D582–D587 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 12 October 2007*
doi:10.1093/nar/gkm853

ButterflyBase: a platform for lepidopteran genomics

Alexie Papanicolaou[1,2,*], Steffi Gebauer-Jung[2], Mark L. Blaxter[1], W. Owen McMillan[3] and Chris D. Jiggins[1,4]

[1]Institute for Evolutionary Biology, University of Edinburgh, King's Buildings, EH9 3JT, UK, [2]Max Planck Institute for Chemical Ecology, Jena, 07745, Germany, [3]Department of Genetics, North Caroline State University, Raleigh, NC 27695-7614, USA and [4]Department of Zoology, University of Cambridge, Downing Street, CB2 3EJ, UK

Received August 15, 2007; Revised September 25, 2007; Accepted September 26, 2007

ABSTRACT

With over 100 000 species and a large community of evolutionary biologists, population ecologists, pest biologists and genome researchers, the Lepidoptera are an important insect group. Genomic resources [expressed sequence tags (ESTs), genome sequence, genetic and physical maps, proteomic and microarray datasets] are growing, but there has up to now been no single access and analysis portal for this group. Here we present ButterflyBase (http://www.butterflybase.org), a unified resource for lepidopteran genomics. A total of 273 077 ESTs from more than 30 different species have been clustered to generate stable unigene sets, and robust protein translations derived from each unigene cluster. Clusters and their protein translations are annotated with BLAST-based similarity, gene ontology (GO), enzyme classification (EC) and Kyoto encyclopaedia of genes and genomes (KEGG) terms, and are also searchable using similarity tools such as BLAST and MS-BLAST. The database supports many needs of the lepidopteran research community, including molecular marker development, orthologue prediction for deep phylogenetics, and detection of rapidly evolving proteins likely involved in host–pathogen or other evolutionary processes. ButterflyBase is expanding to include additional genomic sequence, ecological and mapping data for key species.

INTRODUCTION AND MOTIVATION

The Lepidoptera (butterflies and moths) are remarkably diverse containing more than 100 000 described species. There is a long tradition of research and a number of disciplines use lepidopteran models to investigate fundamental biological phenomena including development and gene regulation, population genetic processes (gene flow, colonization and extinction), adaptation and morphological innovation, speciation and co-evolutionary processes such as host–plant and insect–parasite interactions. As a result, there is a wealth of ecological and genetic knowledge for Lepidoptera.

The silkworm *Bombyx mori* is a model for insect physiology and molecular biology, as well as being an important crop animal. Currently, two whole genome shotgun sequence assemblies are publicly available (1,2) and a joint genome assembly by the Chinese and Japanese teams is expected within 2007. The genomic sequence data are anchored by a number of bacterial artificial chromosome (BAC) libraries, high-density linkage maps of sequence tag sites (STS), cDNA and microsatellite (simple sequence repeats, SSR) markers (3–6) as well as cytogenetic studies (7) which provide a chromosomal framework for genome assembly. Thus the chromosomal framework for genome assembly is in place and as the annotation of the *B. mori* genome progresses, it will facilitate comparative analysis of other species with less complete genomic information (8).

In addition to genomic resources in *Bombyx*, there is increasing amount of EST data for a growing number of Lepidoptera species. Large to moderate-sized EST datasets are becoming easier and less expensive to produce and can be powerful source of markers for comparative mapping, population genetic analysis and studies of adaptive evolution (9). For example, there are large public genomic datasets for the moth pest *Spodoptera frugiperda*, and the butterflies *Bicyclus anynana*, *Heliconius melpomene* and *Heliconius erato*. The generation of sequences for these and other species has led to the discovery that around half of the sequenced genes in Lepidoptera have little or no sequence similarity to proteins from other taxa (8). Species-specific public databases are available for these taxa, but vary widely in accessibility and format (10–13). What is lacking is a central platform for accessing lepidopteran data and more importantly for conducting comparative between species analyses.

To allow the community to benefit from the comparative genomic data available in Lepidoptera, we developed

*To whom correspondence should be addressed. Tel: +493641571561; Fax: +493641571502; Email: alexie@butterflybase.org

an online database and annotation platform, called ButterflyBase. It is available at http://www.butterflybase.org. ButterflyBase is a comparative gene-focused database for all Lepidoptera. ButterflyBase brings together, in a single site, sequence information for all lepidopterans including *B. mori*. ButterflyBase was designed to extend the utility of the publicly available expressed sequence tag (EST) datasets using clustering and protein prediction software, and to provide high-quality annotation for data mining and exploitation, all through a simple and intuitive user interface. With this short article, we hope to introduce users to the utility of the database. Further information regarding technical details can be obtained on request or by browsing the dataset download page.

METHODOLOGY

Datasets

ESTs and full-length cDNA sequences were obtained from public depositions in the EMBL/GenBank/DDBJ database, and clustered using a modified version of the PartiGene suite (14). When the original sequencer chromatograms were available (*H. erato* and *H. melpomene*) we processed them with trace2dbest (14). All other data were pre-processed to remove vector contamination, poly(A) tails and sequences smaller than 150 bp. For some cDNA libraries (where sequence quality was poor), further trimming was performed using a customized version of est_trimmer.pl [provided by Thomas Thiel through the MISA program (15)]. SSR prediction was performed using MISA (15), single nucleotide polymorphisms (SNPs) were predicted using SEAN (16) and databased using custom Perl scripts. A SEAN Java viewer is available as a modified applet, provided by the SEAN author. The methodology of SEAN does not rely on quality information and therefore can be used with our datasets. Instead, it only marks putative SNPs if a single nucleotide change is present in at least two members of the EST cluster and there are no other nucleotide inconsistencies 15 bp up- and downstream of the putative SNP.

PartiGene (14) uses megablast and the CLOBB approach to cluster EST sequences into groups putatively derived from the same mRNA molecule (17). These clusters are subsequently aligned using Phrap (with the *forcelevel* option set to maximum) (18, Green,P., unpublished software). Sequenced organisms in Lepidoptera are often outbred and may, therefore, exhibit substantial allelic variation. Essentially, the presence of low quality, multiple SNPs, sequencing errors, alternative splicing or short indels may allow megablast to generate a cluster of highly similar sequences which is not subsequently aligned by Phrap, thus leading to some clusters containing more than one contig.

ButterflyBase uses a two-letter code to signify the species ID and a third letter to signify molecule type (P for protein, C for nucleotide cluster (or unigene) and in the future B for BAC clone). Each cluster of ESTs and cDNAs has a unique numerical ID, which is stable when additional sequences are added to the dataset. When there is more than one contig per cluster these are indicated by a trailing number. Thus HEC00123_1 is the first contig of a nucleotide cluster from *H. erato* and its protein translation is HEP00123_1. Cluster identifiers are conserved as more sequences are added.

Protein prediction

The protein predictions are ButterflyBase's strongest asset. We use, prot4EST, a protein prediction tool developed specifically for EST data (19). Briefly, this program utilizes a four-tier methodology: first, similarity to known proteins is used in order to detect the open reading frame (ORF) and correct for any potential sequencing errors [using the high-scoring segment pair (HSP) tiling approach], if that fails (e.g. for novel or Lepidoptera-specific genes) ESTSCAN is utilized (20) and if that fails too then DECODER (21) and finally the longest ORF from the six-frame translation. As prior training data (codon usage tables and base composition estimates) for probabilistic prediction of ORFs were not available for many lepidopteran species, we utilized data derived from high-scoring BLAST matches to populate species-specific parameter sets.

Database schema and dataset annotation

The database is driven by PostgreSQL with a customized version of the PartiGene schema. The central entity is a mRNA sequence cluster. Each cluster is annotated with a number of facilities. The most frequently accessed are pre-computed BLAST similarity searches versus a variety of databases: Uniref100; a collection of possible contaminants (e.g. fungi, viruses, bacteria, molecular biology vectors) and phylogenetically selected, nested databases. We chose a number of such databases including *B. mori* nucleotides and proteins; Lepidoptera nucleotides without *B. mori*; proteins from released Arthropoda genomes; Arthropoda sequences without those genomes or Lepidoptera. All BLAST searches have an *E*-value cutoff of $1E{-}4$. Furthermore, predictions enhance the utility of the consensus: a robust protein translation as well as SSR and SNP predictions are currently offered. The protein predictions in turn are annotated with enzyme classification (EC), gene ontology (GO) and Kyoto encyclopaedia of genes and genomes (KEGG) terms. These latter annotations are derived from BLAST searches of annotated protein databases using the annot8r tool (Schmid,R. and Blaxter,M., unpublished software), and a cut-off *E*-value of $1E{-}8$. Furthermore, ButterflyBase provides domain annotations from InterProScan (22) and basic protein statistics to facilitate downstream proteomic and biochemical investigations. Annotations are updated on a 4-month cycle and new sequence data are imported $\sim$2 months after the release of at least 1000 sequences from any lepidopteran species. Communication with the database curators regarding an imminent release will shorten this time. Metadata linked to each mRNA or EST sequence (life cycle stage, tissue, sex, etc.) have also been databased. Original sequence accession numbers are also listed on each cluster page and linked to EMBL, and can

be searched for with the 'Jump to' search box on the left hand side of every page.

A SHORT TOUR

For security and efficiency reasons, the user-interface pages allow the user to explore the data with certain predefined queries (but see access statement below). ButterflyBase permits simple text searches against the sequence annotation. The definition lines of similar sequences are searched, with the option to define a cut-off value for the precomputed BLAST similarity searches. KEGG (23), GO (24) and EC codes and definitions can also be searched. All searches can be limited to a specific organism or cDNA library.

Once a cluster of interest is found, the cluster page shows a range descriptive data, including the raw data (such as sequence traces if available), the number of ESTs in the cluster, the cDNA libraries they belong to, similarity information from BLAST searches against three databases (Uniref, *Drosophila melanogaster* proteins from FlyBase (25) and *B. mori* predicted proteins from ButterflyBase), and links to the output of all the other BLAST similarity searches. The alignment of the constituent sequences to the consensus can be viewed using an interactive image, a Java applet driven by SEAN or a non-Java text view. These alignment views allow the user to pinpoint databased SNPs. The linked protein page contains basic descriptive data, the predicted sequence, the results of BLAST similarity searches and KEGG, EC, GO and InterPro domain annotation.

EST sequences are a key resource for the development of sequence-specific markers for genetic mapping (26). ButterflyBase facilitates marker development by providing sequence information and a tool for designing degenerate or conserved primers. A protein-driven nucleotide alignment of two orthologous lepidopteran clusters is generated and then used for design of primers using Primer3 (27). EST sequences are also of great utility for the design of microsatellite markers (28). Although transcribed microsatellites are often less polymorphic than non-coding ones (15), they are less likely to be multi-copy or mobile (29). In addition, primers are designed on exon sequences, thus reducing the possibility of null alleles. We provide a simple tool to output any microsatellite present in a specific sequence and also a table of all the microsatellite detected in each species' dataset.

ButterflyBase offers also a BLAST server. Three BLAST search modes are available (NCBI-BLASTALL, PSI-BLAST and WU-BLAST-driven MS-BLAST). MS-BLAST (30) allows a user to query protein databases with multiple short peptide sequences derived from high-throughput mass spectrometry data. PSI-BLAST is particularly effective in the detection of distant similarity and will become an important method for detecting lepidopteran homologues of target genes as the database grows. For more complex queries, a database dump file can be downloaded for local replication of the database, as can species-specific FASTA files of the nucleotide cluster consensus and protein predictions, and custom-built annotation databases used in ButterflyBase.

All datasets, including a SQL flatfile of the database are provided for download with their checksum codes. We also provide FASTA files of some of the custom sequence databases used to carry out similarity searches. One drawback of public EST data, however, is the lack of a raw sequence trace repository. PartiGene can utilize these traces to assist the Phrap alignments, but we are also using them to check manually for the quality of specific libraries or clusters of interest. For this reason, all sequence traces we process are publicly available for download from their respective cluster pages along with a short text file on how the sequence was processed by trace2dbest. This is, unfortunately, only available for sequence trace data we have access to, namely *Heliconius* sp. and *B. anynana*. We are, however, encouraging the community to submit to us their raw sequence data.

SUMMARY OF CONTENT AND UTILITY

Website usage is outlined in the online User's Manual but a summary of the content follows. The main webpage provides an up-to-date overview of the content of the database. At the time of print, ButterflyBase has processed 273 077 mRNA sequences from 32 lepidopteran species belonging to a total of 12 families giving circa 71 000 gene and almost as many protein objects. Although most of the sequences are from *B. mori*, there are nonetheless now 17 species with more than 500 sequences, and 12 species with more than 1000, representing a valuable comparative dataset (Table 1). Nearly half of the ButterflyBase clusters have similarity to known proteins outside the Lepidoptera clade. Although identity of sequence does not necessarily translate into identity of function, sequence similarity is a first step towards gene finding in this taxon. Also, ~58% of the genes in ButterflyBase are significantly similar to at least one more ButterflyBase species, thus facilitating annotation and the design of degenerate or conserved markers. What is also apparent is the relatively high proportion of Lepidoptera-specific genes, about one-third of the clusters have hits only in sequences derived from Lepidoptera but in *B. mori* (which is the most complete dataset) the proportion is about half of the gene objects (Table 1). The number of gene objects is an overestimate of the exact number of actual genes due to the nature of EST datasets and the lack of a genome backbone. Thus, two sets of ESTs from the same gene will appear as two unigenes if they do not overlap, however, accuracy will increase as sequence information from more Lepidoptera is provided. Furthermore, the whole of the *B. anynana* dataset and ca. 16% of the *B. mori* dataset contains 3′ sequences. Therefore, these gene objects may contain long untranslated regions (UTRs) which are not conserved. In any case, these observations warrant an in-depth investigation and any putative Lepidoptera-specific genes need to be examined in a phylogenetic context in order to determine if they have evolved novel functions specific to Lepidoptera or if they have retained ancestral functions despite gross sequence divergence on the protein level.

Table 1. The content of ButterflyBase (September 2007)

Species (ButterflyBase Code)	Taxon/Family	Proteins @ NCBI[a]	mRNAs @ Bbase	Gene objects @ BBase	Similar to known proteins[b]	Only exist in Lepidoptera[c]	Found in 2+ ButterflyBase species[d]	Clusters with putative SNPs (total SNPs)[e]
Total: 33	Lepidoptera	6907	273 077	70 867	37 962	25 204	9583 (41 093)	4821 (27 808)
Anagasta (*Ephestia*) *kuehniella* (AKC)	Pyralidae	3	28	23	14	6	5 (14)	5 (14)
*Antheraea polyphemus** (ALC)	Saturniidae	45	22	17	17	0	0 (17)	N/A
Antheraea mylitta (AMC)	Saturniidae	51	3912	1433	943	509	535 (1432)	47 (140)
*Antheraea pernyi** (APC)	Saturniidae	65	40	37	37	0	0 (37)	N/A
Antheraea yamamai (AYC)	Saturniidae	35	610	325	157	82	88 (226)	9 (19)
Bicyclus anynana (BAC)	Nymphalidae	11	9848	5726	2375	1207	1012 (3099)	81 (234)
Bombyx mori (BMC)	Bombycidae	3623	184 577	35 876	17 162	19 174	4776 (17 194)	3756 (22 445)
Bombyx mandarina (BNC)	Bombycidae	54	261	205	105	97	90 (194)	3 (3)
Choristoneura fumiferana (CFC)	Tortricidae	74	652	618	359	82	72 (379)	N/A
Euclidia glyphica (EGC)	Noctuidae	N/A	570	259	138	2	2 (122)	18 (50)
Galleria mellonella (GMC)	Pyralidae	95	93	84	68	8	4 (65)	N/A
Helicoverpa armigera (HAC)	Noctuidae	207	1221	733	634	53	50 (663)	19 (118)
*Hyalophora cecropia** (HCC)	Saturniidae	57	20	16	16	0	0 (16)	N/A
Heliconius erato (HEC)	Nymphalidae	157	17 573	6859	4787	1118	856 (5019)	464 (3236)
Heliconius melpomene (HMC)	Nymphalidae	443	4976	1965	1262	408	422 (1531)	99 (369)
*Heliothis virescens** (HVC)	Noctuidae	152	90	83	83	0	0 (83)	N/A
*Helicoverpa zea** (HZC)	Noctuidae	80	40	38	38	0	0 (38)	N/A
Lonomia obliqua (LOC)	Saturniidae	133	1635	671	503	60	58 (514)	25 (63)
Manduca sexta (MSC)	Sphingidae	582	3683	2291	1256	412	301 (1469)	22 (56)
Ostrinia nubilalis (ONC)	Crambidae	146	1761	543	309	137	133 (418)	40 (162)
*Pieris brassicae** (PBC)	Pieridae	17	5	5	5	0	0 (4)	N/A
Papilio dardanus (PDC)	Papilionidae	14	708	307	236	22	20 (248)	27 (102)
Plodia interpunctella (PIC)	Pyralidae	47	6219	3788	1879	483	414 (2079)	28 (80)
Papilio xuthus (PUC)	Papilionidae	41	25	24	24	0	0 (24)	N/A
Plutella xylostella (PXC)	Plutellidae	188	1286	1021	701	108	124 (747)	3 (11)
Samia cynthia spp.* (SCC)	Saturniidae	49	27	27	27	0	0 (27)	N/A
*Spodoptera exigua** (SEC)	Noctuidae	64	48	42	42	0	0 (42)	N/A
Spodoptera frugiperda (SFC)	Noctuidae	241	31 538	6993	4172	1116	1204 (4741)	149 (528)
Spodoptera litura (SLC)	Noctuidae	66	154	100	85	7	8 (90)	1 (1)
*Spodoptera littoralis** (STC)	Noctuidae	28	23	20	20	0	0 (20)	N/A
Tineola bisselliella (TBC)	Tineidae	1	921	240	170	39	14 (162)	30 (177)
Trichoplusia ni (TNC)	Noctuidae	138	511	498	338	74	61 (379)	N/A

*designates those species with no public ESTs but public full-length mRNA sequences.
[a]Nuclear sequences only, this total includes segmented sequences and is not limited to RefSeq. August 2007. The *B. mori* proteins were limited to 1025 before January 2007.
[b]BLASTx of nucleotide consensus and BLASTp of predicted proteins versus Uniref100 or proteins released by the *Apis mellifera*, *D. melanogaster*, *Tribolium castaneum* and *Anopheles gambiae* genomes or other Arthropoda proteins in EBI with *E*-value cutoff $1E-4$ (source: EBI Jul 2007). We also used in-house clusters of the public EST data for *Aedes aegypti, Anopheles gambiae, Culex pipiens, Drosophila ananassae, Drosophila erecta, Drosophila grimshawi, Drosophila simulans, Drosophila yakuba* and *Tribolium castaneum* (*E*-value cutoff $1E-4$, source: EBI September 2007).
[c]BLASTn of nucleotide consensus versus Lepidoptera nuclear nucleotides, *B. mori* genome from EBI and ButterflyBase EST consensuses but no significant similarity to the databases mentioned above (EBI, Jul 2007, *E*-value cutoffs $1E-4$).
[d]Lepidoptera-specific clusters which were found to have a significant hit in at least one other organism in ButterflyBase using BLASTn for nucleotide consensuses or BLASTp for protein predictions (Jul 2007, *E*-value cutoff $1E-3$). Gene objects present in more than one organism facilitate annotation and marker design. In brackets, a similar count is present for all clusters regardless of similarity to any protein.
[e]Most Lepidoptera cDNA libraries are constructed with relative outbred individuals, thus the relatively high number of SNPs. Even though the number of clusters containing putative SNPs are accurate, the reader has to consider that the total number of SNPs may be inflated as the data here are pooled from all cDNA libraries.

Phylogenetics

The phylogenetic context of Lepidoptera is one of the taxon's strongest advantages for the study of ecology and evolution. Although the amount of public genomic data in Lepidoptera is increasing rapidly, the phylogenetic coverage is limited to the Ditrysia and non-existent for basal clades. A broader phylogenetic sampling, of at least a handful of chosen genes will help improve much of the unresolved lepidopteran phylogeny and also shed more light on the evolutionary dynamics of Lepidoptera-specific genes. Different levels of phylogenetic investigation require different kinds of genes, thus fast-evolving genes are only suited for building phylogenies of closely related species whereas highly conserved genes (such as ribosomal proteins) are best suited for inferring the relationships among the more basal lineages. A broad phylogenetic analysis of ~300 species using up to 26 genes derived from EST sequences is already underway (Leptree.net; Mitter, personal communication) and the tools developed in ButterflyBase will facilitate this and similar research.

Annotation

ButterflyBase is primarily an annotation platform. Currently, the only information provided is similarity

to known sequences, including to other lepidopteran sequences. The aim of the annotation platform is to host enough information to allow researchers to judge if their sequence of interest has a specific annotation identity. This annotation will be essential for annotating novel sequences especially short reads generated in some projects such as cDNA–AFLPs. Currently, we do not provide curated annotation information but in the near future we will publish analysis on orthologue groupings. We plan to allow the community itself to contribute annotations for each ButterflyBase object perhaps by using a Wiki-based annotation platform (31) or the Generic Model Organism Database toolkit (GMOD). In addition, we hope to expand the annotation platform to include both non-EST sequence data and genetic/phenotypic data within 2008. Such an effort will be initialized by a conversion to the more standardized database schema of Chado from the GMOD (32). The major obstacle is however the lack of a fully sequenced genome with which to anchor the genomic data. The quality of the first releases of *B. mori* is not sufficient for the purpose but a joint assembly is expected to be made public within 2007. With a GMOD-compatible database and a *B. mori* genome the capability of ButterflyBase as an annotation tool will be greatly enhanced. Likewise, as additional EST datasets are made public, the quality of the annotation will increase.

Linkage mapping and molecular evolution

ButterflyBase was originally developed for the generation of EST-based molecular markers for *Heliconius* sp. (26,33). Using ButterflyBase data, a researcher may generate conserved, degenerate or species-specific markers of specific single-copy genes. Pringle *et al.* (33) used this approach to provide the first extensive evidence for conserved macro-synteny between *H. melpomene* (a butterfly) and *B. mori* (a moth), two species whose sequence divergence has reached saturation in third codon positions. ButterflyBase provides also predicted SNPs, which have been determined from the clustered alignment. These identified SNPs (and RFLPs) can be verified by visual inspection of the alignment. Such data allow the generation of SNP-based markers to survey natural populations for association mapping projects or estimate the rate of evolution of specific proteins. Researchers using a cDNA approach to acquire SNP information for linkage mapping can also make use of ButterflyBase's services and in the process contribute to the pool of public sequence information for Lepidoptera.

Proteomics

An important function of genomic datasets is to guide future biochemical investigations. In taxa such as Lepidoptera, where much of the proteome is unknown and composed of many previously unidentified genes, *de novo* protein sequencing provides valuable information. In such proteomes, standard methodologies for identifying peptides by mass spectrometric (MS) data are more error-prone and can be misleading. The MS-BLAST server facilitates identification using the ButterflyBase predicted (and often partial) proteins.

Support small-scale sequencing

During the construction of ButterflyBase we used all available Lepidoptera ESTs hosted in the public domain. A fraction of them was unfortunately lacking information, or contained vector contamination and/or low-quality sequence. ButterflyBase provides the facility to host trace information and currently holds raw trace data from *H. erato*, *H. melpomene* and *B. anynana*. In the future, ButterflyBase's pipeline will judge the quality of a cDNA library based on the number of errors as detected from ESTs from other libraries or published full-length mRNAs. This is only possible, however, for species where multiple libraries of sufficient depth exist. In addition, ButterflyBase can offer the service of processing raw traces and generate dbest submission reports to researchers who request so and thus allow for a more standardized collection of Lepidoptera sequence information. In the near future, a new international Advisory Board will guide ButterflyBase and will post a set of recommendations for submissions of data to GenBank.

DATA SUBMISSION AND ACCESS STATEMENT

All ButterflyBase data are freely and publicly accessible. To be included in ButterflyBase, EST and mRNA data should be submitted to EMBL/GenBank/DDBJ (a step which we can handle upon request). We strongly encourage submission of raw trace files (in SCF format) to ButterflyBase. Although the user is limited to predefined queries and can download a copy of the database, we can also run custom queries upon request (email query at butterflybase.org). Our goal for the future is to develop the project guided by the community. Therefore, we welcome requests and contributions.

ACKNOWLEDGEMENTS

We are grateful to the Blaxter Neglected Genomes bioinformatics team for support and use of compute resources, especially Ann Hedley and Ralf Schmid. Hendrik Tilger, Martin Niebergall and Dieter Ruder set up the distributed computing. Walter Traut, Mathieu Joron, Simon Baxter, Jim Mallet, Patricia Beldade and David G. Heckel provided many useful comments. Mathieu Joron created the first EST dataset with which ButterflyBase used for its development. A.P. and S.G.J. are supported by the Max Planck Gesellschaft (Germany), W.O.M. by the American National Science Foundation and NEScent, the National Evolutionary Synthesis Center, C.D.J. by a Royal Society Fellowship (UK) and M.L.B. by the Natural Environment Research Council (NERC, UK). Initial support was provided by the Biotechnology and Biological Sciences Research Council (BBSRC, UK). Author contributions: The initial *Heliconius* EST database was conceived by C.J. and M.B. The extension from the '*Heliconius* ButterflyBase' to 'ButterflyBase' was conceived and developed by A.P. with

additional technical support from S.G.J. Intellectual support and motivation was from W.O.M. This article was drafted by A.P., M.L.B., W.O.M. and C.J. All authors approved the final version of the manuscript. Funding to pay the Open Access publication charges for this article was provided by Max Planck Gesellschaft (Germany).

Conflict of interest statement. None declared.

REFERENCES

1. Xia,Q., Zhou,Z., Lu,C., Cheng,D., Dai,F., Li,B., Zhao,P., Zha,X., Cheng,T. *et al*. (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937–1940.
2. Mita,K., Kasahara,M., Sasaki,S., Nagayasu,Y., Yamada,T., Kanamori,H., Namiki,N., Kitagawa,M., Yamashita,H. *et al*. (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Res*., **11**, 27–35.
3. Wu,C., Asakawa,S., Shimizu,N., Kawasaki,S. and Yasukochi,Y. (1999) Construction and characterization of bacterial artificial chromosome libraries from the silkworm, *Bombyx mori*. *Mol. Gen. Genet*., **261**, 698–706.
4. Yasukochi,Y., Ashakumary,L.A., Baba,K., Yoshido,A. and Sahara,K. (2006) A second-generation integrated map of the silkworm reveals synteny and conserved gene order between lepidopteran insects. *Genetics*, **173**, 1319–1328.
5. Yamamoto,K., Narukawa,J., Kadono-Okuda,K., Nohata,J., Sasanuma,M., Suetsugu,Y., Banno,Y., Fujii,H., Goldsmith,M.R. *et al*. (2006) Construction of a single nucleotide polymorphism linkage map for the silkworm, *Bombyx mori*, based on bacterial artificial chromosome end sequences. *Genetics*, **173**, 151–161.
6. Miao,X.X., Xub,S.J., Li,M.H., Li,M.W., Huang,J.H., Dai,F.Y., Marino,S.W., Mills,D.R., Zeng,P. *et al*. (2005) Simple sequence repeat-based consensus linkage map of *Bombyx mori*. *Proc. Natl Acad. Sci. USA*, **102**, 16303–16308.
7. Yoshido,A., Bando,H., Yasukochi,Y. and Sahara,K. (2005) The *Bombyx mori* karyotype and the assignment of linkage groups. *Genetics*, **170**, 675–685.
8. Beldade,P., McMillan,W.O. and Papanicolaou,A. (2007) Butterfly genomics eclosing. *Heredity* [Epub ahead of print].
9. Bouck,A. and Vision,T. (2007) The molecular ecologist's guide to expressed sequence tags. *Mol. Ecol*., **16**, 907–924.
10. Beldade,P., Rudd,S., Gruber,J.D. and Long,A.D. (2006) A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics*, **7**, 130.
11. Cheng,T.C., Xia,Q.Y., Qian,J.F., Liu,C., Lin,Y., Zha,X.F. and Xiang,Z.H. (2004) Mining single nucleotide polymorphisms from EST data of silkworm, *Bombyx mori*, inbred strain Dazao. *Insect Biochem. Mol. Biol*., **34**, 523–530.
12. Mita,K., Morimyo,M., Okano,K., Koike,Y., Nohata,J., Kawasaki,H., Kadono-Okuda,K., Yamamoto,K., Suzuki,M.G. *et al*. (2003) The construction of an EST database for *Bombyx mori* and its application. *Proc. Natl Acad. Sci. USA*, **100**, 14121–14126.
13. Negre,V., Hotelier,T., Volkoff,A.N., Gimenez,S., Cousserans,F., Mita,K., Sabau,X., Rocher,J., Lopez-Ferber,M. *et al*. (2006) SPODOBASE: an EST database for the lepidopteran crop pest *Spodoptera*. *BMC Bioinformatics*, **7**, 322.
14. Parkinson,J., Anthony,A., Wasmuth,J., Schmid,R., Hedley,A. and Blaxter,M. (2004) PartiGene—constructing partial genomes. *Bioinformatics*, **20**, 1398–1404.
15. Thiel,T., Michalek,W., Varshney,R. and Graner,A. (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare L.*). *Theor. Appl. Genet*., **106**, 411–422.
16. Huntley,D., Baldo,A., Johri,S. and Sergot,M. (2006) SEAN: SNP prediction and display program utilizing EST sequence clusters. *Bioinformatics*, **22**, 495–496.
17. Parkinson,J., Guiliano,D.B. and Blaxter,M. (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics*, **3**, 31.
18. Ewing,B., Hillier,L., Wendl,M. and Green,P. (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*., **8**, 175–185.
19. Wasmuth,J.D. and Blaxter,M.L. (2004) prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics*, **5**, 187.
20. Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., 138–148, http://www.ch.embnet.org/software/ESTScan.html.
21. Fukunishi,Y. and Hayashizaki,Y. (2001) Amino acid translation program for full-length cDNA sequences with frameshift errors. *Physiol. Genomics*, **5**, 81–87.
22. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
23. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.
24. Gene Ontology Consortium. (2006) The gene ontology (GO) project in 2006. *Nucleic Acids Res*., 34, D322–D326.
25. Crosby,M.A., Goodman,J.L., Strelets,V.B., Zhang,P. and Gelbart,W.M. (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res*., **35**, D486–D491.
26. Papanicolaou,A., Joron,M., McMillan,W.O., Blaxter,M.L. and Jiggins,C.D. (2005) Genomic tools and cDNA derived markers for butterflies. *Mol Ecol*., **14**, 2883–2897.
27. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*., **132**, 365–386.
28. Woodhead,M., Russell,J., Squirrell,J., Hollingsworth,P.M., Mackenzie,K., Gibby,M. and Powell,W. (2005) Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. *Mol. Ecol*., **14**, 1681–1695.
29. Zhang,D.-X. (2004) Lepidopteran microsatellite DNA: redundant but promising. *Trends Ecol. Evol*., **19**, 507–509.
30. Shevchenko,A., Sunyaev,S., Loboda,A., Shevchenko,A., Bork,P., Ens,W. and Standing,K.G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem*., **73**, 1917–1926.
31. Salzberg,S.L. (2007) Genome re-annotation: a wiki solution? *Genome Biol*., **8**, 102.
32. Mungall,C.J. and Emmert,D.B. FlyBase Consortium (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
33. Pringle,E.G., Baxter,S.W., Webster,C.L., Papanicolaou,A., Lee,S.F. and Jiggins,C.D. (2007) Synteny and chromosome evolution in the Lepidoptera: evidence from mapping in *Heliconius melpomene*. *Genetics*, **177**, 417–426.