Molecular Ecology (2005) **14**, 2883–2897 doi: 10.1111/j.1365-294X.2005.02609.x

Genomic tools and cDNA derived markers for butterflies

ALEXIE PAPANICOLAOU,* MATHIEU JORON,* W. OWEN MCMILLAN,† MARK L. BLAXTER* and CHRIS D. JIGGINS*

**Institute of Evolutionary Biology, School of Biological Sciences, Ashworth Laboratories, Kings Buildings, West Mains Road, University of Edinburgh, EH9 3JT, Edinburgh, UK, †Department of Biology, Universidad de Puerto Rico-Río Piedras, PO Box 23360, San Juan 00931, Puerto Rico*

Abstract

The Lepidoptera have long been used as examples in the study of evolution, but some questions remain difficult to resolve due to a lack of molecular genetic data. However, as technology improves, genomic tools are becoming increasingly available to tackle unanswered evolutionary questions. Here we have used expressed sequence tags (ESTs) to develop genetic markers for two Müllerian mimic species, *Heliconius melpomene* and *Heliconius erato*. In total 1363 ESTs were generated, representing 330 gene objects in *H. melpomene* and 431 in *H. erato*. User-friendly bioinformatic tools were used to construct a nonredundant database of these putative genes (available at http://www.heliconius.org), and annotate them with BLAST similarity searches, InterPro matches and Gene Ontology terms. This database will be continually updated with EST sequences for the Papilionideae as they become publicly available, providing a tool for gene finding in the butterflies. Alignments of the *Heliconius* sequences with putative homologues derived from *Bombyx mori* or other public data sets were used to identify conserved PCR priming sites, and develop 55 markers that can be amplified from genomic DNA in both *H. erato* and *H. melpomene*. These markers will be used for comparative linkage mapping in *Heliconius* and will have applications in other phylogenetic and genomic studies in the Lepidoptera.

Keywords: cDNA library, colour pattern evolution, ecological genomics, *Heliconius*, marker development, mimicry

Received 24 February 2005; revision accepted 6 April 2005

Introduction

The Lepidoptera have played a pivotal role in the development of ecological and evolutionary theory (for recent reviews see Boggs *et al*. 2003). Thus mimicry was an early example cited by Bates (1863) in support of Darwin's theory of evolution by natural selection, and Poulton carried out some of the earliest evolutionary experiments investigating butterfly warning colours associated with toxicity (Poulton 1884). Studies of crypsis, especially in the context of industrial melanism in moths (Ford 1931; Clarke & Sheppard 1966), were particularly prominent in the rise of ecological genetics (Kettlewell 1955; Ford 1964) as well as in the modern synthesis (Haldane 1956). Lepidoptera also played a central role in discussions of species and their definition, perhaps due to their high morphological diversity and popularity with collectors (for a historical perspective see Mallet 2004). However, while butterflies may have been the model organisms for early evolutionary biologists, they have clearly lost favour compared to taxa more amenable to genetic manipulation. Few geneticists or developmental biologists would consider any butterfly species to be a model organism. Fortunately however, with the increasing ease with which molecular genetic techniques can be applied to nonmodel organisms, the traditional advantages of the Lepidoptera — ease of collection and study in the wild, great diversity of shape and form, and ease of manipulation during development — may lead to a renaissance as old problems are revisited with new genetic techniques (Beldade & Brakefield 2002).

The Neotropical heliconiine butterflies are a model for studying the interaction of ecology and evolution (Turner 1981). Most have evolved strong unpalatability and participate in Müllerian comimicry groups (Benson 1972; Boyden 1976; Chai 1986). Several such groups, or mimicry

Correspondence: Dr Chris Jiggins, Fax: +44(0)131–6506564; E-mail: chris.jiggins@ed.ac.uk

rings, usually coexist in any one locality, each involving several unrelated butterflies converging in visual appearance. However, mimicry rings also change geographically, and several *Heliconius* species have radiated and diverged into diverse colour pattern races (Turner 1976). The same phenomenon of divergence/convergence is repeated between races of a species. This complex interplay of convergence and diversification reaches dramatic levels with the Müllerian comimics *Heliconius melpomene* and *Heliconius erato*, which exhibit about 30 distinct races and about 15 distinct colour patterns. Each race of one species is sympatric and comimetic with a similar-looking race of the other species. The route by which this diversity arose has been extensively discussed (for references and review see Turner & Mallet 1996; Joron & Mallet 1998), and crosses have shown that the genetic loci underlying pattern changes are few and have major effects (Sheppard *et al*. 1985). Furthermore, a major genetic region responsible for colour pattern switching has been pinpointed on a *H. melpomene* linkage group (Jiggins *et al*. 2005), but actual loci remain to be cloned and identified.

Molecular techniques are now readily applicable to non-model organisms as a result of improvements in technology and reductions in cost (Oleksiak *et al*. 2001; Blaxter 2002; Renn *et al*. 2004). In order to begin to characterize the genome of an organism, one of the first steps is to identify coding regions (Parkinson *et al*. 2004b). Expressed sequenced tags (ESTs) are single pass sequences derived from mRNA, such that only actively transcribed genes in a specific tissue and at a specific developmental stage are sampled. Such EST data sets are invaluable for genome annotation (Bergman *et al*. 2002; Haas *et al*. 2002; Brendel *et al*. 2004) and development of genetic markers across the genome (Landais *et al*. 2003). Furthermore, molecular genetic approaches such as EST projects offer the opportunity to uncover the genetic pathways that underlie classic examples of adaptation, and answer questions regarding their evolutionary history and developmental basis that would be otherwise impossible to resolve.

We have developed ESTs from cDNA libraries for two species of *Heliconius* and here demonstrate the utility of this approach in population genetic and gene mapping applications. We wish to develop a suite of conserved physical markers or anchor loci widely spaced across the *Heliconius* genome. These markers can be used to identify regions of conserved synteny and colinearity that will facilitate comparisons of gene location between comimic species. They will also allow comparison with other more complete lepidopteran genomes such as that of *Bombyx mori* (Kazuei *et al*. 2004; Xia *et al*. 2004). It has been estimated that at least 200 markers may be necessary in order to accurately estimate the number of breakpoints representing chromosomal rearrangements between two species (Schoen 2000), and we demonstrate that ESTs can generate a large number of markers very rapidly.

In the longer term we intend to employ ESTs as part of a candidate gene approach to identify colour pattern control genes. The candidate gene approach has proven successful in other organisms, such as higher plants and cichlid fish (Streelman *et al*. 2003; Gupta & Rustgi 2004). The ESTs generated will provide a wealth of new markers for linkage mapping and phylogenetics in the Lepidoptera and represent the first step towards generating a comprehensive database of genes expressed during *Heliconius* development. We have also created an online database that will continue to be populated with butterfly EST sequences as they become available. This will facilitate the analysis of butterfly genomes and the application of genomic data to evolutionary and ecological studies in the Lepidoptera.

Materials and methods

Library construction and sequencing

A directional cDNA library was generated for *Heliconius melpomene* in the plasmid vector pSPORT1. Tissue of pupae from *Heliconius melpomene rosina* was collected from individuals raised from wild-collected eggs in Panama in February 2004, stored in RNAlater (QIAGEN) and transported to the UK. Wing tissue was dissected out and the remaining tissue used for RNA extraction. Future libraries will be developed from the wing tissue. Total RNA was isolated by grinding the tissue in liquid nitrogen, followed by extraction using TRIZOL (Invitrogen). TRIZOL extracts were phase separated with chloroform and isopropyl alcohol to remove protein and DNA. Messenger RNA was selected using the MicroPoly (A) Purist Kit (Ambion) with a single round of poly A selection. The quality of the extracted mRNA was investigated using agarose gel electrophoresis. The mRNA was reverse transcribed to cDNA, cloned using the Superscript II RT kit (Invitrogen) following the manufacturer's instructions. Briefly, this involved ligation of a *Not*I primer adapter to the 5′ end of the mRNA to facilitate reverse transcription. A *Sal*I adaptor was subsequently ligated to the 5′ end of the cDNA to ensure directional cloning. The cDNA was then size fractionated with ChromaSpin drip columns to avoid ligation and cloning of small fragments, degraded RNA, or unbound adaptor. The selected cDNA was ligated into pSPORT1 vector with T4 ligase. Electromax DH5α *Escherichia coli* cells (Invitrogen) were transformed by electroporation with 25 kV/cm, 100 ohms resistance and 25 μfarads in a 0.1 cm cuvette.

Methods for generation of EST sequences from the library are described in more detail in Whitton *et al*. (2004). In brief, the library was plated out on LB agar plates with Xgal blue/white screening and 0.1% ampicillin, and grown overnight. Colonies were picked and grown overnight in 150 μL of LB medium with 100 mg/mL ampicillin. The

colony stock was then divided into three parts; 20 μL were diluted to 100 μL in distilled water and used for PCR amplification (see below); the remainder was divided into two plates and archived in LB broth with 15% glycerol at −80 °C. All archived clones are freely available for non-profit research use upon request.

The diluted cultures were boiled at 100 °C for 5 min to lyse the cells, and the cDNA inserts were amplified using modified M13 primers (M13-forward: CGCCAGGGTTT-TCCCAGTCACGAC and M13-reverse: GGAAACAGCT-ATGACCATG) using a 20-μL total volume PCR mix containing 0.2 mM dNTP mix, 0.2 μL of each of the forward and reverse primers, 2 μL of DNA template, 1× *Taq* buffer and 0.4 unit of QIAGEN *Taq* polymerase. The polymerase chain reaction (PCR) cycling profile was 94 °C for 3 min; 35 cycles of 94 °C for 15 s, 55 °C for 20 s, 72 °C for 3 min; a final cycle of 72 °C for 10 min. Insert size was estimated on agarose gels and positively amplified inserts were then 5′- end sequenced using sequencing reactions containing 2 μL BigDye, 1 μL T7 primer, 2 μL of 5× sequencing buffer, 3 μL of template DNA in a total volume of 10 μL, and visualized on an ABI 3730 sequencer by the SBS Sequencing Service, University of Edinburgh.

A similar directional cDNA library was constructed for *Heliconius erato* in the vector lambda UniZAP using tissue derived from stocks housed at the University of Puerto Rico. First, 300 μg of total RNA from wing-disc tissue samples was collected at the following developmental stages: 5th instar, early prepupal, late prepupal, 24 h after pupation, precolor pupal stages (i.e. pupae > 48 h after pupation, but before the development of wing pigmentation), midpupal stages (i.e. at the beginning of the development of red pigmentation), and 10 days after pupation. All pooled wing-disc samples contained tissue from three geographical races of *H. erato*, *Heliconius erato petiverana*, *Heliconius erato erato*, and *Heliconius erato cyrbia*, with differences in wing colouration and *Heliconius himera*, a closely related sister species (Jiggins *et al*. 1996). Complementary DNA from total RNA was cloned into lambda UniZAP (Stratagene). Sequences were obtained from this library as outlined in Oleksiak *et al*. (2001). Briefly, recombinant phagemids were excised by infecting XLI-Blue MRF′ cells with our phage library in the presence of ExAssist helper phage. These phagemids were then used to transform SOLR cells, which were grown overnight on LB-agar plates containing ampicillin following the manufacturer's instructions. Two millilitre of SB broth containing 150 μg/mL of ampicillin were innoculated with individual recombinant bacterial colonies and incubated with shaking overnight at 37 °C. One microlitre of this bacterial growth was used as template for PCR amplification of cDNA inserts using flanking PucF and PucR primers (see Oleksiak *et al*. 2001). PCR products were cleaned on Sephadex G-50 columns and sequenced using Amersham DYEnamic ET Terminators on a MegaBACE 1000 automated DNA sequencer in the UPR-Rio Piedras Sequencing and Genotyping Facility.

Processing of EST sequences

Publicly available bioinformatic tools developed in Edinburgh (available from http://www.nematodes.org/) were used to process the EST sequences. EST sequences were trimmed of vector and adaptor sequences and poly(A) tails, along with any low quality bases, using trace2dbEST (Parkinson *et al*. 2004a). All ESTs were annotated with similarity information and submitted to the NCBI dbEST database as soon as they were processed.

A nonredundant set of putative gene objects (clusters) was generated using PARTIGENE (Parkinson *et al*. 2004a). PARTIGENE is a Perl pipeline program that uses CLOBB (Parkinson *et al*. 2002), PHRED and PHRAP (Ewing & Green 1998) to assign EST sequences to clusters based on identity. These clusters serve to increase the sequence length for each putative gene and minimize the effect of sequencing errors. The cluster names are prefixed with HEC for *H. erato* and HMC for *H. melpomene*. Each cluster name (e.g. HEC00001) is stable (i.e. will be maintained as additional ESTs are added to the databases) and thus constitutes a unique identifier for the gene it represents. Consensus sequences were then stored in a POSTGRESQL relational database and annotated with sequence similarity information based on BLAST (Altschul *et al*. 1997) searches against a set of custom databases derived from GenBank (November 2004). Putative protein translations were derived using PROT4EST (Wasmuth & Blaxter 2004). Due to the lack of full-length coding sequence data (CDS) for butterflies, the hidden Markov models and codon usage tables required for optimal translation could not be accurately trained. We therefore used codon usage statistics derived from open reading frames from a subset of our own data set that had, according to similarity searches, a full-length open reading frame.

The protein translations were also stored in the postgreSQL database and annotated with InterPro domain matches and Gene Ontology (GO) terms. InterPro searches were performed on the European Bioinformatics Institute web server (http://www.ebi.ac.uk/InterProScan). GO terms are keywords that classify the objects according to recognizable protein domains (The Gene Ontology Consortium, 2000). GO molecular function terms (a subset of higher-level terms called 'GO-slim') terms were assigned by comparing each sequence with the GO function ontology data set available from http://www.geneontology.org. The database can be accessed via a user–friendly web query interface, which uses the searching capabilities of relational databasing and is available at http://www.heliconius.org. The database also includes clustered and annotated data sets for other butterfly species.

Marker development

ESTs of interest were selected for mapping. We aimed to identify conserved genes such that primers would be likely to work over a diversity of lepidopteran species. To determine the taxonomic level across which genes were conserved, BLAST similarity searches were conducted for each cluster consensus sequence against a hierarchical set of databases (using BLASTN for nucleotide databases and BLASTX for protein databases). Matches with a bit score of 80 or greater were considered significant. Bit scores are a conversion of the BLAST search raw scores that are preferable to *E* values because they are not dependent on the size and composition of the database searched. First, we excluded any clusters that were similar to ribosomal RNA genes. Any clusters that showed significant similarity to archaeal or bacterial sequences were classified as highly conserved. Any clusters that lacked similarity to prokaryotic genes were then compared to a database of eukaryotic proteins that excluded arthropod sequences to identify conserved eukaryotic genes. Clusters that remained unannotated were searched against databases comprising (i) arthropodan sequences that contained no hexapod sequences, (ii) hexapod sequences that contained no lepidopteran sequences, (iii) lepidopteran sequences excluding *Heliconius* sp. sequences and (iv) *Heliconius* species sequences only. Thus, we classified our sequences according to the most inclusive taxon in which a known homologue could be identified.

Clusters showing conservation across Eukaryota or in all cellular life were selected for marker development. In each case the *H. melpomene* and/or *H. erato* gene was aligned with a homolog from *Manduca sexta*, *Bombyx mori*, or in some cases other endopterygotes such as *Drosophila melanogaster* or *Anopheles gambiae*, to identify conserved regions. Primers were designed to amplify short (100–300 bp) fragments of coding DNA. Such regions commonly contained introns and therefore larger fragments were amplified from genomic DNA. Amplification was tested using *H. melpomene* genomic DNA using a simple PCR program with no optimization: 94 °C for 2 min; 30 cycles of 94 °C for 20 s, 55 °C for 40 s, 72 °C for 60 s; and finally 72 °C for 10 min. The identity of the PCR products was confirmed by sequencing. Those primer pairs that produced a consistent product in *H. melpomene* were also tested using *H. numata* and *H. erato* genomic DNA under the same conditions.

Chromosomal assignment of markers

Lepidoptera have achiasmatic oogenesis, and thus there is no crossing over in females (Turner & Sheppard 1975). Hence, loci on the same chromosome are inherited in complete linkage from the female parent. This can be used to determine the segregation pattern of female informative markers in a mapping family (called a 'chromosome print', see Yasukochi 1998; Jiggins *et al*. 2005) and assign markers to linkage groups with relative ease. A single F2 mapping family of 46 individuals derived from parental stocks of *H. melpomene malleti* (eastern Ecuador) and *H. melpomene melpomene* (French Guiana) was chosen for the present study. Crosses were carried out between distant populations to maximize the amount of segregating variation. Microsatellite and other genetic markers previously identified were used to determine the chromosome print for 16 of the 21 chromosomes (Jiggins *et al*. 2005). In order to detect segregation of maternal alleles for our markers we first searched for intron size variation by running amplified products on 1.8–2.2% agarose gels stained with ethidium bromide, at 1.5–2.5 V/cm for 12–18 h. If no size variation was identified, amplified products were then digested using one of a set of six restriction enxymes (with 4- or 5-base recognition sites) to identify RFLPs that were subsequently scored on agarose gels.

Analysis of genomic and molecular evolution

Codon usage tables were constructed for the two species using data derived from translations of the ESTs using custom Perl programs. Where orthologous genes were available from the *H. melpomene*, *H. erato* and *B. mori* EST data sets, the consensus sequences were aligned using CLUSTALW (Higgins *et al*. 1994) with manual corrections, guided by BLASTX alignments to identify frame shifts. The synonymous site substitution rate (*Ks*) was calculated for fourfold degenerate sites for each pair of sequences using DNASP (Nei & Gojobori 1986; Rozas *et al*. 2003). The alignments and condon usage tables are available from www.heliconius.org. In addition, sequences of the mitochondrial cytochrome oxidase I (*Co1*) and cytochrome oxidase II (*Co2*) genes were downloaded from GenBank for comparison (Accession nos AF413674; AF413685; AF295564) and Ks similarly calculated. Coding GC content was calculated from all available clusters, excluding those matching rRNA genes.

Results

Library construction and sequencing

cDNA libraries were successfully constructed and used to derive ESTs from *Heliconius erato* and *Heliconius melpomene* (Table 1). Clustering of the ESTs into putative genes yielded 431 *H. erato* clusters and 330 from *H. melpomene*. The high percentage of singletons in both libraries (*c*. 70–80%) suggests that sequencing efforts are a long way from reaching saturation for these libraries. Peptides were predicted from the cluster consensus sequences, and these high-quality translations used for database searches and functional annotation. For the *H. melpomene* library, 68% of

Table 1 Summary of the *Heliconius* EST data sets

Species	*Heliconius erato*	*Heliconius melpomene*
Source of tissue	wing discs from pupae and larvae	pupal body tissue minus wing discs
Number of ESTs	775	588
Accession numbers	CO729474–CO729985; CV126116, CO377782–CO377790	CV525684–CV526458
Average sequence length (bp)	450	500
%GC	44.5%	43.9%
Number of clusters (putative gene objects)	431	330
Number of clusters with one EST (singletons)	338	231
Proportion of clusters with significant BLAST similarity to known proteins*	61%	54.5%
Number of clusters deriving from the mitochondrial genome	11	12

*BLASTP vs. SWISSPROT, November 2004, with a raw score cut-off of 80 bits.

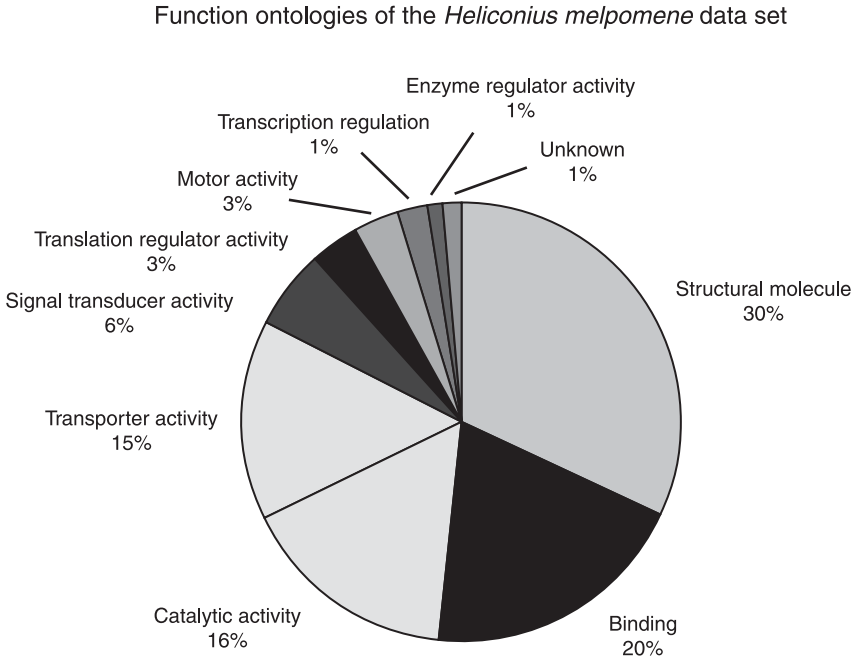

Fig. 1 Assignment of Gene Ontology molecular function terms to the *Heliconius melpomene* gene set. In total 28.5% of the clusters were annotated. Note that one cluster can be associated with more than one GO term.

the clusters were putatively annotated using similarity information using BLAST (Altschul *et al*. 1997) with a cut-off score of 80 bits. Decreasing the cut-off score to 60 bits did slightly increase the number of sequences with annotation, but increased the chance of misannotation. Putative annotation was similarly obtained for 61% of the clusters from the *H. erato* library.

In total 28% of the predicted proteins were annotated with Gene Ontology (GO) molecular function terms (Fig. 1). Thirty-two per cent of annotated proteins were characterized as structural molecules (e.g. ribosomal proteins), 20% showed DNA or protein-binding function and 16% had catalytic activity. In addition, 6% of our data set had signal transduction activity. This pattern is consistent with that seen in other EST data sets, where most of the annotated genes have housekeeping functions (Landais *et al*. 2003; Mita *et al*. 2003; Fei *et al*. 2004). Determination of housekeeping function will become more accurate once the number of cDNA libraries increases to cover more distinct developmental stages and tissues. As GO and other public databases are populated with additional data, we will be able to provide annotation for a higher fraction of the data set.

Marker development

Of the annotated *H. melpomene* clusters with similarity to other proteins, our searches showed that 22.3% of these had significant similarity to proteins from Archeabacteria and Eubacteria and 48% to Eukaryota excluding Arthropoda. Only 2% had similarity only to nonhexapodan

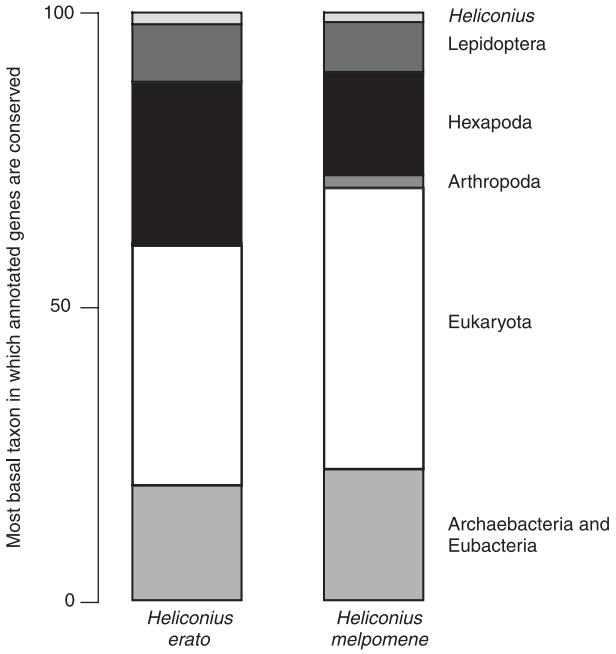

Fig. 2 Breakdown of the degree of conservation of *Heliconius melpomene* and *Heliconius erato* clusters. Each cluster was characterized according to the most inclusive taxon in which a significant match was found. Markers were developed from clusters with similarity to loci conserved in all cellular life (Archaebacteria, Eubacteria and Eukaryota) or all Eukaryota.

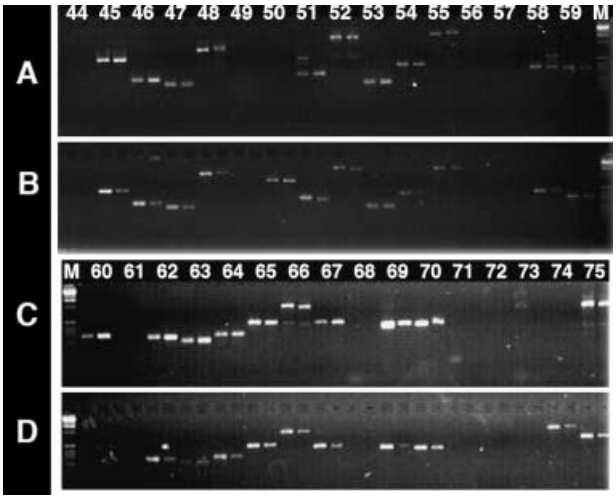

Fig. 3 PCR amplification from genomic DNA of markers developed from ESTs. We tested *Heliconius erato* (A, C), and *Heliconius melpomene* (B, D). Size differences are likely due to indel formation in intron sequences. Numbers labelling each lane (locus amplified) relate to the primer code in Table 2.

arthropods. About 17.6% were similar to hexapod non-lepidopteran proteins and 8.5% were similar only to other lepidopteran proteins. Similar results were found for the *H. erato* data set (Fig. 2).

To develop markers from the EST sequences, primers were designed for 82 genes that were highly conserved in Eukaryota, particularly ribosomal proteins. Without any PCR optimization, 62 of these primer pairs produced successful amplification of a single fragment from *H. melpomene* genomic DNA, of which 48 had fragment sizes significantly larger than expected from the EST sequence, implying the presence of one or more introns within the fragment. Of these 82 primer pairs, 60 amplified a single clear band from *H. erato* genomic DNA (for example see Fig. 3), and 55 worked well in both species. Results are also shown for *Heliconius numata* (Table 2). Sequences were obtained for 20 of these successfully amplified genes and their identity confirmed by BLAST (Table 2).

Chromosomal assignments

Segregating variation was identified for 13 of the 35 genes tested. Of these, eight were unambiguously assigned to previously identified linkage groups using the method of forbidden recombinants, based on linkage with microsatellite markers already mapped in *H. melpomene* (Jiggins *et al*. 2005). Seven of these markers are ribosomal proteins that also amplify from *H. erato*. Ribosomal protein L15 (*RpL15*) maps to linkage group three (LG3), *RpS14* to LG6, *RpL13* and *RpS11* to LG10, *RpL30* to LG11 and *RpL7* and *RpS12* to LG12. The remaining marker, O-glycosyltransferase gene (*Ogt*) mapped to linkage group 18. This linkage group also contains the gene *Cubitus interruptus*, which is linked to a major colour pattern locus in *H. erato* (Tobler *et al*. 2005). The remaining markers showing segregating variation but could not be assigned to linkage groups, and are likely located on the five linkage groups that do not yet have gene or microsatellite markers.

Synonymous divergence between Heliconius *species*

The nucleotide synonymous site divergence between *H. erato* and *H. melpomene* was calculated for 20 nuclear and two mitochondrial genes (Table 3). The mean corrected (Jukes–Cantor model) *Ks* is 0.164 (median 0.171) for nuclear genes and 0.340 for mitochondrial *Co1* and *Co2*. The corrected *Ks* values between *B. mori* and *H. melpomene* were 0.403 (*Co1* and *Co2*) and 1.661 (nuclear), and between *B. mori* and *H. erato* were 0.386 (*Co1* and *Co2*) and 1.632 (nuclear). GC content for *H. melpomene* was 43.9% and for *H. erato* was 44.5%.

Discussion

From 1363 EST sequences, 431 and 330 putative genes have been identified for *Heliconius erato* and *Heliconius melpomene*, respectively, two Müllerian comimic butterfly species. These

Table 2 Markers developed from EST sequences for *Heliconius* species. Primer sequences are shown for each locus, and the estimated product size based on the EST sequence. Product size is shown for amplifications from genomic DNA of *Heliconius melpomene*, *Heliconius numata* and *Heliconius erato*. A significant discrepancy in size compared to that expected indicates the presence of introns in the amplified fragment. Eighty-two primers were designed in total, but only primers amplifying successfully at least one species are included. 'Align' shows the taxonomic group to which the sequence was aligned in order to design primers, where O is Obtomectera, E is Endopterygota, H is *Heliconius*, X is *Xenopus* and G is *Gallus*. Abbreviations are the gene abbreviation from FLYBASE for the top hit in the *Drosophila* genome for each marker

Code on gels	Putative locus name	Designed with cluster	Align	Abbrev.	Primer pair sequences (5′–3′)	Size in kb			
						Expected size	*H. melpomene*	*H. numata*	*H. erato*
1&	Ribosomal protein L13	HMC00007	O	*RpL13*	F: CAATCAGCCTGCTCGGAAACACC R: TTCCTCCTCGGACGCTTCACC	0.306	1	1*	0.7/1.0
2&	Ribosomal protein L14	HMC00255	O	*RpL14*	F: CCTGGTTGCTGAAGGTTCC R: CCGCTTTAACTTTCAGACTCTTG	0.324	0.35	0.35	0.35
3&	Ribosomal protein L15	HMC00048	O	*RpL15*	F: CGATGTTATGCGCTTCCTT R: AATCTTGTGCGACCCAGTA	0.289	0.7	0.8	0.5
5	Ribosomal protein L22	HMC00223	O	*RpL22*	F: AAGTTCGCCATTGACTGCA R: CCTCATTGTCGCTTTCAGC	0.276	1/0.9	1.2	0.8
6&	Ribosomal protein L23a	HEC00323	H	*RpL23a*	F: AAACACCTGCACCTGCTCC R: GCCCTTTAATGCCTGTCTTCT	0.255	0.25	0.3	0.25
7&	Ribosomal protein L24	HMC00066	O	*RpL24*	F: AGAAAGGGCAGGAAGAAGA R: CTGCCTTTTGGCTTACTTT	0.228	*	1.6	0.6
8&	Ribosomal protein L30	HMC00257	O	*RpL30*	F: TTGGCTCTAGTAATGAAATCTG R: CTCAGGCAAGGTCGTAATG	0.239	0.6	0.6	n/a
9&	Ribosomal protein L32	HMC00074	O	*RpL32*	F: AGACCGCAAATCGTCAAAA R: CGGCGTTGGTCACCCTGAT	0.315	2	2	n/a
10&	Ribosomal protein L35	HMC00099	O	*RpL35*	F: AAAATGGGGAAGGTCAAGT R: GATCTCCTTTTGGGTCTTG	0.326	1.6	1.6	0.8
12&	Ribosomal protein L38	HMC00180	O	*RpL38*	F: AAGAAGAACCCTGAGAATG R: TCTTTAACCTGGAGACCTG	0.091	1.6	1.4	1.2
14&	Ribosomal protein L44	HMC00006	O	*RpL44*	F: AAAATGGTGAACGTACCAAAAC R: GCCCAATTCAAAATGCTTG	0.239	0.6	0.6	0.6
15&	Ribosomal protein L7	HMC00059	O	*RpL7*	F: TCCGCAAAGTGAAGGGACAAC R: TCCTCACGGTTACCAAAGTCA	0.207	0.7	0.6	0.65
16	Ribosomal protein L8	HMC00346	O	*RpL8*	F: TCTTTAGATTATGCTGAACG R: GTCTTGAACTTGTATGGGT	0.084	0.2	0.2	0.2
17&	Ribosomal protein L9	HMC00311	H	*RpL9*	F: CCCAGAGGGATTAACTGTTCA R: CCTGGTGCCATCTTTACTC	0.312	0.6	0.6	0.6
18&	Ribosomal protein P2	HMC00202	O	*RpP2*	F: GTTACGTGGCCGCGTATTT R: CTCCGACTCTTCCTTCTTGG	0.265	1.2	1.2	n/a
19&	Ribosomal protein SA/P40	HEC00108	P	*RpP40*	F: CAAAGAGCCGTTCTCAAGT R: TAGACGCAGAACTTCACG	0.270	1	1	1
20&	Ribosomal protein S10	HMC00201	O	*RpS10*	F: CAAAACCGTGTGTCTATCTATGA R: GAGTGTTGCAGGCACAATTTC	0.218	0.65	0.65	0.65
21&	Ribosomal protein S11	HMC00299	O	*RpS11*	F: CTTATATTGACAAGAAGTGCC R: TCTCCTTGTTTCACGTCCC	0.173	1	n/a	n/a

Table 2 *Continued*

Code on gels	Putative locus name	Designed with cluster	Align	Abbrev.	Primer pair sequences (5′–3′)	Size in kb			
						Expected size	*H. melpomene*	*H. numata*	*H. erato*
22	Ribosomal protein S12	HMC00254	O	*RpS12*	F: AACCCCGTCTTGAGCGG R: TGGCCTTGCCATCTTTGTC	0.257	2	n/a	1.2
23	Ribosomal protein S14	HMC00250	O	*RpS14*	F: TCTCCTTACGCCGCTATGTT R: TCGTCTACCACCCTTCCTG	0.196	0.6	n/a	n/a
25&	Ribosomal protein S17	HMC00263	O	*RpS17*	F: GCGAAAATCATTATTGAGAAG R: GTCTGCCACCATAACCTCC	0.321	1	1	1
27&	Ribosomal protein S6	HMC00173	O	*RpS6*	F: CGGGATGTCAGAAGTTATT R: GGGCGAGTACAGACAAGTTAG	0.264	0.5	0.5	0.4
29&	COP-9	HMC00210	E	*CSN8*	F: GAAAAACAAGAACTCGAGG R: ATTTTGAGGTATTCTTTTCCA	0.092	1	0.9	0.8
30&	Calcium ATPase	HMC00272	O	*Ca-P60A*	F: TTCACGGAGGACGAAGACAC R: CGGCAGTACCAGAACCCATA	0.214	1.7	2	1.7
31	Glutathione S-transferase	HEC00329/ HMC00275	H	*CG6781*	F: TAGTGAACAAGTATGGCAAAGG R: CGAGGAAGGTATCCAGGAGT	0.144	0.5	0.65	0.9
32	O-glycosyltransferase	HEC00418	E	*Ogt*	F: TGATAGCGTCGGGGCAAGTG R: GGGTTCACCAACAGCAGGGA	0.248	0.25	0.7	0.6
33	SUI1/eIF1	HMC00176	E	*CG17737*	F: GACCCATTCGCCGATGCTAT R: GATTTGGTGAGCCACTGGCAA	0.223	0.25	0.25	0.25
34	Signal peptide subunit	HMC00238	E	*CG5677*	F: CAACTGGAATGTAAAACAG R: TCCATCATCCCAGAAGTA	0.130	0.15	0.15	0.15
35	Translationally controlled tumour protein	HMC00167	O	*CG4800*	F: TCGGTGACAAGAAATCCTTCA R: ATATCCCTGTATTCCAACAT	0.091	1.6	1.6	1.6
39	Eukaryotic translation initiation factor 4E	HMC00052	O	*CG10124*	F: TACCCTCTGAGCTTCGTCAA R: AGTGGCAGAACTGTGCTTGA	0.345	1.2	n/a	n/a
40	Eukaryotic initiation factor 4A	HEC00026	O	*eIF-4a*	F: GCGCAAGCTCAGTCAGGAAC R: GGAGGGCACGACGGGTAATC	0.254	0.35	0.35	0.35
41	Eukaryotic initiation factor 3B	HEC00036	O	*eIF3-S9*	F: CAGAAGGATTCGTGGATGAT R: TCCAAGTAAATAAAGCCAGT	0.206	0.5	0.45	0.5
42	SSR3-signal sequence receptor	HEC00346	O	*CG5885*	F: CCGAAACACGAAATTCCAACT R: AGAGCCAGAAGACCAGATGC	0.261	1	0.9	1
43	Ribosomal protein L10	HMC00413	O	*RpL10*	F: AGCGACCGTTGACGACTTTCC R: GAGCCTCGATGACCTGTGCC	0.296	1	1*	1
45	Ribosomal protein L21	HMC00248	O	*RpL21*	F: GGCGACATCGTTGACATTAG R: GGGATGGGTGCAAGTAAGAC	0.315	0.5	0.6	0.7
46	Ribosomal protein L28	HMC00407	O	*RpL28*	F: GCCCAACAATGTGACTAACC R: ATCGCTGATGCACGACGAAG	0.229	0.25	0.25	0.25
47	Ribosomal protein L34	HMC00356	O	*RpL34*	F: GTTAGAACACCAGGTGGACG R: TGAGGAAAGCTCTGACAATG	0.179	0.2	0.2	0.2

Table 2 *Continued*

Code on gels	Putative locus name	Designed with cluster	Align	Abbrev.	Primer pair sequences (5′–3′)	Size in kb			
						Expected size	*H. melpomene*	*H. numata*	*H. erato*
48	Ribosomal protein L35	HMC00545	O	*RpL35*	F: TGCTAAGGTTACAGGAGGAG R: AGTCAAGAGGCTTGTATTTC	0.111	1.2	1.4	1.2
51	Ribosomal protein L3	HMC00018	O	*RpL3*	F: TGGGAGGTTTCCCTCATTAT R: CCTTATCAGCAGGTGTTTGG	0.169	0.4	0.3	0.6/0.4
52	Ribosomal protein S15A	HMC00443	O	*RpS15A*	F: ATGGTGCGTATGAACGTATT R: AATACCCCCACTGGTTGTAAG	0.288	1.6	1.6	1.6
53	Ribosomal protein S27	HMC00531	O	*RpS27*	F: CGCAGTTGACTTATTGCACCC R: TTAACCTGGCACGACCACCA	0.173	0.2	0.2	0.2
54	Ribosomal protein L27a	HMC00280	O	*RpL27a*	F: ACATGGACAAGTACCACCCT R: TAGTAACCCGCCTTAACAAT	0.158	0.5	0.4	0.5
55	Ribosomal protein L6	HMC00417	O	*RpL6*	F: TTGGTGGTGAGAAGAATGGAGG R: GACGCAGTGGGCATGAGTTG	0.230	1.6	1.6	1.6
57	Ribosomal protein S19	HMC00421	O	*RpS19*	F: CTGATGGCTCCGCAGACCAC R: CGTCTGCCTTGGGTGGTGAG	0.196	n/a	0.7	0.6
58	Histidine triad protein member 5	HEC00057	E	*CG2091*	F: TATCCAGCCACAGACAAACA R: TCTTTTGTAAGTCCGTCCCA	0.207	0.5	0.6/0.4	0.6/0.4
59	Protein disulphide isomerase	HEC00025	O	*ERp60*	F: TTTCATCTTCGGACTTCGTGTTAG R: AGGATTCGGCGAGTTCTTGT	0.127	0.35	0.35	0.4
60	GTP-binding nuclear protein RAN	HEC00063	E	*ran*	F: GAAACGATACGTCGCTACAC R: TGTCCTTGATGTCCACTTTG	0.240	n/a	n/a	0.3
61	Hyphantrin — Apolipoprotein D	HEC00068	G	*GLaz*	F: GGATTGAAACGTGTAATGGA R: CGTTCTAGCACAGTGTAGGC	0.257	0.4	n/a	0.4
62	Catalase 2	HEC00074	O	*CG9314*	F: TCAAAGAATTGTGCACGCTA R: AAGAAAACTGGCAAATGGTT	0.215	0.25	0.25	0.25
63	Poly A binding protein, cytoplasmic	HEC00085	O	*pAbp*	F: AGGATCATGTGGTCCCAACG R: ATTTGCAGCTTCTTCTGTTT	0.175	0.2	0.2	0.2
64	Kisir larval gene	HEC00099	O	*kisir*	F: TTGGACATCAGAACATACTT R: ACTTACTGCTTGTGCCCATC	0.159	0.35	0.3	0.3
65	Calreticulin	HEC00123	O	*Crc*	F: ATGGGAATCCAAATGGGTGT R: TTCGTGCTTGACTGAGAACTG	0.169	0.5	0.5	0.5
66	Proliferating cell nuclear antigen (PCNA)	HEC00144	O	*mus209*	F: TTACGCAGCTCCATCTTGAA R: GACGGTATCTGCATTGTCTT	0.241	1	0.9	1
67	Drosophila CG15081; Human repressor of oestrogen	HEC00166	E	*CG15081*	F: TGGCCCAAAGTAAATTAAATGAT R: TTGGCTACAACCGATTTCAACAC	0.398	0.45	0.45	0.45
69	Muscular protein 20	HMC00060	E	*Mp20*	F: CCAATCTACGGATTTGTGGG R: CTGGGACGCCATCTTGTTCG	0.179	0.5	0.45	0.4
70	Ribosomal protein L17	HMC00061	O	*RpL17*	F: AAAATGGGTCGTTATTCTCG R: GCCTTTGTAGTCAGCGTTAG	0.283	0.4	0.4	0.4

Table 2 *Continued*

Code on gels	Putative locus name	Designed with cluster	Align	Abbrev.	Primer pair sequences (5′–3′)	Size in kb			
						Expected size	*H. melpomene*	*H. numata*	*H. erato*
71	Calcyclin-binding protein	HMC00102	E	*CG3226*	F: AAATGGATATGGTTGGGATCA R: GCAAGAAAGATAACTACCATGTC	0.199	n/a	0.2	n/a
72	Seven-up alpha	HMC00164	O	*svp*	F: TCGTTTGCGGTGATAAGTCC R: CTTCCCTCCTCATGCCCATT	0.166	0.15	n/a	0.15
73	Bent (LD10678p)	HMC00216	E	*bt*	F: AGAAGCCGCAGTTGCGTCAA R: CAGCCTCGATACGGCACTCG	0.358	n/a	1.6	0.7
74	Prophenol oxidase activating enzyme precursor (PPAE)	HMC00218	O	*CG3066*	F: GACAAACTCTAAGCGGAAAG R: CATCTTCAGCGAATACTCCA	0.112	1.2	1.2	n/a
75	Eukaryotic translation elongation factor 1 gamma	HMC00436	O	*Ef1γ*	F: ATGTAGAGCGTGCGAAGTCAGA R: TGGTTGGCAAGGGTCAGGAA	0.171	0.7	0.7	0.7
76	Psma3 — Proteasome subunit alpha type 3	HMC00235	X	*Prosα7*	F: CAAAACAGAAATTGAAAAGCT R: TTCGCTTGGTTCTCTGCTTC	0.166	0.5	0.5	2
77	Karyopherin alpha-3	HMC00318	E	*Kap-α3*	F: GCTGGACTCTTACCCAAAAT R: ATTAATAACCTGCGTATCCT	0.151	0.2	0.2	0.2
78	Peptidylprolyl isomerase protein 3	HMC00319	G	*CG11777*	F: AAGGGATTCATAGTACAAACA R: CCATCAATTATTCTTCCAAA	0.194	n/a	n/a	0.5
79	Chickadee	HMC00425	O	*chic*	F: AAGATAGTGGCTGGTTTCG R: GTTACTTATTTCCCTGTGGA	0.279	1	n/a	1
80	T-complex protein 1, delta subunit	HMC00440	O	*CG5525*	F: CTATCCTAAAACAAATGAGTG R: GACATGCCTTCAACAACCTG	0.191	0.7	n/a	0.7
81	Eukaryotic translation elongation factor 2	HMC00462	O	*Ef2a*	F: ATTCCCACGACGAGAAGATG R: TGTACTACAGCGAAGGGTTTG	0.305	1.2	1	1.2
82	Mitotic checkpoint control protein (Bub3) gene	HMC00474	O	*Bub3*	F: ATGAGCATAGTAGGCGAAAAG R: TACCTATGGAACTGACAAAG	0.440	2	1.2	1

n/a signifies instances where there was no amplification.
*marks more than one band visualized on a gel. Where a number is provided as well (e.g. RpL13), then the secondary band(s) were significantly fainter than the main band.
/separates size fragments that had equal intensity when visualized on agarose, suggesting two alleles with very large intron size variation, or a multiple copy gene.
& next to the code number, signifies that the *Heliconius melpomene* PCR product derived with these markers has been sequenced and its identity verified.

Table 3 Silent substitution rate for nuclear (a) and mitochondrial (b) genes between *Heliconius melpomene* (HMC), *Heliconius erato* (HEC) and *Bombyx mori* (Bombyx). Both absolute (*Ks*) and Jukes–Cantor corrected (JC) values of *Ks* are shown, with the sample size in number of synonymous sites (bp)

(a)	HMC-HEC			HMC-Bombyx			HEC-Bombyx		
Gene	*Ks*	JC	bp	*Ks*	JC	bp	*Ks*	JC	bp
EF1a	0.180	0.205	122.5	Not calculated			Not calculated		
EF1g	0.235	0.281	89.5	0.645	1.475	141.8	0.652	1.524	85.2
P0	0.194	0.224	103.3	0.691	1.912	103.4	0.615	1.286	109.8
RpL10a	0.090	0.096	100.0	0.604	1.227	99.3	0.615	1.286	139.8
RpL11	0.150	0.167	46.7	0.720	2.414	56.3	0.713	2.257	111.5
RpL13a	0.155	0.174	90.2	0.692	1.920	122.8	0.629	1.369	116.8
RpL17	0.160	0.180	112.7	0.700	2.038	129.9	0.642	1.456	112.1
RpL19	0.179	0.205	111.5	0.514	0.866	111.0	0.502	0.830	117.5
RpL27a	0.109	0.118	100.8	0.728	2.659	100.9	0.700	2.024	101.5
RpL4	0.156	0.175	96.0	0.646	1.481	96.5	0.613	1.275	148.2
RpL5	0.112	0.121	98.6	0.802	N/A	101.3	0.802	N/A	101.3
RpL8	0.107	0.116	65.3	0.604	1.227	64.6	0.587	1.145	138.0
RpL9	0.074	0.077	108.8	0.661	1.597	131.7	0.736	2.966	108.8
RpS14	0.087	0.092	115.4	0.666	1.644	115.6	0.649	1.506	115.5
RpS23	0.078	0.083	102.3	0.557	1.016	102.4	0.508	0.848	102.4
RpS3a	0.238	0.286	122.1	0.721	2.448	122.0	0.707	2.140	133.0
RpS4	0.265	0.326	98.3	0.795	N/A	96.9	0.073	2.770	160.7
RpS25	0.062	0.064	81.0	0.609	1.251	82.2	0.645	1.475	82.2
RpS8	0.070	0.074	85.3	0.762	N/A	107.6	0.702	2.056	130.2
RpS9	0.194	0.225	118.4	0.634	1.401	119.8	0.590	1.161	140.6
Sum			1968.6			2006.0			2254.8
Mean	0.145	0.164	98.4	0.671	1.661	105.6	0.615	1.632	118.7
Median	0.153	0.171	100.4	0.666	1.539	103.4	0.642	1.465	115.5
Standard Deviation	0.061	0.078	18.9	0.076	0.528	21.6	0.150	0.608	20.9

(b)	HMC-HEC			HMC-Bombyx			HEC-Bombyx		
Gene	Ks	JC	bp	Ks	JC	bp	Ks	JC	bp
Co1	0.396	0.562	182	0.431	0.642	183	0.418	0.612	183
Co2	0.347	0.466	144	0.405	0.583	78.3	0.371	0.512	78.8
Sum			326			261.3			261.8
Mean	0.372	0.514		0.418	0.613		0.395	0.562	
Standard Deviation	0.035	0.068		0.018	0.042		0.033	0.071	

sequences are now available on a public online database that has been developed specifically for the analysis of butterfly EST data. Furthermore, a large percentage of these genes are conserved across a broad taxonomic scale and can be amplified from genomic DNA. This demonstrates the ease with which marker loci can be developed for nonmodel organisms such as *Heliconius*. These markers from the active transcriptome can be used for comparative linkage mapping within *Heliconius* and across the Lepidoptera. In addition, phylogenetic studies across the Lepidoptera will benefit from the availability of markers from across the genome.

The ESTs have been converted to a nonredundant data set and annotated with open source bioinformatic tools (see http://www.nematodes.org) that needed only minimal adaptation for this project. The resulting database includes a stable cluster nomenclature and detailed summary of information for each cluster, including precomputed BLAST searches against a variety of databases, sequence and trace file download, and access to and search via all the annotation methods described in this study. Sequences for other species of Papilionideae will be added as they become available, and it is hoped that this resource will assist the growing application of genomic technology to entomological research (Wang *et al*. 2005; Berenbaum 2002; Tautz 2002; Evans & Gundersen-Rindal 2003; Heckel 2003; Landais *et al*. 2003).

The analysis of these clusters demonstrates how BLAST searches against a taxonomically hierarchical set of

sequence databases can be used to identify the degree to which putative genes are conserved across the diversity of life. For example, coupled with the *Bombyx mori*-clustered EST database (Peregrín-Alvarez *et al*. 2005), Lepidoptera-specific genes can be identified. These might include ancestral genes that are undergoing rapid evolution in Lepidoptera, developmental switches unique to Lepidoptera and potential target genes for taxon-specific pesticides (Domazet-Loso & Tautz 2003).

This study was mainly aimed at developing markers for comparison of synteny and linkage between species, and therefore required conserved markers that would amplify consistently from different species of Lepidoptera. Therefore genes that showed significant similarity to homologues conserved in all cellular life or in all eukaryotes were selected. Half of the genes chosen were ribosomal proteins. There was an overall success rate of 67% (55 of 82 loci) for amplification in both *H. erato* and *H. melpomene* (e.g. Figure 3). Since only a single primer pair was designed for each locus, and no PCR optimization was attempted, this high success rate demonstrates the ease with which new markers can be generated from EST sequences where primers are located in conserved coding sequence. Most of the markers amplified (75%) had introns, as indicated by an amplified fragment significantly larger than expected from the exon sequence. Intron sequences are likely to contain large amounts of genetic variation (Beltran *et al*. 2002), facilitating scoring of segregating alleles in within-species crosses (Jiggins *et al*. 2005). So far we have assigned eight genes to seven linkage groups. The marker development method described here can therefore produce a set of anchor loci that will soon cover each of the 21 chromosomes of these species and facilitate a broad genome-wide analysis of linkage conservation between *Heliconius* species and other lepidopteran species such as *B. mori*. As the assembly of the *Bombyx* genome sequence improves, shared markers will facilitate gene finding in *Heliconius* based on the complete *Bombyx* sequence.

Identification of segregating variation in our mapping families was relatively straightforward. An initial screen of 35 markers revealed segregating variation for 13 loci, several of which showed length variation that was readily scored on agarose gels. The remainder showed restriction site variation after an initial screen of six common restriction enzymes. Obviously more variable sites would be found by screening different mapping families, or by using single nucleotide polymorphism (SNP) technology. It is also worth noting that a lack of amplification could represent null alleles specific to the individuals tested, so some of the 'failed' primers might nonetheless prove useful in different individuals or populations.

The sampling of two divergent lineages within *Heliconius* permits analysis of genetic divergence across multiple genes. We have estimated a silent substitution rate (*Ks*) between *H. erato* and *H. melpomene* of 15.5% across 20 loci, with considerable variance between genes. For comparison, this lies somewhere between the distance separating *Drosophila melanogaster* and *Drosophila simulans* (mean 8.8% across 13 genes, Takano 1998; 11.7% across 5826 genes, D. Pollard, unpublished) and that between *D. melanogaster* and *Drosophila yakuba* (mean 23.4% across 13 genes, Takano 1998; median 28.7% across 5826 genes, D. Pollard, unpublished). *Drosophila melanogaster* and *D. yakuba* diverged 12.8 ± 2.7 million years ago (Ma) (i.e. during the Miocene period; Powell 1997; Tamura *et al*. 2004). While the *Drosophila* molecular clock is unlikely to apply to butterflies (Zakharov *et al*. 2004), this comparison allows the considerable morphological diversity seen in *Heliconius* to be placed into the context of divergence in the *Drosophila melanogaster* subgroup.

Although the genes sampled here represent a tiny fraction of the complete *Heliconius* transcriptome, there are nonetheless several potentially interesting loci with respect to butterfly biology. Abundant functional groups identified in the clustered data sets include many ribosomal proteins (25% of *H. melpomene* but only 5.6% of *H. erato*) that have proved to be excellent conserved markers for linkage mapping. There is also an abundance of cuticle precursor proteins (10–12% of both libraries) and chitinases (*c*. 1% in *H. erato*). There are three sets of clusters of unknown function that are shared between the two species' data sets. Clusters HMC00304, HEC00007 and HEC00389 have no significant similarity to other known proteins but are very similar to each other. Each predicted peptide translation contains a signal peptide and transmembrane domain and other C-terminal conserved domains. HMC00304 is also predicted to have an RNA-binding domain. Similarly, HMC00361 and HEC00056 are almost identical to each other but unique to *Heliconius*, as are HMC00009 and HEC00337.

Genes for mapping were selected from those with significant functional annotation. For example, cluster HMC00218 is similar to prophenoloxidase-activating enzyme precursor (PPAE). Prophenoloxidase, a complex enzyme, is a major innate immune response in invertebrates and plays a major role in the processes of sclerotization and melanization of cuticle in insects (Pentz *et al*. 1990; Lee *et al*. 2004). InterPro searches were also used to identify putative protein domains after the clusters were translated to the most likely protein object. For instance, cluster HMC00318 encodes for armadillo/beta-catenin repeats. BLAST similarity searches identify HMC00318 as a protein of the karyopherin alpha family, which plays a central role in nucleocytoplasmic transport. Genes with armadillo domains are involved in transducing signals from the *wingless* (WNT) pathway (Mason *et al*. 2003). Given the known role of the WNT pathway in butterfly wing pattern formation (Carroll *et al*. 1994), this locus is a possible candidate for involvement in wing pattern signalling. Another

cluster of interest, HEC00398, has a high mobility group (HMG) domain. HMG domains complex with other transcriptional regulators such as Hox proteins. Hox genes encode homeodomain-containing transcriptional regulators that are crucial to patterning of the animal body plan, but might also play a role in wing patterning.

There has been considerable interest in pigment protein evolution in butterflies (Hsu *et al*. 2001), and there is some evidence that light sensitivity varies between *Heliconius* species with different colour patterns (Struwe 1972). This would accord with the known role of colour patterns in mate recognition (Jiggins *et al*. 2001). It would be particularly interesting to determine whether radiation and convergence of colour patterns is correlated with similar evolutionary changes in colour perception. Cluster HMC00077 is similar to an arrestin (retinal S-antigen). Arrestin is a major protein of the retinal rod and interacts with rhodopsin (represented in our data sets by HMC00207 and HMC00187). Transcripts related to olfactory or other chemosensory proteins are also present in the database (e.g. HMC00381, HMC00194 and HMC00109): these are potentially important in mate recognition (Gilbert 1976). In summary, we have singled out a few genes of potential interest, but the value of the database will increase greatly as a larger proportion of the genome is sampled.

Conclusions

The EST data presented here and ongoing sequencing efforts in *Heliconius* provide a valuable resource for molecular ecologists studying the Lepidoptera in general, and butterflies in particular. Furthermore the online database (www.heliconius.org) provides easy access to the data and will continue to be updated as additional EST sequences become available for the Papilionideae. This will provide a publicly available nonredundant set of those genes present in the butterflies that have been sampled to date. Here the utility of these analyses has been demonstrated through the straightforward development of 64 conserved markers that can be used for linkage mapping in *Heliconius*. Although we here concentrate on designing conserved markers, one can imagine a variety of uses for the gene sequences. However, the success of this approach depends on populating the database with ESTs from many species, and we therefore encourage all those involved in generating EST sequences for the Lepidoptera to submit them to the public domain as early as possible.

Acknowledgements

We would like to thank Andrew Gilles and Jill Lovell for help with sequencing, Claire Whitton and Katelyn Fenn for sharing technical expertise, James Wasmuth for prepublication access to prot4EST, Ralf Schmid for help with PARTIGENE, Ann Hedley for PHP scripting, and Alistair Anthony for UNIX introductory lessons. We would like to acknowledge the financial support of BBSRC (AP and MJ), NERC (MLB), the Royal Society (CDJ), NSF (WOM) and the Smithsonian Tropical Research Institute (CDJ).

References

Altschul SF, Madden TL, Schaffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Bates HW (1863) *The Naturalist on the River Amazons; A Record of Adventures, Habits of Animals, Sketches of Brazilian and Indian Life, and Aspects of Nature under the Equator during Eleven Years of Travel*. John Murray, London.

Beldade P, Brakefield PM (2002) The genetics and evo-devo of butterfly wing patterns. *Nature Reviews Genetics*, **3**, 442–452.

Beltran M, Jiggins CD, Bull V *et al.* (2002) Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating *Heliconius* butterflies. *Molecular Biology and Evolution*, **19**, 2176–2190.

Benson WW (1972) Natural selection for Müllerian mimicry in *Heliconius erato* in Costa Rica. *Science*, **176**, 936–939.

Berenbaum MR (2002) Postgenomic chemical ecology. *Journal of Chemical Ecology*, **28**, 873–896.

Bergman C, Pfeiffer B, Rincón-Limas D *et al.* (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biology*, **3**, 1.

Blaxter M (2002) Opinion piece. Genome sequencing: time to widen our horizons. *Briefings in Functional Genomics and Proteomics*, **1**, 7–9.

Boggs CL, Watt WB, Ehrlich PR (2003) *Butterflies: Ecology and Evolution Taking Flight*, 1st edn. University of Chicago Press, Chicago.

Boyden TC (1976) Butterfly palatability and mimicry experiments with *Ameiva* lizards. *Evolution*, **30**, 73–81.

Brendel V, Xing L, Zhu W (2004) Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics*, **20**, 1157–1169.

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

Chai P (1986) Field observations and feeding experiments on the responses of rufous-tailed jacamars (*Galbula ruficauda*) to free-flying butterflies in a tropical rainforest. *Biological Journal of the Linnean Society*, **29**, 161–189.

Clarke CA, Sheppard PM (1966) A local survey of the distribution of industrial melanic forms in the moth *Biston betularia* and estimates of the selective values of these in an industrial environment. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **165**, 424–439.

Domazet-Loso T, Tautz D (2003) An evolutionary analysis of orphan genes in *Drosophila*. *Genome Research*, **13**, 2213–2219.

Evans JD, Gundersen-Rindal D (2003) Beenomes to *Bombyx*: future directions in applied insect genomics. *Genome Biology*, **4**, 107.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Research*, **8**, 186–198.

Fei Z, Tang X, Alba RM *et al.* (2004) Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. *Plant Journal*, **40**, 47–59.

Ford EB (1931) *Mendelism and Evolution*. Methuen, London.

Ford EB (1964) *Ecological Genetics*. Methuen, London.

Gilbert LE (1976) Postmating female odor in *Heliconius* butterflies: a male-contributed antiaphrodisiac? *Science*, **193**, 419–420.

Gupta PK, Rustgi S (2004) Molecular markers from the transcribed/expressed region of the genome in higher plants. *Functional and Integrative Genomics*, **4**, 139–162.

Haas B, Volfovsky N, Town C *et al.* (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biology*, **3**, 1–12.

Haldane JBS (1956) The theory of selection for melanism in Lepidoptera. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **145**, 303–306.

Heckel DG (2003) Genomics in pure and applied entomology. *Annual Review of Entomology*, **48**, 235–260.

Higgins D, Thompson J, Gibson T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–1680.

Hsu R, Briscoe AD, Chang BSW, Pierce NE (2001) Molecular evolution of a long wavelength-sensitive opsin in mimetic *Heliconius* butterflies (Lepidoptera: Nymphalidae). *Biological Journal of the Linnean Society*, **72**, 435–449.

Jiggins CD, McMillan WO, Neukirchen W, Mallet J (1996) What can hybrid zones tell us about speciation? The case of *Heliconius erato* and *H. himera* (Lepidoptera: Nymphalidae). *Biological Journal of the Linnean Society*, **59**, 221–242.

Jiggins CD, Naisbit RE, Coe RL, Mallet J (2001) Reproductive isolation caused by colour pattern mimicry. *Nature*, **411**, 302–305.

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

Joron M, Mallet J (1998) Diversity in mimicry: paradox or paradigm? *Trends in Ecology & Evolution*, **13**, 461–466.

Kazuei M, Kasahara M, Shimada T, Morishita S, Sasaki T (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Research*, **11**.

Kettlewell HBD (1955) Selection experiments on industrial melanism in the Lepidoptera. *Heredity*, **9**, 323–342.

Landais I, Ogliastro M, Mita K *et al.* (2003) Annotation pattern of ESTs from *Spodoptera frugiperda* Sf9 cells and analysis of the ribosomal protein genes reveal insect-specific features and unexpectedly low codon usage bias. *Bioinformatics*, **19**, 2343–2350.

Lee MH, Osaki T, Lee JY *et al.* (2004) Peptidoglycan recognition proteins involved in 1,3-beta-D-glucan-dependent prophenoloxidase activation system of insect. *Journal of Biological Chemistry*, **279**, 3218–3227.

Mallet J (2004) Perspectives Poulton, Wallace and Jordan. How discoveries in *Papilio* butterflies led to a new species concept 100 years ago. *Systematics and Biodiversity*, **1**, 441–452.

Mason DA, Mathe E, Fleming RJ, Goldfarb DS (2003) The *Drosophila melanogaster* importin α3 locus encodes an essential gene required for the development of both larval and adult tissues. *Genetics*, **165**, 1943–1958.

Mita K, Morimyo M, Okano K *et al.* (2003) The construction of an EST database for *Bombyx mori* and its application. *Proceedings of the National Academy of Sciences, USA*, **100**, 14121–14126.

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, **3**, 418–426.

Oleksiak MF, Kolell K, Crawford DL (2001) Utility of natural populations for microarray analysis: isolation of genes necessary for functional genomic studies. *Marine Biotechnology*, S203–S211.

Parkinson J, Guiliano D, Blaxter M (2002) Making sense of EST sequences by CLOBBING them. *BMC Bioinformatics*, **3**, [online].

Parkinson J, Anthony A, Wasmuth J *et al.* (2004a) PARTIGENE — constructing partial genomes. *Bioinformatics*, **20**, 1398–1404.

Parkinson J, Mitreva M, Whitton C *et al.* (2004b) A transcriptomic analysis of the phylum Nematoda. *Nature Genetics*, **36**, 1259–1267.

Pentz ES, Black BC, Wright TR (1990) Mutations affecting phenol oxidase activity in *Drosophila*: quicksilver and tyrosinase-1. *Biochemical Genetics*, **28**, 151–171.

Peregrín-Alvarez JM, Yam A, Sivakumar G, Parkinson J (2005) PARTIGENEDB: collating partial genomes. *Nucleic Acids Research*, **1**, 33.

Poulton EB (1884) The essential nature of the colouring of phytophagous larvae (and their pupae); with an account of some experiments upon the relation between the colour of such larvae and that of their food-plants. *Proceedings of the Royal Society of London*, **38**, 269–315.

Powell JR (1997) *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press, New York.

Renn S, Aubin-Horth N, Hofmann H (2004) Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. *BMC Genomics*, **5**, 42.

Rozas J, Sanchez-DeI, Bario JC, Messeguer X, Rozas R (2003) DNASP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.

Schoen DJ (2000) Comparative genomics, marker density and statistical analysis of chromosome rearrangements. *Genetics*, **154**, 943–952.

Sheppard PM, Turner JRG, Brown KS, Benson WW, Singer MC (1985) Genetics and the evolution of muellerian mimicry in *Heliconius* butterflies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **308**, 433–610.

Streelman JR, Albertson RC, Kocher TD (2003) Genome mapping of the orange blotch colour pattern in cichlid fishes. *Molecular Ecology*, **12**, 2465–2471.

Struwe G (1972) Spectral sensitivity of the compound eye in butterflies *Heliconius*. *Journal of Comparative Physiology*, **79**, 191–196.

Takano TS (1998) Rate variation of DNA sequence evolution in the *Drosophila* lineages. *Genetics*, **149**, 959–970.

Tamura K, Subramanian S, Kumar S (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular Biology and Evolution*, **21**, 36–44.

Tautz D (2002) Insects on the rise. *Trends in Genetics*, **18**, 179–180.

Tobler A, Kapan D, Flanagan NS *et al.* (2005) First-generation linkage map of the warningly colored butterfly *Heliconius erato*. *Heredity*, **94**, 408–417.

Turner JRG (1976) Adaptive radiation and convergence in subdivisions of the butterfly genus *Heliconius* (Lepidoptera: Nymphalidae). *Zoological Journal of the Linnean Society*, **58**, 297–308.

Turner JRG (1981) Adaptation and evolution in *Heliconius*: a defense of NeoDarwinism. *Annual Review of Ecology and Systematics*, **12**, 99–121.

Turner JRG, Mallet JLB (1996) Did forest islands drive the diversity of warningly coloured butterflies? Biotic drift and the shifting balance. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **351**, 835–845.

Turner JRG, Sheppard PM (1975) Absence of crossing-over in female butterflies *Heliconius*. *Heredity*, **34**, 265–270.

Wang J, Xia Q, He X *et al.* (2005) SilkDB: a knowledge base for silkworm biology and genomics. *Nucleic Acids Research*, **33 (Database Issue)**, D399–402.

Wasmuth J, Blaxter M (2004) PROT4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics*, **5**, 187.

Whitton C, Daub J, Thompson M, Blaxter M (2004) Expressed sequence tags: medium-throughput protocols. *Methods Molecular Biology*, **270**, 75–92.

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.

Yasukochi Y (1998) A dense genetic map of the silkworm, *Bombyx mori*, covering all chromosomes based on 1018 molecular markers. *Genetics*, **150**, 1513–1525.

Zakharov EV, Caterino MS, Sperling FAH (2004) Molecular phylogeny, historical biogeography, and divergence time estimates for swallowtail butterflies of the genus *Papilio* (Lepidoptera: Papilionidae). *Systematic Biology*, **53**, 193–215.

This work is part of AP's master's thesis and the beginning of long-term projects in CJ's and WOM's labs which aim to identify the molecular basis of butterfly colour pattern evolution and mimicry. This paper demonstrates the ease of development of cDNA and genomic tools in molecular ecology of otherwise neglected taxa, a core interest of MB's lab. The authors suggest that evolutionary theory will benefit from collecting genomic information from the diversity of life outside the molecular genetics laboratory.
