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 17, 2001; Accepted September 18, 2001

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. Following on the success of the *Drosophila* genome project, 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. The current cycle of reannotation focuses on establishing a comprehensive data set of gene models (i.e. transcription units and CDSs). There are many points of entry to the genome within FlyBase, most notably through maps, gene 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. A landmark event in the study of *D.melanogaster*, the publication of the whole genome shotgun assembly sequence of the genome and the initial assessment of its structure and encoded products, occurred in March 2000 (1,2). Along with its other responsibilities, FlyBase is committed to maintaining up-to-date annotations of the genome.

Overview

The taxonomic scope of FlyBase is the family Drosophilidae. Currently, the vast majority of data concern the one species *D.melanogaster*, although this may change as a result of anticipated comparative sequencing of *Drosophila pseudoobscura*. 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 nearing completion, 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 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 the database. 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 seven-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,

*Correspondence should be addressed to: William M. Gelbart, Harvard University, The Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 02138, USA. Tel: +1 617 495 2906; Fax: +1 617 496 1354; Email: gelbart@morgan.harvard.edu

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

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, crosslinks 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, which provides CVs for the description of the molecular functions, biological processes and cellular components of gene products (3). FlyBase indexes gene records with GO terms and the supporting evidence. In the last year all of the annotations to GO that were computed during the Celera annotation jamboree (1,2) have been reviewed and amended where necessary. 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 (4). 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

For the last 3 years, FlyBase has produced Annotated Reference Gene Sequences (ARGS), which present an integrated and synthesized view of published sequence and physical map features, incorporating data from the primary literature and from community sequence submissions. These FlyBase annotations are contributed to the NCBI Reference Sequence (RefSeq) project.

These annotations include a wide range of data types, including locations of mutational lesions, aberration breakpoints, transposon insertion sites and rescue fragments, as well as transcript structures and predicted proteins. The backbone sequence (or 'reference sequence') for FlyBase gene annotations is now the finished Release 3 genomic sequence, based upon the Berkeley *Drosophila* Genome Project (BDGP), European *Drosophila* Genome Project (EDGP) and Celera sequencing effort. ARGS do not include gene prediction data; these reports include only experimentally validated sequence-level features. FlyBase is continuing to create such integrated annotations for genes for which community data are available.

Only a small subset of gene reports include such a rigorous gene annotation; it can be accessed from the top page of the gene report by clicking on the thumbnail presentation labelled 'FlyBase gene annotation'. Linked to the thumbnail is a full annotation graphic, which in turn is linked to a textual report of the reference sequence and its features expressed in standardized FlyBase format and linked to appropriate FlyBase reports.

Currently, for the vast majority of genes, there is insufficient experimental information to produce ARGS. Instead their gene structures are based on computational analyses merging EST and similarity data with the output of gene prediction programs. Those annotations based on the whole genome shotgun assembly (1) of the genome (termed genome sequence Releases 1 and 2) are accessed through GadFly, the FlyBase genome annotation database.

The BDGP has finished sequencing the euchromatin of the *D.melanogaster* genome, and this finished version of the genome is referred to as Release 3. Release 3 sequences have been reanalyzed computationally using gene prediction programs, alignments to known fly sequences (recently augmented by >200 000 ESTs and by full-length cDNA sequence of the *Drosophila* Gene Collection), and by cross-species similarities identified by BLAST searches. Hundreds of user-submitted Release 1 and 2 error reports are included in the re-evaluation of the annotations. Computational annotation has been performed by several collaborating groups and these data are presented to FlyBase curators for the final step of expert evaluation.

The results of all of these analyses are stored in the Release 3 version of GadFly. Individual gene annotations can be viewed, including the evidence on which the annotation is based. Each page also includes links to interactive BLAST similarity searches and InterPro protein domain analysis of the predicted proteins. Where available, matches to cDNA clones from the BDGP cDNA/EST project are provided. The scope of this project is such that errors are inevitable, and error report entry forms are provided for community input and correction.

This Release 3 reannotation has focused on improving the protein-coding genes' exon–intron structure. Future rounds will tie more features to the genomic sequence, including P-element insertions and features identified by comparative genomic analysis of the expected *D.pseudoobscura* genomic sequence. Improvement of the annotations will depend increasingly 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 data sets. Much of the curated data employ 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 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/.

Users are also referred to the BDGP and GO Consortium web sites for additional related information: BDGP, http://www.fruitfly.org/; GO Consortium, http://www.geneontology.org/.

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

When referencing this article, we suggest that the following URL is included: http://flybase.bio.indiana.edu/.

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) and the Medical Research Council, London. The BGDP 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. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
2. Rubin,G.M., Yandell,M.D., Wortman,J.R., Gabor Miklos,G.L., Nelson,C.R., Hariharan,I.K., Fortini,M.E., Li,P.W., Apweiler,R., Fleischmann,W. *et al.* (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.
3. Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
4. Drysdale,R. (2001) Phenotypic data in FlyBase. *Brief. Bioinform.*, **2**, 68–80.