Chapter 1

Using FlyBase, a Database of *Drosophila* Genes and Genomes

Steven J. Marygold, Madeline A. Crosby, Joshua L. Goodman, and The FlyBase Consortium*

Abstract

For nearly 25 years, FlyBase (flybase.org) has provided a freely available online database of biological information about *Drosophila* species, focusing on the model organism *D. melanogaster*. The need for a centralized, integrated view of *Drosophila* research has never been greater as advances in genomic, proteomic, and high-throughput technologies add to the quantity and diversity of available data and resources.

FlyBase has taken several approaches to respond to these changes in the research landscape. Novel report pages have been generated for new reagent types and physical interaction data; *Drosophila* models of human disease are now represented and showcased in dedicated Human Disease Model Reports; other integrated reports have been established that bring together related genes, datasets, or reagents; Gene Reports have been revised to improve access to new data types and to highlight functional data; links to external sites have been organized and expanded; and new tools have been developed to display and interrogate all these data, including improved batch processing and bulk file availability. In addition, several new community initiatives have served to enhance interactions between researchers and FlyBase, resulting in direct user contributions and improved feedback.

This chapter provides an overview of the data content, organization, and available tools within FlyBase, focusing on recent improvements. We hope it serves as a guide for our diverse user base, enabling efficient and effective exploration of the database and thereby accelerating research discoveries.

Key words FlyBase, *Drosophila*, Database, Genetics, Genomics, Translational research

1 Introduction

Since its inception in 1992, FlyBase has provided an online repository of biological data about *Drosophila* species, focusing on the model organism *D. melanogaster*. Data in FlyBase are either curated manually from the primary research literature or are incorporated computationally from various sources, with the two input streams being integrated into a series of ‘report’ pages and other portals on

*The members of the FlyBase Consortium are listed in the Acknowledgements.

Christian Dahmann (ed.), *Drosophila: Methods and Protocols*, Methods in Molecular Biology, vol. 1478, DOI 10.1007/978-1-4939-6371-3_1, © Springer Science+Business Media New York 2016

the website. Many links to related data and resources at external databases are also incorporated. As a result, FlyBase serves as a nexus for all *Drosophila*-related information.

While the core purpose of FlyBase has not changed significantly over the years, we have had to continually review our data integration and presentation strategies to reflect the changing nature of *Drosophila* research. For example, when FlyBase was founded, the genomic sequence of *D. melanogaster* (or any other metazoan) was not yet known, and DNA microarray and RNA-Seq technologies did not yet exist. Since then, the volume of relevant data has increased massively as whole-genome and high-throughput studies have become increasingly common, and the number of new datasets and novel resource collections has expanded. Furthermore, the FlyBase user base has diversified as more researchers from other disciplines take advantage of the *Drosophila* system to conduct their experiments—particularly those interested in modeling human diseases.

Much of the information in FlyBase is partitioned into 20 different data classes, each with an associated report on the website (Table 1). Traditional reports, such as those for genes, alleles, or references, have been supplemented with several new ones in recent years to reflect the different reagents being used and new data types being produced. New or improved tools have also been developed in order to effectively search and analyze these novel data. A list of all current FlyBase tools, along with a brief description of their functionalities, is provided in Table 2 and specific use cases will be presented in context below.

Table 1
Data classes and reports in FlyBase

Data class/report	Number of records[a]		Factor change	Comment
	FB2007_01	FB2015_04		
Genetic/genomic data				
Genes (sequenced)	58,101 (27,377)	246,273 (192,537)	4.2 (7.0)	Large increase owing to incorporation of non-*D. melanogaster* genomes
Transcripts	39,018	293,576	7.5	Large increase owing to incorporation of non-*D. melanogaster* genomes
Polypeptides	34,302	256,047	7.5	Large increase owing to incorporation of non-*D. melanogaster* genomes
Alleles	108,697	226,157	2.1	
Aberrations	32,412	22,608	0.70	Decrease owing to removal of 'potential' DrosDel deletions

(continued)

Table 1 (continued)

Data class/report	Number of records[a] FB2007_01	FB2015_04	Factor change	Comment
Balancers	540	608	1.1	
Recombinant constructs	24,693	111,788	4.5	Large increase owing to incorporation of several genome-wide construct collections
Insertions	72,782	174,464	2.4	
Natural transposons	1189	1472	1.2	
Sequence features	n/a	184,028	n/a	New in FB2009_05
Physical interactions	n/a	16,075	n/a	New in FB2011_01
Integrated data				
Large dataset metadata	n/a	709	n/a	New in FB2009_05
Gene groups	n/a	278	n/a	New in FB2015_02
Human disease models	n/a	44	n/a	New in FB2015_04
Reagents				
Stocks	85,022	140,101	1.6	
Strains	n/a	284	n/a	New in FB2015_01
Cell lines	n/a	188	n/a	New in FB2009_06
Clones	304,240	723,550	2.4	
Other				
References (papers)	184,744 (78,122)	212,340 (94,736)	1.1 (1.2)	
Images	870	1258	1.4	

[a] The number of records at the time of writing (FB2015_04, September 2015) is compared to that of the previous review [1] (FB2007_01, August 2007)

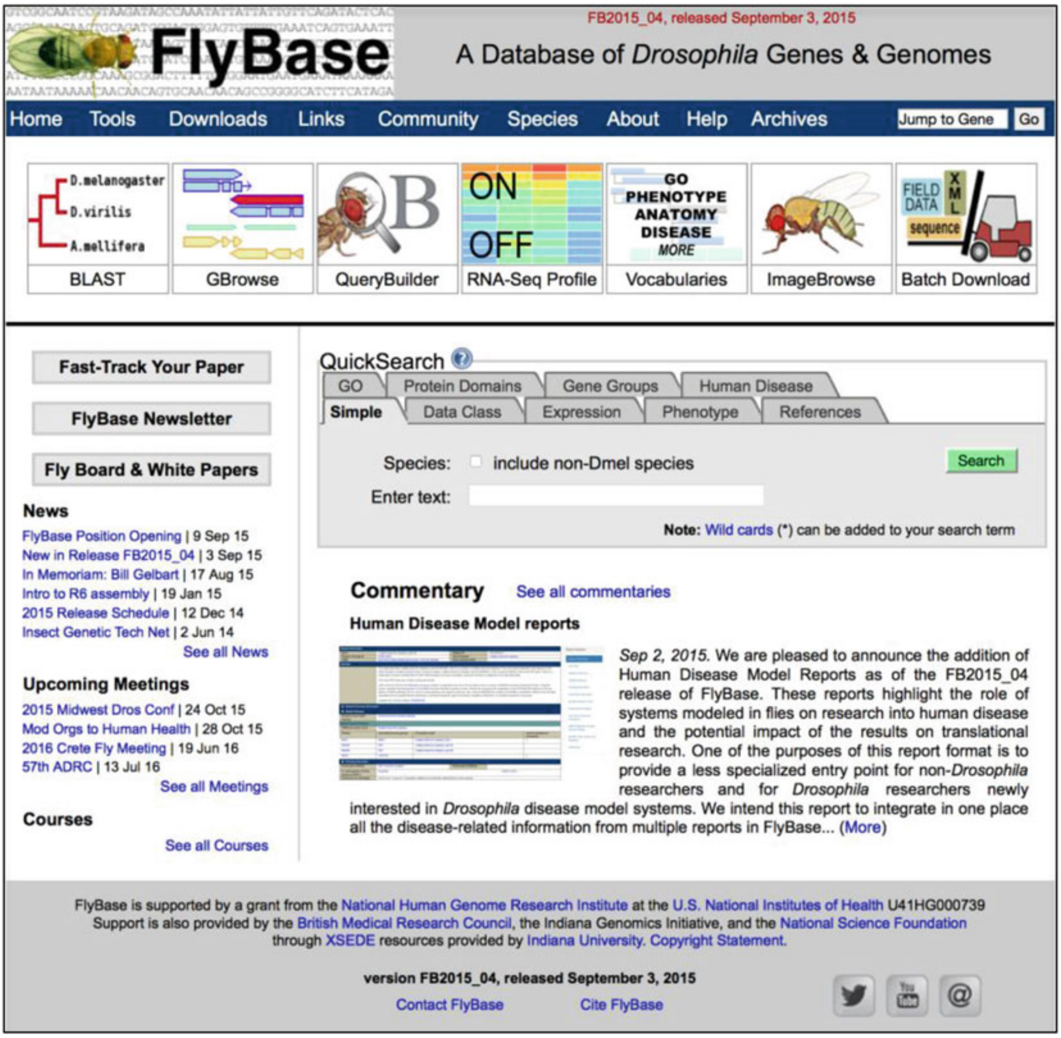
An overview of the FlyBase database was last presented in 2008 [1]. The main purpose of the current chapter is to provide a primer on how users can best use FlyBase today, with an emphasis on new and updated features. Inevitably, some areas are only mentioned briefly: the reader is referred to documentation on the FlyBase website and cited publications for more details.

Table 2
Tools in FlyBase

Tool name	Function/usage
Query tools and portals	
QuickSearch	Simplified searches on various data classes
QueryBuilder	Advanced search on a field-by-field level for most data classes
Vocabularies	Search or browse all controlled vocabularies used to annotate records
CytoSearch	Search for genetic objects mapped via cytology-based data
RNA-Seq Profile	Find genes with specific patterns of expression across modENCODE developmental stage, tissue, treatment, or cell line RNA-Seq data
Interactions Browser	Explore genetic and physical interactions via static images
ImageBrowse	Browse through *Drosophila* images by organ system, life-cycle, tagma, or germ layer
Genomic/Map tools	
BLAST	NCBI BLAST for finding nucleotide and protein sequences in *Drosophila* and dozens of related species
GBrowse	Graphical or tabular representation of the 12 sequenced *Drosophila* genomes
CytoSearch	Search for genetic objects mapped via sequence-based data
Feature Mapper	Search for sequence-mapped features that overlap a specific region or gene
RNA-Seq By Region	Evaluate the expression levels of exons, introns and/or intergenic regions from modENCODE developmental and tissue RNA-Seq data
Coordinates Converter	Convert coordinates between different genome releases (e.g. *D. melanogaster* R5 to R6)
Retrieve/convert tools	
Batch Download	Bulk download of individual report fields, FASTA or XML files in a variety of formats
Coordinates Converter	(See Coordinates Converter above)
Upload/Convert IDs	Update lists of old IDs, convert lists (e.g. genes to alleles), or upload IDs into a hit list

2 The Homepage

The main purpose of the homepage is to provide an indication of, and intuitive access to, all available data and tools in FlyBase (Fig. 1). In addition, the homepage highlights new features within FlyBase and advertises topical issues of interest to the fly community.



News

FlyBase Position Opening | 9 Sep 15
New in Release FB2015_04 | 3 Sep 15
In Memoriam: Bill Gelbart | 17 Aug 15
Intro to R6 assembly | 19 Jan 15
2015 Release Schedule | 12 Dec 14
Insect Genetic Tech Net | 2 Jun 14

Upcoming Meetings

2015 Midwest Dros Conf | 24 Oct 15
Mod Orgs to Human Health | 28 Oct 15
2016 Crete Fly Meeting | 19 Jun 16
57th ADRC | 13 Jul 16

Courses

Note: Wild cards (*) can be added to your search term

Commentary

Human Disease Model reports

Sep 2, 2015. We are pleased to announce the addition of Human Disease Model Reports as of the FB2015_04 release of FlyBase. These reports highlight the role of systems modeled in flies on research into human disease and the potential impact of the results on translational research. One of the purposes of this report format is to provide a less specialized entry point for non-*Drosophila* researchers and for *Drosophila* researchers newly interested in *Drosophila* disease model systems. We intend this report to integrate in one place all the disease-related information from multiple reports in FlyBase... (More)

FlyBase is supported by a grant from the National Human Genome Research Institute at the U.S. National Institutes of Health U41HG000739
Support is also provided by the British Medical Research Council, the Indiana Genomics Initiative, and the National Science Foundation through XSEDE resources provided by Indiana University. Copyright Statement.

version FB2015_04, released September 3, 2015

Fig. 1 The FlyBase homepage

2.1 Overview

The 'Navigation Bar' (NavBar) along the top of this (and every FlyBase) page incorporates drop-down menus containing links to key pages. For example, the 'Tools' menu lists all FlyBase tools, grouped by usage, and includes a 'Tools Overview' page to help users with search strategies. 'Downloads' contains links to all bulk data files that are available to download (*see* Subheading 10.2), also with an overview page describing them. The 'Links' menu has direct links to major external sites, such as the Berkeley *Drosophila* Genome Project (BDGP) and modENCODE, along with a comprehensive list of ~250 network and ~75 reagent resources of interest to *Drosophila* researchers. A new 'Community' menu has been added to group features that facilitate interactions between researchers and FlyBase, such as Fast-Track Your Paper (FTYP) and the FlyBase Community Advisory Group (FCAG) (*see* Subheading 11).

The 'About' menu collates general information about the FlyBase database and consortium, and includes links to FlyBase release notes and to a listing of FlyBase-authored publications. The ubiquitous NavBar also features the 'Jump to Gene' box (*see* Subheading 2.3).

Immediately below the NavBar on the homepage are prominent pictographs providing direct links to the most popular tools. Direct links to important community features are shown on the left-hand side of the homepage, including FTYP and the 'FlyBase Newsletter'. Beneath these are links to general 'News' items and upcoming 'Meetings' and 'Courses'. New or improved features within FlyBase are usually accompanied by an extended 'Commentary'—an abbreviated teaser section is shown on the homepage, and all recent teasers cycle in order that several concurrent improvements can be brought to the attention of users. A link is also provided to view all current and previous Commentaries.

The FlyBase website is updated with new data (and often new features) about 6 times per year (see the 'Release Schedule' link under the 'About' menu of the NavBar). The header and footer of the homepage, and indeed every FlyBase page, state the version number and date of release of the FlyBase instance currently being viewed (Fig. 1). For example, 'FB2015_04' indicates the 4th release of 2015 and was used for compiling the data and screenshots in this chapter. It is important that users take note of this release number when using FlyBase data to direct experiments, and quote it when referring to FlyBase data in publications. Summary statistics for the current release and a record of general changes to FlyBase in each release are provided as 'Release Notes' and 'New in this release' pages, respectively, accessible via the NavBar 'About' menu. If needed, previous FlyBase releases (and accompanying 'Release Notes') are available via the 'Archives' menu of the NavBar.

2.2 QuickSearch

QuickSearch, located at the heart of the homepage, is the primary search tool on FlyBase and can be used to access all data types (Fig. 1). It has been significantly improved in recent years to make it as intuitive and flexible as possible [2]. The 'Simple' tab provides a Google™-type functionality in that any text can be entered and a search is performed across the entire database for possible matches. Using the 'Data Class' tab restricts the search to the specified class and, optionally, just to symbols, names and IDs (rather than 'All text').

Other QuickSearch tabs offer dedicated search options for particular data classes or associated data. The 'References' tab allows field-specific searching of the comprehensive set of *Drosophila*-specific publications in the FlyBase bibliography [3]. Similarly, the 'Human Disease' and 'Gene Groups' tabs facilitate searching of these particular classes of integrated data (*see* Subheading 9). Rather than searching by data class, the remaining QuickSearch tabs allow searching for entities that share Expression, Phenotype, Protein Domain, or Gene Ontology (GO) annotations.

If a query in QuickSearch, or any other FlyBase tool, results in multiple possible matches, then a tabular 'hit list' is produced [4]. The hit list serves as a core list-management tool in FlyBase and can be manipulated in several ways. The entire list, or a selected subset, can be: sorted by the entries in any of the columns; analyzed or refined by criteria appropriate to the given data; converted to a related data class (e.g., alleles converted to genes); downloaded as a list; or exported to other FlyBase tools for further refinements or to download specified data.

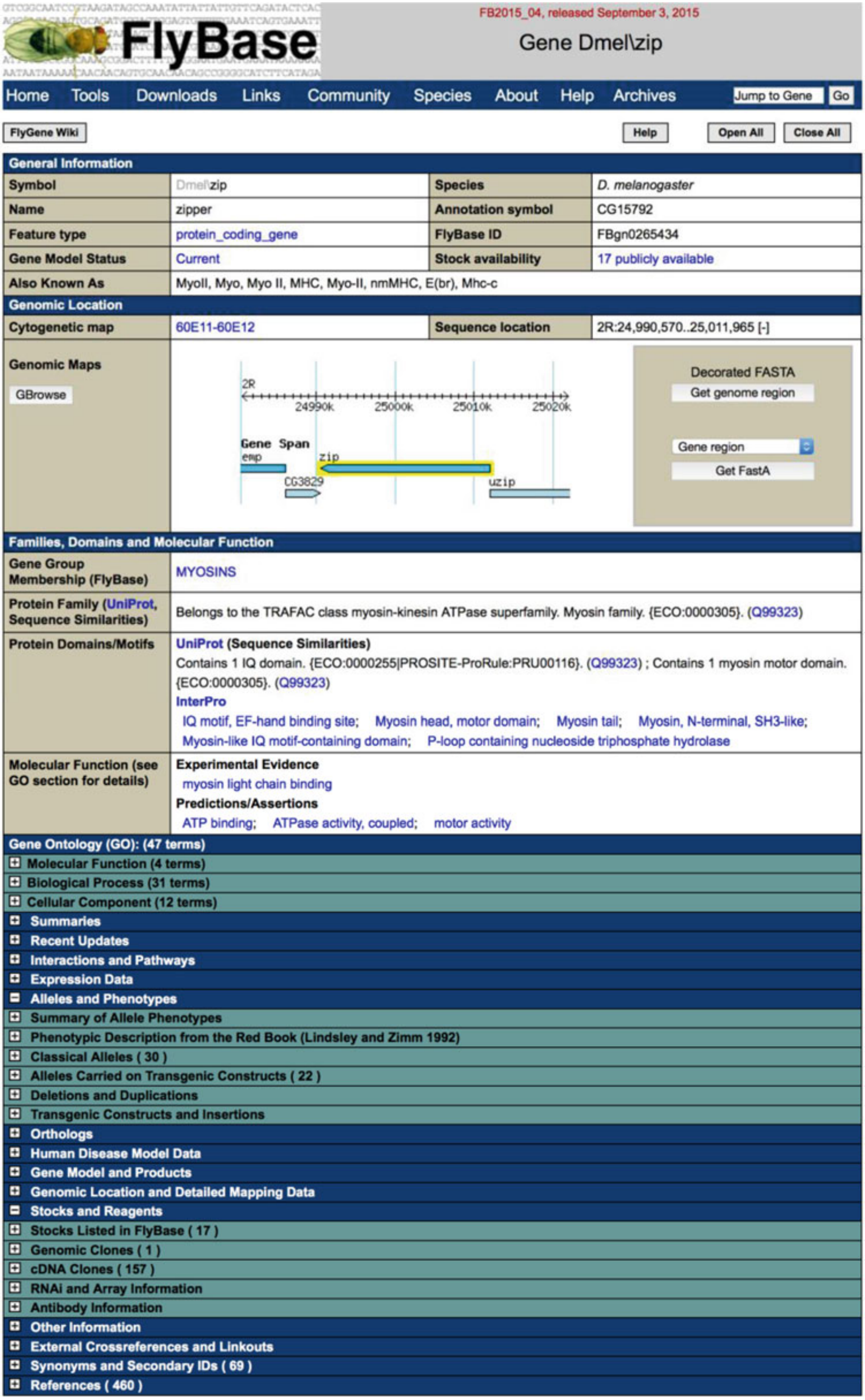
2.3 Jump to Gene

'Jump to Gene' is a navigation tool located in the NavBar and is thus available from all FlyBase pages. It is useful for quickly moving between Gene Reports where the current FlyBase gene symbol or ID (FBgn number) is known. (Greek characters must be spelled out, e.g., 'α-Est1' as 'alpha-Est1'.) In fact, 'Jump to Gene' also accepts current FlyBase IDs for all data classes, thus providing a direct route to other reports of interest. Note that input not recognized as a valid symbol or ID is treated as a gene synonym and a best match to a gene will be attempted. As the best match may not be to the intended gene, 'Jump to Gene' should not be used as a query tool—the 'Simple' or 'Data Class' tabs of QuickSearch should be used when the current FlyBase symbol or ID is not known for certain or when conducting a search.

3 The Gene Report

The Gene Report is the best place to start for users interested in a particular gene [2] (Fig. 2). In addition to listing gene-centric data (e.g., genomic location, expression data, orthologs), this report also summarizes data more fully described in separate reports (e.g., mutant alleles, phenotypes of those alleles, protein interaction data) and provides many links both within FlyBase and to external sites where additional information or details can be found. As such, Gene Reports act as hubs from which to explore all that is known about *Drosophila* genes.

The upper 'always open' part of the Gene Report has recently been reorganized to emphasize functional information, including a new section labeled 'Families, Domains and Molecular Function' (Fig. 2). This section highlights membership of the gene to any FlyBase 'gene groups' (*see* Subheading 9.2) or UniProt protein families, any predicted UniProt or InterPro domains/motifs, and summarized molecular function data based on FlyBase GO annotations. At the top of the report, the 'General Information' section includes gene identifiers (FlyBase ID, symbol, name, and CG number), the 'Feature type' (e.g., 'protein coding gene', 'miRNA gene', or 'pseudogene'), and the 'Gene Model Status', which helps to distinguish genes currently localized to the sequenced genome



FB2015_04, released September 3, 2015

Gene Dmel\zip

General Information			
Symbol	Dmel\zip	Species	*D. melanogaster*
Name	zipper	Annotation symbol	CG15792
Feature type	protein_coding_gene	FlyBase ID	FBgn0265434
Gene Model Status	Current	Stock availability	17 publicly available
Also Known As	MyoII, Myo, Myo II, MHC, Myo-II, nmMHC, E(br), Mhc-c		
Genomic Location			
Cytogenetic map	60E11-60E12	Sequence location	2R:24,990,570..25,011,965 [-]
Genomic Maps			
Families, Domains and Molecular Function			
Gene Group Membership (FlyBase)	MYOSINS		
Protein Family (UniProt, Sequence Similarities)	Belongs to the TRAFAC class myosin-kinesin ATPase superfamily. Myosin family. {ECO:0000305}. (Q99323)		
Protein Domains/Motifs	**UniProt (Sequence Similarities)** Contains 1 IQ domain. {ECO:0000255|PROSITE-ProRule:PRU00116}. (Q99323) ; Contains 1 myosin motor domain. {ECO:0000305}. (Q99323) **InterPro** IQ motif, EF-hand binding site; Myosin head, motor domain; Myosin tail; Myosin, N-terminal, SH3-like; Myosin-like IQ motif-containing domain; P-loop containing nucleoside triphosphate hydrolase		
Molecular Function (see GO section for details)	**Experimental Evidence** myosin light chain binding **Predictions/Assertions** ATP binding; ATPase activity, coupled; motor activity		

Gene Ontology (GO): (47 terms)
- ⊞ Molecular Function (4 terms)
- ⊞ Biological Process (31 terms)
- ⊞ Cellular Component (12 terms)

⊞ Summaries

⊞ Recent Updates

⊞ Interactions and Pathways

⊞ Expression Data

⊟ Alleles and Phenotypes
- ⊞ Summary of Allele Phenotypes
- ⊞ Phenotypic Description from the Red Book (Lindsley and Zimm 1992)
- ⊞ Classical Alleles (30)
- ⊞ Alleles Carried on Transgenic Constructs (22)
- ⊞ Deletions and Duplications
- ⊞ Transgenic Constructs and Insertions

⊞ Orthologs

⊞ Human Disease Model Data

⊞ Gene Model and Products

⊞ Genomic Location and Detailed Mapping Data

⊟ Stocks and Reagents
- ⊞ Stocks Listed in FlyBase (17)
- ⊞ Genomic Clones (1)
- ⊞ cDNA Clones (157)
- ⊞ RNAi and Array Information
- ⊞ Antibody Information

⊞ Other Information

⊞ External Crossreferences and Linkouts

⊞ Synonyms and Secondary IDs (69)

⊞ References (460)

Fig. 2 The Gene Report, using *zipper* as an example. The 'Alleles and Phenotypes' and 'Stocks and Reagents' sections have been opened to show the first tier of subsections

from those whose gene model is incomplete/withdrawn or genes defined only by mutations. Below this is a section summarizing 'Genomic Location' information. It gives the cytogenetic position and sequence coordinates of the gene, alongside a graphical snapshot of its genomic location and convenient options to download FASTA files associated with the gene model.

The remaining data are organized into separate sections that are closed by default and can be opened by clicking on their title bars (Fig. 2). Several of these have been updated in recent years. For example, the 'Gene Ontology (GO)' section that displays the full set of GO annotations [5] now clearly distinguishes terms based on experimental evidence from those based on predictions/assertions. The 'Summaries' section that follows has also been revised and now contains additional entries including a description of the FlyBase gene group to which the gene belongs and a functional summary imported from UniProt. Further down the page, the 'Orthologs' section now contains three discrete subsections. The first, labeled 'OrthoDB Orthologs', displays the orthologs of the given gene as computed by OrthoDB [6], arranged into taxonomic groups. The second subsection, 'Human Orthologs' specifically highlights the orthologous human gene(s) (again inferred by OrthoDB). Here, links are provided to the corresponding Ensembl [7], HUGO Gene Nomenclature Committee (HGNC [8]) and Online Mendelian Inheritance in Man® (OMIM® [9]) gene and phenotype (disease) reports. Finally, the 'External Data' subsection features link-outs to species-specific orthologs determined by the integrative ortholog prediction tool, DIOPT [10]. Note that link-outs such as these appear in several sections of the Gene Report and are consolidated in the 'External Cross-references and Linkouts' section toward the bottom of the report, facilitating navigation between databases.

Several other Gene Report sections have seen more significant updates, including expression data, physical interactions, and models of human disease. These are described separately below.

Querying genes and gene-related data: Gene Reports themselves are best found via the 'Jump to Gene' box on the NavBar or via the 'Simple' or the 'Data Class (genes)' tabs of QuickSearch. GO and Protein Domain annotations can be searched via their dedicated QuickSearch tabs. Strategies for searching for data in other sections or fields of the Gene Report are detailed below, or may be addressed using QueryBuilder [11] or the 'Simple' tab of QuickSearch.

4 Alleles and Phenotypes

The generation and study of mutants have been central to *Drosophila* research ever since its very beginnings [12] and remain a major component of FlyBase today. These data are summarized on the relevant Gene Report in the 'Alleles and Phenotypes' section

(Fig. 2), and are described in full in separate Allele Reports (not shown). FlyBase makes extensive use of controlled vocabularies (CVs or 'ontologies') in recording allelic and phenotypic data [13, 14]. These are collections of related terms (e.g., 'allele class', 'phenotypic class', 'fly anatomy') arranged into parent–child hierarchies. They allow a single, defined term to be used across all FlyBase annotations, which in turn allows users to search with that term (or any of its synonyms) to retrieve all records annotated with it or its children. CV terms and associated annotations can be queried using the Vocabularies tool [11] (formerly TermLink), accessible via the pictograph on the homepage or the Tools menu of the NavBar.

4.1 Classical Alleles, Transgenic Constructs, and Insertions

Classical alleles are defined as mutations that affect a gene at its endogenous locus. They are presented in Allele Reports and also appear in the 'Classical Alleles' subsection on the Gene Report (Fig. 2). They traditionally include point mutations, insertional mutations and intragenic deletions, though more recent additions include lesions induced by various recombination-mediated techniques and site-specific cleavage events. Where known, mutations are annotated with an 'origin of mutation' term (e.g., 'ethyl methanesulfonate' or 'CRISPR/Cas9'), an 'allele class' term (e.g., 'amorphic allele') and details of their molecular lesion. If the mutation is caused by an insertion of a transposable element, then both an Allele Report and an Insertion Report are created to completely describe the lesion, with a prominent link forged between them.

The molecular details and uses of transgenic constructs appear in dedicated Recombinant Construct Reports. In addition, an associated Allele Report is created in such cases in order to properly and fully capture phenotypic data. That is, the Alleles data class in FlyBase comprises both 'classical alleles' and 'alleles carried on transgenic constructs', as indicated within the 'Alleles and Phenotypes' section of the Gene Report (Fig. 2). The allele 'origin of mutation' CV has been expanded to accommodate this convention and so includes terms such as 'in vitro construct—RNAi'. The relationship between constructs and their associated alleles is clearly indicated and reciprocally linked in their respective report pages. Similarly, any specific insertions of a transgenic construct are captured in Insertion Reports and are reciprocally linked to their corresponding Recombinant Construct Report.

Recombinant Construct or Insertion Reports, as appropriate, are also made for reporters (e.g., lacZ or GFP) or binary drivers (e.g., GAL4). Again, an associated Allele Report is made in all these cases so that phenotypic (and expression) data dependent on their use can be stored and presented in a consistent manner across the database. Note that FlyBase uses a species prefix to distinguish genes originating from 'foreign' (non-*D. melanogaster*) species, and so these examples appear as alleles of '*Ecol\lacZ*', '*Avic\GFP*', and '*Scer\GAL4*' on the website.

Transgenic techniques and resources are constantly expanding [15]. FlyBase responds to the former by devising suitable curation strategies and/or revising CVs as appropriate. One response to the latter is the creation of the Large Dataset Metadata Report that collates the metadata and membership of large-scale collections of constructs, insertions, etc. (*see* Subheading 9.1). The report page of each member contains basic descriptive information about the collection, together with a link to the respective metadata report.

Querying alleles, constructs, and insertions: Specific alleles, transgenic constructs or insertions are best searched via the 'Data Class' tab of QuickSearch. As mentioned, the Vocabularies tool is useful to find instances of particular allele classes or mutagenic techniques, which can then be refined further using other FlyBase tools. CytoSearch, FeatureMapper, or GBrowse (*see* Subheading 7) are the preferred methods if you are looking for reagents that are mapped to specific genomic regions.

4.2 Phenotypes

Phenotypic data are attached to alleles or allele combinations using terms from the 'phenotypic class' and 'fly anatomy' CVs. The phenotype CV comprises ~190 terms that are commonly used to describe *Drosophila* phenotypes, such as 'lethal', 'sterile', 'homeotic' or 'Minute' [13]. The anatomy CV is much larger, comprising >8800 terms that can be used to comprehensively describe *Drosophila* anatomy [14]. Both types of CV term can be refined through the use of 'qualifier' terms that restrict the meaning of the term to a specific developmental stage, sex or other experimental/genetic condition—these appear after a 'pipe' symbol on the website, for example 'small body | larval stage'. Any additional genotypic components that are necessary for the given phenotype, such as alleles *in trans* or GAL4 drivers, are included in the phenotype annotation and appear with the prefix 'with' on the website. CV-based phenotypic annotations are often supplemented with free text clarifications or extra details. All phenotypic statements in FlyBase are curated from the published literature.

CV-based phenotype annotations are shown with their associated allele in a table on the Gene Report in the 'Summary of Allele Phenotypes' subsection (Fig. 2). Clicking on an individual allele takes you to the corresponding Allele Report that additionally contains any free text description of the phenotype together with the source reference(s).

Querying phenotypes: The dedicated 'Phenotype' tab in QuickSearch facilitates searching of alleles by phenotypic class and/or anatomy terms, with an option to refine the search through the use of qualifiers. The Vocabularies tool offers a browsable view of the same data, while QueryBuilder can be used to compose more complex combinatorial queries.

4.3 Disease Model Annotations

Drosophila alleles or allele combinations that generate phenotypes stated to be models of human disease are additionally annotated using appropriate terms from the Disease Ontology [16, 17]. As for regular phenotypes, disease model annotations may be associated with either classical alleles (of fly genes orthologous to human 'disease genes') or with transgenic alleles (where disease-causing forms of human genes, or their *Drosophila* orthologs, are expressed via transgenic constructs). The criterion for disease annotation in FlyBase is that the phenotype must recapitulate some aspect of the disease pathology, though this can range from anatomical defects or behavioral abnormalities to cellular or molecular changes. Additional alleles may be described as modifying a disease model, either 'ameliorating' or 'exacerbating' it, if genetic interactions are observed. This information is presented in tabulated form in both the Gene Report and Allele Report in the section titled 'Human Disease Model Data' as well as in the new Human Disease Model Report (*see* Subheading 9.3).

Querying disease model annotations: The 'Human Disease' tab of QuickSearch and the Vocabularies tool can both be used to find a Disease Ontology term and view records annotated with it.

5 Expression Data

Separate Gene Expression Reports have been retired from FlyBase in favor of integrating these data into the Gene Report in a dedicated 'Expression Data' section (Fig. 2). Expression data may derive from either 'low-' or 'high-throughput' studies.

5.1 Low-Throughput Expression Data

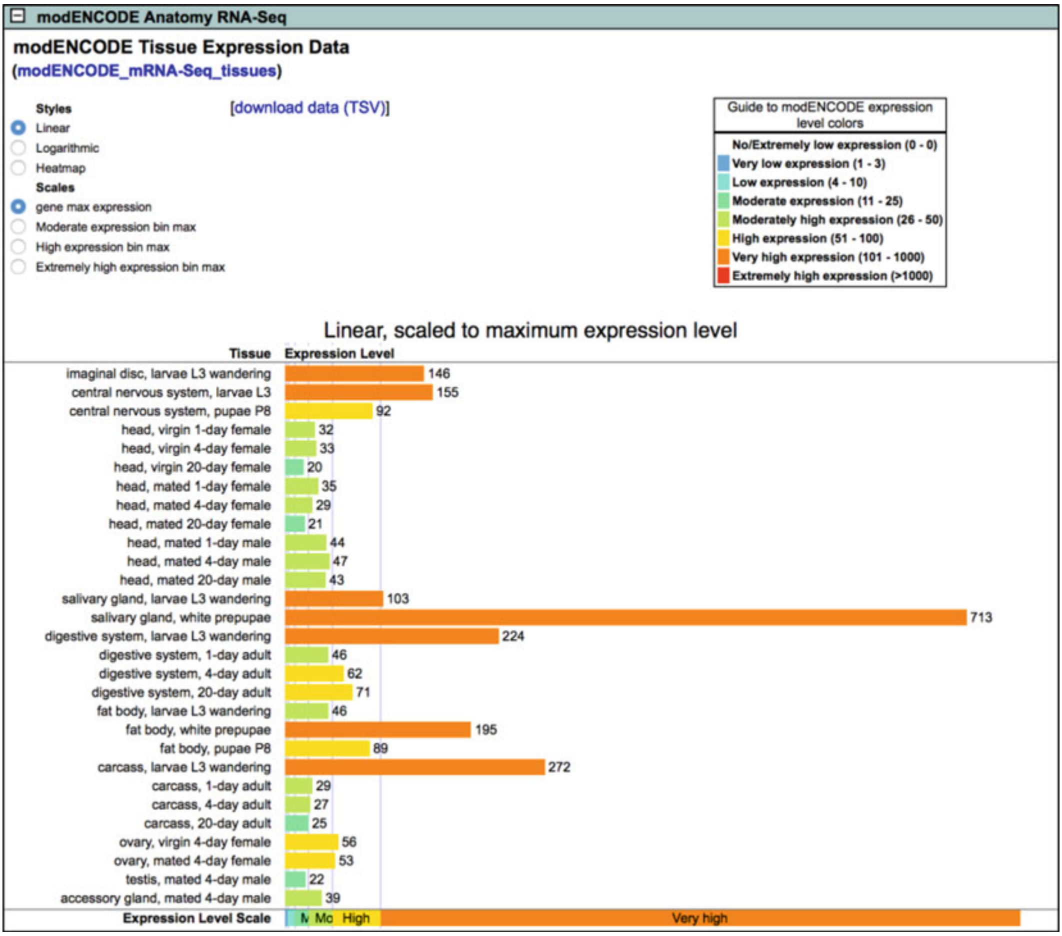
Traditional expression assays, such as in situ hybridization or immunolocalization to embryos or tissues, and stage- or tissue-specific Northern blots, are presented in the first three subsections of the 'Expression Data' section: 'Transcript Expression', 'Polypeptide Expression', and 'Expression Deduced from Reporters'. These are data from the published literature, captured in a highly controlled format using the FlyBase anatomy and developmental stage CVs. For nuanced aspects of an expression pattern additional free-text descriptions are provided. Embryonic transcript expression data include data from the BDGP [18], which comprises in situ hybridizations using cDNA probes for over 7000 genes, plus descriptions conforming to the FlyBase anatomy CV. The actual in situ images can be viewed by following the 'BDGP expression data' link in the subsection 'External Data and Images' (*see* Subheading 5.3).

Querying low-throughput expression data: The use of hierarchical CVs allows expression data captured at a very detailed level to be queried using more general terms. The QuickSearch 'Expression' tab provides a dedicated interface for this type of query. It also allows combinatorial queries, typically to specify both stage and tissue. The initial hit list returned by this query is of endogenous genes for which the expression pattern is observed; reporter constructs or insertions can be retrieved by selecting one of the alternative result options at the top of the initial hit list.

5.2 High-Throughput Expression Data

Within the 'Expression Data' section of the Gene Report, the 'High-Throughput Expression Data' subsection includes expression plots of high-throughput mRNA RNA-Seq data from modENCODE [19] and mRNA microarray data from FlyAtlas [20]. FlyBase has produced quantitative views of these data in different stages, tissues, and cell culture types presented as bar graphs (Fig. 3). For the RNA-Seq data, RPKM counts [21] (reads per kilobase per million reads) have been calculated, averaged over the exonic extents of the gene. A further subsection, 'Expression Clusters', provides links to datasets consisting of genes possessing similar mRNA RNA-Seq expression dynamics, as determined by modENCODE [19, 22].

Querying high-throughput expression data: Several new FlyBase tools that use the calculated RPKM data have been developed. The



⊟ modENCODE Anatomy RNA-Seq

modENCODE Tissue Expression Data
(modENCODE_mRNA-Seq_tissues)

[download data (TSV)]

Styles
● Linear
○ Logarithmic
○ Heatmap

Scales
● gene max expression
○ Moderate expression bin max
○ High expression bin max
○ Extremely high expression bin max

Guide to modENCODE expression level colors
No/Extremely low expression (0 - 0)
🟦 Very low expression (1 - 3)
🟩 Low expression (4 - 10)
🟩 Moderate expression (11 - 25)
🟨 Moderately high expression (26 - 50)
🟨 High expression (51 - 100)
🟧 Very high expression (101 - 1000)
🟥 Extremely high expression (>1000)

Linear, scaled to maximum expression level
Tissue
Expression Level
imaginal disc, larvae L3 wandering 146
central nervous system, larvae L3 155
central nervous system, pupae P8 92
head, virgin 1-day female 32
head, virgin 4-day female 33
head, virgin 20-day female 20
head, mated 1-day female 35
head, mated 4-day female 29
head, mated 20-day female 21
head, mated 1-day male 44
head, mated 4-day male 47
head, mated 20-day male 43
salivary gland, larvae L3 wandering 103
salivary gland, white prepupae 713
digestive system, larvae L3 wandering 224
digestive system, 1-day adult 46
digestive system, 4-day adult 62
digestive system, 20-day adult 71
fat body, larvae L3 wandering 46
fat body, white prepupae 195
fat body, pupae P8 89
carcass, larvae L3 wandering 272
carcass, 1-day adult 29
carcass, 4-day adult 27
carcass, 20-day adult 25
ovary, virgin 4-day female 56
ovary, mated 4-day female 53
testis, mated 4-day male 22
accessory gland, mated 4-day male 39
Expression Level Scale
N
Mo
High
Very high

Fig. 3 High-throughput expression data. Bar graph of modENCODE RNA-seq tissue expression data for the *zipper* gene, as shown within the 'Expression Data' section of the Gene Report. In this example, the view is configured to be linear and scaled to the gene's maximum expression level

versatile ‘RNA-Seq Profile’ tool allows retrieval of genes with user-defined RNA-Seq expression patterns and levels (values are binned); this tool can be accessed from a pictograph on the homepage, from the Tools menu of the NavBar, or from the QuickSearch ‘Expression’ tab. A section of the QuickSearch 'Expression' tab also provides options to ‘Search for similarly expressed genes', using the modENCODE RNA-Seq datasets. The ‘RNA-Seq by Region’ tool returns the average RPKM over a specified genomic region and also offers the option of a gene-specific query that returns an exon-by-exon RPKM count; this tool can be accessed from the Tools menu or from the ‘High-Throughput Expression Data’ subsection of the Gene Report.

5.3 External Link-Outs

The final ‘Expression Data’ subsection on the Gene Report is ‘External Data and Images’, with gene-specific links to other databases that include expression data for *Drosophila*. There are links to the original BDGP in situ data, as well as FlyExpress [23] analyses that use the BDGP data and allow a ‘Find Similar Patterns’ option. The FlyAtlas link provides the underlying microarray data used for the bar graphs described above. For the FlyExpress and SliceSeq [24] databases, sample images are shown.

6 Interactions

Both genetic and physical interaction data are presented in FlyBase. The former are primarily recorded in Allele Reports, while the latter are given in dedicated Physical Interaction Reports; both are summarized in the ‘Interactions and Pathways’ section of the relevant Gene Report page. Both types of interaction can be viewed either as a graphical ‘network diagram’ provided by esyN [25] or within the FlyBase Interactions Browser tool that includes additional viewing and configuration options [4]. The ‘External Data’ subsection of the ‘Interactions and Pathways’ part of the Gene Report provides link-outs to relevant pages at third-party interaction databases, including BioGRID [26], DroID [27], and InterologFinder [28].

6.1 Genetic Interactions

Genetic interaction data are recorded at the allele level using phenotypic class and anatomy CV terms (and optional qualifiers), similar to phenotype annotations but with the addition of terms such as ‘enhanced by’ or ‘suppressor of’ to indicate the nature of the interaction, together with the interacting allele. (Negative results, e.g., ‘not enhanced by’, are also captured.) In addition to enhancer/suppressor-type interactions, synthetic phenotypes that are present in a mutant combination but absent in single mutant conditions are also captured. All these interaction statements appear in the ‘Interactions’ section of the Allele Report, alongside free text clarifications where necessary. All genetic interaction statements in FlyBase are curated from the published literature.

The allele-level genetic interaction statements are used to compute a gene-level summary of these data, and this appears as a table within the 'Summary of Genetic Interactions' subsection in the Gene Report. This table shows the interacting genes, the nature of the interaction (limited to enhancer/suppressor-type interactions), and the reference(s) supporting the interaction. These gene-level interactions are used to power the esyN network diagram.

Querying genetic interactions: Both allele-level and gene-level genetic interaction data can be queried directly via the Interactions Browser tool. More specific and/or combinatorial searches may be conducted using QueryBuilder.

6.2 Physical Interactions

The Physical Interaction Report displays pairwise physical interaction data for gene products, either protein–protein or RNA–protein. Each report includes the experimental assays used (e.g., co-immunoprecipitation, peptide mass fingerprinting), the role of each protein in an assay (e.g., bait or prey; whether a tagged or endogenous protein was used), the esyN network diagram, and a link to the Interactions Browser tool. With the goal of producing a set of high-confidence pairwise interactions, our current focus is on smaller-scale physical interaction data curated from the literature, which usually include multiple types of support for a described interaction. High-throughput interaction datasets are curated only when the authors take care to filter out false positives—criteria for curation may include: multiple negative control purifications, accounting for protein abundance in assessing the likelihood that a purified factor is a contaminant, a calculation of the confidence level, and an explicit cut-off to separate high confidence and lower confidence interactions. Examples are the DPiM dataset [29], the Hippo Pathway Interactome [30], and the ECIA extracellular interactome [31].

Within the 'Summary of Physical Interactions' subsection of the Gene Report, all pairwise physical interactions involving that gene product are presented in tabulated form, with assays used, attributed publications, and links to the corresponding Physical Interaction Report.

Querying physical interactions: The 'Simple' or the 'Data Class (physical interactions)' tabs of QuickSearch can be used to find interactions involving a given gene or to search for assay terms present in the Physical Interaction Report.

7 Genomic Data

Genomic data in FlyBase comprise gene model annotations (i.e., the exon–intron structure and transcription and translation start/termination sites of genes) and any other sequence-based features that can be mapped to specific genomic coordinates, whether endogenous (e.g., regulatory regions, origins of replication) or describing a lesion/reagent (e.g., insertion sites, RNAi amplicons). All these data are viewable through the FlyBase implementation of GBrowse [32] (Fig. 4). Many are also associated with discrete Sequence Feature Reports and are searchable through the FeatureMapper tool [11] (Fig. 5). Note that FlyBase currently uses release 6 of the sequenced *D. melanogaster* genome [33]—the Coordinates Converter tool, accessible from the Tools menu of the NavBar, can be used to convert data from release 3, 4, or 5 coordinates [32].

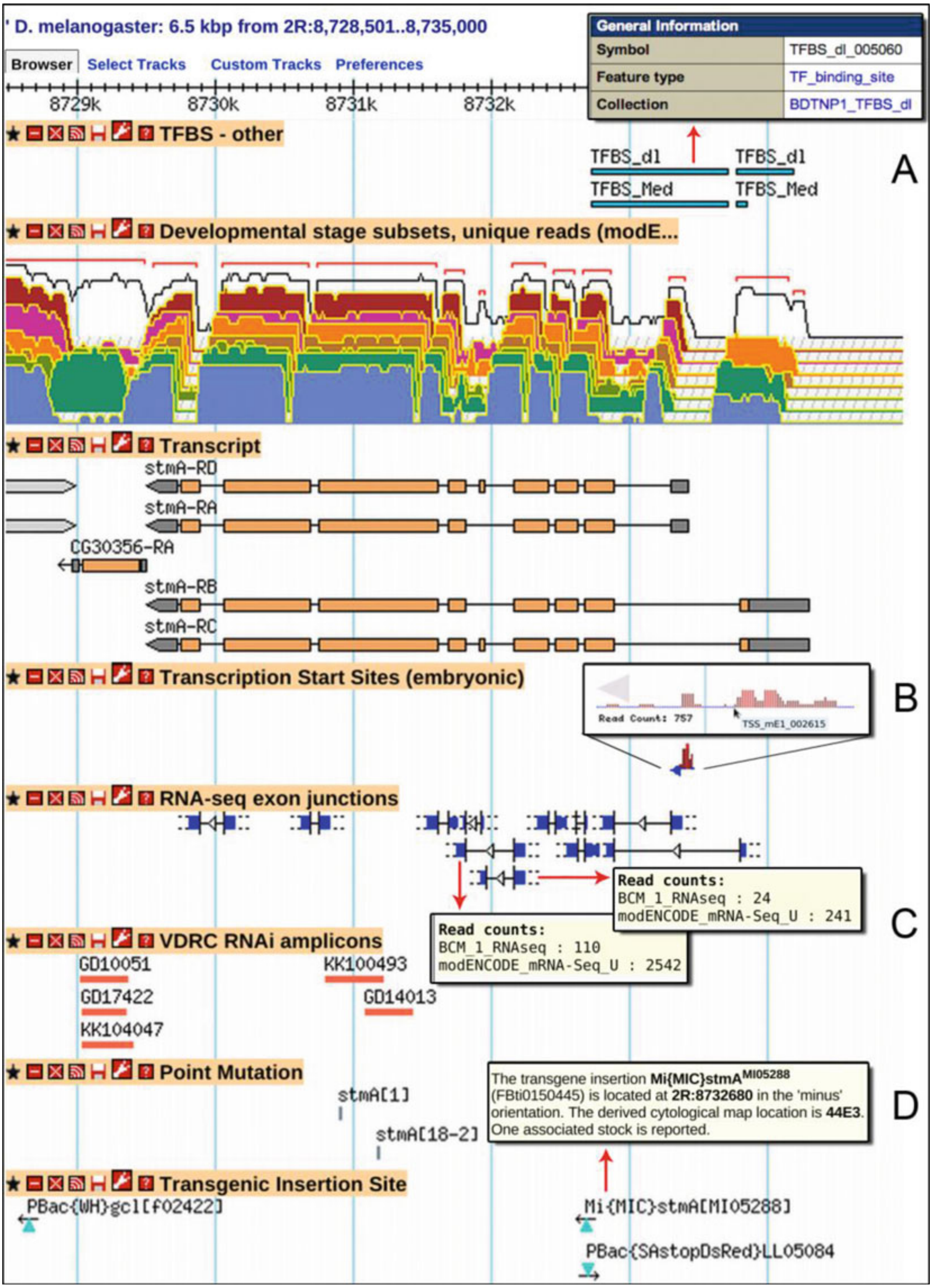
7.1 Gene Model Annotations in FlyBase

For *D. melanogaster*, FlyBase has produced manually annotated gene models for over a decade [34]. Since 2010, RNA-Seq data [19, 35] and new transcription start site data [33, 36] have supported many major changes in the gene model annotations for this species. This prompted a comprehensive review of all existing gene models and the annotation of several thousand new genes, primarily long non-coding RNA genes [34]. Transcript and protein data are tabulated in the 'Gene Model and Products' section of the Gene Report (Fig. 2), with links therein to more detailed reports. *D. melanogaster* gene models continue to be updated regularly based on new high-throughput and literature-based data. An updated gene model set is submitted to GenBank approximately once a year and serves as the NCBI RefSeq set for this species.

For eight of the other sequenced *Drosophila* species (*D. ananassae, D. erecta, D. pseudoobscura pseudoobscura, D. simulans, D. yakuba, D. mojavensis, D. virilis*, and *D. willistoni*), the long-standing CAF1-generated gene model annotations [37] have been replaced recently by sets generated by NCBI as part of their GNOMON annotation pipeline [38]. Gene model annotations for three other species (*D. grimshawi, D. persimilis*, and *D. sechellia*) have not been updated owing to poor genome assembly quality or to lack of RNA-Seq data, which provides the primary basis for robust annotation by the GNOMON pipeline.

Querying gene model data: Gene models can be searched directly in GBrowse (Subheading 7.2) or via the CytoSearch or FeatureMapper tools (both accessible via the Tools menu of the NavBar). The FlyBase BLAST tool allows sequence-based queries against annotated transcripts or proteins from the 12 *Drosophila* species mentioned above. For *D. melanogaster*, Sequence Ontology terms and controlled comments have been used extensively to describe gene models and transcripts [34, 39]. These enable queries for exceptional cases, such as all genes with dicistronic transcripts or all transcripts annotated with non-canonical translation

Fig. 4 (continued) *(C)* Mousing over an individual RNA-Seq junction produces a pop-up that provides read counts; relative read counts of the two selected junctions indicate that the small alternative exon is not present in the majority of *stmA* transcripts. (*D*) Mousing over a Transgenic Insertion Site produces a pop-up with additional information, including whether there is a publicly available stock. For other genomic reagents, such as the Point Mutations and RNAi amplicons shown, availability of stocks can be determined by clicking through to the full reports



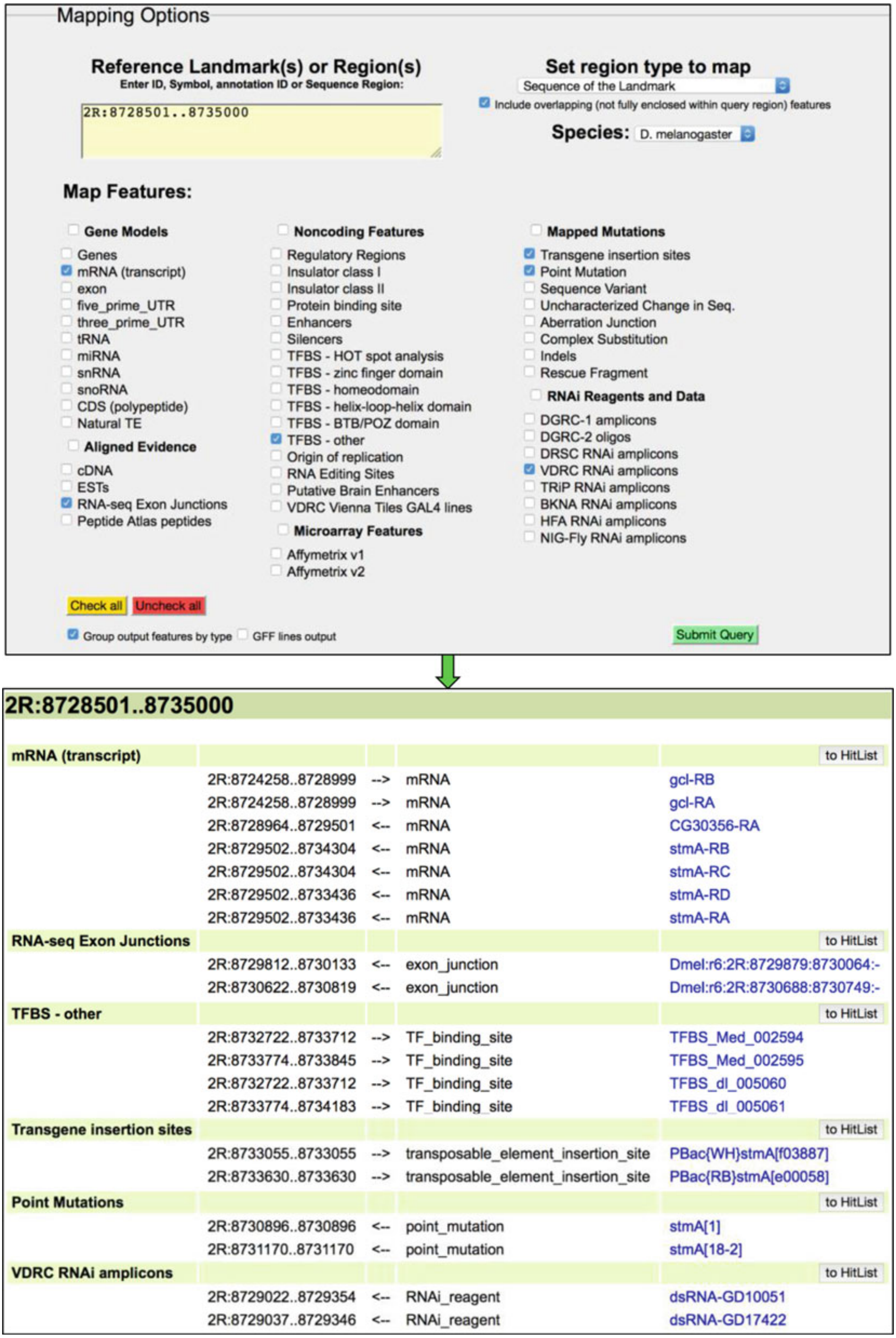
General Information	
Symbol	TFBS_dl_005060
Feature type	TF_binding_site
Collection	BDTNP1_TFBS_dl

```
Read counts:
BCM_1_RNAseq : 24
modENCODE_mRNA-Seq_U : 241
```

```
Read counts:
BCM_1_RNAseq : 110
modENCODE_mRNA-Seq_U : 2542
```

The transgene insertion **Mi{MIC}stmA**^{MI05288} (FBti0150445) is located at **2R:8732680** in the 'minus' orientation. The derived cytological map location is **44E3**. One associated stock is reported.

Fig. 4 GBrowse. This view of sequence features and genomic data in the region of the *stmA* gene has been customized by: (1) using the 'Select Tracks' option at the upper left to turn several default tracks off and new data tracks on; (2) using track-specific options accessed from the wrench/spanner icon in the title bar to select a subset of the RNA-Seq dataset shown; and (3) dragging tracks to preferred positions vertically. (*A*) Clicking on most objects in GBrowse links to the full FlyBase report for that feature, as shown here for a transcription factor binding site (TFBS) sequence feature. (*B*) By zooming in, the details of a defined Transcription Start Site (TSS) can be seen, including a bar graph of TSS distribution within the defined region and the total number of reads.



Mapping Options

Reference Landmark(s) or Region(s)

Enter ID, Symbol, annotation ID or Sequence Region:

```
2R:8728501..8735000
```

Set region type to map Sequence of the Landmark

☑ Include overlapping (not fully enclosed within query region) features

Species: D. melanogaster

Map Features:

□ **Gene Models**

- □ Genes
- ☑ mRNA (transcript)
- □ exon
- □ five_prime_UTR
- □ three_prime_UTR
- □ tRNA
- □ miRNA
- □ snRNA
- □ snoRNA
- □ CDS (polypeptide)
- □ Natural TE

□ **Aligned Evidence**

- □ cDNA
- □ ESTs
- ☑ RNA-seq Exon Junctions
- □ Peptide Atlas peptides

□ **Noncoding Features**

- □ Regulatory Regions
- □ Insulator class I
- □ Insulator class II
- □ Protein binding site
- □ Enhancers
- □ Silencers
- □ TFBS - HOT spot analysis
- □ TFBS - zinc finger domain
- □ TFBS - homeodomain
- □ TFBS - helix-loop-helix domain
- □ TFBS - BTB/POZ domain
- ☑ TFBS - other
- □ Origin of replication
- □ RNA Editing Sites
- □ Putative Brain Enhancers
- □ VDRC Vienna Tiles GAL4 lines

□ **Microarray Features**

- □ Affymetrix v1
- □ Affymetrix v2

□ **Mapped Mutations**

- ☑ Transgene insertion sites
- ☑ Point Mutation
- □ Sequence Variant
- □ Uncharacterized Change in Seq.
- □ Aberration Junction
- □ Complex Substitution
- □ Indels
- □ Rescue Fragment

□ **RNAi Reagents and Data**

- □ DGRC-1 amplicons
- □ DGRC-2 oligos
- □ DRSC RNAi amplicons
- ☑ VDRC RNAi amplicons
- □ TRiP RNAi amplicons
- □ BKNA RNAi amplicons
- □ HFA RNAi amplicons
- □ NIG-Fly RNAi amplicons

☑ Group output features by type □ GFF lines output

2R:8728501..8735000

mRNA (transcript)				to HitList
	2R:8724258..8728999	-->	mRNA	gcl-RB
	2R:8724258..8728999	-->	mRNA	gcl-RA
	2R:8728964..8729501	<--	mRNA	CG30356-RA
	2R:8729502..8734304	<--	mRNA	stmA-RB
	2R:8729502..8734304	<--	mRNA	stmA-RC
	2R:8729502..8733436	<--	mRNA	stmA-RD
	2R:8729502..8733436	<--	mRNA	stmA-RA
RNA-seq Exon Junctions				to HitList
	2R:8729812..8730133	<--	exon_junction	Dmel:r6:2R:8729879:8730064:-
	2R:8730622..8730819	<--	exon_junction	Dmel:r6:2R:8730688:8730749:-
TFBS - other				to HitList
	2R:8732722..8733712	-->	TF_binding_site	TFBS_Med_002594
	2R:8733774..8733845	-->	TF_binding_site	TFBS_Med_002595
	2R:8732722..8733712	-->	TF_binding_site	TFBS_dl_005060
	2R:8733774..8734183	-->	TF_binding_site	TFBS_dl_005061
Transgene insertion sites				to HitList
	2R:8733055..8733055	-->	transposable_element_insertion_site	PBac{WH}stmA[f03887]
	2R:8733630..8733630	-->	transposable_element_insertion_site	PBac{RB}stmA[e00058]
Point Mutations				to HitList
	2R:8730896..8730896	<--	point_mutation	stmA[1]
	2R:8731170..8731170	<--	point_mutation	stmA[18-2]
VDRC RNAi amplicons				to HitList
	2R:8729022..8729354	<--	RNAi_reagent	dsRNA-GD10051
	2R:8729037..8729346	<--	RNAi_reagent	dsRNA-GD17422

Fig. 5 FeatureMapper. (1) Query interface. In this example, a single sequence range has been entered and several mapped features have been selected (corresponding to the GBrowse view shown in Fig. 4). The default output is to group features by type in HTML format. (2) Mapping results. The sequence coordinates, strand and

start sites, using QueryBuilder or the 'Simple' tab of QuickSearch. 'Transcripts' and 'Polypeptides' are data class options in QuickSearch and QueryBuilder, thus allowing class-specific and field-specific queries of gene products, respectively.

7.2 GBrowse

GBrowse, a genome annotation viewer that is part of the Generic Model Organism Database (GMOD) tool suite [40], is used by FlyBase to show gene models and supporting data, such as cDNAs, ESTs, RNA-Seq data, transcription start sites, gene predictions, and aligned proteins [32] (Fig. 4). In addition, this versatile tool allows representation of many other types of sequenced-based data and reagents—essentially anything that maps to the genome can be represented on GBrowse. By using the 'Select Tracks' option, the user can choose to view mapped genetic variants such as mutational lesions, transgenic insertions, aberration extents, and aberration breakpoints; regions carried on transgenic constructs such as rescue fragments and RNAi reagents; microarray oligonucleotides and RNAi amplicons; or high-throughput mapping of transcription-factor binding sites, insulator elements, and RNA-editing sites. When zoomed in to a range of 100–200 bp, the tracks indicating forward/reverse translation and 'DNA/GC content' switch to the nucleotide or protein sequence. The current version of this tool is GBrowse2 [41], which allows rapid customization options: for example, a selected track can be moved by simply dragging the track title bar vertically, and tracks can be closed, opened, or removed using the icons in the track title bar. Moreover, navigation within a genomic region has been facilitated by limited smooth-track panning (side-to-side sliding) and by a function that allows the user to lasso a smaller region and zoom in.

RNA-Seq expression data [19, 42] are particularly informative when viewed in GBrowse. In 2010, FlyBase debuted a new topographical presentation of these data for GBrowse that allows visual assessment of many RNA-Seq tracks at once (Fig. 4). By clicking on the wrench/spanner icon in the track title, the presentation can be changed from *log2* to linear, and from tilted to vertical; specific tracks corresponding to different tissues and/or development stages can be shown or hidden. RNA-Seq exon junction data [19, 35], presented in a separate track on GBrowse, are extremely useful for judging alternative splicing and isoform-specific expression.

GBrowse supports a number of download options, accessible from the drop-down menu on the upper right of the page, including a FASTA file of the sequence shown and an HTML table view or a GFF file of all the mapped genes and features selected. The sequence of a lassoed genomic region can also be viewed and downloaded.

Fig. 5 (continued) symbol of each sequence feature are presented in a table. Links to hit lists are shown for each group to enable further analyses or downloads. (Note that only a subset of hits of each type is shown in this example.)

GBrowse can be accessed from one of the pictograph buttons at the top of the homepage or via the Tools menu on the NavBar. In addition, there is a link to the appropriate genomic region in GBrowse on the reports for every localized gene and mapped sequence feature. A genomic BLAST hit obtained using the FlyBase BLAST tool also includes a link to the relevant region in GBrowse.

7.3 Sequence Features and Other Genomic Data Tracks

'Sequence features' are defined as regions of DNA/RNA that can be mapped to the genome sequence and to which a discrete function can usually be ascribed. They include endogenous regions such as enhancers, insulators, transcription factor binding sites, transcription start sites and origins of replication, as well as experimental reagents that map to the genome, such as RNAi reagents and putative enhancer element constructs. Sequence features appear within discrete tracks on GBrowse and are associated with dedicated report pages. (Note that certain GBrowse tracks, including point mutations, transgenic insertions, and aberration extents, are not classed as 'sequence features' and are instead associated with specific Allele, Insertion, or Aberration Reports.) Most sequence features currently in FlyBase were generated in response to the modENCODE project [22] and similar large-scale experiments [43, 44].

The Sequence Feature Report is flexible, in order to accommodate many different types of genome-associated data. The typical report includes a link to the Large Dataset Metadata Report (*see* Subheading 9.1) to which it belongs, the sequence itself and its genomic location, a genome snapshot showing the alignment of that feature alongside other sequence features included within that region, and links to any relevant external websites/databases. Clicking on the 'GBrowse' link near the top of the page goes to a full genome view of the respective region in GBrowse.

Querying sequence features and other genomic data tracks: Limited querying can be performed within GBrowse itself by specifying a 'Landmark or Region' and selecting particular tracks for display. A better approach is to use FeatureMapper (Fig. 5), which provides an intuitive interface for retrieval of specified genome features in one or more genomic regions, with results presented in a convenient table that includes an option to export to a hit list where possible. The CytoSearch tool allows retrieval of genes, aberrations, and transgenic insertions mapped to the genomic sequence. Sequence features are also included as a specific option in the 'Data Class' tab of QuickSearch.

8 Reagents

There are several ways to find reagents associated with a specific gene or genomic region. The 'Stocks and Reagents' section of the Gene Report is a good place to start. Here, subsections list publicly available fly stocks, genomic and cDNA clones, cell-based RNAi reagents and antibodies described in the published literature (Fig. 2). Other reagents are best found by searching a genomic region of interest using GBrowse, FeatureMapper, or CytoSearch. For example, the Janelia/GMR [45] and VDRC [44] putative enhancer collections are not associated with specific genes, while some classes of transgenic insertions are not listed in the Gene Report. Moreover, a visual representation of the location of a sequence-based reagent relative to the gene of interest is often informative when planning experiments.

8.1 Stocks

Stock Reports display the stock list genotype and the source collection, together with the stock number hyperlinked to the specific record at the appropriate stock center to facilitate ordering. There are links to Stock Reports from other appropriate reports (primarily alleles, aberrations, transgenic constructs, and insertions) throughout FlyBase. The Bloomington *Drosophila* Stock Center is the most widely represented source, though many others are included—a complete list can be found in the 'Links' menu on the NavBar.

Querying stocks: Stocks can be searched specifically by selecting 'stocks' in the 'Data Class' tab of QuickSearch.

8.2 Strains

FlyBase Strain Reports contain data about wild type strains such as 'Oregon-R', significant mutant strains such as 'iso-1' (the *D. melanogaster* strain sequenced by the BDGP [33]), as well as the 200 or so inbred lines generated by the *Drosophila* Genetics Reference Panel [46]. The reports include information on the origin and history of the strain alongside any known genetic or phenotypic components (e.g., the 'iso-1' strain harbors several mutations). Where relevant, links are also provided to Large Dataset Metadata Reports (Subheading 9.1) that describe strain collections, and to Stock Reports to facilitate ordering. (Note that stocks are instances of strains in theory, but they are effectively distinct in time and place and may have characteristics that differ from the strains from which they descended.)

Querying strains: Strains can be searched using the QuickSearch 'Simple' tab.

8.3 Cell Lines

Cell Line Reports display data obtained from the *Drosophila* Genomics Resource Center (DGRC) on cell lines, such as 'Kc167' or 'S2R+'. The reports include the source and development stage of each line, its sex and karyotype (where known), and any parental or descendent lines. A link back to the DGRC is also provided for additional data and ordering information.

Querying cell lines: Cell lines can be searched specifically by selecting 'cell lines' in the 'Data Class' tab of QuickSearch.

8.4 cDNAs

cDNAs are shown in GBrowse and appear in the 'Stocks and Reagents' section of the Gene Report of the aligned gene(s). Links from GBrowse go to the GenBank report; links from the Gene

Report go to the FlyBase Clone Report. The Clone Report includes the sequence, links to GenBank, and fields for ‘Known Problems’ and ‘FlyBase assessment’. Examples of known problems are clones that are chimeric or that contain genomic DNA or transposon sequences. The FlyBase assessment field displays a note if the clone has been replaced, for example “Caution: This cDNA clone replaced by FI01005”. There is also a link to the DGRC where clones are available from that resource.

Querying cDNAs: cDNA clones can be searched specifically by selecting ‘clones’ in the ‘Data Class’ tab of QuickSearch. FeatureMapper should be used to find cDNAs associated with a specific gene or genomic region.

9 Integrated Reports

As the amount of *Drosophila* data and resources increase in FlyBase, it has become both necessary and useful to organize and integrate related data into discrete sets or collections. This has multiple benefits, including the ability to associate metadata across a range of related entities, and to present related data to users in new ways that aid comprehension. To date, FlyBase has developed three types of such integrated reports.

9.1 Large Dataset Metadata

Large Dataset Metadata Reports, previously named Library/Collection Reports, provide information on large datasets and reagent collections that apply to the set as a whole. Examples of datasets are the protein interaction network defined by the *Drosophila* Protein interaction Mapping (DPiM) project [29], the set of RAMPAGE transcription start sites [36], and datasets generated by the modENCODE project [22]. Examples of collections are the set of dsRNA amplicons used for RNAi-knockdown assays in cell culture by the *Drosophila* RNAi Screening Center [49], the set of defined X-chromosome duplications made by the Bloomington Stock Center [50], and several large construct and insertion collections. Metadata describing cDNA libraries are also captured in this format. The Large Dataset Metadata Report includes the type of dataset or collection, a brief description of the set, a summary of the experimental details, and a link to download all the associated features. Links to external data repositories and reagent sources are provided where relevant. The ‘Description’ field of the dataset report is propagated to each member report; reciprocal links are provided.

Querying large dataset metadata: The ‘Simple’ or the ‘Data Class (large dataset metadata)’ tabs of QuickSearch can be used to find datasets and collections of interest.

9.2 Gene Groups

Gene Group Reports have been introduced to allow easy access to, and analysis of, related sets of *D. melanogaster* genes and their associated data [47]. Examples of gene groups include members of a gene family (Actins, Wnts, etc.), subunits of a protein complex (proteasome, ribosome, etc.), or other functional groupings (protein kinases, Ubiquitin E3 ligases, etc.). All gene groups in FlyBase are based on published literature and the basis for the membership of each group is clearly attributed. The main feature of these reports is a 'Members' table that lists the genes comprising the group, arranged into a series of subgroups where appropriate. Buttons are provided to facilitate the downloading of associated data (phenotypes, expression data, protein interactions, etc.) using Batch Download (Subheading 10.2), or to further refine or analyze the gene set by exporting it to a standard hit list. Also shown are links to equivalent gene groups for other organisms, including nematodes (WormBase [48]) and humans (HGNC [8]). To aid navigation, the 'Families, Domains and Molecular Function' section of the Gene Report contains a link to any associated gene group(s) (Fig. 2).

Querying gene groups: Gene groups can be retrieved by entering the symbol/name of a group or any member gene in the 'Gene Groups' tab of QuickSearch. This tab also includes a link to a browsable list of all current gene groups in FlyBase.

9.3 Human Disease Models

Human Disease Model Reports provide a less specialized entry point into FlyBase for researchers interested in *Drosophila* models of human disease [17]. Data from numerous outside sources, including OMIM, and from recent reviews are presented in a general 'Disease Summary' section, followed by information on orthology between a human gene implicated in the disease and the related *Drosophila* gene(s). For many diseases, multiple causative genes have been implicated; OMIM describes these as different disease subtypes and groups them into 'phenotypic series'. In the Human Disease Model Report, such a phenotypic series of subtypes is presented in a table titled 'Related Diseases', which includes links to other relevant Human Disease Model Reports and provides a quick view of which disease subtypes have been modeled in flies.

The major portion of the disease report is devoted to 'Experimental Findings' in *Drosophila*, focusing on disease-related implications and results. Descriptions of specific experiments are meant to be generally accessible, with links to Allele Reports with more detailed information. Results may include data using both fly genes and human genes introduced into flies. The 'Experimental Findings' section initiates with a FlyBase-authored summary that presents a concise review, including phenotypes, interactions, and suitability of the model for drug assays; in addition, new findings and emerging mechanistic themes are highlighted. At the end of this section, a link to the FlyBase Disease Wiki is provided; comments and contributions from users are encouraged, especially those with expertise in the specific disease model. The last sections of the report draw relevant data from other sections of FlyBase, including physical interaction data for the orthologous *Drosophila* gene(s), a table of genetic reagents and stocks useful for investigations of human disease, and a table of Disease Ontology-based annotations of alleles used for that disease model (*see* Subheading 4.3).

There are links to relevant Human Disease Model Reports in the 'Human Disease Model Data' section of Gene Reports (Fig. 2). Note that many such links are found in FlyBase Gene Reports for human genes (e.g., *Hsap\SNCA* and *Hsap\TARDBP*).

Querying Human Disease Model Reports: These reports can be found by using the 'Human Disease' or 'Simple' tabs of QuickSearch, or by searching the Disease Ontology within the Vocabularies tool.

10 Bulk Data Analysis and Downloads

Users increasingly want to be able to process data in bulk. They may have generated a hit list of genes (or any other data class) within FlyBase, or have a list of IDs from elsewhere to upload, and wish to analyze/refine this list or obtain associated data. Alternatively, users may wish to directly obtain bulk data files corresponding to a particular data type for processing off-line.

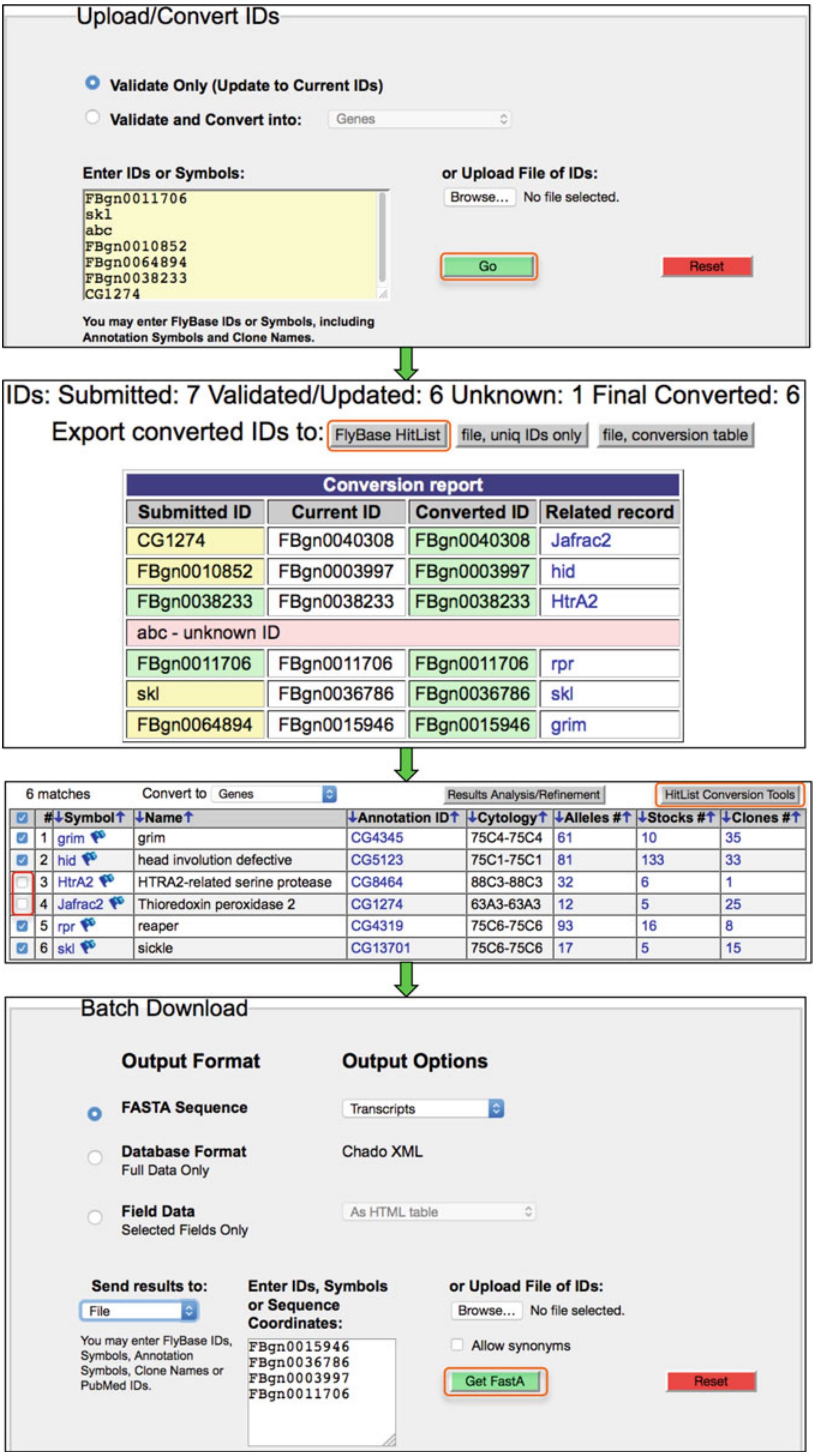
10.1 Uploading and Analyzing Data

A list of IDs (e.g., gene symbols or CG numbers, allele or insertion symbols, FlyBase identifiers) can be pasted or uploaded into the Upload/Convert IDs tool (Fig. 6; accessed via the Tools menu on the NavBar). This tool will then validate the list, updating any obsolete IDs to the current version where possible, and generate a 'Conversion report' clearly indicating if any of the submitted IDs failed verification. The user can choose to correct these cases, or ignore them before proceeding to convert the list into a standard FlyBase hit list (*see* Subheading 2.2). This list can then be further analyzed/refined before being exported or downloaded as required.

10.2 Downloading Data

Batch Download is a powerful tool for generating customized output files in various formats for most data types in FlyBase [11]. Users may arrive at Batch Download via a hit list (as described above), by navigating to it from the Tools menu of the NavBar, or by clicking on its pictograph on the homepage. If the first, then the input list will be pre-filled (Fig. 6); otherwise the user can paste in or upload a list of symbols or IDs directly. Depending on the

Fig. 6 (continued) (*first column, red box*). The 'HitList Conversion Tools' button (*orange box*) is then clicked and 'Export to Batch Download' is selected (not shown). (4) The Batch Download interface shows the search box pre-populated with the four final gene IDs. In this example, transcript sequences in FASTA format have been selected for download with the results being sent to a 'File'



Upload/Convert IDs

● Validate Only (Update to Current IDs)

○ Validate and Convert into: Genes

Enter IDs or Symbols:

```
FBgn0011706
skl
abc
FBgn0010852
FBgn0064894
FBgn0038233
CG1274
```

or Upload File of IDs: Browse... No file selected.

You may enter FlyBase IDs or Symbols, including Annotation Symbols and Clone Names.

IDs: Submitted: 7 Validated/Updated: 6 Unknown: 1 Final Converted: 6

Export converted IDs to: FlyBase HitList | file, uniq IDs only | file, conversion table

Conversion report

Submitted ID	Current ID	Converted ID	Related record
CG1274	FBgn0040308	FBgn0040308	Jafrac2
FBgn0010852	FBgn0003997	FBgn0003997	hid
FBgn0038233	FBgn0038233	FBgn0038233	HtrA2
abc - unknown ID			
FBgn0011706	FBgn0011706	FBgn0011706	rpr
skl	FBgn0036786	FBgn0036786	skl
FBgn0064894	FBgn0015946	FBgn0015946	grim

6 matches Convert to Genes

☑	#	Symbol	Name	Annotation ID	Cytology	Alleles #	Stocks #	Clones #
☑	1	grim	grim	CG4345	75C4-75C4	61	10	35
☑	2	hid	head involution defective	CG5123	75C1-75C1	81	133	33
□	3	HtrA2	HTRA2-related serine protease	CG8464	88C3-88C3	32	6	1
□	4	Jafrac2	Thioredoxin peroxidase 2	CG1274	63A3-63A3	12	5	25
☑	5	rpr	reaper	CG4319	75C6-75C6	93	16	8
☑	6	skl	sickle	CG13701	75C6-75C6	17	5	15

Batch Download

Output Format / Output Options

● FASTA Sequence — Transcripts

○ Database Format Full Data Only — Chado XML

○ Field Data Selected Fields Only — As HTML table

Send results to: File

You may enter FlyBase IDs, Symbols, Annotation Symbols, Clone Names or PubMed IDs.

Enter IDs, Symbols or Sequence Coordinates:

```
FBgn0015946
FBgn0036786
FBgn0003997
FBgn0011706
```

or Upload File of IDs: Browse... No file selected.

□ Allow synonyms

Fig. 6 Batch upload and download. (1) The Upload/Convert IDs tool is used to type/paste in a mixture of gene identifiers. (2) The resulting validation report shows that six of the seven entries were validated/updated. Note that in two cases a secondary FBgn ID was entered and updated successfully. The 'FlyBase HitList' button (*orange box*) is then clicked to export the IDs. (3) The resulting hit list shows the six validated genes in a table with columns appropriate to the gene data class. At this stage, two of the genes have been de-selected

nature of the input and the desired outcome, the output format can then be specified as ‘FASTA Sequence’ (with the option to further specify introns, UTRs, CDS, etc.), ‘Database Format’ (XML), or as ‘Field Data’ (with output options of an HTML table, a tab-separated value (tsv) file, or in the same format as the precomputed files described below). If the ‘Field Data’ option is selected, the user can then specify any combination of data fields (appropriate to the given data class) from a page styled in the same format as a standard FlyBase report page.

Bulk files of FlyBase data can be downloaded using our FTP site (ftp://ftp.flybase.org/releases/) or the ‘Downloads’ menu of the NavBar on the website (see the ‘Overview’ page under the Downloads menu for more details). ‘Precomputed files’ contain particular slices of FlyBase data that users or collaborators have requested over the years or are otherwise difficult to obtain in bulk (Table 3). Notable recent additions include *D. melanogaster* unique protein isoforms, RPKM gene expression values, gene groups, and physical

Table 3
Precomputed bulk data files available from FlyBase

File name	Brief description
Genetic/genomic data	
gene_map_table_*	Localization information for *Drosophila* genes
gene_orthologs_*	Dmel genes and orthologs in sequenced *Drosophila* species
gene_association.fb	Gene Ontology terms assigned to Dmel genes
gene_summaries_*	Automated gene summaries as shown on Gene Reports
gene_rpkm_report_*	Dmel gene expression values based on RNA-Seq
dmel_unique_protein_isoforms_*	Dmel genes and their unique protein isoforms
allele_phenotypic_data_*	CV phenotypic data associated with alleles
allele_human_disease_model_data_*	Disease model data associated with alleles
gene_genetic_interactions_*	Summary of Dmel gene-level genetic interactions
allele_genetic_interactions_*	Allele-level genetic interactions with CV terms
physical_interactions_*	Dmel gene pairs whose products physically interact
insertion_mapping_*	Localization information for Dmel insertions
Integrated data	
dataset_metadata_*	All dataset/collections and all associated features
gene_group_data_*	All gene groups, relationships and members

(continued)

Table 3
(continued)

File name	Brief description
Reagents	
genomic_clone_data_*	Genomic clone IDs, names, and accession numbers
cDNA_clone_data_*	cDNA/EST IDs, names, library, and accession numbers
Other data files	
fbrf_pmid_pmcid_doi_fb_*	All IDs for references in FlyBase that have a PMID
species.ab	Data on all species in FlyBase, including abbreviations
*Correspondence table*s	
fb_synonym_fb_*	Symbols, names and synonyms for most features in FlyBase
fbgn_NAseq_Uniprot_*	FlyBase gene IDs ⇔ nucleotide and protein accessions
fbgn_annotation_ID_*	Current and secondary FBgn and annotation IDs for genes
fbgn_fbtr_fbpp_*	FlyBase gene IDs ⇔ FlyBase transcript and polypeptide IDs
fbal_to_fbgn_*	FlyBase allele IDs ⇔ FlyBase gene IDs
cyto-genetic-seq	Dmel cytogenetic map ⇔ genetic map ⇔ genomic coord.
Ontology files	
fly_anatomy	Fly anatomy ontology (FBbt)
fly_development	Fly developmental stage ontology (FBdv)
flybase_controlled_vocabulary	FlyBase miscellaneous ontology (FBcv)
flybase_stock_vocabulary	FlyBase stock ontology (FBsv)
go-basic	Gene ontology (GO)
image	FlyBase image ontology (FBbi)
so	Sequence ontology (SO)
doid	Human disease ontology (DO)

Note that only a subset of the available files is shown here
*Represents the release number, for example 'fb_2015_04'; 'Dmel' = *D. melanogaster*

interactions. Also included are several useful correspondence tables and the ontology files used in FlyBase (Table 3). In addition, Chado XML (database format) files are provided for all FlyBase data classes and comprehensive sets of FASTA, GFF, and GTF files are available for the 12 originally sequenced and annotated *Drosophila* species (*see* Subheading 7.1). The FASTA files comprise many different cuts of genomic data, including annotation categories such as small RNA classes and pseudogenes, components of gene model annotations such as exons, introns, UTRs and predicted translations, as well as other genome features such as transposons and intergenic sequences. As described above, Batch Download can also be used to obtain specified subsets of data in precomputed file, Chado XML or FASTA format by selecting the appropriate output options.

Most bulk files are regenerated for every release of FlyBase. Those corresponding to the current or previous (archived) versions of FlyBase are found under the appropriate submenus/subfolders on the web/FTP site. The release version used for a particular file is indicated in the file name and in the header lines of the file itself.

11 The FlyBase Community

FlyBase engages with our user community through multiple approaches. The primary method for users to get in touch with FlyBase about any matter remains our 'Contact FlyBase' page, accessible via the 'Help' menu on the NavBar or the link in the footer of any FlyBase page. All other community resources are grouped under the 'Community' menu of the NavBar and/or are found on the homepage.

If a user wants to specifically alert us to a *Drosophila* publication or data therein to be added to FlyBase, then the 'Fast Track Your Paper' (FTYP) tool should be used [51]. This tool allows the user to indicate the key genes studied and flag data types present in a paper. The resulting gene-to-publication links are submitted directly to the FlyBase database while the data type information is used to prioritize the paper for more detailed curation. We actively solicit FTYP submissions using our 'EmailAuthor' pipeline, whereby the corresponding author of a *Drosophila* publication is automatically sent an email that includes a link to a personalized FTYP form [51]. Approximately 50% of authors respond to this request, thereby reducing by half the amount of manual triaging to be done by FlyBase curators.

Our recently launched 'FlyBase Community Advisory Group' (FCAG) is a worldwide group of over 500 volunteers (lab heads, postdocs, students, technicians) who use FlyBase for a range of purposes. We contact this group up to six times per year with a survey on a variety of subjects to get feedback about how data collection, presentation, and searching on FlyBase can be improved. By consulting this relatively large, diverse group of researchers, we hope to implement changes to FlyBase that are helpful for the greatest number of people.

Users may also help improve FlyBase by contributing to the Human Disease Wiki (described in Subheading 9.3) or the FlyGene Wiki. There is a link to the latter at the top and within the 'Summaries' section of each Gene Report. This is pre-seeded with the automatically generated FlyBase summary and users are encouraged to modify or add to this text to build up a more complete and readable summary of each gene's main features and functions.

The FlyBase Forum is a Google™ Group that provides an alternative, more open platform for users to interact both with FlyBase and with each other. The forum has two areas: one for general questions and discussions about FlyBase and *Drosophila* protocols, etc., and the other for relevant job postings.

Users are made aware of new or changed features in FlyBase through any of several means. First, there are the 'News' and 'Commentary' sections of the FlyBase homepage (Fig. 1). Second, users can sign up to receive an occasional Newsletter via email by clicking the link on the homepage. The Newsletter contains release announcements, significant website updates, and other important *Drosophila* community news. Third, to obtain more frequent updates, users can follow FlyBase on Twitter™ by clicking on the icon in the footer of any FlyBase page. Fourth, users can choose to subscribe to any FlyBase record (a specific gene, transgenic construct, reference, etc.) and receive automatic updates through a feed reader by clicking the icon in the 'Recent Updates' section of any report page. Finally, users have the opportunity to see and hear about FlyBase updates in person at the Annual *Drosophila* Research Conference in the USA and the biennial European *Drosophila* Research Conference, where FlyBase representatives give presentations and are available to answer questions. Previous conference presentations and pamphlets can be obtained via the 'FlyBase Guides' link under the 'Help' menu in the NavBar.

Acknowledgements

FlyBase is funded by the National Human Genome Research Institute at the U.S. National Institutes of Health (#U41HG000739, PI W.M. Gelbart) and the U.K. Medical Research Council (#G1000968, PI N.H. Brown). Support is also provided by the Indiana Genomics Initiative and the National Science Foundation through XSEDE resources provided by Indiana University. At the time of writing, the FlyBase Consortium included William Gelbart, Norbert Perrimon, Cassandra Extavour, Kris Broll, Madeline Crosby, Gilberto dos Santos, David Emmert, Sian Gramates, Kathleen Falls, Beverley Matthews, Susan Russo Gelbart, Andrew Schroeder, Christopher Tabone, Pinglei Zhou, Mark Zytkovicz; Nicholas Brown, Giulia Antonazzo, Helen Attrill, Marta Costa, Steven Marygold, Gillian Millburn, Laura Ponting, Alix Rey, Nicole Staudt, Raymund Stefancsik, Jose-Maria Urbano; Thomas Kaufman, Josh Goodman, Gary Grumbling, Victor Strelets, Jim Thurmond; Richard Cripps, Maggie Werner-Washburne, Phillip Baker.

Dedication

We wish to mark the passing of our colleague and one of FlyBase's founders, Dr. William Gelbart, who continued in his role of PI until his death. Bill's leadership, enthusiasm, insight and humor will be greatly missed.

References

1. Drysdale R, FlyBase Consortium (2008) FlyBase: a database for the *Drosophila* research community. Methods Mol Biol 420:45–59
2. McQuilton P, St Pierre SE, Thurmond J et al (2012) FlyBase 101—the basics of navigating FlyBase. Nucleic Acids Res 40:D706–D714
3. Marygold SJ, Leyland PC, Seal RL et al (2013) FlyBase: improvements to the bibliography. Nucleic Acids Res 41:D751–D757
4. Wilson RJ, Goodman JL, Strelets VB et al (2008) FlyBase: integration and improvements to query tools. Nucleic Acids Res 36:D588–D593
5. Tweedie S, Ashburner M, Falls K et al (2009) FlyBase: enhancing *Drosophila* Gene Ontology annotations. Nucleic Acids Res 37:D555–D559
6. Kriventseva EV, Tegenfeldt F, Petty TJ et al (2015) OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res 43:D250–D256
7. Cunningham F, Amode MR, Barrell D et al (2015) Ensembl 2015. Nucleic Acids Res 43:D662–D669
8. Gray KA, Yates B, Seal RL et al (2015) Genenames.org: the HGNC resources in 2015. Nucleic Acids Res 43:D1079–D1085
9. OMIM®: Online Mendelian Inheritance in Man®. http://omim.org
10. Hu Y, Flockhart I, Vinayagam A et al (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics 12:357
11. St Pierre SE, Ponting L, Stefancsik R et al (2014) FlyBase 102—advanced approaches to interrogating FlyBase. Nucleic Acids Res 42:D780–D788
12. Morgan TH (1910) Sex limited inheritance in *Drosophila*. Science 32:120–122
13. Osumi-Sutherland D, Marygold SJ, Millburn GH et al (2013) The *Drosophila* phenotype ontology. J Biomed Semantics 4:30
14. Costa M, Reeve S, Grumbling G et al (2013) The *Drosophila* anatomy ontology. J Biomed Semantics 4:32
15. Mohr SE, Hu Y, Kim K et al (2014) Resources for functional genomics studies in *Drosophila melanogaster*. Genetics 197:1–18
16. Kibbe WA, Arze C, Felix V et al (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res 43:D1071–D1078
17. Millburn GH, Crosby MA, Gramates LS et al (2016) FlyBase portals to disease model research in *Drosophila*. Dis Model Mech 9:245–252
18. Tomancak P, Beaton A, Weiszmann R et al (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. Genome Biol 3, RESEARCH0088
19. Graveley BR, Brooks AN, Carlson JW et al (2011) The developmental transcriptome of *Drosophila melanogaster*. Nature 471:473–479
20. Robinson SW, Herzyk P, Dow JA et al (2013) FlyAtlas: database of gene expression in the tissues of *Drosophila melanogaster*. Nucleic Acids Res 41:D744–D750
21. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628
22. modENCODE Consortium, Roy S, Ernst J et al (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330:1787–1797
23. Kumar S, Konikoff C, Van Emden B et al (2011) FlyExpress: visual mining of spatiotemporal patterns for genes and publications in *Drosophila* embryogenesis. Bioinformatics 27:3319–3320
24. Combs PA, Eisen MB (2013) Sequencing mRNA from cryo-sliced *Drosophila* embryos to determine genome-wide spatial patterns of gene expression. PLoS One 8:e71820
25. Bean DM, Heimbach J, Ficorella L et al (2014) esyN: network building, sharing and publishing. PLoS One 9, e106035
26. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R et al (2015) The BioGRID interaction database: 2015 update. Nucleic Acids Res 43:D470–D478
27. Murali T, Pacifico S, Yu J et al (2011) DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene

interactions for *Drosophila*. Nucleic Acids Res 39:D736–D743

28. Wiles AM, Doderer M, Ruan J et al (2010) Building and analyzing protein interactome networks by cross-species comparisons. BMC Syst Biol 4:36
29. Guruharsha KG, Rual JF, Zhai B et al (2011) A protein complex network of *Drosophila melanogaster*. Cell 147:690–703
30. Kwon Y, Vinayagam A, Sun X et al (2013) The Hippo signaling pathway interactome. Science 342:737–740
31. Ozkan E, Carrillo RA, Eastman CL et al (2013) An extracellular interactome of immunoglobulin and LRR proteins reveals receptor-ligand networks. Cell 154:228–239
32. dos Santos G, Schroeder AJ, Goodman JL et al (2015) FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res 43:D690–D697
33. Hoskins RA, Carlson JW, Wan KH et al (2015) The Release 6 reference sequence of the *Drosophila melanogaster* genome. Genome Res 25:445–458
34. Matthews BB, Dos Santos G, Crosby MA et al (2015) Gene model annotations for *Drosophila melanogaster*: impact of high-throughput data. G3 (Bethesda) 5:1721–1736
35. Daines B, Wang H, Wang L et al (2011) The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing. Genome Res 21:315–324
36. Batut P, Dobin A, Plessy C et al (2013) High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res 23:169–180
37. Drosophila 12 Genomes Consortium, Clark AG, Eisen MB et al (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218
38. The NCBI Eukaryotic Genome Annotation Pipeline. http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/
39. Crosby MA, Gramates LS, Dos Santos G et al (2015) Gene model annotations for *Drosophila melanogaster*: the rule-benders. G3 (Bethesda) 5:1737–1749
40. GMOD: the Generic Model Organism Database project. http://gmod.org/wiki/Main_Page
41. Stein LD (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. Brief Bioinform 14:162–171
42. Berger C, Harzer H, Burkard TR et al (2012) FACS purification and transcriptome analysis of *Drosophila* neural stem cells reveals a role for Klumpfuss in self-renewal. Cell Rep 2:407–418
43. Pfeiffer BD, Jenett A, Hammonds AS et al (2008) Tools for neuroanatomy and neurogenetics in *Drosophila*. Proc Natl Acad Sci U S A 105:9715–9720
44. Kvon EZ, Kazmar T, Stampfel G et al (2014) Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. Nature 512:91–95
45. Jenett A, Rubin GM, Ngo TT et al (2012) A GAL4-driver line resource for *Drosophila* neurobiology. Cell Rep 2:991–1001
46. Mackay TF, Richards S, Stone EA et al (2012) The *Drosophila melanogaster* Genetic Reference Panel. Nature 482:173–178
47. Attrill HL, Falls K, Goodman JL et al (2016) FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. Nucleic Acids Res 44:D786–D792 doi:10.1093/nar/gkv1046
48. Harris TW, Baran J, Bieri T et al (2014) WormBase 2014: new views of curated biology. Nucleic Acids Res 42:D789–D793
49. Flockhart IT, Booker M, Hu Y et al (2012) FlyRNAi.org—the database of the *Drosophila* RNAi screening center: 2012 update. Nucleic Acids Res 40:D715–D719
50. Cook RK, Deal ME, Deal JA et al (2010) A new resource for characterizing X-linked genes in *Drosophila melanogaster*: systematic coverage and subdivision of the X chromosome with nested, Y-linked duplications. Genetics 186:1095–1109
51. Bunt SM, Grumbling GB, Field HI et al (2012) Directly e-mailing authors of newly published papers encourages community curation. Database (Oxford) 2012:bas024