Nucleic Acids Research, 1997, Vol. 25, No. 1 63–66

FlyBase: a *Drosophila* database

The FlyBase consortium*

FlyBase, Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 02138, USA

Received September 19, 1996; Accepted September 20, 1996

ABSTRACT

FlyBase is a database of genetic and molecular data concerning *Drosophila*. FlyBase is maintained as a relational database (in Sybase) and is made available as html documents and flat files. The scope of FlyBase includes: genes, alleles (and phenotypes), aberrations, transposons, pointers to sequence data, clones, stock lists, *Drosophila* workers and bibliographic references. The *Encyclopaedia of Drosophila* is a joint effort between FlyBase and the Berkeley *Drosophila* Genome Project which integrates FlyBase data with those from the BDGP.

BACKGROUND

Drosophila melanogaster is one of the most studied eukaryotic organisms. Introduced to modern biology in the early years of this century, research with *D.melanogaster* has been at the forefront of most areas of biology, from genetics to ecology, from neurobiology to evolution. *Drosophila* geneticists have been well served by a series of catalogs of mutations, the first of which was published in 1925, and by regular publication (again, dating from 1925) of bibliographies of the *Drosophila* literature. The last conventional catalog of the genes and mutations of *D.melanogaster* was published in 1992 (3), although data collection ceased in late 1989.

Beginning in October 1992, the National Center for Human Genome Research of the NIH has funded the FlyBase project with the objective of designing, building and releasing a database of genetic and molecular information concerning this insect. FlyBase also receives support from the Medical Research Council, London.

SCOPE

The core of FlyBase is data concerning the genes and mutations of *Drosophila*:

- Gene name; gene symbol; synonym(s) for name; synonym(s) for symbol; genetic map position; polytene chromosome map position; nature of gene product(s); molecular data; gene expression pattern data; similar genes in other organisms; database cross-references.
- Allele(s) name; allele(s) symbol; allele name synonym(s); allele symbol synonym(s); origin of allele; phenotypic information; molecular data.
- Chromosome aberrations.
- Clones (cosmids, P1s, YACs).
- Molecular data.
- P-elements.
- Transposon constructs and their components.
- Bibliographic references.
- Stock lists.
- People.
- Allied data.
- FlyBase identifier numbers.

In their present form these data are a mixture of information in a highly structured, parseable form and in free text. All of the data are available as flat (ASCII) files, the majority being the output of selected data sets from the relational database implementation of FlyBase.

The taxonomic scope of FlyBase is the family Drosophilidae. Data from all species is now curated, but the historical data is only reasonably complete for *D.virilis* (with the help of H. Kress), *D.ananassae* (from Y. N. Tobari) and *D.buzzatii* (from J. S. F. Barker).

FlyBase identifier numbers

All data classes listed below have unique identifiers in FlyBase. These allow them to be 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. Identifier codes now used are:

FBgn	gene identifier (e.g. FBgn0001234)
FBal	allele identifier
FBab	aberration identifier
FBrf	bibliographic reference identifier
FBsp	species identifier

*The current members of the FlyBase Consortium are: W. M. Gelbart, M. Crosby, B. Matthews, W. P. Rindone, J. Chillemi, S. Russo Twombly and D. Emmert, Biological Laboratory, Harvard University, Cambridge, MA, USA; M. Ashburner, R. A. Drysdale, E. Whitfield, G. H. Millburn and A. de Grey, Department of Genetics, Downing Street, Cambridge CB2 3EH, UK; T. Kaufman, K. Matthews, D. Gilbert and V. Strelets, Department of Biology, Indiana University, Bloomington, IN, USA; C. Tolstoshev, NCBI, Bethesda, MD, USA

*Correspondence should be addressed to M. Ashburner, Department of Genetics, Downing Street, Cambridge University, Cambridge, CB2 3EH, UK. Tel: +44 1223 333969; Fax: +44 1223 333992; Email: m.ashburner@gen.cam.ac.uk

FBmc	construct identifier
FBba	balancer identifier
FBtp	engineered transposon identifier
FBti	transposon insertion identifier
FBtr	transcript identifier
FBpr	protein identifier
FBms	molecular component identifier

Genes, alleles and aberrations

In November 1996 FlyBase included information on 14 622 genes (11 375 *D.melanogaster*), 35 384 alleles (32 311 *D.melanogaster*) and 13 366 chromosome aberrations (13 031 *D.melanogaster*). Except for historical data, inherited from Lindsley and Zimm (3) for example, all data are attributed to a single publication (including personal communications to FlyBase; these are archived and made accessible to users). Search tools allow data on genes, their alleles and aberrations to be retrieved by a simple query. The data are a mixture of free text and controlled syntax. FlyBase uses a standard controlled vocabulary of terms to describe, for example, mutagens and anatomical parts of *Drosophila*.

Nomenclature and synonyms

The genetic nomenclature of *D.melanogaster* is chaotic (though perhaps not in the technical sense of this word). FlyBase has written and maintains a document on nomenclatural standards for the community (flybase/nomenclature/nomenclature.txt, also available as an html document from FlyBase WWW servers. This is updated from FlyBase, 1995). The synonymy of *Drosophila* gene, allele and aberration names is very extensive. FlyBase attempts to record all synonyms (47 280 as of November 1996) and search tools are designed to allow the recovery of records by synonym.

Map data

All map data are stored in FlyBase in a common form, regardless of whether these data are genetic, cytogenetic or molecular. This allows FlyBase to output integrated maps in a variety of formats.

A major project on the analysis of map data is now complete and this allows the automatic generation of genetic and cytogenetic maps and the identification of data conflicts. Tools have been written to output map information in a variety of forms, including graphical and tabular. For example, the CytoSearch tool allows users to query the map data in a number of different ways (e.g. 'output all of the genes known to map between 35B1 and 35C1 on the polytene chromosomes', 'output all of the deletions that uncover 35B1', 'output all of the cosmid clones on the X chromosome'). Another option is a graphical map tool that allows the display of selected classes of object (genes, aberrations, clones, transposon insertions) from an image that represents the chromosomes.

Bibliographic references

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). As of November 1996 the number of bibliographic records was 81 435, including 4745 theses and 18 206 abstracts.

Nature of gene product(s)

In 1996 FlyBase substantially revised the ways in which the nature of a gene's product is described. FlyBase now classifies the nature of a gene's product in two ways: 'structure' and 'function'. For 'structure' FlyBase relies on cross-references to the PROSITE database. If no such cross-reference exists then FlyBase uses a vocabulary modelled on that of PROSITE to give an indication of the structural feature(s) of a gene's product. For 'function' FlyBase uses the EC name and EC number for those products that are enzymes (and are included in the ENZYME database). For products that are not enzymes (or are not included in the ENZYME database) FlyBase uses a controlled vocabulary to summarise the (molecular) nature of a gene's product.

FlyBase is now working with others to construct a hierarchical classification of biological processes that can be used to classify gene functions. It is hoped that a preliminary version of this classification will be released early in 1997.

Database cross-references

FlyBase extensively cross-references its objects with those in other genetic and molecular databases. FlyBase receives daily updates of new and revised records from the EMBL/DDBJ/GenBank databases and stores their primary accession numbers and Protein Identifier Numbers (PIDs) in the gene, allele or aberration records. FlyBase also stores cross-references (by accession number) to both SwissProt and PIR, as well as to the Eukaryotic Promoter Database (EPD), dbSTS, dbEST, TRANSFAC, PDB, NRL_3D and GCR databases. FlyBase now includes over 6455 accession number cross-references to the EMBL/DDBJ/GenBank database, 1050 to SwissProt and 1903 to PIR. FlyBase also cross-links to other genetic databases (see below). FlyBase provides these external databases with flat file tables of their accession numbers linked to FlyBase accession numbers, encouraging reciprocal DBXREF links.

Molecular data

FlyBase curates information on the molecular organization of genes and detailed information on transcripts and protein products and their expression. Expression pattern curation uses a controlled vocabulary for the description of anatomy and life stages. These data can be accessed via FlyBase gene reports.

FlyBase collaborates very closely with both the Berkeley and European *Drosophila* Genome Projects. An integrated list of P1, cosmid and YAC clones from these projects is available and can be searched by cytological location.

FlyBase curates the structure of artificial constructs (including plasmids, vectors and constructs used for transformation). These data are reported via graphical maps of transposons and plasmids linked to sequence data and, where appropriate, to mutant alleles and publications. For each transposon and vector there are links to the components used in its construction.

Similar genes in other organisms

One of the most urgent needs for those building genetic databases is a stable mechanism to cross-reference genes (and other objects) between organisms. In the absence of such a mechanism FlyBase now simply includes the gene symbol and organism of loci said, by investigators, to encode a similar (or homologous) product. These cross-references (1659 as of November, 1996) include the gene symbol approved by the appropriate community (e.g. HUGO) and, where possible, the gene's accession number in the appropriate database (OMIM, GDB, MGD, ECOGENE, *Saccharomyces* Genome Database). Some of these links (e.g. with GDB) are now reciprocal.

Stock lists

FlyBase provides access to the stock lists of the three major stock centers for *D.melanogaster* (Bloomington, Mid-America and Umea) and for that of the *Drosophila* species stock center at Bowling Green. It also provides access to the stock lists of individual laboratories, if these are provided to FlyBase. FlyBase works with the major *Drosophila* stock centers to ensure consistency of nomenclature.

People directory

FlyBase maintains a directory of names, addresses, telephone and fax numbers and email addresses of people in the *Drosophila* community. Those with particular roles in the community (e.g. principal investigators, stock-keepers, members of the *Drosophila* Board) are tagged. There are now over 5500 records in this directory.

Allied data

FlyBase cannot, and should not, be wholly comprehensive. We encourage others to build specialised databases. At present FlyBase offers help in linking these to FlyBase (by the use of FB identifiers, for example) and in making these available through the FlyBase servers. Several databases of allied data are now available through FlyBase: these include a complete list of valid species in the family Drosophilidae (Dr G. Baechli, Zurich), a *Drosophila* codon usage table and the *Drosophila* records of the Transcription Termination Signal Database. All *Drosophila* records of the Environmental Mutagen Information Center are also available.

FlyBase has a depository for images (flybase/allied-data/images).

Although not allied data, FlyBase makes the complete unchanged text of Lindsley and Zimm (3) available (by permission of Academic Press) and keeps a file of errors in this book that have been noticed. The text of the earlier Lindsley and Grell (1968) is also available on FlyBase. There is also a file of errors noticed in Ashburner's *Drosophila. A Laboratory Handbook and Manual* (Cold Spring Harbor, 1989).

IMPLEMENTATION

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. The schema is now being extended to accommodate physical maps and sequences from the major *Drosophila* genome projects.

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

ACCESS

FlyBase provides users with a variety of modes of access: http, gopher, e-mail, ftp of flat files and via the *Encyclopaedia of Drosophila*.

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	e-mail access

Mirror sites are available in Europe, Japan, Australia and the USA. The major mirror sites are now:

http://www.embl-ebi.ac.uk/flybase/	http access
gopher.embl-ebi.ac.uk 7071	gopher access
ftp.embl-ebi.ac.uk (in /pub/databases/flybase)	ftp access
http://www.angis.su.oz.au:7081/	http access
http://shigen.lab.nig.ac.jp:7081/	http access
shigen.lab.nig.ac.jp 7071	gopher access
http://cbbridges.harvard.edu:7081/	http access
cbbridges.harvard.edu 7071	gopher access

The *Encyclopaedia of Drosophila* is available from:

http://shoofly.berkeley.edu/

The flat files derived from the Sybase tables are often available in several formats, as well as being WAIS indexed for queries. For example, the bibliography is available in Unix REFER format (which can be used by many bibliographic packages) as well as in text and 'comma-separated-values' formats. The genetic data are available in readable text formats and in a format in which different fields are coded (the latter allow users to write simple code to construct their own queries on the data). The FlyBase gene file is indexed at many SRS sites.

FlyBase publishes a subset of the data in printed form as special issues of *Drosophila Information Service*. Two such issues were published in June 1994: *DIS* ***73*** includes data on gene loci, gene function and gene and allele synonyms; *DIS* ***74*** is a bibliography of the *Drosophila* literature for the period 1982–1993. Issues to include genetic maps, transposon data and an update of the bibliography are planned.

FlyBase and the Berkeley *Drosophila* Genome Project jointly produce *The Encyclopaedia of Drosophila*. This presents a merge of the information in FlyBase with the data of the Berkeley project. This is a collaborative project with G. M. Rubin, S. Lewis, C. Harmon and G. Helt.

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. FlyBase has recently released its first direct user data submission tool, for the submission of data

concerning GA4/UAS transposons. It is expected that further direct submission tools for other classes of data will be made available in the coming year.

DOCUMENTATION

A complete FlyBase Reference Manual and a shorter User Manual are available from FlyBase servers in a variety of formats (html, rtf, Postscript and text). A brief introduction, 'Getting started with FlyBase' is also available.

Announcements of major database updates and concerning the release of new tools are made through postings to the bionet.drosophila bulletin newsgroup. FlyBase users are encouraged to use this newsgroup to track changes to FlyBase.

ADDRESSES

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 (1997) FlyBase—The *Drosophila* Database. Available from http://flybase.bio.indiana.edu/ *Nucleic Acids Res.* **25,** 63–66.

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 Center for Human Genome Research) and the Medical Research Council, London, UK. John Merriam (UCLA) was a member of the consortium until July 1994. We thank him for his invaluable contributions. The *Encyclopaedia of Drosophila* is supported by a grant from National Institutes of Health (National Center for Human Genome Research) to G. M. Rubin.

REFERENCES

1 FlyBase (1995) *Drosophila melanogaster.* In A. Stewart (ed.), Genetic Nomenclature Guide. *Trends Genet.* Supplement to March 1995, pp. 26–29.
2 Lindsley,D.L. and Grell,E.H. (1968) Genetic Variations of *Drosophila melanogaster.* Carnegie Institution, Washington, DC.
3 Lindsley,D.L. and Zimm,G.G. (1992) *The Genome of Drosophila melanogaster.* Academic Press, San Diego, CA.