172–175 *Nucleic Acids Research, 2003, Vol. 31, No. 1*
DOI: 10.1093/nar/gkg094

The FlyBase database of the *Drosophila* genome projects and community literature

The FlyBase Consortium*

FlyBase, The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA

Received September 16, 2002; Accepted September 27, 2002

ABSTRACT

FlyBase (http://flybase.bio.indiana.edu/) provides an integrated view of the fundamental genomic and genetic data on the major genetic model *Drosophila melanogaster* and related species. FlyBase has primary responsibility for the continual reannotation of the *D. melanogaster* genome. The ultimate goal of the reannotation effort is to decorate the euchromatic sequence of the genome with as much biological information as is available from the community and from the major genome project centers. A complete revision of the annotations of the now-finished euchromatic genomic sequence has been completed. There are many points of entry to the genome within FlyBase, most notably through maps, gene products and ontologies, structured phenotypic and gene expression data, and anatomy.

SCOPE

The fruit fly, *Drosophila melanogaster*, is one of the most studied eukaryotic organisms and a central model for the human genome project. During the calendar years 2001 and 2002, the euchromatin of the *D. melanogaster* genome has been finished to high genome project sequencing standards by the Berkeley Drosophila Genome Project (BDGP: http://www.fruitfly.org/). FlyBase has re-evaluated the annotations of the genome based on the finished sequence and on a great deal of newly-available cDNA and protein similarity data. Along with its other responsibilities, FlyBase is committed to maintaining up-to-date annotations of the *D. melanogaster* genome.

Overview

The taxonomic scope of FlyBase is the family Drosophilidae. Currently, the vast majority of data concern the one species *D. melanogaster*. A considerable amount of comparative genomic data on *Drosophila pseudoobscura* is expected to appear from the whole genome shotgun sequencing currently being undertaken by the Baylor College of Medicine (BCM-HGSC) Human Genome Sequencing Center (http://hgsc.bcm.tmc.edu/projects/drosophila/update.html.) FlyBase represents abstracted and value-added curated genetic and genomic data from the *Drosophila* 'literature', i.e. from the genome centers, published scientific literature, accessions from nucleic acid, protein and other databases, written personal communications and bulk submissions. All information in FlyBase is attributed, meaning it is attached to a specific bibliographic citation. It is an important function of FlyBase to attempt to integrate and standardize this literature, particularly in the area of usage of structured terminology (ontologies and nomenclature).

While the genome sequence of the euchromatin of *D. melanogaster* is essentially complete, the predictions of the transcription units and protein products will continue to be in flux for some time, and the relationship of these molecularly defined genes to those defined solely by phenotypic criteria is necessarily incomplete. For example, while current annotations predict about 13 500 protein-coding genes, there are ~9000 genes known only through phenotypic or expression criteria that have yet to be connected to any gene model.

The central data sets

FlyBase organizes genetic and genomic data on chromosomal sequences and map locations, on the structure and expression patterns of encoded gene products, on mutational and transgenic variants and their phenotypes. The publicly-funded stock centers are included, as are numerous crosslinks to sequence databases and to homologs in other model organism databases. Images and graphical interfaces within FlyBase

*Correspondence should be addressed to William M. Gelbart, FlyBase, The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA. Tel: +1 6174952906; Fax: +1 6174961354; Email: gelbart@morgan.harvard.edu

Current members of the FlyBase Consortium are W. Gelbart, L. Bayraktaroglu, B. Bettencourt, K. Campbell, M. Crosby, D. Emmert, P. Hradecky, Y. Huang, S. Letovsky, B. Matthews, S. Russo, A. Schroeder, F. Smutniak, P. Zhou, and M. Zytkovicz (Biological Laboratories, Harvard University, Cambridge, MA); M. Ashburner, R. Drysdale, A. de Grey, R. Foulger, G. Millburn and C. Yamada (Department of Genetics, University of Cambridge, Cambridge, UK); T. Kaufman, K. Matthews, D. Gilbert, G. Grumbling and V. Strelets (Department of Biology, Indiana University, Bloomington, IN); C. Shemen (NCBI, Bethesda); G. Rubin, B. Berman, E. Frise, M. Gibson, N. Harris, J. Kaminker, S. Lewis, B. Marshall, S. Misra, C. Mungall, S. Prochnik, J. Richter, C. Smith, S. Shu, J. Tupy and C. Wiel (University of California, Berkeley, CA, and Lawrence Berkeley National Laboratories)

include interactive and static regional maps, as well as anatomical drawings and photomicrographs.

FlyBase identifier numbers

Many data classes in FlyBase have unique identifiers in FlyBase. These allow FlyBase data objects to be cross-referenced, both within FlyBase and externally. FlyBase identifiers are of the form: FBxxnnnnnnn, where xx is a two-letter code signifying the type of identifier, and nnnnnnn is a 7-digit number padded with leading zeros. For example, a gene identifier would take the form of FBgn0001234. Annotations of the genome within the GadFly (Genome Annotation Database) section of FlyBase are identified in the form FBan0012345.

FlyBase attribution

A key feature of FlyBase is a comprehensive bibliography of conventional and unconventional publications (e.g. films, archival material and even newspaper articles) on the family Drosophilidae, covering all aspects of its study.

Organization of FlyBase data

Some classes of FlyBase data are completely structured. These include nomenclature, map data, cross-links to external databases and a variety of controlled vocabularies (CVs); lists of the CVs are available in the Documents section of FlyBase. Free text data are included for data classes not easily constrained into CVs.

Gene Ontology (GO) data in FlyBase

FlyBase is one of the founding participants in the GO consortium (http://www.geneontology.org/), which provides CVs for the description of the molecular functions, biological processes and cellular components of gene products. FlyBase indexes gene records with GO terms and the attributed evidence. From the FlyBase home page, GO indices are accessed via a link termed 'Function, Location, Process, Structure'. Both searching and browsing of the GO terms in FlyBase links one to the gene records that include each GO term. The FlyBase Genes Search provides GO term search options for Product_function, Product_process and Cellular_component.

Phenotypic data in FlyBase

Phenotypic data are attached to the alleles of genes (1). These data are represented by a combination of free text and CVs describing 'Phenotypic class' and 'Anatomy'. Mutant phenotype data are now partitioned into that which pertains to mutant alleles of one gene, or that which pertains to genetic interactions (that is, the phenotypes of multiply-mutant genotypes). The new Genetic Interaction data class uses the same controlled vocabularies as the 'Phenotypic class' and 'Anatomy' data classes, but combines the terms with a conditional genotype syntax. This syntax expresses the salient features of the interaction, namely whether the interaction is suppressive or enhancing, and the identity of the interacting allele. The Allele Search form is designed to maximize the efficiency of interrogation of phenotypic data. Alternative routes into mutant phenotype make use of the Anatomy CVs (see below).

Anatomy

Information relating to anatomy can be found in the 'Anatomy&Images' section of FlyBase. Data that are associated with anatomical features can be retrieved by searching or browsing the anatomical CV. These CV terms are linked to reports that include other FlyBase objects annotated with the specified term, thus providing, for example, an integrated view of alleles with phenotypes affecting the same anatomical entity or reporter genes expressed in that tissue or structure. A collection of anatomical images annotated with many of these terms can be browsed, allowing access to anatomy-related data without knowledge of specific anatomical terms.

FlyBase Genome Annotations

The backbone sequence (or 'reference sequence') for FlyBase gene annotations is now the finished Release 3 BDGP/BCM-HGSC/Celera euchromatin genomic sequence of *D. melanogaster*. FlyBase has recently completed a comprehensive re-annotation of Release 3. Several major sources of improved or additional data have been used in this effort: the finished high-quality Release 3 genomic sequence itself, the greatly expanded BDGP EST/cDNA collection (2), additional recent *Drosophila* cDNA data submitted to GenBank by the community, error reports from FlyBase users, ARGS literature-enhanced annotations from FlyBase, additional protein sequence information extracted from the major protein databases, and GENSCAN (in addition to Genie) gene prediction data (3,4). The newly annotated sequence includes a complete set of predicted tRNA genes (5), as well as the members of other classes of non-protein-coding genes (snRNAs, snoRNAs, microRNAs) that have been publicly reported. Unlike Release 2, Release 3 contains accurate sequence for all transposable elements in the sequenced $y^1;cn^1bw^1sp^1$ strain. During re-annotation, every gene model was subjected to visual evaluation, using the graphical genomic feature editor Apollo. Several manuscripts describing various aspects of Release 3 and the reannotation project have been submitted and are expected to be published by the end of 2002.

The view of the genome that has emerged from the reannotation data has approximately the same number of predicted protein-coding genes, but there are considerable differences in gene structures. Forty three percent of the approximately 16 000 unique peptides in Release 3 fail to exactly match a peptide in Release 2. Approximately 10% of existing gene models underwent major changes, such as division into two genes, or the merger of previously separate genes. In many cases, exons were added or deleted, and the majority of gene models were improved at the level of precision of exon/intron boundaries or extension to include terminal untranslated regions (UTRs) of transcripts. Thanks to the vastly increased set of cDNA and EST data, 75% of genes now contain annotated UTR sequences, and 20% of genes include two or more alternative transcripts. Some atypical gene models were also uncovered, for example, dicistronic transcripts and pairs of genes with overlapping UTRs on the same strand.

The updated gene models can be viewed on new Genome Annotation pages, which can be accessed from the FlyBase gene reports (click on the annotation number listed under 'Genomic sequence analysis') or from the annotation query page (http://www.fruitfly.org/cgi-bin/annot/query). In addition to a detailed view of the transcripts produced by a given gene, this viewer presents the data supporting the gene model, any comments added by the annotator and a thumbnail view of the immediate genomic region. There are links to the reports for adjacent genes, to FASTA files of protein and transcript sequences, to the GenBank entry for the genomic region and to the FlyBase interactive Genome Browser view of the surrounding region. In addition to access through FlyBase, these annotations are contributed to the NCBI Reference Sequence (RefSeq) project and are accessible through NCBI.

FlyBase will continue to improve gene annotations, but future changes will be on a gene-by-gene basis, rather than a comprehensive survey of the entire genome. Several large-scale data sets will be added to our analysis: other insect genomes (the existing *Anopheles gambiae* and anticipated *D. pseudoobscura* and *Apis mellifera* genomic sequences) and the recently completed *Drosophila* Gene Collection (DGC) cDNA sequences (2). In addition, a high priority will be placed on extracting additional sequence features from GenBank entries and the published literature, such as regulatory elements and mutational lesions, expanding the existing ARGS curation pipeline. Improvement of the annotations will continue to depend on community input through the literature and personal communication to FlyBase.

Interrogating FlyBase

FlyBase data are organized into a variety of data classes for ease of access. Query tools that permit field-specific searches, combinatorial queries and menu-driven selection of CVs are available. Organized lists of 'hits' to a given query are produced, and single or multiple items from these hit lists can be retrieved. A Synopsis report is first produced and the user is provided with several options for more extensive reports.

IMPLEMENTATION

FlyBase currently consists of multiple, tightly coupled relational datasets. Much of the literature-curated data are housed in a Sybase RDBMS. Most FlyBase data sets are accessible in computable formats and numerous options for downloading bulk data are available. GadFly annotations in GFF or XML format or sequences in multiple FASTA format are also available for downloading. The entire GadFly database (as MySQL) and software used to generate, analyze and display GadFly data are freely available as Open Source software. FlyBase is currently undertaking an integration of the literature and genome annotation databases into a single RDBMS.

FlyBase is a member of the GMOD collaboration to produce a generic set of tools for construction of new model organism databases (http://www.gmod.org/) and all future software and schema development will occur within the GMOD framework.

FlyBase provides users with a variety of modes of access including the major web interface and FTP for downloading of files. The primary FlyBase server has the following addresses: http access, http://flybase.bio.indiana.edu/; ftp access, ftp://flybase.bio.indiana.edu/flybase/.

FlyBase mirror sites are available in Europe, Asia, Australia and the USA; a complete listing can be found in the FlyBase Mirrors section. While network problems can be unpredictable, in general, FlyBase recommends that users connect to a mirror site that provides the most rapid response time.

DOCUMENTATION

Complete Nomenclature and FlyBase Reference Manuals are available from FlyBase servers in html format.

Announcements of major FlyBase updates are made through postings to the bionet.drosophila bulletin newsgroup. FlyBase users are encouraged to use this newsgroup to track changes to FlyBase.

ADDRESSES

Interaction with the user community is vital for the success of FlyBase. We encourage the submission of new data, the correction of errors, and ideas for making this database of even greater use to the community.

Requests for help and questions about FlyBase should be addressed to flybase-help@morgan.harvard.edu. Reports of errors in FlyBase or data updates, should be addressed to flybase-updates@morgan.harvard.edu. Mail may be addressed to FlyBase, Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.

REFERENCING FLYBASE

We suggest that FlyBase be referenced as follows: FlyBase (2003). The Fly Base database of the Drosophila genome projects and community literature. Available from http://flybase.bio.indiana.edu/. *Nucleic Acids Res.*, **31**, 172–175.

We suggest that the abbreviation FB be used for FlyBase, regardless of the particular FlyBase product.

ACKNOWLEDGEMENTS

FlyBase is supported by grants from the National Institutes of Health (National Human Genome Research Institute) to W. M. Gelbart and the Medical Research Council, London to M. Ashburner. The Berkeley *Drosophila* Genome Project is supported by grants from the Department of Energy and from the National Institutes of Health (National Human Genome Research Institute) to G. M. Rubin.

REFERENCES

1. Drysdale,R. (2001) Phenotypic data in FlyBase. *Brief. Bioinform.*, **2**, 68–80.

2. Stapleton,M. *et al.* (2002) The *Drosophila* Gene Collection: identification of putative full-length cDNAs for 70% of *D. melanogaster* genes. *Genome Res.*, **12**, 1294–1300.
3. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
4. Reese,M.G. *et al.* (2000) Genie—Gene finding in *Drosophila melanogaster*. *Genome Res.*, **10**, 529–538.
5. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.