FlyBase: History, Present, Future and Funding

A report from FlyBase 6/28/2021

Note: Part of the text below is from Wikipedia as well as from the 2021 FlyBase report to the FlyBoard.

FlyBase: History

Drosophila melanogaster has been an experimental organism since the early 1900s, and has since been placed at the forefront of many areas of research. As this field of research spread and became global, researchers working on the same problems needed a way to communicate and monitor progress in the field. This niche was initially filled by community newsletters such as the Drosophila Information Service (DIS), which dates back to 1934 when the field was starting to spread from Thomas Hunt Morgan's lab. Material in these pages presented regular 'catalogs' of mutations, and bibliographies of the Drosophila literature. These lists of genes, mutant phenotypes and bibliographies were further compiled into more formal publications, notably Herskowitz's* Bibliographies of *Drosophila* and the "Red Books" of Lindsley and Grell and Lindsley and Zim**.

As the compilations of Drosophila data became more and more extensive it became clear that this method of summarization of the aggregate knowledge of the field was no longer scalable. It was therefore proposed in 1991 to utilize the growing computing infrastructure to shift the newsletters and books into an online database. In October 1992, the National Center for Human Genome Research (NHGRI) of the NIH funded the FlyBase project with the objective of designing, building and releasing a database of genetic and molecular information concerning *Drosophila melanogaster*. Through this funding and support from the UK Medical Research Council a variety of data were made available over the Internet through the nascent FlyBase website, including: the gene and chromosomal aberration lists of the RedBook; an accumulated bibliography; lists of stocks; and lists of clones. As the sequencing of the *Drosophila melanogaster* genome progressed FlyBase collaborated with the informatics groups of the Berkeley Drosophila Genome Project (BDGP) and European Drosophila Genome Project (EDGP) to unify the genome sequence and associated gene predictions with genetic data.

Following the transfer of existing data into FlyBase, mechanisms for biocuration of the wealth of new data being steadily produced by the community were developed. This included: regular updating of the FlyBase bibiliography; curation of new alleles, transgenic constructs and phenotypes from the majority of the primary research literature; using new data sources (such as large scale RNA sequencing) to update gene models; annotating gene expression using controlled vocabularies (ontologies) to ensure searches return comprehensive results; and annotation of Drosophila models of human disease. FlyBase members were amongst the founders of the Gene Ontology Consortium, which utilizes a shared ontology to annotate gene function in many organisms.

FlyBase is a mature project with an experienced staff of long-term employees and many of our activities are continuous. Currently, the FlyBase project is carried out by a consortium of Drosophila researchers and computer scientists at Harvard University, University of Cambridge (UK), Indiana University, and the University of New Mexico. For a list of current FlyBase team members see: https://wiki.flybase.org/wiki/FlyBase:About#FlyBase_Consortium.

FlyBase: Present

Information in FlyBase originates from a variety of sources ranging from the primary research literature to large-scale genome projects. FlyBase staff currently curate a wealth of data types including: molecular descriptions of new mutant alleles and other abberations; details of transgenic constructs and transposon insertions; stock information on these new genetic tools; sequence-level gene models; mutant phenotypes and inferred gene function with the Gene Ontology; Genetic and Protein-Protein interactions; and high-throughput and manually curated expression data on genes and Gal4 drivers. FlyBase query tools allow navigation through DNA or protein sequence, by gene or mutant name, or through terms from the several ontologies used to capture functional, phenotypic, and anatomical data. The database offers several different query tools in order to provide efficient access to the data available and facilitate the discovery of significant relationships within the database. Links between FlyBase and external databases provide opportunity for further exploration into other model organism databases and other resources of biological and molecular information. Altogether, FlyBase over the years has matured from database to knowledgebase. See for example a recent publication (Larkin et al., 2021).

FlyBase has three main goals:

1. To continue curation of literature and reagents relevant to Drosophila research. This is an essential goal that ensures that Drosophila researchers can continue to rely on FlyBase to find the latest innovations in the field and the reagents for experimental design. FlyBase prioritizes curation of data on previously uncharacterized genes, as well as those revealing new information on gene function, signaling pathways, and human diseases. Over the years, FlyBase curation has continuously evolved to integrate machine learning tools with human expert curation to generate high quality information. This goal standard data is an essential training set for the development of new AI tools to assist essential human curation. Flybase continuously develops new ways to display this information in an intuitive, integrated, readily searchable format.

2. To improve FlyBase's utility to the human genetics and population genetics communities, by curating and integrating relevant data sets, and developing tools that enable better access to this wealth of data. As a member of The Alliance of Genome Resources (the “Alliance”), FlyBase works closely with other Model Organism Databases (MODs) to integrate data sets and develop tools to enable cross-species analyses. This effort has a major impact on the fly community, accelerating the development of models of human diseases.

3. To facilitate more integrative analyses and approaches, FlyBase integrates and

displays large-scale studies, transcriptomic and proteomic data sets. In addition, FlyBase provides access and displays tools available within the community, and incorporates the most useful data sets and tools for retrieving and displaying complex data sets to enable more researchers to take a global approach to their genetic research.

Overall, FlyBase has become an invaluable tool used to conduct research with *Drosophila*. One measure of success is that the average number of sessions and page views in 2020 were 137k and 662k, respectively. It is also worth reflecting how much researcher time and money is saved by having an up-to-date compendium of research findings on each Drosophila gene; assembling this information on just one gene through searching the literature is a very lengthy process.

FlyBase Future

As noted, FlyBase has become an invaluable tool used to conduct research with Drosophila. In the next five years, FlyBase efforts will focus on three main areas:

1. ***FlyBase contribution to the specific needs of the Drosophila community.*** FlyBase provides a crucial openly-accessible centralized resource for Drosophila genetic and genomic data to enable researchers and educators worldwide, both in the Drosophila community and broader biomedical sciences community, to further their research. In addition to the curation activities already mentioned, more recent innovations include: rapid appearance of relevant papers on each gene report thanks to author curation; Gene Group and Signalling Pathway curation; import of graphical abstracts in FlyBase references; identification of new lncRNAs, anti-sense lncRNAs and smORFs; incorporation of available transcription start site data into FlyBase; addition of new anatomy terms, especially new neuron types; review and improvement of phenotypic class ontologies; annotation of all Drosophila cell types and curation of scRNAseq data sets; updating genomic sequences of Drosophila Genetic Reference Panel (DGRP) strains.
2. ***FlyBase contribution to the Alliance.*** FlyBase is involved in direct data-sharing collaborations with a number of external data resources and is a member of the Alliance. The Alliance consortium is organized to design a web portal that gathers and integrates the data from several model organism databases (MODs) (*D. melanogaster*, *C. elegans*, *S. cerevisiae*, *D. rerio*, *M. musculus* and *R. norvegicus),* in collaboration with the Gene Ontology Consortium (GOC), so that at one site a researcher can find out what is known about a particular gene's function in all these model organisms. This data integration and harmonization also enables clinical researchers to efficiently access and translate findings in model organisms to new disease diagnoses and treatments. The primary goals of the Alliance are: **a.** To provide unprecedented support for comparative genomics *via* unified user interfaces and APIs for data types shared by the MODs, and **b.** To promote sustainability and operational efficiencies by building a “knowledge commons” based on shared, modular infrastructure. FlyBase helps the Alliance to develop the next generation of model organism knowledgebases capable of adapting to the rapidly changing data science landscape.

3. FlyBase outreach. FlyBase will continue to meet user community needs by: **a.** Enabling accelerated incorporation of published data by identifying new opportunities for direct user data submissions and improving existing tools; **b.** Maintain project-wide help desk *via* email; **c.** Providing in-person training through presentations and demonstrations at Drosophila and other conferences; **d.** Providing on-line training through maintenance and development of documentation and video tutorials; **e.** Publicizing updates and new features through various media, including an email-based Newsletter, our Twitter feed and peer-reviewed publications**; f.** Enhancing community-driven webpages and portals; **g.** Soliciting and responding to feedback from research community *via* the FlyBase Community Advisory Group; **h.** Collaborating with Model Organism Database and biocuration communities to find common solutions to shared goals; and **i.** Illustrating and supporting the exceptional value of FlyBase and *Drosophila* in the classroom as teaching tool for the next generation of wet bench researchers and data scientists.

FlyBase Funding

Most of FlyBase funding in the past has come from a large grant from the National Human Genome Research Institute (NHGRI) and a smaller grant the UK Medical Research Council. This situation changed in 2016 when NHGRI decided to reduce FlyBase support in an effort to promote the Alliance. Our projected budget for our final year of current 5-year grant cycle and for our next renewal in 2023 will be 50% of what it was in 2016. Additional funds for FlyBase are an NHGRI supplement for the Alliance and an NSF grant, which altogether will bring FlyBase's funding closer to 60-65% of 2016. Our efforts to explore mechanisms that would bring in more international funding have so far been unsuccessful. We therefore went to the community to request they contribute to help FlyBase keep up to date. We have been touched by the diversity of these contributions, such as from the hat passed round at the end of an advanced genetics class. However, since contributions to the community correspond to approximately 5% of the current FlyBase needs, these valuable contributions cannot be the sole solution to the funding deficit.

We plan to continue the user-fee collection to supplement FlyBase funding and are grateful for the strong support from our community. Nonetheless, the projected shortfall will severely restrict the ability of FlyBase to keep pace with the developments in the Drosophila field. Not only has the pace of research continued to increase the amount of data per publication, but the increasing diversity of the field, for example into physiology and metabolism, requires substantial new effort to curate, integrate and display these new data types. Overall, the steady decrease in FlyBase support since 2016 is a concern as it negatively impact on biomedical discovery.

References:

Herskowitz, I. H., 1953 Bibliography on the genetics of Drosophila, pp. 212.
Herskowitz, I. H., 1958 Bibliography on the genetics of Drosophila, pp. 296.
Herskowitz, I. H., 1963 Bibliography on the genetics of Drosophila, pp. 344.
Herskowitz, I. H., 1969 Bibliography on the genetics of Drosophila, pp. 376.

Herskowitz, I. H., 1974 Bibliography on the Genetics of Drosophila, pp. 159-218.

Lindsley, D. L., and E. H. Grell, 1968 Genetic variations of Drosophila melanogaster.

Lindsley, D. L., and G. G. Zimm, 1992 The Genome of Drosophila melanogaster.

Larkin, A., Marygold, S. J., Antonazzo, G., Attrill, H., Dos Santos, G., Garapati, P. V., Goodman, J. L., Gramates, L. S., Millburn, G., Strelets, V. B., Tabone, C. J., Thurmond, J. and FlyBase Consortium. (2021) FlyBase: updates to the Drosophila melanogaster knowledge base. Nucleic Acids Res. **49**(D1):D899-D907. PMID: 33219682