemis. In addition to searches using the MAKER/Kaikogaas predicted gene set, 2 kb sections of BAC sequence that overlapped by 500 bp from the 7g12, 11j7 and 29b7 BACs were used in BLASTn searches against the *H. m. cythera* raw 454 sequence reads (cutoff e = 0.001). This analysis recovered expressed noncoding sequence, and generated additional gene models. The final number of 454 reads in each gene model is given in Table 1.

Genes for which splice variants were identified from 454 transcriptome sequence (see Results) were verified for *H. m. cythera* and *H. m. malleti* by reverse transcription–PCR (RT–PCR) amplification and sequencing, using the same cDNA used to generate the 454 libraries as a template. Full details of these spliced gene models will be published elsewhere (G. Wu, personal communication). 5′ and 3′ rapid amplification of CDNA ends (RACE) (Clontech) primers were also designed to amplify longer transcripts than could be obtained from 454-EST sequence reads for *HM00023/HM00024*. In the case of *HM00023* both 3′- and 5′-RACE primers were designed in each of the exons (coding and noncoding) identified from both *H. m. cythera* and *H. m. malleti*. Those which amplified clear products are given in Table S2 (Supporting information). The cDNA used for 454 sequencing was again used as a template. Thus, the final *H. melpomene* gene prediction set is based on automated gene prediction with support and modification based on *H. melpomene* transcript sequence, amino acid homology with other insects, and in some cases, additional RT–PCR.

To assess synteny around the *HmYb* locus, the genomic location of *B. mori* homologues was recorded from SilkDB (Xia *et al.* 2004), and *H. melpomene* genomic location defined as the number of nucleotides from the start of the AEHM-46 m10 BAC, with continual numbering over the entire BAC tilepath. The relative locations of *B. mori* and *H. melpomene* genes were plotted by taking the mid-point of each gene.

Results

Linkage mapping of HmYb *and* HmSb

The *HmYb* locus controls the presence or absence of a yellow bar on the hindwing of *Heliconius melpomene*, whilst the tightly linked locus *HmSb* similarly controls a white margin on the hindwing. We have used molecular markers to define the genomic location of these loci. A total of 371 individuals from four broods were scored for the segregation of the hindwing bar, and, because of genetic dominance, 175 of these were also scored for the hindwing margin (Fig. 1). From the 175 individuals scored for both pattern elements, two individuals were identified with recombinant phenotypes between *HmYb* and *HmSb*. These individuals were confirmed as genetic recombinants using molecular markers, offering the first molecular confirmation of recombination between tightly linked pattern loci in *Heliconius* (Fig. 3b). Assuming a genome-wide recombination rate of 180 kb/cM (Jiggins *et al.* 2005), the two recombinants indicate that *HmYb* and *HmSb* are 150 kb apart, although this is an approximate estimate given the variation in recombination rate known to occur across the genome (e.g. see Fig. 3).

We constructed a BAC tilepath across the *HmYb/Sb* loci, and carried out fine-scale linkage mapping of markers developed from both coding and noncoding DNA. The BAC tilepath consisted of eleven clones, 1.42 Mb in total length, representing 1.15 Mb of nonredundant sequence. All clones have been sequenced by the Wellcome Trust Sanger Institute to HTG phase 3 quality (Fig. 3a). Linkage mapping of 32 markers across the region (Fig. 3b) demonstrated that the *HmYb* phenotype was completely linked to molecular markers across a 323 kb region, encompassing a portion of the AEHM-7g12 BAC, and all of the clones AEHM-11j7, 29b7 and 21b20. This inclusive region for the *HmYb* pattern locus was defined by recombination events at a gene in the AEHM-7g12 BAC (*HM00004/Trehalase 1B*, see below) and a noncoding BAC-end marker in AEHM-21b20 (21b20-T7, Fig. 3b). The *HmSb* phenotype was completely linked to markers across an overlapping region of approximately 271 kb and included part of AEHM-21b20, all of AEHM-22A15 and AEHM-24o2, and one end of AEHM-31B4. The *HmSb* pattern locus was defined by recombination events at the BAC-end markers 36a11-SP6 and 31b4-SP6.

Characterizing the H. melpomene *transcriptome*

Expression libraries were created from developing wing discs of two phenotypically distinct *H. melpomene* strains. The first, *H. m. cythera*, displayed the *HmB* red

Table 1 Annotated genes at the *HmYb* locus with putative homologues from *Bombyx mori* and *Drosophila melanogaster*

Heliconius melpomene gene	Putative gene name	No. *H. melpomenel* 454-EST reads		Best *B. mori* BLASTp hit	% ID	Best *D. melanogaster* BLASTp hit	% ID	Identified in butterflies	
		cy	*Ma*					Bic.	Me.
HM00001	Acylpeptide hydrolase	0	1	BGIBMGA005667	46	NS			
HM00002	Acylpeptide hydrolase	3	4	BGIBMGA005668	54	NS			
HM00003	HM00003	0	0	BGIBMGA005547	66	CG10949	26		
HM00004 [*HmYb*]	Trehalase-1B	5	3	BGIBMGA005665	68	CG9364	40	X	X
HM00006 [*HmYb*]	Trehalase-1A	3	5	BGIBMGA005664	58	CG9364	40	X	X
HM00007 [*HmYb*]	B9 protein	0	0	BGIBMGA005663	67	CG14870	37		
HM00008 [*HmYb*]	HM00008	0	0	BGIBMGA005548	81	CG5098	22		
HM00010 [*HmYb*]	WD40 repeat domain 85	27	4	BGIBMGA005662	60	CG3184	40		
HM00011 [*HmYb*]	CG18292	1	1	BGIBMGA005661	91	CG18292	62		
HM00012 [*HmYb*]	CG2519	0	0	BGIBMGA005549	54	CG2519	41		
HM00013 [*HmYb*]	Unkempt	0	0	BGIBMGA005660	74	Unkempt	52		
HM00014 [*HmYb*]	Histone H3	55	38	BGIBMGA005550	100	His3.3A	89	X	
HM00015 [*HmYb*]	HM00015	25	12	BGIBMGA005659	52	CG30373	33	X	X
HM00016 [*HmYb*]	HM00016	1	0	BGIBMGA005551	58	CG5280	26		
HM00017 [*HmYb*]	RecQ Helicase	15	9	BGIBMGA005666	69	mus309	32	X	X
HM00018 [*HmYb*]	HM00018	3	5	BGIBMGA005553	80	NS			
HM00019 [*HmYb*]	BmSuc2	0	0	BGIBMGA005555	55	NS			X
HM00020 [*HmYb*]	CG5796	67	61	BGIBMGA005556	69	CG5976	48		X
HM00021 [*HmYb*]	HM00021	72	37	BGIBMGA005657	28	NS			
HM00022 [*HmYb*]	Enoyl-CoA hydratase	94	58	BGIBMGA005656	80	CG6543	57		X
HM00023 [*HmYb*]	ATP binding protein*	x	x	BGIBMGA005557	50	CG10581	32		X
HM00024 [*HmYb*]	HM00024	42	25	BGIBMGA005655	42	Sur-8	19	X	X
HM00025 [*HmYb*]	HM00025	34	61	BGIBMGA005652	57	cort	15	X	X
HM00026 [*HmSb*]	Poly(A)-specific ribonuclease (parn)	27	22	BGIBMGA005650	60	NS			
HM00027 [*HmSb*]	CG31320	37	23	BGIBMGA005649	67	CG31320	20		
HM00028 [*HmSb*]	ARP-like	60	23	BGIBMGA005559	84	ARP-like	72	X	X
HM00029 [*HmSb*]	CG4692	37	28	BGIBMGA005648	87	CG4692	85	X	X
HM00030 [*HmSb*]	Proteasome 26S non ATPase subunit 4	71	42	BGIBMGA005560	93	Pros54	65	X	X
HM00031 [*HmSb*]	HM00031	5	5	NS		NS			
HM00032 [*HmSb*]	Zinc phosphodiesterase	104	59	BGIBMGA005646	65	jhI-1	42		X
HM00033 [*HmSb*]	Serine/threonine-protein kinase (LMTK1)	3	10	BGIBMGA005561	45	ror†			
HM00034	WD repeat domain 13 (Wdr13)	4	3	BGIBMGA005645	81	NS			
HM00035	Tyrosine phosphatase (truncated)	3	7	BGIBMGA005642	50	Dome	18	X	X
HM00036	WAS protein family homologue 1	0	0	BGIBMGA005643	62	Wash	27		
HM00037	Tyrosine phosphatase (full length)	38	32	BGIBMGA005642	59	Dome	18	X	X
HM00038	Lethal (2) k05819 CG3054	1	0	BGIBMGA005562	72	Lethal (2) k05819	40		X
HM00039	Mitogen-activated protein kinase (MAPKK)	70	44	BGIBMGA005641	89	Licorne	64	X	X
HM00040	DNA excision repair protein ERCC-6	26	19	BGIBMGA005640	64	Hel89B	20	X	X

Table 1 *Continued*

		No. *H. melpomenel* 454-EST reads						Identified in butterflies	
Heliconius melpomene gene	Putative gene name	*cy*	*Ma*	Best *B. mori* BLASTp hit	% ID	Best *D. melanogaster* BLASTp hit	% ID	Bic.	Me.
HM00041	Penguin	137	77	BGIBMGA005637 BGIBMGA005638	70 66	Penguin	31	X	X
HM00042	Thymidylate kinase	0	0	NS	-	CG5757	48		X
HM00043	Caspase-activated DNase	1	0	BGIBMGA005639	71	Rep4	34		
HM00044	Regulator of ribosome biosynthesis	161	102	BGIBMGA005564	80	CG32409	57	X	X
HM00045	CG12659	8	6	BGIBMGA005565	82	CG12659	40	X	
HM00046	CG33505	42	17	BGIBMGA005636	77	CG33505	47		X
HM00047	Sr protein	83	43	BGIBMGA005635	62	SC35	72		
HM00048	HM00048	0	0	NS		CG5435	27		
HM00049	HM00049	2	1	BGIBMGA005566	29	NS			
HM00050	Shuttle craft	2	1	BGIBMGA005634	58	Shuttle craft	35		
HM00051	HM00051	0	0	BGIBMGA005567	44	NS			
HM00052	HM00052	0	0	BGIBMGA005633	48	NS		X	
HM00053	Lethal (2) giant larvae	2	0	BGIBMGA005570	80	Lethal (2) giant larvae	45		
HM00054	Zn finger protein	63	35	BGIBMGA005571	45	Crooked legs	18	X	X
HM00055	Cohesion subunit	1	0	BGIBMGA005632	71	CG41265	21		
HM00056	Tetratricopeptide repeat protein	0	0	BGIBMGA005572	83	CG4525	48		
HM00057	Ecdysone oxidase	0	0	BGIBMGA005711	40	CG9512	30	X	X
HM00058	Cuticle protein (Cpr64Ac)	81	25	BGIBMGA005631	59	Cuticular protein 64Ac	40	X	X
HM00059	CG6734	2	0	BGIBMGA005629	62	CG6734	36		
HM00060	Uridine 5′-monophosphate synthase	87	21	BGIBMGA005628	71	Rudimentary-like	52	X	X
HM00061	HM00061	3	1	BGIBMGA005626	24	CG7368	22	X	X

Putative gene names are based on BLASTp homology with a bitscore cut-off of 200. *One combination of exons at *HM00023* shows homology with a *Bombyx mori* predicted gene which encodes a putative ATP binding protein. There are, however, many noncoding splice variants at this locus with no homology to other insect sequences. †The *Heliconius melpomene* HM00033 predicted protein is approximately four and a half times longer than *Drosophila melanogaster* ror (3061 amino acids vs. 685), but the length is conserved with *B. mori* BGIBMGA005561 (2841 amino acids). Four genes (*HM00013/unkempt*; *HM00041/ penguin*, *HM00053/ Lethal (2) giant larvae*, *HM00054/* Zn finger *protein*) have putative *D. melanogaster* homologues with wing mutant phenotypes. Presence of putative homologues in *Bicyclus anynana* (*Bic.*) and *Melitaea cinxia* (*Me.*) at a BLASTn bitscore cut-off of 80 is also shown. *cy*, *H. m. cythera*; *ma*, *H. m. malleti*; NS, nonsignificant sequence identity at a BLASTp bitscore cut-off of 80; EST, expressed sequence tags.

forewing band, yellow hindwing bar (*HmYb*) and hindwing margin (*HmSb*), whilst the second, *H. m. malleti*, displayed the red-rayed *HmD* hindwing phenotype, but was somewhat polymorphic for the shape of the forewing yellow band phenotype. Each DNA pool was normalized to reduce the transcript level of highly expressed genes and increase the rate of gene discovery. After trimming a total of 195 159 sequence reads were obtained for *H. m. malleti* representing 48 Mb from one run of 454 FLX sequencing, and 482 027 reads, representing 103 Mb, for *H. m. cythera* from 1.5 runs (Accession no. SRA008857). The average read length after trimming was 214 bp.

To choose the best *de novo* assembly method for our data, we assembled sequence reads using both the Roche Newbler pipeline and the open source MIRA assembler (Chevreux *et al*. 2004; Margulies *et al*. 2006). The results are illustrated by alignment of two reference genes to the AEHM-11j7 BAC clone (Fig. 2). Both assemblers performed well on the relatively simple *Histone H3* gene. However, the MIRA assembler significantly under-assembled the data for the *Enoyl Co-A Hydratase* gene, and was more prone to indel errors (Fig. 2). The Newbler pipeline produced two contigs for this gene as compared with the six generated by MIRA. Addition of the published Sanger ESTs led to a further improvement in the assembly, such that the two Newbler contigs were joined into a single contig. On the basis of this comparison, we used the Newbler assembly to generate a reference contig set for further analysis. In combination with the 4971 Sanger EST sequences available on dbEST (September 2008) for *H. melpomene*, an assembly of all the *H. melpomene* data generated 25 585 contigs, representing 13 Mb, and 56 404 singletons. Of these, there were 8546 'large' contigs with an average contig length of 1130 bp representing 9.7 Mb. This provides an excellent resource for annotation of genomic sequence and future expression profiling experiments.

Our coverage of the transcriptome was assessed by similarity searches of the joint 454 + EST assembly (*c.* 82 000 contigs and singletons) with reference proteomes. Overall, our transcriptome assembly contains 11% of the Uniref100 database with 8370 *H. melpomene* proteins identified and 9107 (55%) of the *Bombyx mori* reference proteome (with 10 075 *H. melpomene* proteins identified). To identify proteins yet to be curated, we searched with tBLASTX the assemblies of two butterflies with substantial public EST data: *Bicyclus anynana* and *Melitaea cinxia*. The *H. melpomene* assembly identified 6019 and 6346 of these proteins respectively. It is expected that as more butterfly transcriptomes are

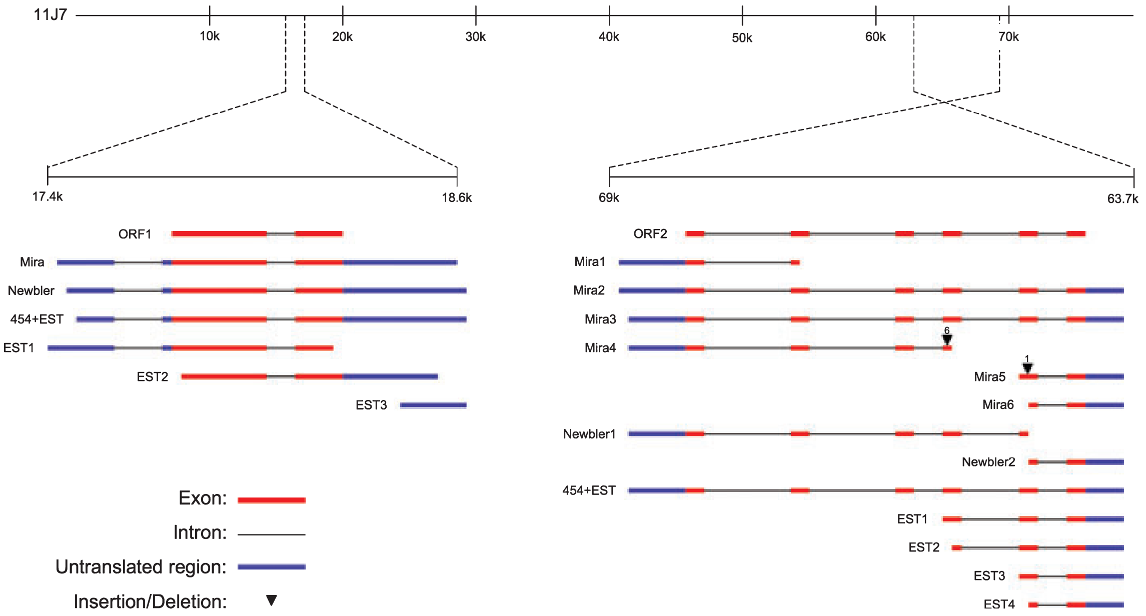

Fig. 2 Performance of sequence assembly strategies at the exon level. TOP: Two regions of the BAC clone AEHM-11j7 (GenBank identifier: 160950684) were analyzed, indicated by dashed lines. LEFT: A 411-base open reading frame (ORF1) homologous to *Histone H3* was compared. MIRA and Newbler assemblies performed equally well, as both had 100% coverage of ORF1, and predicted similar untranslated regions. RIGHT: A 894-base ORF (ORF2) homologous to *enoyl-CoA hydratase* was found containing six exons spanning a total of 4948 bases. Six MIRA contigs match this region, covering 11–100% of ORF2; two nonoverlapping Newbler contigs account for 99.2% of the ORF; when Sanger expressed sequence tags (EST) were included in the Newbler assembly (i.e. 454 + EST), 100% coverage was attained. All contigs are free of insertion/deletion (indel) polymorphisms within the coding regions except for Mira4 and 5, which have six (five single base insertions and one deletion) and one indels, respectively.

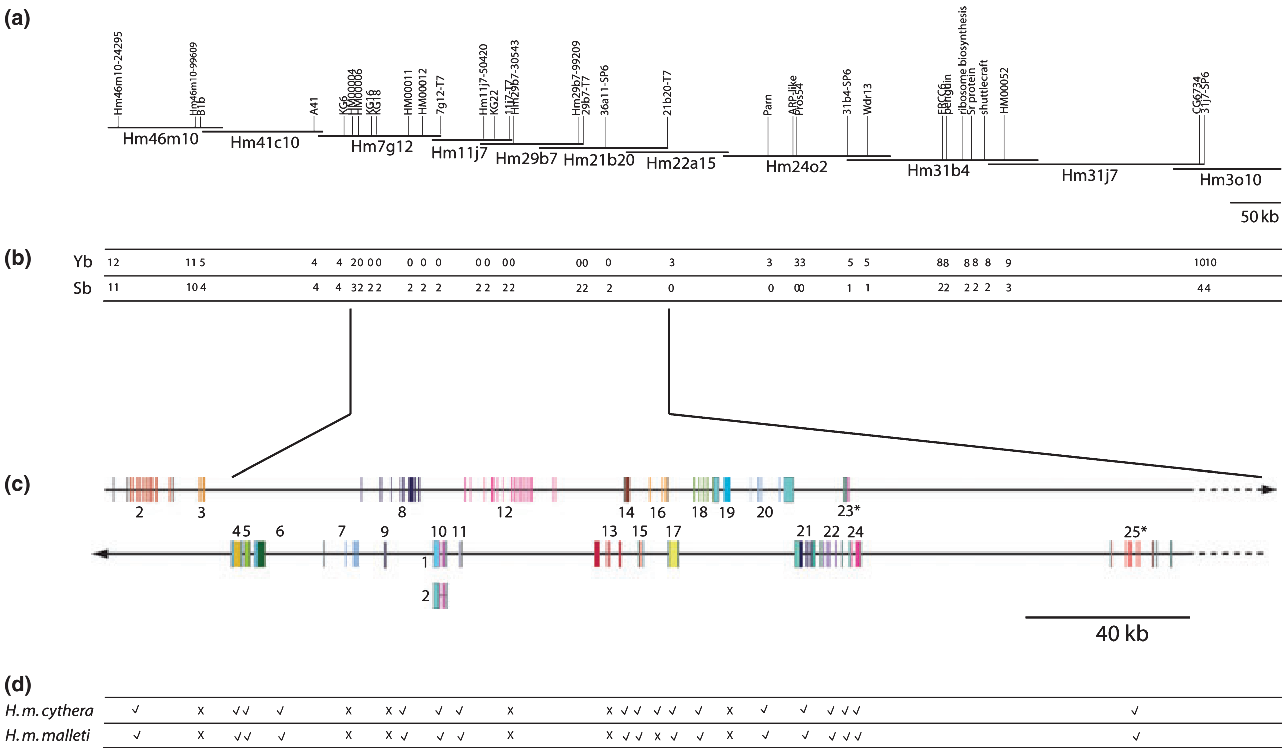

Fig. 3 The *HmYb/Sb* loci. (a) The BAC clone tilepath spanning the *HmYb/Sb* loci showing markers used for mapping. (b) Fine-scale linkage mapping shows the number of recombinants between molecular markers and the two pattern loci in 371 individuals (*HmYb*) and 175 individuals (*HmSb*) respectively. (c) The BAC clones containing *HmYb* locus genes (AEHM-7g12, AEHM-11j7, AEHM-29b7) are shown in detail. BAC AEHM-21b20 is not predicted to contain any genes and is therefore shown as a dotted line not to scale. The *HM000* prefix is not shown on gene numbers for clarity. Genes with splice variants are indicated by *; see Results and Fig. S3 (Supporting information) for details. Untranslated regions (UTRs) are shown in turquoise. *HM00010* has alternative 5′-UTRs, one shared with *HM00011* (-1) and one unique (-2). (d) Presence or absence of gene objects in transcriptome data sets from two *Heliconius melpomene* pattern races shows that the majority of genes in the region are expressed in developing wings of both *H. m. cythera* and *H. m. malleti.*

becoming available, we will be able to accurately identify a higher number of butterfly proteins.

A customized set of 347 *H. melpomene* repeat sequences (available upon request) was combined with a database of insect repeats (see Methods) for BAC sequence annotation. One putative *H. melpomene* repeat showed high similarity to a zinc-finger transcription factor (BLASTX, e-value 3e-29) and this sequence was removed from the database for subsequent analysis. RepeatMasker (v. open-3.2.6) (A.F.A. Smit, R. Hubly, P. Green, unpublished data. v. open 3.2.6.) was then used to mask repeats within BAC clones prior to automated gene annotation. In total, the masked sequence represented between 16.5% (AEHM-11j7) and 34.9% (AEHM-41c10) of the total BAC clone sequence.

Annotation of the HmYb *region*

To identify putative functional sites in both coding and noncoding sequence, we carried out a detailed annotation of the BAC sequences linked to *HmYb/Sb*. Building on a previous study, where eight putative genes were identified in three *H. melpomene* BAC sequences linked to *HmYb* (AEHM-41c10, 7g12 and 11j7) (Papa *et al.* 2008), we here annotate the entire *HmYb/Sb* candidate region. In total 59 gene objects were found in the BAC tilepath spanning the clones AEHM-7g12- 31j7, 45 of which had clear homologues in other insect species (Table 1). Linkage mapping narrowed the *HmYb* locus to a region containing 20 genes (*HM00004–HM00025*), and the *HmSb* locus to eight gene objects (*HM00026–HM00033*).

Across the annotated tilepath region, there were three instances of gene duplication: Acylpeptide hydrolases *HM00001* and *HM00002*; Trehalases *HM00004* and *HM00006* (*HmYb* region); and tyrosine phosphatases *HM00035* and *HM00037*. The Trehalase duplication has been noted previously (Papa *et al.* 2008), although the copy numbers should be referred to as *Trehalase-1A* (*HM00006*) and *Trehalase-1B* (*HM00004*) to reflect direction of transcription and probable orthology to *B. mori Trehalase* genes (Fig. S2, Supporting information). *H. melpomene HM00004* has a premature stop codon at amino acid position 137. A single *H. m. cythera* 454 sequence read covering the region also encoded the stop codon, confirming that this was unlikely to be a sequencing error. Tyrosine phosphatases *HM00035* and *HM00037* had 87% amino acid identity, but *HM00035* was truncated by approximately 600 amino acids at the C terminus.

Ten additional genes showed alternative splicing; *HM00010*, *HM00011*, *HM00023* and *HM00025* in the *HmYb* region, *HM00026, HM00032* in the *HmSb* region, and *HM00039, HM00040, HM00047* and *HM00054* outside the candidate regions. Two of these, *HM00023* and *HM00054* had a large number of apparently noncoding RNA transcripts. At one of these, *HM00023*, RT–PCR was used to confirm gene models, which demonstrated a large number of splice variants for this gene (Fig. S3, Supporting information). In total, 18 putative exons were identified from transcriptome sequence, and these were joined in many combinations. One exon combination (as annotated in Fig. 3c) had homology to a *B. mori* gene in the corresponding chromosomal location with predicted ATP binding activity, but the *H. melpomene* protein product was truncated at both ends relative to *B. mori* (Table 1, Fig. S3, Supporting information). Other transcripts contained further premature termination codons, and typically did not encode an open reading frame longer than a single exon because of insertions/deletions at intron/exon boundaries. Annotation of *HM00025* (a putative cort/fizzy family homologue) also revealed a number of transcript variants. The most common transcript had nine exons, encoding a protein product of 365 amino acids (Fig. 3). RT–PCR confirmed an alternative transcript that lacked exon 5, and identified a number of alternative 5′-untranslated regions (UTRs).

In addition to this splicing in coding regons, in one case a single 5′-UTR was associated with two genes (*HM00010*, *HM00011*, *HmYb* region) (Fig. 3c). Annotation of *HM00010*, a gene characterized by WD40 repeat domains, revealed two alternative 5′-UTR regions: one shared with *HM00011* (-1), and one adjacent to the first coding exon (-2, Fig. 3c). The expression of both isoforms was confirmed using RT–PCR. No similar alternative splicing was found in either *B. mori* or *Drosophila melanogaster* EST databases, and in *D. melanogaster* putative *HM00011* and *HM00010* homologues (Table 1) are separated by ~4.5 Mb.

Previous studies have shown conserved synteny between moths and butterflies at a chromosomal level (Pringle *et al.* 2007; Sahara *et al.* 2007; Beldade *et al.* 2009). More specifically, the eight *HmYb* linked genes previously identified also show conserved microsynteny with *B. mori* (Papa *et al.* 2008). Here, with more detailed annotation, gene order remains strongly conserved right across the entire *HmYb/Sb* loci with respect to *B. mori* (Fig. 4). Despite the considerable evolutionary distance between the species (≈100 million years Pringle *et al.* 2007) and large difference in estimated genome size (*B. mori* = 500 Mb Xia *et al.* 2004; *H. melpomene* = 292 Mb (Jiggins *et al.* 2005), chromosomal location, gene order, direction of transcription and relative spacing between genes were all highly conserved (Fig. 4). Seven predicted genes were present in *B. mori* and not *H. melpomene* in the *HmYb* region (*BGIBMGA005653, –5654, –5558, 5554, –5552, –5658, –5546*), although according to the *B. mori* genome annotation, only the last three of these had support from EST sequence or BLAST homology. Overall, there was an almost uniform increase in *B. mori* intergenic and intronic sequence relative to *H. melpomene*. As described previously (Papa *et al.* 2008), a single gene (a putative homologue of *RecQ Helicase, HM00017*) was found to have been translocated within the sequenced region (Fig. 4), and a small inversion was found outside the candidate regions covering genes *HM00041–44*. The first and last genes (*HM00001, HM00061*) were also found in opposite orientations between the species. The truncated duplicate *HM00035* was not found in *B. mori*, and gene *HM00057/ecdysone oxidase* was not found within the homologous *B. mori* scaffold contig (nscaf2829), but several putative homologues were found in other genomic regions.

Discussion

Genetic loci controlling *Heliconius* wing patterns are 'hotspots' for evolutionary change (Joron *et al.* 2006a; Papa *et al.* 2009). The tightly linked *HmSb* and *HmYb* loci control the presence of a white wing margin and yellow hindwing bar respectively in *Heliconius melpomene*. In the co-mimic species, *Heliconius erato*, this genomic region controls parallel pattern elements, and in both species the loci also have pleiotropic effects on various aspects of wing phenotype (Baxter *et al.* 2009). The *HmYb* locus also differentiates closely related species such as *Heliconius cydno* and *Heliconius pachinus* and hence contributes to reproductive isolation and speciation (Gilbert 2003; Naisbit *et al.* 2003). In *Heliconius numata*, the homologous region is a 'supergene' that controls virtually all aspects of pattern variation, and facilitates mimicry of distantly related *Melinaea* (Ithomiini) species (Joron *et al.*, 2006c). Here, for the first time we have sequenced, mapped and annotated the region responsible for these pattern changes in heliconius. Furthermore, we have characterized the wing transcriptome of *H. melpomene* using 454 pyrosequencing, facilitating sequence annotation and generating hypotheses for identification of functional variants at *HmYb*.

Considering that the transcriptome data are derived from a single tissue (wings) isolated over a relatively short period of the lifespan (late larval through pupal stages), the next-generation sequencing approach we

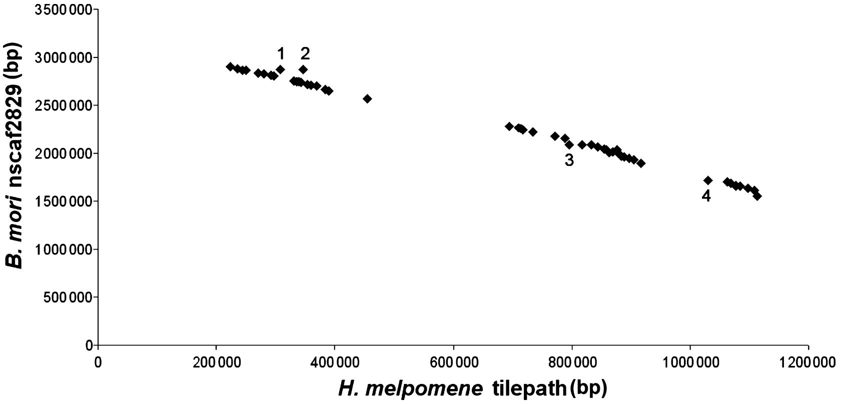

Fig. 4 Comparison of microsynteny between *Heliconius melpomene* and *Bombyx mori* across the *HmYb/Sb* tilepath. The chromosomal location, relative order and spacing of putative homologues between *B. mori* and *H. melpomene* across BAC clones AEHM-7g12-31j7 is highly conserved. Gene location is given relative to the start of the AEHM-46m10 BAC sequence for *H. melpomene*, and within genomic scaffold contig nscaf2829 for *B. mori*. The increase in genome size between *B. mori* and *H. melpomene* has occurred at a consistent rate across both coding and noncoding regions. The numbered points are (i) CG2519, (ii) recQ helicase, (iii) the novel Tyrosine phosphatase gene HM00035 and (iv) Lethal (2) giant larvae.

employed has obtained a remarkable number of genes, indicating a highly complex wing transcriptome. Comparison with the *Bombyx mori* proteome suggests that a minimum of 9107 (55%) unique genes have been sampled in our libraries. For comparison, a microarray study of pupal wing development in *Drosophila melanogaster* showed expression of 9394 genes in total (Ren *et al.* 2005), suggesting that we may have nearly saturated gene coverage of the butterfly 'wing' transcriptome. As in previous studies, the high sequence coverage of many genes allows some correction of the higher error rates encountered in pyrosequencing, by alignment of high sequence depth at each base position (Cheung *et al.* 2006; Huse *et al.* 2007; Vera *et al.* 2008). Nonetheless, alignment shows that the coverage of *B. mori* amino acid positions in our data is 37%, suggesting that overall sequence coverage of each gene remains incomplete. The transcriptome coverage obtained here is significantly greater than that from another recent survey of a butterfly, *Melitaea cinxia*, in which around 600 K 454 sequence reads showed similarity to 6289 *B. mori* proteins (Vera *et al.* 2008). The increased coverage from *Heliconius* is surprising given the broader coverage of tissues sampled in *Melitaea*, but the discrepancy is most probably because of advances in technology, such that the 454FLX technology used here gives significantly longer read lengths (250 bp rather than 110 bp).

A detailed annotation of the candidate *HmYb/Sb* locus is essential for developing and testing hypotheses regarding the nature of the wing pattern genes (Papa *et al.* 2008). In addition, alignment of the transcriptome to high-quality genome sequence allows for verification of the assembly methods used and of the coverage obtained in the transcriptomic survey. Thus the combination of a fully-sequenced BAC tilepath and deep coverage of expressed sequence can generate testable hypotheses about the nature of patterning loci. In total, eight putative genes were identified as candidates for *HmSb*, and 20 for *HmYb*. In both instances, a high proportion were present in our wing transcriptome data (8/8 and 17/20 respectively, Fig. 3c, d), further suggesting that the overall coverage is relatively complete. Similarly, many of these genes were also present in other butterfly transcriptomes (Table 1). Interestingly, the 24 000 *Heliconius* EST sequences currently in NCBI's dbEST were insufficient for even rough genome annotation. Only 4 out of the 20 *HmYb* genes were represented in published EST sequences (BLASTn with bitscore cutoff of >500; significant sequence identity for *HM000 06*, *14*, *21* and *22*). Hence, the *H. melpomene* 454 transcriptome libraries we have generated are a powerful tool for annotation of the *Heliconius* genome, and the deep gene coverage obtained will facilitate identification of the developmental pathways expressed during butterfly wing development.

Strikingly, no obvious candidate genes identified from previous studies of butterfly wing patterning are located in the genomic region studied here. The conserved developmental pathways wingless and TGF-β are deployed in pattern specification in other butterflies (Carroll *et al.* 1994; Monteiro *et al.* 2006), with transcription factors such as *engrailed*, *hedgehog* and *distalless* expressed in patterns correlated with wing eyespots (Keys *et al.* 1999; Brunetti *et al.* 2001; Reed & Gilbert 2004), none of which are represented here. We have previously speculated that polymorphism in *Heliconius* might similarly be controlled by transcription factor(s) that integrate spatial information and regulate downstream effector genes such as pigmentation pathways (Joron *et al.*, 2006c). We here identify four genes with mutant wing phenotypes in the fruitfly *D. melanogaster* (*unkempt*, *crooked legs*, *lethal giant larvae* and *penguin*), the first two of which are also transcription factors. One of these, *HM00013/unkempt* (Mohler *et al.* 1992), was located within the candidate region for the two wing patterning loci, and although not recovered from the transcriptome libraries (Table 1), has been shown to be expressed in wings using RT–PCR (Fig. S4, Supporting information). The remaining three genes, *crooked legs*, *lethal giant larvae* and *penguin* (Prout *et al.* 1997; D'Avino & Thummel 2000), are located outside the candidate regions but are all present in the transcriptome libraries. These genes should not be entirely ruled out as candidates, since pattern switching could result from *cis*-regulatory control of these genes from within the candidate regions. Once identified, it will be interesting to determine whether the molecular mechanisms of pattern determination are derived in *Heliconius*, or shared with other butterflies. The recent mapping of the Bigeye mutant in *Bicyclus anynana* to the same chromosome as the *HmYb* locus provides an intriguing hint that the latter may be the case (Beldade *et al.* 2009).

Alternative splicing is an under-studied source of evolutionary and phenotypic diversity, which is likely to have important implications in evolutionary and ecological studies (Marden 2008). Here, we have located nine genes with patterns of alternative splicing including alternate exons and UTRs that might have functional significance in wing development. In one case, *HM00025*, there were differences between the two colour pattern races in the isoforms present, making this gene a good candidate for further investigation. As described previously for the *Melitaea* butterfly, alternative splicing is not well characterized in *de novo* EST assemblies, or by automated gene prediction pipelines (Vera *et al.* 2008). These alternative splicing patterns would not have been discovered without manual alignment of the transcriptome to the genome.

A striking characteristic of wing patterning loci in *H. melpomene* is the tight genetic clustering of genes with related phenotypic effects. Notably, *HmB* and *HmD* loci both regulate red pattern elements and are grouped on linkage group 18, whilst *HmYb*, *HmSb* and *HmN* all regulate yellow elements and are grouped on linkage group 15. Nonetheless, the evidence that these genetically linked elements are indeed multiple loci is based on rare phenotypic recombinants, which could also be explained by developmental aberration or mutation at unlinked loci (Sheppard *et al.* 1985; Mallet 1989). Our molecular genotyping provides the first evidence from linked genetic markers that at least two of these loci are indeed distinct elements, with *HmYb* and *HmSb* separated on the same chromosome by around 150 kb. Furthermore, recent work has also linked mate preference to colour pattern loci in between the sister species, *H. cydno* and *H. pachinus* (Kronforst *et al.* 2006), and in genetic crosses between *H. cydno* and *H. melpomene* (R. Merrill, personal communication). It seems likely therefore that, as in other taxa, multiple traits related to divergence and speciation are tightly linked in the genome (Hawthorne & Via 2001; Noor *et al.* 2001). Our annotation could therefore form the basis for study of many speciation-related traits.

Conclusion

Heliconius colour patterns offer an excellent opportunity to study genes involved in both local adaptation and speciation. One of the most intriguing aspects of the *HmYb* locus is its homology with loci identified in both the phenotypically convergent *Heliconius erato* and the divergent species *Heliconius numata* (Joron *et al.*, 2006c). Here, we have mapped this region in detail and sequenced 11 BAC clones representing 1.2 Mb of contiguous high-quality genomic sequence across the *HmYb* and *HmSb* loci. The molecular identification of the region described here in *Heliconius melpomene* has already formed the basis for parallel analysis in both *H. erato* and *H. numata* by our collaborators (B. Coulterman, M. Joron, in preparation). Furthermore, we have generated extensive sequence surveys of both genomic and transcriptomic sequence, which have facilitated both identification of repeat sequences and genome annotation. The strength of the approach described here is that a combination of high-quality genome sequence alongside high-throughput transcriptomics allows us to generate detailed gene models for all genes expressed during wing development, identify patterns of alternate splicing and characterize noncoding transcripts that would be otherwise difficult to identify. The mapping of the transcriptome to genomic sequence is obviously routine in organisms with a complete genome sequence (Weber *et al.* 2007), but is more rarely applied in nonmodel systems. In some ways, we have only scratched the surface of the potential utility of the transcriptomic data, which can also be used for single nucleotide polymorphism identification, comparative genomic mapping, microarray design and quantitative transcriptomic experiments. This highlights the exciting ecological and evolutionary questions that can now be addressed by taking advantage of rapid developments in sequencing technology (Hudson 2008).

Acknowledgements

This work was funded by a NERC studentship to LF, a BBSRC grant to CJ and RffC, a NERC New Investigator Grant to CJ, a Royal Society URF to CJ, and a BBSRC PhD studentship to NC. We thank all the Sanger sequencing operations staff for library construction, DNA sequencing, loading and sequence improvement, and Ian Korf for the *Bombyx* HMM for gene prediction.

Conflicts of interest

The authors have no conflict of interest to declare and note that the sponsors of the issue had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Baxter SW, Papa R, Chamberlain N *et al.* (2008a) Convergent evolution in the genetic basis of Mullerian mimicry in *Heliconius* butterflies. *Genetics*, **180**, 1567–1577.

Baxter SW, Johnston SE, Jiggins CD (2009) Butterfly speciation and the distribution of gene effect sizes fixed during adaptation. *Heredity*, **102**, 57–65.

Beldade P, Brakefield PM, Long AD (2002) Contribution of distal-less to quantitative variation in butterfly eyespots. *Nature*, **415**, 315–318.

Beldade P, Saenko SV, Pul N, Long AD (2009) A gene-based linkage map for *Bicyclus anynana* butterflies allows for a comprehensive analysis of synteny with the lepidopteran reference genome. *Plos Genetics*, **5** (2) e1000366.

Brunetti CR, Selegue JE, Monteiro A *et al.* (2001) The generation and diversification of butterfly eyespot color patterns. *Current Biology*, **11**, 1578–1585.

Cantarel BL, Korf I, Robb SMC *et al.* (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, **18**, 188–196.

Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, **134**, 25–36.

Carroll SB, Gates J, Keys DN *et al.* (1994) Pattern formation and eyespot determination in butterfly wings. *Science*, **265**, 109–114.

Carver T, Berriman M, Tivey A *et al.* (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, **24**, 2672–2676.

Cheung F, Haas BJ, Goldberg SMD *et al.* (2006) Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. *Bmc Genomics*, **7**, 272.

Chevreux B, Pfisterer T, Drescher B *et al.* (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, **14**, 1147–1159.

Colosimo PF, Peichel CL, Nereng K *et al.* (2004) The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biology*, **2**, E109.

D'Avino PP, Thummel CS (2000) The ecdysone regulatory pathway controls wing morphogenesis and integrin expression during *Drosophila* metamorphosis. *Developmental Biology*, **220**, 211–224.

Ferguson L, Jiggins CD (2009) Shared and divergent expression domains on mimetic *Heliconius* wings. *Evolution & Development*, **11** (5): 498–512.

Gilbert LE (2003) Adaptive novelty through introgression in *Heliconius* wing patterns: evidence for shared genetic 'tool box' from synthetic hybrid zones and a theory of diversification. In: *Ecology and Evolution Taking Flight: Butterflies as Model Systems* (eds Boggs CL, Watt WB, Ehrlich PR). University of Chicago Press, Chicago 281–318.

Hawthorne DJ, Via S (2001) Genetic linkage of ecological specialization and reproductive isolation in pea aphids. *Nature*, **412**, 904–907.

Hoekstra HE (2006) Genetics, development and evolution of adaptive pigmentation in vertebrates. *Heredity*, **97**, 222–234.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**, R143.

Jiggins CD, Mavarez J, Beltran M *et al.* (2005) A genetic linkage map of the mimetic butterfly *Heliconius melpomene*. *Genetics*, **171**, 557–570.

Joron M, Jiggins CD, Papanicolaou A, McMillan WO (2006a) *Heliconius* wing patterns: an evo-devo model for understanding phenotypic diversity. *Heredity*, **97**, 157–167.

Joron M, Papa R, Beltran M *et al.* (2006b) A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biology*, **4**, e303.

Jurka J, Kapitonov VV, Pavlicek A *et al.* (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, **110**, 462–467.

Keys DN, Lewis DL, Selegue JE *et al.* (1999) Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science*, **283**, 532–534.

Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

Kronforst MR, Young LG, Kapan DD *et al.* (2006) Linkage of butterfly mate preference and wing color preference cue at the genomic location of wingless. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 6575–6580.

Larkin MA, Blackshields G, Brown NP *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V (2002) Apollo: a sequence annotation editor. *Genome Biology*, **3**, 1–14.

Li WZ, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Mallet J (1989) The genetics of warning colour in Peruvian hybrid zones of *Heliconius erato* and *H. melpomene*. *Proceedings of the Royal Society of London B*, **236**, 163–185.

Mallet J, Beltran M, Neukirchen W, Linares M (2007) Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *Bmc Evolutionary Biology*, **7**, 28.

Marden JH (2008) Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. *Heredity*, **100**, 111–120.

Margulies M, Egholm M, Altman WE *et al.* (2006) Genome sequencing in microfabricated high-density picolitre reactors (vol 437, pg 376, 2005). *Nature*, **441**, 120–120.

Mohler J, Weiss N, Murli S *et al.* (1992) The embryonically active gene, unkempt, of *Drosophila* encodes a Cys3his finger protein. *Genetics*, **131**, 377–388.

Monteiro A, Glaser G, Stockslager S, Glansdorp N, Ramos D (2006) Comparative insights into questions of lepidopteran wing pattern homology. *BMC Dev Biol*, **6**, 52.

Mundy NI, Badcock NS, Hart T *et al.* (2004) Conserved genetic basis of a quantitative plumage trait involved in mate choice. *Science*, **303**, 1870–1873.

Nadeau NJ, Burke T, Mundy NI (2007) Evolution of an avian pigmentation gene correlates with a measure of sexual selection. *Proceedings of the Royal Society B-Biological Sciences*, **274**, 1807–1813.

Naisbit RE, Jiggins CD, Mallet J (2003) Mimicry: developmental genes that contribute to speciation. *Evol Dev*, **5**, 269–280.

Nijhout HF (1991) *The Development and Evolution of Butterfly Wing Patterns*. Smithsonian Institution Press, Washington.

Noor MAF, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 12084–12088.

Otto TD, Gomes LHF, Alves-Ferreira M, de Miranda AB, Degrave WM (2008) ReRep: computational detection of repetitive sequences in genome survey sequences (GSS). *Bmc Bioinformatics*, **9**, 366.

Papa R, Morrison CM, Walters JR *et al.* (2008) Highly conserved gene order and numerous novel repetitive elements in genomic regions linked to wing pattern variation in *Heliconius* butterflies. *BMC Genomics*, **9**, 345.

Papa R, Martin A, Reed RD (2009) Genomic hotspots of adaptation in butterfly wing pattern evolution. *Current Opinion in Genetics and Development*, **18**, 559–564.

Papanicolaou A, Joron M, Mcmillan WO, Blaxter ML, Jiggins CD (2005) Genomic tools and cDNA derived markers for butterflies. *Molecular Ecology*, **14**, 2883–2897.

Pringle EG, Baxter SW, Webster CL *et al.* (2007) Synteny and chromosome evolution in the lepidoptera: evidence from mapping in Heliconius melpomene. *Genetics*, **177**, 417–426.

Prout M, Damania Z, Soong J, Fristrom D, Fristrom JW (1997) Autosomal mutations affecting adhesion between wing surfaces in *Drosophila melanogaster*. *Genetics*, **146**, 275–285.

Prud'homme B, Gompel N, Rokas A *et al.* (2006) Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, **440**, 1050–1053.

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, **35**, D61–D65.

Reed RD, Gilbert LE (2004) Wing venation and distal-less expression in *Heliconius* butterfly wing pattern development. *Development Genes and Evolution*, **214**, 628–634.

Ren N, Zhu CM, Lee H, Adler PN (2005) Gene expression during drosophila wing morphogenesis and differentiation. *Genetics*, **171**, 625–638.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365–386.

Rutherford K, Parkhill J, Crook J *et al.* (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

Sahara K, Yoshido A, Marec F *et al.* (2007) Conserved synteny of genes between chromosome 15 of Bombyx mori and a chromosome of *Manduca sexta* shown by five-color BAC-FISH. *Genome*, **50**, 1061–1065.

Shapiro MD, Marks ME, Peichel CL *et al.* (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, **428**, 717–723.

Sheppard PM, Turner JRG, Brown KS, Benson WW, Singer MC (1985) Genetics and the evolution of Mullerian mimicry in *Heliconius*. *Philos Trans R Soc Lond B*, **308**, 433–613.

Steiner CC, Weber JN, Hoekstra HE (2007) Adaptive variation in beach mice produced by two interacting pigmentation genes. *PLoS Biology*, **5**, e219.

Turner JR, Sheppard PM (1975) Absence of crossing-over in female butterflies (Heliconius). *Heredity*, **34**, 265–269.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiology*, **144**, 32–42.

Wittkopp PJ, Vaccaro K, Carroll SB (2002) Evolution of yellow gene regulation and pigmentation in *Drosophila*. *Current Biology*, **12**, 1547–1556.

Xia Q, Zhou Z, Lu C *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937–1940.

The authors have a broad interest in the genetic basis of evolutionary adaptation. Laura Ferguson, Siu Fai Lee, Nicola Nadeau and Simon Baxter, Nicola Chamberlain, Richard ffrench-Constant and Chris Jiggins are specifically working on the genetics of wing patterning in Heliconius butterflies. The work was carried out as part of Laura's PhD, and the other group members were postdoctoral researchers. Mathieu Joron is a senior researcher in Paris focusing on the genetics of the polymorphic Heliconius species, H. numata. Paul Wilkinson, Alexie Papanicolaou, Sujai Kumar and Thuan-Jin Kee are interested in developing bioinformatic tools for analysis of evolutionary problems. Richard Clark, Claire Davidson, Rebecca Glithero and Helen Beasley are interested in genomic sequencing and analysis. Heiko Vogel is an expert in preparation of transcriptome libraries, and has a general interest in the genetic aspects of chemical ecology.

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Extension of the BAC tilepath

Table S2 Primers used to map and characterize the *HmYb/Sb* loci

Table S3 BAC clone accession numbers

Fig. S1 Example of BAC extension from fingerprint contig 105.

Fig. S2 Neighbour-joining tree of Trehalase gene orthology in insect species.

Fig. S3 Splice variants identified from the *HM00023* locus in *H. m. malleti* and *cythera*.

Fig. S4 Amplification of *unkempt* transcripts from *Heliconius melpomene* races with different *HmYb* alleles.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.