Nucleic Acids Research, 1999, Vol. 27, No. 1 85–88

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 October 5, 1998; Accepted October 8, 1998

ABSTRACT

The FlyBase Drosophila genetics database and the public interfaces of the Berkeley Drosophila Genome Project (BDGP) and European Drosophila Genome Project (EDGP) are in the process of integrating. At present, the data of these projects are available from independent, but hyperlinked, WWW sites (FlyBase URL, http://flybase.bio.indiana.edu/ ; BDGP URL, http://fruitfly.berkeley.edu/ ; EDGP URL, http://edgp.ebi.ac.uk/). Because of the considerable overlap of data classes between the contributions of the Drosophila genome projects and the Drosophila community, the new and enlarged FlyBase consortium views the implementation of a single integrated Drosophila genomics/genetics server as essential to the scientific community. This integration will occur in a stepwise fashion over the next 1–2 years. In this report, the salient features of the current databases and how to interrogate and navigate the extensive data sets are discussed.

BACKGROUND

The fruit fly, *Drosophila melanogaster*, is one of the most studied eukaryotic organisms and a central model for the human genome project. Introduced as an experimental organism in the early years of this century, the genetic prowess of *D.melanogaster* has placed this fly at the forefront of many areas of research, notably gene regulation, chromosome behavior, developmental biology, cell biology, neurobiology, population biology, ecology and evolution. Beginning in 1992, data on the genetics and genomics of *D.melanogaster* and related species have been electronically available over the Internet through the funded FlyBase, BDGP and EDGP informatics groups. These groups recognize that many (perhaps all) genome project and community data types overlap considerably, and that it would be of great value to present the scientific community with an integrated view of these data. Beginning in 1998, the BDGP, EDGP and FlyBase have begun the process of merging their data distribution efforts into a single project. In the short term, however, the data will be accessible only through independent servers, and so it is the purpose of this report to describe current aspects of the FlyBase, BDGP and EDGP public databases. However, it should be realized that the merger is in progress, and the specific aspects of database access will be in flux for the next several months. Some new merged efforts are also discussed.

SCOPE

The classes of data

The taxonomic scope of FlyBase is the family Drosophilidae. However, the vast majority of data concerns the one species *D.melanogaster*.

FlyBase curation of community literature

FlyBase represents abstracted and value-added curated genetic and genomic data from the Drosophila 'literature', i.e., from the published scientific literature, accessions from nucleic acids, protein and other databases, written personal communications, and bulk submissions. All information in FlyBase is attributed, meaning that it is attached to a specific bibliographic citation. The current view of the *D.melanogaster* genome and the functions of individual genes is necessarily incomplete. Given this incompleteness, and the heterogeneity of the literature, it is an important function of FlyBase to attempt to integrate the literature, particularly by the use of structured terminology (controlled vocabularies and nomenclature).

The genetic data sets. FlyBase organizes genetic data in terms of chromosomal locations in the genome, and in terms of the products that the genes encode. Gene descriptions therefore include recombinational, cytogenetic, physical and sequence level information on map locations, functional, expression pattern and structural information on gene products, and biological roles

*The current members of the FlyBase Consortium are: W. M. Gelbart, M. Crosby, B. Matthews, J. Chillemi, S. Russo Twombly, D. Emmert, L. Bayraktaroglu, F. Smutniak and S. Kossida (Biological Laboratories, Harvard University, Cambridge, MA, USA); M. Ashburner, R. A. Drysdale, E. J. Whitfield, G. H. Millburn and A. de Grey (Department of Genetics, University of Cambridge, Cambridge, UK); T. Kaufman, K. Matthews, D. Gilbert, V. Strelets and G. Grumbling (Department of Biology, Indiana University, Bloomington, IN, USA); C. Tolstoshev (NCBI, Bethesda, MD, USA); G. M. Rubin, S. Lewis, G. Helt, S. Misra, N. Harris, P. Brokstein, A. Loraine and D. Simas (University of California, Berkeley, CA, USA); T. Benos (European Bioinformatics Institute, Hinxton, Cambridge, UK).

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

based upon inferences from mutant phenotypes. Information on specific mutant alleles, chromosomal aberrations, transposons and transgenic insertions are also included, as are the strain lists of the publicly-funded stock centers and a few private collections. Because of the importance of related information from other systems, FlyBase includes numerous crosslinks to other databases, such as to accessions in nucleic acids and protein databases, and records of homologs in other species databases. As the FlyBase data set expands, it is becoming increasingly important to provide graphical data summaries wherever possible. Images within FlyBase include regional cytogenetic, physical and sequence level maps, as well as anatomical drawings and photomicrographs.

Ancillary data sets. In addition to these core data sets, FlyBase provides other services to the scientific community. Contact information for the *Drosophila* community is available. Sets of data that are not themselves curated by FlyBase, but that may be of interest to the community are maintained in an Allied Data division of the FlyBase server. These data are posted directly in the form submitted to FlyBase. Most notably, in its Allied Data section, FlyBase mirrors 'The Interactive Fly', a database of gene products and pathways produced by Tom and Judy Brody. In collaboration with the Brodys, FlyBase maintains a hierarchically organized index of cross-links between The Interactive Fly and FlyBase gene reports.

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, a bibliographic reference of the form FBrf0098765, and an RNA transcript of the form Fbtr0005678.

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. These controlled vocabularies include those describing the function and structure of gene products, subcellular and anatomical terms for describing phenotypes and gene expression patterns, mutagens, taxonomic species, transposon properties, etc. Other classes of FlyBase data are captured as free text. In general, the intention is that the structured data classes in FlyBase are the best approaches for interrogating the database and identifying relevant information, and the free text imparts additional information concerning these relevant items. While free text searches are available in FlyBase, their results are likely to be somewhat idiosyncratic, since by their nature, use of terminology is not constrained.

A new FlyBase data class—annotated gene reference sequences. As part of the expanded project, FlyBase is now curating reference sequences of the genome. These reference sequences are being curated on a gene by gene basis. In these reference sequence reports, typically the reference sequence is based on finished genome project clone sequence. It synthesizes sequence level and physical map information about the locations of transcripts, polypeptides, regulatory elements, mutant alleles, transgenic rescue fragments and other features that can be tethered onto the reference sequence. These data are captured from primary nucleic acids database accessions, from the primary literature and from personal communications. These reference sequence reports will be submitted to GenBank as FlyBase-authored annotated sequences, and will be cross-referenced to FlyBase feature tables.

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. This bibliography includes the complete texts of all of the published Drosophila bibliographies, and information from major external resources, such as MEDLINE, BIOSIS, the Zoological Record and the Environmental Mutagen Information Center (by permission). The bibliography is updated from these and other sources. To ensure consistency there is a satellite file of all 'multi-publication' sources, e.g., journals and edited publications, which includes full names, dates and places of publication, volume number ranges, and ISBNs or ISSNs and CODENS. By far the greater part of these data have been checked on the Library of Congress and British Library online catalogs. Bibliographic records are coded as to type (e.g., journal article, abstract, review, thesis, book, film).

FlyBase maintains a collection of offprints of publications on Drosophila, housed in Cambridge, UK. This collection is cross-referenced with the bibliography, and copies of obscure publications can be supplied on request.

Berkeley Drosophila Genome Project

Genomic sequence, clone libraries and physical maps. The BDGP is currently sequencing the euchromatic sequence of chromosomes 2 and 3. Partially sequenced clones are submitted weekly to the Phase 1 High Throughput Genomic (HTG) division of GenBank. Finished clones are submitted to the phase 3 division first as unannotated sequences, and in the near future as annotated entries. Finished clones are being analyzed computationally followed by human curation of those computed results. Sequence data on these clones are available on the Berkeley Fly Database (BFD) server; all available sequence data are in the public nucleic acids sequence databanks (GenBank/EMBL/DDBJ).

Libraries of clones serve both as primary sequencing templates and as reagents to the community. At this time, complete reports about these clones can be obtained only from the genome project Web sites, most completely using the Berkeley Fly Database search tools. There, data are available on the BDGP bacteriophage P1/BAC (bacterial artificial chromosome) clones and physical map, the EDGP cosmid/BAC clones and maps, and YAC (yeast artificial chromosomes) clones localized to the genome by *in situ* hybridization.

Data on STS's, also deposited in the dbSTS section of GenBank, are available on the genome project servers. Sequence tagged sites (STSs) are short sequences that were used to detect clone overlap and thereby, to construct the physical maps. Most come from sequencing of the ends of genomic clones, but others come from the sequence of known genes or regions flanking P element insertions.

Expressed sequence tags (EST) and full length cDNAs. The EST section of FlyBase links to data from two different BDGP projects: the large BDGP/HHMI project sequencing the ends of cDNAs from libraries with a high percentage of full-length

clones, and a small-scale project analyzing cDNAs derived from membrane-associated polysomes.

The BDGP/HHMI full-length cDNA EST project's goal is to sequence the 5′ ends of 80 000 cDNA clones made from high quality Drosophila cDNA libraries. The long-term goal is to generate the full-length sequence of representative clones for each gene and compare them to the genomic sequence to generate a transcript map of the genome. These full-length sequences will be available for querying and are being submitted to GenBank.

The membrane-associated embryonic cDNA clones, which are in general not full length, have been analyzed by whole embryo *in situ* hybridization. Data from this library can be queried for clones expressed in particular tissues using the FlyBase controlled vocabularies.

Insertion lines. In order to correlate genetic and molecular maps of the genome the BDGP is characterizing Drosophila strains containing insertions of two types of P transposable elements. The first are insertions of P elements that have been selected because the insertion results in a readily detected phenotype such as lethality. The second are insertions of elements that allow controlled misexpression of a gene at the site of insertion. Flanking sequence data and phenotypic data, and in some cases expression data, are collected from these lines. The flanking sequence data are compared with both the genomic and EST sequences to precisely map insertion sites and to associate the phenotype produced by an insertion with a particular open reading frame.

European Drosophila Genome Project

The EDGP is sequencing the distal tip of the X chromosome from cosmid and BAC clones. Finished clone sequences are deposited in the EMBL DNA Sequence Data Library, first as unannotated Phase 3 sequences and then as annotated sequences, where genes and other sequence features have been predicted by computation. As soon as sequences are annotated, these are curated by FlyBase for identified genes and the relevant genetic information curated in FlyBase gene records.

INTERROGATING FlyBase DATA

FlyBase community literature

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 controlled vocabularies are available. A variety of tools permit graphical or textual querying by map location as well. Organized lists of 'hits' to a given query are produced, and single or multiple items from these hit lists can be retrieved. Unless the full report of an item is itself relatively short, a summary (brief) report is first produced and the user is provided with several options for more extensive reports. Very large reports, particularly of some extensively-studied genes, exceed 1 Mb in size and might take some time to download. FlyBase is investigating approaches for breaking up such large reports.

Anatomical information (querying for phenotypes or gene expression patterns) will soon be accessible through graphical interfaces in which anatomical diagrams are used to enable queries of expression patterns or phenotypes by anatomical term. Hierarchical views of these anatomical terms can also be used to support expression pattern queries.

Berkeley Drosophila Genome Project

There are three methods available for searching and locating reports on the different BDGP data classes: query formulation, graphical browsing or preformatted tables. Simple queries can be made using a standard html form that allows some qualifications on the query. The Java-based graphical views of the genome enable users to search based on cytological or sequence location. Further, users can dynamically enlarge or reduce the view, and select customized subsets of the data classes to be displayed. The tabular reports provide summaries of information on sequencing status.

Reports on all genomic clones (P1, BAC, cosmid or YAC), STSs, and BDGP P element insertions can be found using the BDGP BFD textual or graphical query tools. These reports also link to information and graphical displays of the physical contigs made up of these clones. When available, the sequences of the genomic clones are presented as well as those of large contiguous regions formed by joining the sequences of individual clones.

The 'BFD Map Viewers' item on the BDGP home page links to Java displays which allow interactive browsing of the genome at progressively higher resolution. At any level of viewing, from whole genome to the DNA sequence, users can select an individual graphic entity to retrieve a text report describing that data item. In some cases, these text reports operate through links to the FlyBase report of that entity. ChromoView and ArmView show the physical map at any cytological position in the genome, with physical contigs shaded as to whether sequence information is available. These contigs link to the physical map viewer, CytoView, which diagrams the STS links between genomic clones that make up physical contigs. Clones and STSs link to text reports, and the Clone reports link to the CloneView display of the annotated genomic sequence for that particular clone. For those clones inspected by a human curator, the most relevant results for each predicted and previously known gene identified in the sequence will be summarized and labelled with formal gene names or valid FlyBase symbols.

Users typically identify ESTs they are interested in by sequence similarity to a query sequence, for example by using the FlyBase BLAST server. The results of such a BLAST search will hyperlink to 'clot' reports for the positive match, a clot being a group of cDNA clones with similar EST sequences. A Java alignment viewer in the clot report allows one to inspect the alignment of EST sequences in detail. Since the clots are also analyzed for sequence similarities using BLAST, the results of these searches can also be browsed by querying by key word. Finally, the EST Query page also allows one to access details about an individual clone using its clone or accession number.

European Drosophila Genome Project

The EDGP and BDGP jointly maintain BLAST servers for querying Drosophila sequence data. These servers are available from the EDGP and BDGP home WWW pages. EDGP cosmid and BAC clone data are integrated with the BDGP clone data, and can be accessed from the BFD search page.

FUTURE PLANS

The independent FlyBase, BDGP and EDGP servers are gradually merging into a single public database. Integrated reports of genome project and community information will be provided. The details of the layouts and formats of these pages, and the textual and graphical querying and browsing tools will be evolving.

IMPLEMENTATION

Community literature

FlyBase is built with a relational database management system (Sybase). The present schema has been implemented for most of the data and most files accessed via the FlyBase servers are the products of the Sybase tables.

FlyBase data are maintained by curators working from the literature and filling in standard forms that are parsed into the Sybase tables.

Berkeley Drosophila Genome Project

The BDGP public database (the Berkeley Fly Database) utilizes Informix as its database server. This database is directly accessed from the web via CGI scripts to respond to queries and generate reports. The EST portion of the data is still maintained in an earlier Illustra database and is accessed by the same techniques.

The BDGP data are the culmination and synthesis of output from various laboratory information management systems that rely upon a variety of different technologies reflective of each project's history, these include: Flydb (an ACeDB variant), Filemaker, 4D, Illustra, and Informix.

European Drosophila Genome Project

The EDGP production data are maintained in ACeDB. Downloads of these data are sent weekly to the BDGP and the BDGP reciprocates. These weekly exchanges keep the two BLAST servers and ftp sites in synchrony.

ACCESS

Community literature

The primary FlyBase server has the following addresses:

http://flybase.bio.indiana.edu/	http access
flybase.bio.indiana.edu 72	gopher access
flybase.bio.indiana.edu (in /flybase)	ftp access
flybase-gopher@indiana.edu	Email access

Mirror sites are available in Europe, Asia, Australia and the USA. The FlyBase Views section indicates available mirrors. While network problems can be unpredictable, in general FlyBase recommends that users connect to a mirror site that provides the most rapid response time.

FlyBase WWW access now provides users with the option of customized report formats and database interconnections.

Berkeley Drosophila Genome Project

http://fruitfly.berkeley.edu/	http access
fruitfly.berkeley.edu	ftp access

All sequence data are mirrored at the EDGP site.

European Drosophila Genome Project

Progress, and direct access to sequence data and annotations, are available from http://edgp.ebi.ac.uk/ . All sequence data are mirrored at the BDGP site.

DOCUMENTATION

A FlyBase Reference Manual is available from FlyBase servers in html format. A brief introduction, 'Getting started with FlyBase', is included as the first section of the Reference Manual. Announcements of major database updates and new tools are made through postings to the bionet.drosophila newsgroup. FlyBase users are encouraged to use this newsgroup to track changes to FlyBase.

Descriptions of the BDGP graphical displays, frequently asked questions about the laboratory projects and methods are available in html format at the BDGP web site. Announcements of database updates and new features are posted on the home page.

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. All such mail is automatically sent to all members of the FlyBase collaboration, including the newly incorporated genome projects.

REFERENCING FLYBASE

We suggest that FlyBase be referenced as follows: FlyBase (1999). The FlyBase Database of the Drosophila Genome Projects and Community Literature. Available from http://flybase.bio.indiana.edu/ . *Nucleic Acids Res.*, **27**, 85–88.

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 Berkeley Drosophila Genome Project is supported by a grant from National Institutes of Health (National Human Genome Research Institute), the Department of Energy to G. M. Rubin. Funding for the BDGP EST project is provided by the Howard Hughes Medical Institute. The European Drosophila Genome Project is supported by a contract from the European Union (coordinated by D. M. Glover).