



Finding information about uncharacterized *Drosophila melanogaster* genes

Stephanie E. Mohr ^{1,*} Ah-Ram Kim,¹ Yanhui Hu ¹ Norbert Perrimon^{1,2,*}

¹Department of Genetics, Blavatnik Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

²Howard Hughes Medical Institute, Boston, MA 02115, USA

*Corresponding author: Department of Genetics, Blavatnik Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA. Email: stephanie_mohr@hms.harvard.edu; *Corresponding author: Department of Genetics, Blavatnik Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA. Email: perrimon@genetics.med.harvard.edu

Genes that have been identified in the genome but remain uncharacterized with regards to function offer an opportunity to uncover novel biological information. Novelty is exciting but can also be a barrier. If nothing is known, how does one start planning and executing experiments? Here, we provide a recommended information-mining workflow and a corresponding guide to accessing information about uncharacterized *Drosophila melanogaster* genes, such as those assigned only a systematic coding gene identifier. The available information can provide insights into where and when the gene is expressed, what the function of the gene might be, whether there are similar genes in other species, whether there are known relationships to other genes, and whether any other features have already been determined. In addition, available information about relevant reagents can inspire and facilitate experimental studies. Altogether, mining available information can help prioritize genes for further study, as well as provide starting points for experimental assays and other analyses.

Keywords: *Drosophila*; knowledgebases; databases; information mining; uncharacterized genes; novel genes; data curation

Introduction

Decades of experimental and bioinformatics analyses have resulted in an excellent understanding of the number of protein-coding genes in the *Drosophila* genome (~13,900) and their specific locations along the genome. Moreover, for many *Drosophila* genes, their biochemical and/or biological functions have been elucidated. Nevertheless, functional information remains lacking for a large proportion of *Drosophila* genes. In evidence of this, for ~60% of protein-coding genes (CGs) in the *Drosophila* genome (8,653 of 13,986), there were 10 or fewer publications associated with the gene as of July 2023 (Fig. 1). About 70% of genes in this category have only a CG identifier. These identifiers act as placeholders, while a gene awaits functional characterization and assignment of a text-based name approved by FlyBase, a curated database of *Drosophila* gene information (Thurmond et al. 2019; Larkin et al. 2021).

A goal of the *Drosophila* research community is to gain functional information for a larger proportion of *Drosophila* genes. One reason this goal is important is that functional characterization of conserved genes in *Drosophila* can provide insights into the human gene “unknown,” i.e. relatively uncharacterized human genes (Rocha et al. 2023). An important way in which insights into the functions of previously uncharacterized genes can be gained is through unbiased, large-scale phenotypic screens. There is something thrilling about the moment a large-scale screen is completed and a list of genes identified in the study becomes available. The names of some well-characterized genes might show up on the list. However, as the ultimate goal of a research

study is to gain novel insights into the topic under investigation, the real excitement arguably lies in the genes listed only by their CG identifiers and other relatively uncharacterized genes. Unfortunately, that excitement might be tempered by the fact that uncharacterized genes can also be challenging to study. If there is no information about a gene and its product, one might be left wondering where to start. However, due to the efforts of individual research projects and large-scale collaborative studies, there is information available for CGs and other relatively uncharacterized *Drosophila* genes. Through mining information already obtained in studies performed in *Drosophila* or in other species, as well as by applying predictive algorithms, a lot more can be ascertained or predicted about uncharacterized *Drosophila* genes than might initially be apparent.

Several knowledgebases and meta-databases provide summary overviews of gene and protein information, access to phenotype data, lists of predicted paralogs or orthologs, disease-related information for human orthologs, and lists of relevant published papers, including the Alliance for Genome Resources (Alliance; Alliance of Genome Resources Consortium 2020), FlyBase (Larkin et al. 2021), FlyMine (Lyne et al. 2007), Gene2Function (G2F; Hu, Comjean, Mohr et al. 2017), MARRVEL (Wang, Al-Ouran et al. 2017), Monarch Initiative Explorer (Monarch; Shefchek et al. 2020), NCBI Gene, and UniProt (UniProt Consortium 2023). These databases and more specialized resources also provide access to a wealth of detailed experimental data and predictions about genes and their orthologs that can inform prioritization for follow-up studies and guide the design of experiments.

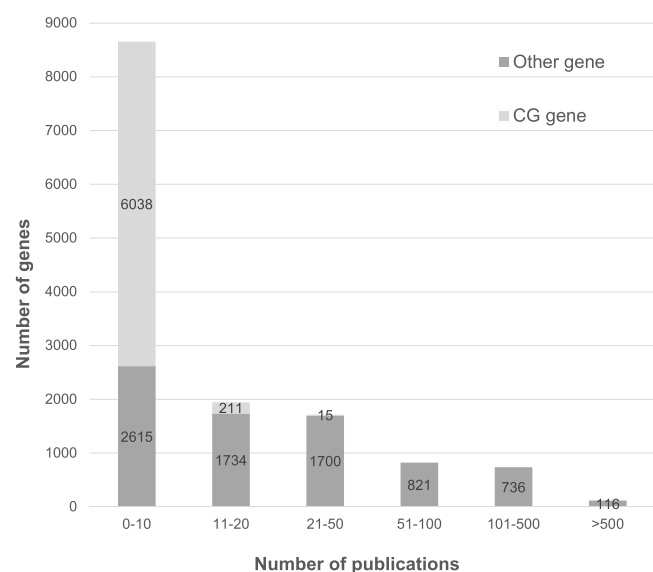


Fig. 1. Number of publications per genome-annotated *Drosophila melanogaster* protein-coding gene. On the x-axis, genes are binned based on the number of publications associated with the gene. The y-axis shows the number of CG genes (lighter gray) or other genes (darker gray) in each bin. Publications associated with >100 genes were removed prior to the analysis. Data were obtained from NCBI on 26 July 2023, at <https://ftp.ncbi.nlm.nih.gov/gene/DATA/>. For about 8,653 of the 13,986 protein-coding genes in the *Drosophila* genome, 10 or fewer publications are associated with the gene. As shown, this group is enriched for genes with only a CG systematic identifier (no text-based name).

In Fig. 2, we present a recommended information-mining workflow (Fig. 2). What follows is a corresponding guide to accessing information about uncharacterized *Drosophila* genes from multiple online resources, including meta-databases and more specialized sites. This guide updates and adds to information covered in our previous review of resources available to the *Drosophila* research community (Mohr et al. 2014) and supplements the more focused information we provided in a review of resources relevant to spatial mapping following single-cell RNA-seq studies in *Drosophila* (Mohr et al. 2021). Moreover, the information below should be viewed as additional to the many excellent navigation guides, demo videos, and other tutorials provided by experts associated with individual databases. We refer to some of these tutorials in specific sections below. We also want to bring attention to the availability of several primers on *Drosophila* research including Greenspan (2004) and Hales et al. (2015), publications in the FlyBook series from Genetics (<https://academic.oup.com/genetics/pages/flybook>), a searchable online *Drosophila* Protocols Portal (<https://www.flymai.org/tools/protocols/web/>), and a “New to Flies” section in the FlyBase Wiki (https://wiki.flybase.org/wiki/FlyBase:New_to_Flies).

Step 1: what is known about the gene in *Drosophila*?

Step 1, part 1: getting a quick overview of gene information

Summary information in FlyBase Gene Reports

The first thing a seasoned *Drosophila* researcher is likely to do when encountering a CG identifier or an unfamiliar gene name is to find a corresponding information page at FlyBase (<https://flybase.org/>). FlyBase is a comprehensive knowledgebase of *Drosophila* gene and genome annotations; gene, protein, and

phenotype information; *Drosophila* research community resources; and more (Thurmond et al. 2019; Larkin et al. 2021). At FlyBase, each annotated gene is associated with a webpage known as a Gene Report. A FlyBase Gene Report provides basic information about the gene, such as gene identifiers, chromosomal location, gene products, gene summaries, expression data, phenotype data, and associated publications and displays or links to more detailed information curated from the literature or based on large-scale public datasets. To quickly access a FlyBase Gene Report for a CG# or other *Drosophila* gene, find the search box at the top right-hand side of any FlyBase webpage; choose “Jump to Gene” (J2G) as the search type; enter the CG#, gene symbol, or other gene identifier; and click to perform a search. A Google search with “FlyBase” and the CG#, gene name, or gene symbol is also likely to retrieve the FlyBase Gene Report for the gene. Each FlyBase Gene Report displays “General Information” about the gene at the top and includes a list of “Report Sections” on the right with hyperlinks to specific sections within the Gene Report. Clicking on the plus signs within sections or subsections expands the page to reveal detailed information.

Help navigating FlyBase

FlyBase has a long track record of helping *Drosophila* researchers navigate Gene Reports and other features of the knowledgebase. Recent publications from the FlyBase team describing how to navigate FlyBase include Gramates et al. (2022); Jenkins et al. (2022); Marygold and FlyBase (2023). FlyBase has also provided specific guidance on navigating “gene groups” (curated gene families; Rey et al. 2018) and *Drosophila* models of human diseases (Millburn et al. 2016). Additional resources include demo videos at the FlyBase YouTube channel (https://wiki.flybase.org/wiki/FlyBase:FlyBase_Help_Index#Video_Tutorials) and the FlyBase “tweutorial” series, which can be accessed through the FlyBase Twitter account (https://wiki.flybase.org/wiki/FlyBase:FlyBase_Help_Index#Tweutorials). FlyBase has also established a presence on the open source social media platform Mastodon (<https://mstdn.science/@FlyBase>).

Additional databases that provide quick summary overviews

FlyBase is a perennial favorite of *Drosophila* researchers and is considered by other databases to be the authority on *Drosophila melanogaster* gene and genome annotations. Notably, in addition to including a largely text-based resource in the form of Gene Reports, FlyBase also provides a visual summary of gene features, expression data, annotated features such as transcription binding sites, and more, via the FlyBase instance of the JBrowse viewer (Buels et al. 2016), which can be accessed from Gene Report pages. Nevertheless, FlyBase is not the only site to consider as a “go-to” resource for an initial search for summary information about a *Drosophila* gene. Alternative meta-databases are listed in Table 1 and include the Alliance for Genome Resources (Alliance; Alliance of Genome Resources Consortium 2020), which offers a standardized format across model organisms and an updated web design; G2F (Hu, Comjean, Mohr et al. 2017), which provides summary information for a *Drosophila* gene and its orthologs in a single tabular view; FlyMine (Lyne et al. 2007), which describes itself as a “data warehouse” that can be used to mine integrated data; MARRVEL (Wang, Al-Ouran et al. 2017), in which model organism gene information is displayed in the context of information about human orthologs; Monarch (Shefchek et al. 2020), which emphasizes phenotype information; NCBI Gene, which is standardized across the large number of species supported,

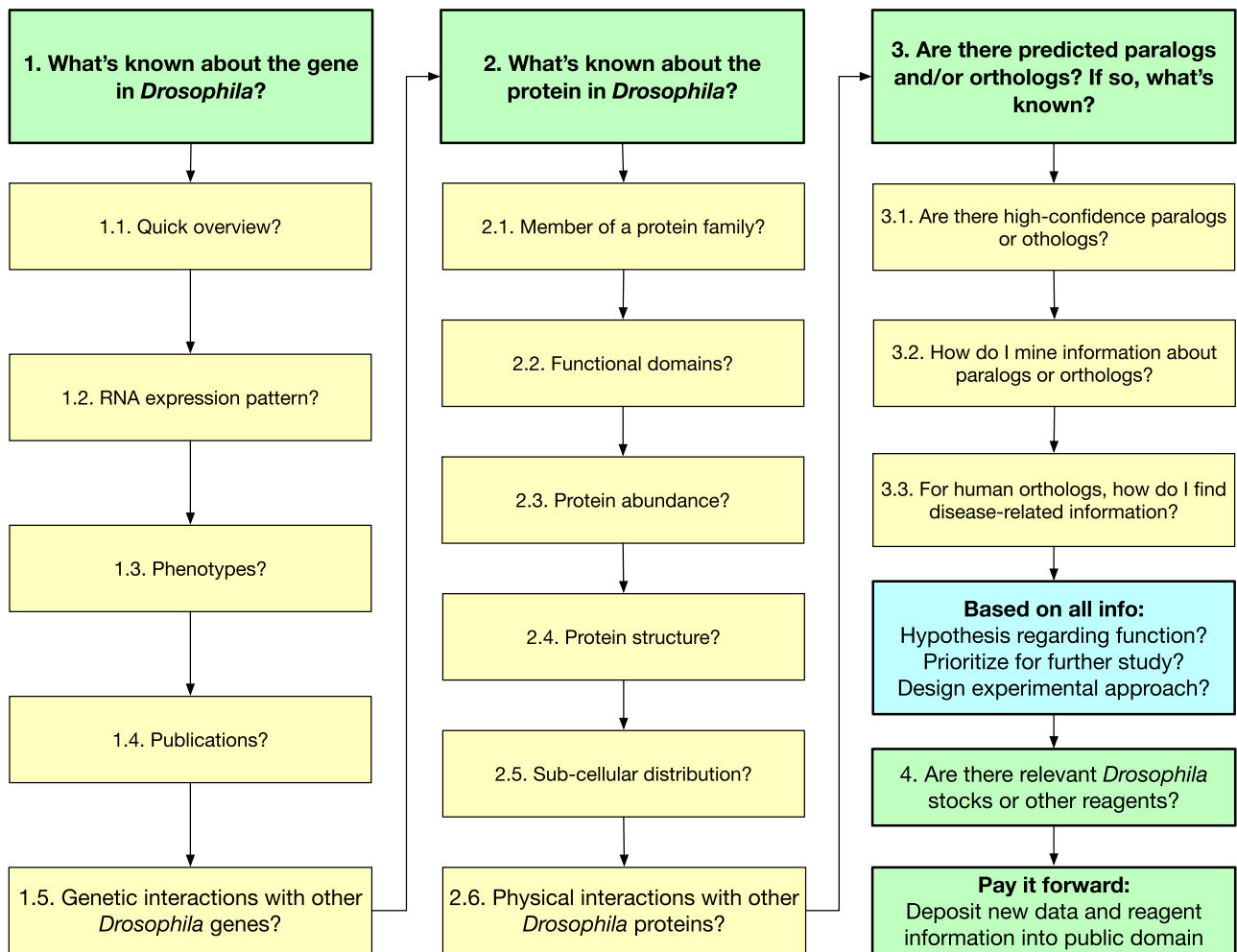


Fig. 2. Workflow for mining information about uncharacterized *Drosophila* genes. Top boxes, main workflow questions to be addressed through information mining. Subsequent numbered boxes, individual queries that help address the main question. "Based on ..." box, point of integration of information, when a decision is made to include or exclude a gene from further studies. Genes identified at steps 1.5 or 2.6 can enter the pipeline at step 1. As a final step, we suggest "paying it forward" by aiding curators and depositing information into relevant databases, e.g. using FlyBase Fast Track Your Paper and/or data submissions to NCBI. Figure made using OmniGraffle.

including model and nonmodel species; and UniProt (UniProt Consortium 2023), which emphasizes protein-related information. A simple web browser search with "*Drosophila*" and the CG# or other gene symbol can also provide a good starting point for finding relevant information and resources. Factors that might influence the choice of a "go-to" starting knowledgebase include the type of information being sought out; the importance of viewing ortholog and/or human disease-related information at an early stage; and the appearance of the user interface (UI), which differs significantly among these sites. These alternative sites import information from FlyBase but might not be in sync with the most recent FlyBase genome and gene annotation releases. An additional meta-database, GeneMANIA (<https://genemania.org>; Franz et al. 2018), provides a more limited amount of summary information. At Genemania, information is present in the form of a network drawn using Cytoscape (<https://cytoscape.org/>; Shannon et al. 2003), with different types of features indicated as edges of different colors.

Navigating additional resources

These resources additional to FlyBase have similarly made efforts to help users navigate the information they provide. Multimedia tutorials for the Alliance are compiled on a central webpage

(<https://www.alliancegenome.org/tutorials>); the parent organization of FlyMine, InterMine (Kalderimis et al. 2014), provides a user documentation page with links to a "getting started" page and video demos (<http://intermine.org/intermine-user-docs/>); G2F provides guidance and a demo video at the G2F About/Help page (<https://www.gene2function.org/search/help>); MARRVEL provides guidance and a demo video at the MARRVEL FAQs page (<https://marrvel.org/faq>), as well as in Wang, Liu et al. (2019) and Wang, Mao et al. (2019); the Monarch homepage "examples" tab includes gene pages for non-*Drosophila* model species that demonstrate the type of information a *Drosophila* gene search would provide (<https://monarchinitiative.org>); NCBI provides tutorials that include relevant "How to: Find ..." pages at (<https://www.ncbi.nlm.nih.gov/home/tutorials/>); and UniProt provides help with navigation in Lussi et al. (2023) and Zaru et al. (2023). At least 2 of these resources have a presence at Mastodon: the Alliance (<https://mstdn.science/@AllianceGenome@genomic.social>) and Monarch (https://mstdn.science/@monarch_initiative@genomic.social).

Gathering basic information about a gene

The top of a FlyBase Gene Report or equivalent page at another meta-database typically displays summary information about

Table 1. Online resources that provide comprehensive summary information about *Drosophila* genes.

Online resource	URL
Alliance (Gene Page)	https://www.alliancegenome.org/
FlyBase (Gene Report)	https://flybase.org/
FlyMine (Gene Page)	https://www.flymine.org
Gene2Function (Gene Search Results)	https://www.gene2function.org/search/
MARRVEL ("Model organism gene" search)	https://marrvel.org/
Monarch ("Explore ... genes")	https://monarchinitiative.org/
NCBI Gene (Gene Report)	https://www.ncbi.nlm.nih.gov/gene/
UniProt ("Find Your Protein" Results)	https://www.uniprot.org/

the gene model, including its genomic location and a list of transcripts (isoforms) encoded by the gene, a concise presentation of known or inferred functional information (e.g. the "gene summary" at FlyBase or the "automated description" at the Alliance), and gene ontology (GO) terms associated with the gene. In some cases, this information alone will be sufficient to decide if a gene should be included or excluded from follow-up studies. In many cases, however, additional relevant data will be needed.

Step 1, part 2: finding RNA expression patterns

Accessing RNA expression and localization data

Identifying when and where a *Drosophila* gene is expressed is a common early goal, as this information can provide insights into the potential relevance of the gene to the process under study and guide experimental design. Over time, *Drosophila* researchers have been able to conduct large-scale transcriptomics and in situ hybridization studies at increasing scale, precision, and resolution. Large-scale transcriptomics datasets include those from the modENCODE project (bulk RNAseq; [modEncode Consortium et al. 2010](#)), FlyAtlas 2 (bulk RNAseq; [Leader et al. 2018](#)), and Fly Cell Atlas (FCA; single-cell RNAseq; [Li et al. 2022](#)). Large-scale RNA localization datasets include the Berkeley *Drosophila* Genome Project's (BDGP) in situ library ([Tomancak et al. 2002](#); [Tomancak et al. 2007](#); [Hammonds et al. 2013](#)), the Fly-FISH project ([Lecuyer et al. 2007](#); [Wilk et al. 2016](#)), and the Dresden Ovary Table (DOT; [Jambor et al. 2015](#)). [Figure 3](#) summarizes questions that might motivate a search for RNA expression and localization data, displays corresponding data resources, and shows at what online resources researchers can navigate to one or more of the datasets. In [Table 2](#), we provide URLs to relevant meta-databases, knowledgebases, and focused databases that include RNA expression and localization data.

We also note the availability of expression data in the form of insertions or fusions of fluorescent proteins, or insertions of GAL4, under the control of endogenous promoters. For GAL4 insertions, expression can be visualized in combination with upstream activation sequence (UAS)-fluorescent protein constructs. Fluorescent-tagged resources include the FLYtRAB resource (http://rablibrary.mpi-cbg.de/cgi-bin/rab_overview.pl), which displays expression data for yellow fluorescent protein (YFP) reporters knocked into endogenous loci encoding 27 *Drosophila* members of the Rab GTPase protein family ([Dunst et al. 2015](#)). In addition, expression of fluorescent-tagged proteins is available for a number of MiMIC insertion strains generated as part of the *Drosophila* Gene Disruption Project (GDP; [Venken et al. 2011](#); [Nagarkar-Jaiswal, DeLuca et al. 2015](#); [Nagarkar-Jaiswal, Lee et al. 2015](#)). With regard to GAL4

insertions, the GDP has generated insertions of T2A-GAL4 or Kozak-GAL4 generated by conversion of MiMIC or by clustered regularly interspaced short palindromic repeats (CRISPR)-mediated integration ([Lee et al. 2018](#); [Kanca et al. 2019](#); [Kanca et al. 2022](#)). Image data for GDP insertion strains can be accessed online by searching for a gene and then clicking on the hyperlinked text at "line" to view available images; for MiMIC fly stocks, start at the URL <https://flypush.research.bcm.edu/pscreen/rmce/> and for CRIMIC fly stocks, start at the URL <https://flypush.research.bcm.edu/pscreen/crimic/crimic.php>. Expression data from the GDP are particularly relevant for researchers interested to view expression in the third larval instar brain.

As noted in [Fig. 3](#), FlyBase provides access to 3 major RNA expression datasets of 2 types, namely, the modENCODE and FlyAtlas bulk RNAseq datasets, and the FlyCellAtlas single-cell RNAseq dataset. FlyBase also provides hyperlinks from Gene Report pages to other datasets, e.g. the BDGP in situ database (e.g. see "External Data and Images" in the "Expression Data" section) and displays computationally integrated RNAseq data covering multiple tissues and stages (e.g. at FlyBase JBrowse, choose an aggregated dataset at "Oliver lab SRA Aggregated RNA-Seq" in the "RNAseq" subsection of the "Expression" tracks; [Hu Qian et al. 2023](#)). However, no single database contains all of the available RNA expression and localization data, so visiting more than one site might be necessary ([Fig. 3](#), [Table 2](#)). It is also notable that some resources offer very different visualizations of the same datasets, including single-gene heatmaps at several resources, multigene heatmaps at DGET ([Hu, Comjean, Perrimon et al. 2017](#)), topology-like "TopoView" visualizations along the genome in JBrowse at FlyBase ([Gramates et al. 2022](#)), coexpression networks at GeneMANIA ([Franz et al. 2018](#)), gene expression maps at FlyExpress ([Kumar et al. 2017](#)), and anatomy-based "Anatograms" at the EBI Single Cell Expression Atlas ([Thakur et al. 2023](#)).

Step 1, part 3: finding phenotype information

Accessing phenotype information

FlyBase Gene Reports include a "phenotypes" section that lists curated in vivo phenotypes associated with the gene. For relatively uncharacterized genes, these are likely to be limited to information about the viability and/or fertility of any available mutant alleles and results reported in large-scale RNAi screens. Most other meta-databases listed in [Table 1](#) rely on FlyBase as a source for phenotype data. However, some only display selected subsets of the data from FlyBase, and some additional types of data, such as cell-based screen data, are not available at all sites. A search with [CG33054](#) provides an instructive example of how phenotype information differs at different meta-databases. Partial results are shown in [Fig. 4](#), and a full analysis is presented below.

Example differences in phenotype data at different sites

At FlyBase, [CG33054](#) is associated with the phenotypes "partial lethal," "some die at pupal stage," "visible phenotype," and "fertile" ([Fig. 4a](#)). Moreover, FlyBase provides detailed information about an anatomical phenotype (i.e. an effect on chaetae) and displays the fly stock reagents associated with the phenotype annotations. At the Alliance, only results from the mutant allele (insertion strain) are shown in the "phenotype" section, and 2 links are provided ([Fig. 4b](#)). One of these connects to the Gene Report for [CG33054](#) at FlyBase, and the other connects to a record for CRISPR cell screen results displayed at BioGRID ([Oughtred et al. 2019](#)), where we can see that [CG33054](#) was not a "hit" (positive result) in a *Drosophila* cell-based screen for essential genes reported

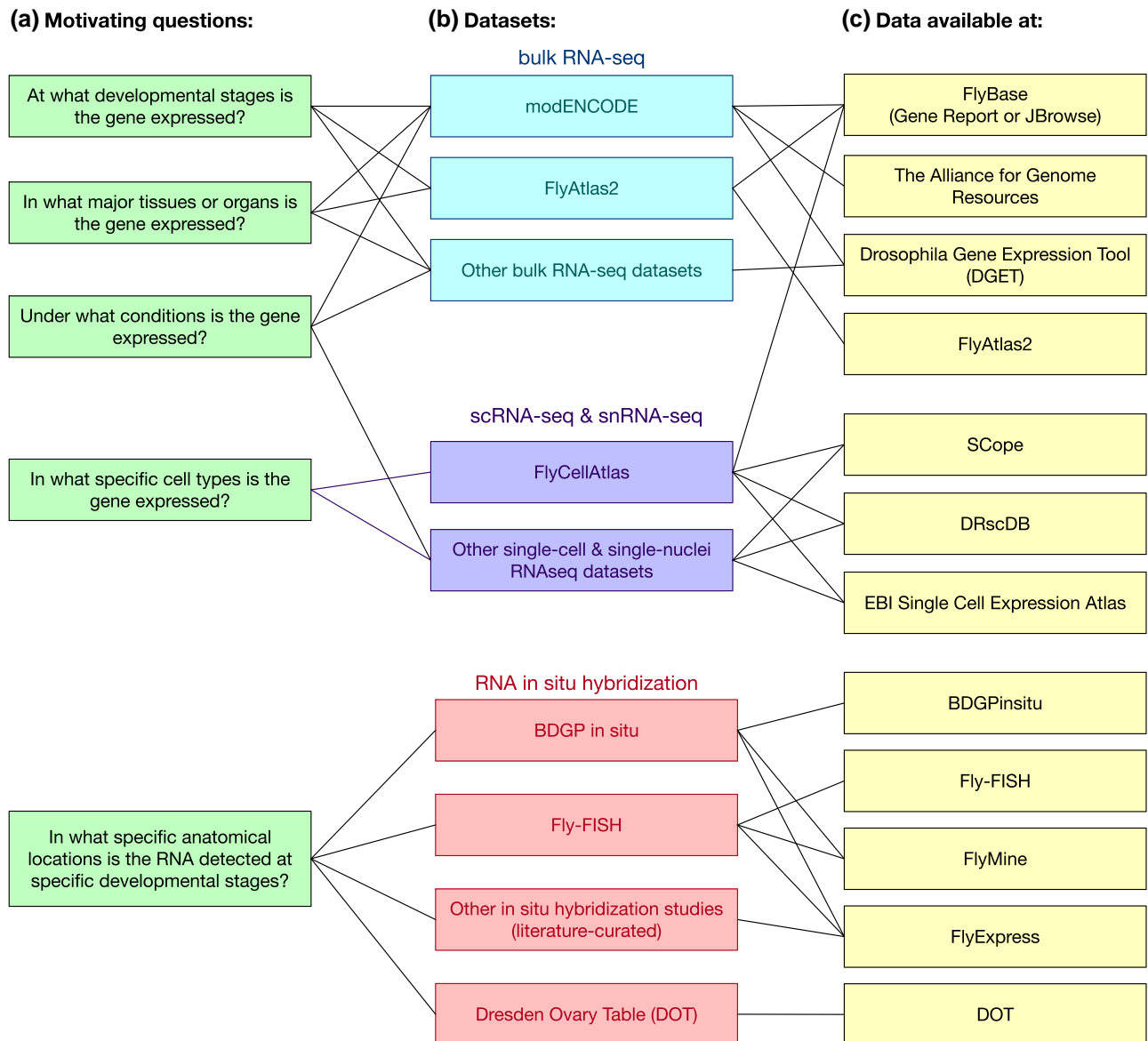


Fig. 3. Navigating information about RNA expression in *Drosophila*. a) Left-hand boxes, questions that can be addressed using RNA expression or localization data. b) Major public datasets. Center top boxes, bulk RNAseq datasets; center-middle boxes, single-cell or single-nucleus RNA-Seq datasets; center-bottom boxes, in situ hybridization datasets. Lines between boxes in a) and b) connect motivating questions with datasets that can help address the question. c) Right-hand boxes, databases at which the datasets shown in b) can be accessed. Lines between boxes in b) and c) connect datasets to meta-databases or specialized databases at which the datasets can be accessed. Note that representative images of in situ hybridization data are available at BDGP in situ, Fly-FISH, FlyExpress, and DOT, whereas FlyMine displays results in text form. Notably, additional expression information is available based on insertions of fluorescent proteins or GAL4 under endogenous control, e.g. as developed by the GDP. Figure made using OmniGraffle.

by (Viswanatha et al. 2018). At FlyMine, phenotype information is listed at a gene page in the “Disease” section, with RNAi phenotypes listed in a dedicated table. At G2F, in vivo phenotype information is imported via InterMine for *Drosophila* and other species. In addition, the G2F results table for CG33054 includes a column “RNAi Cell Data” that shows that CG33054 was a hit in 20 cell-based RNAi screens. Clicking on the hyperlinked number retrieves a list of the specific cell-based RNAi screens in which CG33054 was a hit. At MARRVEL, in the “phenotype” subsection of the “model organisms” section of the record for OARD1 (the human ortholog of CG33054), a summary visualization indicates that CG33054 is associated with phenotype(s) that affect growth, the nervous system, and the integument and 2 phenotypes in an “other” category (Fig. 4c). Clicking on the highlighted box at

“nervous system” displays the term “chaeta” (along with a link to a definition of the term at FlyBase), and clicking on other highlighted boxes reveals more information and links to FlyBase. At Monarch, both the mutant allele and RNAi results from FlyBase are displayed (Fig. 4d). NCBI Gene does not include phenotype data but does have external links to *Drosophila* cell RNAi screen data at NCBI PubChem BioAssay (Wang, Cheng et al. 2017; Kim et al. 2023). UniProt includes a “Phenotypes and Variants” subheader and displays phenotype information for well-characterized genes (e.g. *wg* and *InR*) but not for CG33054 (as of access in August 2023).

The meta-databases that include in vivo mutant phenotype data and/or RNAi phenotype data import the information from FlyBase. However, different databases import different subsets

Table 2. Online resources that provide access to *Drosophila* RNA expression and localization datasets.

Online resource	Dataset(s) included	URL
BDBP in situ database	BDGP in situ project data	https://insitu.fruitfly.org/cgi-bin/ex/insitu.pl
Drosophila Gene Expression Tool	multiple bulk RNAseq datasets	https://www.flyrnai.org/tools/dget/web/
Dresden Ovary Table (DOT)	DOT project in situ data	http://tomancak-srv1.mpi-cbg.de/DOT/main
DRscDB	multiple scRNAseq datasets	https://www.flyrnai.org/tools/single_cell/web/
EBI Single Cell Expression Atlas	multiple scRNAseq datasets	https://www.ebi.ac.uk/gxa/sc/home
FlyAtlas 2	FlyAtlas 2 bulk RNAseq data	https://motif.mvls.gla.ac.uk/FlyAtlas2
FlyBase	multiple bulk RNAseq and scRNAseq datasets	https://flybase.org/
FlyBase (JBrowse) ^a	multiple bulk RNAseq and scRNAseq datasets, and an integrated RNAseq resource	https://flybase.org/
FlyExpress database	BDGP, Fly-FISH, and other in situ datasets	http://www.flyexpress.net
Fly-FISH database	Fly-FISH in situ project data	https://fly-fish.ccb.utoronto.ca/
FlyMine	multiple bulk RNAseq and in situ datasets	https://www.flymine.org
SCoPE	FlyCellAtlas and other scRNAseq datasets	https://scope.aertslab.org/#/FlyCellAtlas/*/welcome

^a Expression data are available in topological-like TopoView format at the JBrowse genome browser, which can be accessed from a FlyBase Gene Report.

of the FlyBase data and present the data in different ways (Table 3, Fig. 4). In addition, G2F, Monarch, and MARRVEL display phenotype information for orthologs in humans and other model organism species alongside or nearby phenotype data for *Drosophila*. Moreover, although phenotype data curated by FlyBase are comprehensive with regard to the published literature, it does not include some large-scale in vivo and cell-based screen datasets that are available at other sites (see below and Table 3).

Specialized resources for in vivo RNAi and CRISPR phenotype data

Phenotype data from tissue-specific RNAi and CRISPR studies present a special case, as datasets from studies based on use of the GAL4-UAS system (Brand and Perrimon 1993) are best viewed in context, i.e. with both the UAS-controlled reagent and the GAL4 driver used in the study indicated. FlyBase displays RNAi phenotype data together with information about what GAL4 driver was used in the study (Fig. 4a, blue arrow). However, not all meta-databases that display FlyBase RNAi phenotype data make it obvious what GAL4 driver is associated with a given phenotype (Table 3). Although FlyBase curates small-scale, high-confidence RNAi phenotype results, e.g. as reported as part of the main text of a research report, FlyBase does not typically curate phenotype information from large-scale in vivo RNAi screen datasets, e.g. as reported as supplemental tables. However, data from large-scale in vivo RNAi screens, as well as some unpublished screens, have been curated and imported into the Transgenic RNAi Project (TRiP) RNAi Stock Validation and Phenotype (RSVP) Plus database (Perkins et al. 2015), which also includes curated RNAi phenotype data from FlyBase. At RSVP Plus, all available results for a given gene can be viewed by entering the CG number or other identifier and leaving “Any Driver” as the default in the “Drivers” field, or results can be limited to specific drivers. Notably, RSVP Plus includes UAS-RNAi fly stocks from all 3 major RNAi stock collections, the TRiP, Vienna *Drosophila* RNAi Center (VDRC), and NIG-Japan collections. In addition, RSVP Plus includes results for CRISPR studies in which the GAL4-UAS system is used to either control expression of Cas9 (including modified forms of Cas9, such as dead Cas9 fusions used for CRISPR activation) or control expression of sgRNA(s). Currently, there are much more phenotype data available for UAS-RNAi studies than for UAS-CRISPR system studies;

however, the number of UAS-CRISPR system phenotype datasets is likely to increase in the future.

Specialized resources for cell-based RNAi and CRISPR phenotype data

Cell-based datasets are a complementary resource that provides phenotype data relevant to cellular function. Cell-based RNAi and CRISPR datasets are available from some meta-databases and from specialized databases (Table 3). The *Drosophila* RNAi Screening Center (DRSC) makes results from genome-wide arrayed RNAi screens in *Drosophila* cells done using DRSC reagent libraries searchable at multiple online tools, including Gene Lookup (Hu et al. 2021) and Updated Targets of RNAi Reagents (UP-TORR; Hu et al. 2013). In addition, as noted, cell RNAi screen data are included in outputs at the DRSC’s G2F resource. Moreover, a significant subset of DRSC *Drosophila* cell RNAi screen datasets is also available at NCBI PubChem BioAssay. DRSC screen datasets, datasets from screens supported by the Sheffield RNAi Screening Center or the DKFZ (Boutros lab), and other genome-wide *Drosophila* cell RNAi screen datasets are searchable at GenomeRNAi (Schmidt et al. 2013). As mentioned previously, BioGRID has results from a genome-wide pooled CRISPR knockout screen in *Drosophila* cells (see Fig. 4b), and it is likely that more such data will be available in the near future.

Step 1, part 4: finding relevant publications

Although it is unlikely that a CG or another uncharacterized *Drosophila* gene is associated with an extensive published literature, it may still have some associated references. FlyBase associates fly genes with publications including “research reports,” “personal communications to FlyBase,” and other types of publications and displays these in the “References” section of the corresponding FlyBase Gene Reports. Some, but not all, meta-databases shown in Table 1 also provide links to publications. Table 4 summarizes publication results retrieved with CG33054 from these online resources as well as results retrieved at the literature mining tool BioLitMine (Hu et al. 2020), the BioRxiv preprint server, NCBI PubMed and related sites, and Google Scholar. The FlyBase Gene Report for CG33054 lists 13 research papers associated with the gene and a total of 28 publications. At G2F, the table of results includes a count of publications associated with the gene identifier that hyperlinks to retrieval of those publications at PubMed. Similarly, navigating to model organism results at MARRVEL links to a

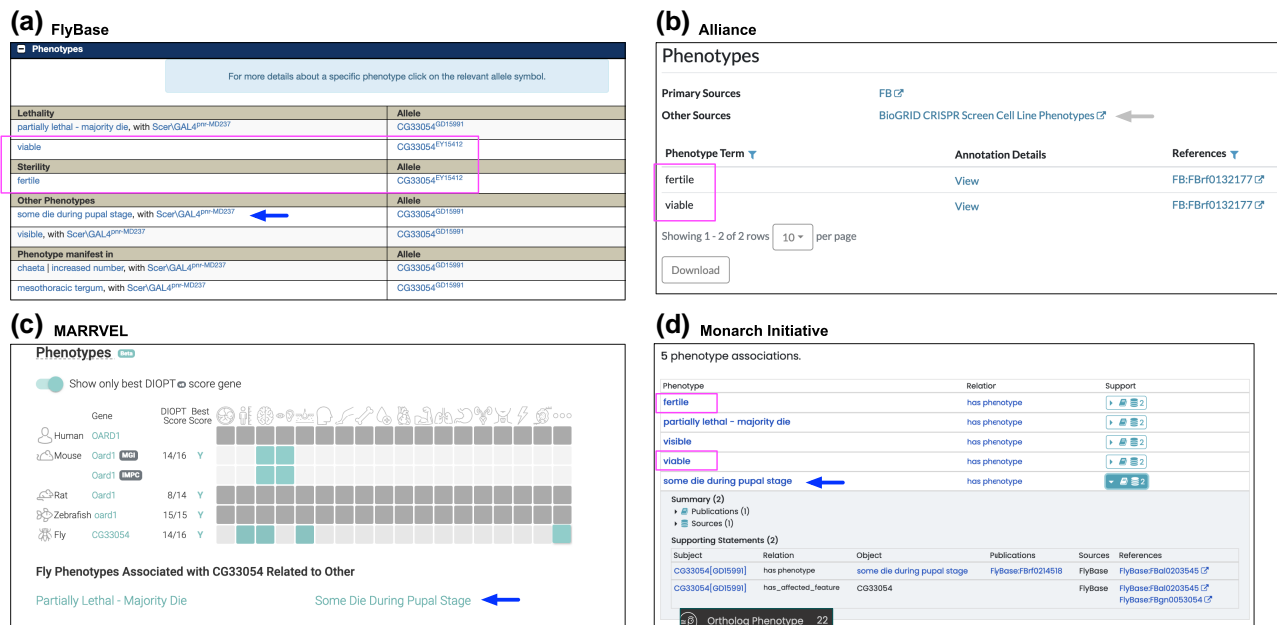


Fig. 4. Display of phenotype data for *CG33054* at 4 meta-databases. a) FlyBase. Phenotype data are displayed on the Gene Report page for *CG33054*. FlyBase is the source for curated phenotype information displayed at most other sites. Note that RNAi phenotypes include “some die during pupal stages.” This is also displayed at MARRVEL and Monarch (blue arrows in a, c, and d). b) The Alliance. Only phenotypes associated with the mutant allele are shown (magenta box); the RNAi results are not shown. In addition to displaying a subset of FlyBase data, the Alliance also provides a link to cell-based CRISPR screen data at BioGRID. c) MARRVEL. Phenotype information for *CG33054* and its orthologs is summarized. Clicking on a highlighted box in the grid retrieves more detailed phenotype information (blue arrow). d) Monarch. Both mutant allele (open boxes) and RNAi data (e.g. arrow) are included. Rows can be expanded to reveal source, as shown for the phenotype “some die ...”. Phenotype information for orthologs is also available (inset at bottom: 22 phenotypes are associated with orthologs of *CG33054*).

set of publications at PubMed. Monarch cites specific research articles (publication “is about” the gene) and other sources, such as FlyBase reports (“source” relationship to the gene). One reason the number of publications retrieved differs is that different resources apply different filters for exclusion of papers that list large numbers of genes (e.g. full-genome sequence reports).

Step 1, part 5: finding genetic interaction partners

Genetic interactions can help develop informed hypotheses regarding function based on a “guilt by association” model. Most meta-databases listed in Table 1 display genetic interactors on the gene report, gene page, or protein page: Alliance gene pages include a “Genetic Interactions” section; FlyBase gene reports include an “Summary of Genetic Interactions” in the “Interactions” section; FlyMine displays genetic interactions in a network visualization in an “Interactions” section, distinguishing physical and genetic interactions by edge color; G2F displays a “count” of genetic interactions that links to an information table; and both Monarch and UniProt include “interaction” sections that do not clearly distinguish physical from genetic interactions but link to source information. In addition, most of the interaction-specific databases discussed below in the context of physical interactions include genetic interactions. Once identified, interacting genes might be subjected to the same information mining workflow (Fig. 2).

Step 2: what is known about the *Drosophila* protein?

Step 2, part 1: is the protein a member of a protein family?

Some proteins, such as enzymes, can easily be recognized based on comparison of the primary amino acid sequence to the sequences

of members of well-characterized protein families. At least 3 related types of information can help ascertain if a protein is a member of a specific protein family: (1) GO annotations (in particular, GO “molecular function” annotations), (2) membership in an annotated group, such as a “Gene Group” annotated by FlyBase (Attrill et al. 2016; Rey et al. 2018) or a gene list annotated in the Gene List Annotation for *Drosophila* (GLAD) resource (Hu et al. 2015), and (3) inclusion in a list of known or predicted members protein family as annotated by resources such as Pfam or Panther (Table 5). For *CG33054*, GO annotations, Gene Group membership, and other annotations consistently identify the corresponding protein as a protein with “ADP-ribosylglutamate hydrolase” activity. Some resources combine display of protein family membership with display of protein domains, which are discussed below.

Step 2, part 2: does the protein contain conserved protein domains or other features?

Whether or not they are members of a specific protein family, many proteins contain primary sequences, which, based on amino acid similarity, have been annotated as conserved protein domains of unknown function (DUFs) or associated with specific biochemical functions (e.g. kinase domains), interaction (e.g. WD40 domains), subcellular localization, or other features. Availability of conserved protein domain information at meta-databases and other online resources is summarized in Table 6. As a supplement to identification of conserved protein domains, sequence analysis tools can be used to detect motifs such as secretion signals, membrane-spanning domains, protein cleavage sites, and nuclear localization signals. Commonly used resources include the SignalP online resource (Teufel et al. 2022), Phobius (Kall et al. 2007), DeepTMHMM (Jeppe et al. 2022), ProP (Duckert et al. 2004), and NucPred (Brameier et al. 2007; Table 7). In addition, iProteinDB (Hu et al. 2019) displays

Table 3. Online resources that provide access to *Drosophila* phenotype information.

Online resource [dataset(s)]	Phenotype information source	URL
<i>Drosophila</i> in vivo mutant phenotypes		
Alliance	FlyBase curation (mutant allele phenotypes)	https://www.alliancegenome.org/
FlyBase	FlyBase curation	https://flybase.org/
Monarch	FlyBase curation	https://monarchinitiative.org/
<i>Drosophila</i> in vivo RNAi phenotypes		
FlyBase	FlyBase curation (includes GAL4 driver info)	https://flybase.org/
FlyMine (at "Disease")	FlyBase curation	https://www.flymine.org
G2F	"Disruption phenotype" from UniProt	https://www.gene2function.org
MARRVEL	FlyBase curation	https://marrvel.org/
Monarch	FlyBase curation	https://monarchinitiative.org/
RSVP Plus	FlyBase curation (RNAi phenotypes), TRiP-curated screen datasets, user submissions (includes GAL4 driver)	https://www.flyrnai.org/cgi-bin/RSVP_search.pl
Cell-based RNAi screen phenotypes		
DRSC Gene Lookup	Genome-wide cell RNAi screen data (DRSC)	https://www.flyrnai.org/cgi-bin/DRSC_gene_lookup.pl
G2F	Cell RNAi screen data (DKFZ, DRSC, and other sources from GenomeRNAi)	https://www.gene2function.org
GenomeRNAi	Cell RNAi screen data (DRSC, DKFZ, and other sources)	http://www.genomernai.org/
NCBI PubChem	Cell RNAi screen data (DRSC)	https://pubchem.ncbi.nlm.nih.gov/
BioAssay ^a		
UP-TORR	Genome-wide cell RNAi data (DRSC)	https://www.flyrnai.org/up-torr
Cell-based CRISPR knockout screen phenotypes		
G2F	Cell CRISPR screen data (DKFZ, DRSC, and other sources from GenomeCRISPR)	https://www.gene2function.org
GenomeCRISPR	Cell CRISPR screen data	http://genomecrispr.dkfz.de/
BioGRID	CRISPR cell screen data	https://thebiogrid.org/
Additionally displays phenotypes associated with genes from other species		
G2F	OMIM and InterMine APIs	https://www.gene2function.org
MARRVEL	OMIM and model organism databases (MODs)	https://marrvel.org/
Monarch	OMIM and MODs	https://monarchinitiative.org/

^a A search for "DRSC" at the PubChem "Explore Chemistry" main page retrieves the DRSC as a data source and displays a list of 48 "BioAssays" (screens) associated with ~39,000 DRSC "Substances" (double-stranded RNA reagents).

Table 4. Example search results: publications associated with CG33054 at selected databases and search tools.

Online resource	Publication results retrieved
Gene or protein info databases	
Alliance	N/A
FlyMine	N/A
FlyBase	List of 28 publications, including 13 research papers
G2F	Link to 29 PubMed records
MARRVEL	Link to 29 PubMed records
Monarch	List of 30 publications, including 25 research papers
NCBI Gene	Link to 29 PubMed records
UniProt	N/A
Publication search sites	
BioLitMine (literature mining tool)	List of 27 PubMed records associated with 16 medical subject heading (MeSH) terms, with hyperlinks
BioRxiv (preprint server)	1 (a "now published" preprint)
NCBI PubMed	0 ("term not found")
NCBI PubMed Central	2 PMC records
Europe PMC	2 PMC records
Google Scholar	9 publications, including 5 research papers

Search results reported above are those obtained on a single day in June 2023 using "CG33054" as the search term.
N/A, not applicable (these resources do not include publication or reference sections on gene/protein pages).

experimental phosphorylation site data from *D. melanogaster* and other *Drosophila* species, as well as other related information that can be used to identify or predict phosphorylation sites on *D.*

melanogaster proteins. iProteinDB also includes information about posttranslational modifications such as acetylation, oxidation, and carbamidomethylation, although available data for these modifications are much more limited as compared with phosphorylation data.

Step 2, part 3: what is known about abundance of the protein in specific developmental stages or in specific tissues or organs?

Protein abundance data are available for a limited subset of *Drosophila* proteins. Protein detection information curated from the literature by FlyBase is included in the "Polypeptide expression" subsection of the "Expression" section of a gene report, e.g. based on immunolocalization studies, and includes links to published sources. With regard to protein abundance specifically, FlyBase displays data from a large-scale developmental proteome study (Casas-Vila et al. 2017) in the "High-Throughput Expression" subsection of the "Expression" section of a gene report. As of June 2023, we were unable to identify any other meta-databases that include protein abundance data.

Step 2, part 4: what is the predicted structure of the protein?

Viewing or analyzing 3D structures can provide insights into biochemical function, subcellular localization, and other features of a protein. Table 8 summarizes resources that display protein structures, including experimentally determined structures in the Protein Data Bank (PDB; Bittrich et al. 2023), and/or structures predicted using increasingly sophisticated approaches, including the proteome-wide predictions made using the AlphaFold approach (Jumper et al. 2021) or the ESM Fold approach (Lin et al. 2023). In

Table 5. Accessing protein function and family annotations for *Drosophila* proteins.

Online resource	Provides	URL
Search by gene symbol or CG number		
Alliance	GO annotations (“Function—GO annotations” section of a gene page)	https://www.alliancegenome.org
FlyBase	Gene Group membership and GO annotations (“Function” section of a gene report)	https://flybase.org/
FlyMine	GO annotations (“Gene Ontology” section of a gene page)	https://www.flymine.org
G2F	GO annotations (“GO Function Count” links to a table of terms and other information)	https://www.gene2function.org
GLAD	Gene list annotations (“find group membership” search)	https://www.flyrnai.org/tools/glad/web/
MARRVEL	GO annotations (for the fly protein and orthologs; “Gene Ontology” subsection of “Model Organisms” section of the corresponding human ortholog page)	https://marrvel.org
Monarch	GO annotations (“function” section of a gene page, “relation” is “enables”)	https://monarchinitiative.org
NCBI Gene	GO annotations (“General gene information” section)	
UniProt	GO annotations (“Function” section of a protein page) and results from multiple protein family annotations (“Family & Domains” section of a protein page)	https://www.uniprot.org/
Search by amino acid sequence		
InterPro	Protein family annotations from Pfam (search with amino acid sequence)	https://www.ebi.ac.uk/interpro/

Table 6. Accessing protein domain annotations for *Drosophila* proteins.

Online resource	Content and navigation tips	URL
Search by gene symbol or CG number		
FlyBase	Domain annotations from Pfam and SMART, accessible in the “Gene Model & Transcripts” section of a gene report	https://flybase.org/
FlyMine	Domain annotations from multiple resources, accessible from the “Gene → Protein + Domains” subsection of the “Protein” section of a gene page	https://www.flymine.org
DIOPT	DIOPT search results include a link to an alignment page that includes protein domain annotations	https://www.flyrnai.org/diopt
G2F	Domain annotations, accessible from the links in the “Protein Alignment” column of a search results table	https://www.gene2function.org
GeneMANIA	Network view of proteins with shared protein domains from Pfam and Interpro	https://genemania.org
NCBI Gene	“Relevant information” links include a link to relevant “conserved domains” at NCBI CDD	https://www.ncbi.nlm.nih.gov/gene
UniProt	Results from multiple protein domain annotations (“Family & Domains” section of a protein page), also includes annotation of predicted signal sequences, propeptides, and posttranslational modifications (“PTM/Processing” section of a protein page)	https://www.uniprot.org/
Search by amino acid sequence		
HMMER	Hidden Markov Model-based search for domains, with a protein sequence as the input	https://www.ebi.ac.uk/Tools/hmmer/
NCBI CDD	Domain annotations in the NCBI Conserved Domains Database (CDD), with an option to view results in concise, standard, or full mode	https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
ProScan	Domain annotations (motifs) from PROSITE	https://prosite.expasy.org/

addition, partial structures for large proteins or other new predictions can be done using the Google Colab notebook ColabFold (Mirdita et al. 2022) or RoseTTAFold (Baek et al. 2021). This is a rapidly developing area. One place to go to find updated Google notebooks supporting protein structure prediction algorithms is this GitHub page: <https://github.com/sokrypton/ColabFold>.

Step 2, part 5: what is the known or predicted subcellular localization of the protein?

Similar to protein abundance data, subcellular localization data are also available for only a limited subset of *Drosophila* proteins. In some cases, however, subcellular localization can be predicted based on features of the protein, e.g. signal sequences or membrane-spanning domains, and/or based on the subcellular localization of related proteins in other species. An indirect source of known or predicted information regarding subcellular localization is GO “cellular component” (GO-CC) annotations, which are available at most of the major meta-databases that include *Drosophila* information (Table 1). In addition, some meta-databases include additional information relevant to

known or predicted subcellular information, such as relevant GO terms, and the DeepLoc online resource can be used to predict subcellular localization (Thumulari et al. 2022; Table 9).

Step 2, part 6: does the uncharacterized protein interact with other proteins?

Identifying physical interactions between an uncharacterized protein and other proteins can help provide insights into function. For example, if available evidence suggests that an uncharacterized protein interacts with proteins in a given pathway or complex, you might hypothesize that the protein is a component or regulator of that pathway or complex. Physical interactions have been experimentally detected in low- or high-throughput modes using methods such as co-immunoprecipitation, mass spectrometry (MS), and yeast 2-hybrid screening (Y2H). Large-scale, high-throughput datasets include the Curagen Y2H dataset (Giot et al. 2003), the FlyBi Y2H dataset (Tang et al. 2023), additional Y2H datasets included in DroID (Yu et al. 2008; Murali et al. 2011), and the *Drosophila* Protein interaction Map (DPiM) MS dataset

(Guruharsha et al. 2011). Online resources that provide access to protein–protein interaction (PPI) and protein complex information are listed in Table 10. They include meta-databases already discussed as well as specialized resources, such as COMPLEAT (Vinayagam et al. 2013), the EBI Complex Portal (Meldal et al. 2015, 2022), the Database of Interacting Proteins (DIP; Xenarios et al. 2000), the Drosophila Interactions Database (DroID; Yu et al. 2008; Murali et al. 2011), the EBI IntAct database (Kerrien et al. 2012), the Molecular Interaction Search Tool (MIST; Hu et al. 2018), and STRING (Szklarczyk et al. 2023). MIST includes “interologs” or predicted PPIs based on interactions among related proteins in other species. In addition, several of the resources include an option to view genetic interactions as well as PPIs.

Step 3: are there paralogs or orthologs of the *Drosophila* gene and if so, what is known about them?

Step 3, part 1: are there high-confidence paralogs or orthologs of the gene?

For CG genes or other relatively uncharacterized genes for which there are paralogs (closely related genes within *Drosophila*) and/or orthologs (closely related genes in another species),

Table 7. Example resources for prediction of specific protein features.

Resource	Useful to predict	URL
DeepTMHMM	Membrane topology	https://dtu.biolib.com/DeepTMHMM
iProteinDB	Phosphorylation sites, other PTMs	https://www.flyrnai.org/tools/iproteindb/web/
NucPred	Nuclear localization signal sequences	https://nucpred.bioinfo.se/cgi-bin/single.cgi
Phobius	Membrane topology and signal peptides	https://phobius.sbc.su.se/
ProP	Furin cleavage sites	https://services.healthtech.dtu.dk/services/ProP-1.0/
SignalP	Signal peptides	https://services.healthtech.dtu.dk/services/SignalP-6.0/

PTMs, posttranslational modifications.

Table 8. Accessing experimentally derived protein structures and structure predictions.

Source	Navigation path	URL
Experimentally derived structures from PDB		
FlyMine	Search>Proteins>Protein Visualizer	https://www.flymine.org
G2F	Search by Gene>click result in “3D Structure” column	https://www.gene2function.org
RCSB PDB	Search>Use “refinements” to filter by species	https://www.rcsb.org/
Structure predictions from AlphaFold		
EBI AlphaFold Structure Prediction Database	Search>Use “filter” to filter by species	https://alphafold.ebi.ac.uk/
FlyBase	Gene report>Gene Model and Products>Structure	https://flybase.org/
SWISS-MODEL	Click fly icon at home>Search>Click ID in UniProtKB column>Click AlphaFold ID at “Available Structures”	https://swissmodel.expasy.org/
UniProt	Protein page>Structure	https://www.uniprot.org
Other structure predictions		
ColabFold	Generate a new prediction from a protein sequence	https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
RoseTTAFold at Robetta ^a	Click Submit from Structure Prediction	https://rosetta.bakerlab.org/submit.php
SWISS-MODEL	Click fly icon at home>Search with gene name	https://swissmodel.expasy.org/
ESM Fold Sequence	Generate a new prediction from a protein sequence	https://esmatlas.com/resources?action=fold

^a Requires users to register.

information about the function of those paralogs/orthologs can be used to develop hypotheses about the function of the uncharacterized gene. Conservation of a *Drosophila* gene in the human genome is often used as a filter for prioritization, with genes conserved in humans given higher priority than those that are not, e.g. when the overall goal of a study is to inform understanding about some aspect of human health or disease. Many algorithms and approaches have been developed to identify putative paralogs and orthologs, a goal that is similar to but distinct from identification of protein family membership and protein domains (Table 6). For information mining, it can be appropriate to “cast a wide net” and find all potential paralogs or orthologs. By contrast, with other goals in mind, such as modeling of human gene variants, it can be more appropriate to focus on the highest confidence predictions.

Overview of paralog and ortholog resources

The DRSC Integrative Ortholog Prediction Tool (DIOPT; Hu et al. 2011) integrates ortholog predictions from several algorithms and curated databases and provides a comprehensive view of predicted ortholog relationships between *Drosophila* genes and genes in humans or common model organisms (Table 11). DIOPT results include a “voting system” score based on the number of algorithms that predict a given paralog or ortholog relationship, and this score serves as a useful proxy for confidence. Paralog and ortholog relationships integrated by DIOPT can be searched and viewed at the DIOPT portal. DIOPT is also the source of ortholog and/or paralog information at many other resources, including at the Alliance (subset of DIOPT ortholog results), FlyBase (DIOPT-based orthologs and paralogs), G2F (DIOPT-based ortholog and paralog pairs), GuideXpress (DIOPT-based mosquito-*Drosophila* ortholog mapping; Viswanatha et al. 2021), Paralog Explorer (DIOPT-based paralogs; Hu et al. 2022), and MARRVEL (DIOPT-based mapping to human genes; Table 11). Advantages of using the DIOPT online portal directly are that you can easily search more than one gene at a time and download results as a tab-delimited table and that updated versions of the resource become immediately available, whereas other resources reliant on DIOPT may have a delay before the updated version is integrated. Advantages of other meta-databases that display DIOPT-based orthologs include concurrent display of other information (e.g. information included at FlyBase gene

Table 9. Accessing information about known or predicted subcellular localization of *Drosophila* proteins.

Online resource	Provides	URL
Alliance	"Function-GO Annotations" includes GO Cellular Component (GO-CC) annotations in ribbon diagram form	https://www.alliancegenome.org/
FlyBase	"Cellular Component" subsection in the "Gene Ontology" subsection of the "Function" section; "immunolocalization" data ^a in the "Polypeptide Expression" subsection of the "Expression" section of a gene report	https://flybase.org/
DeepLoc	Prediction of subcellular localization for a given protein sequence	https://services.healthtech.dtu.dk/services/DeepLoc-2.0/
G2F	Count of "Cellular Component" GO terms included in gene search results links to associated GO-CC terms	https://www.gene2function.org/search/
MARRVEL	"Gene Ontology" subsection of "Model Organisms" information on a human gene page includes GO-CC terms	https://marrvel.org/
NCBI Gene	"Gene Ontology" subsection of "General Gene Information" section on a gene page includes GO-CC terms	https://www.ncbi.nlm.nih.gov/gene/
UniProt	"GO Annotations" subsection of a protein page includes GO-CC in ribbon and text displays and links to full annotations at QuickGO (Binns et al. 2009); when available, a cell diagram is highlighted to display subcellular localization information	https://www.uniprot.org/

^a Immunolocalization data annotations are primarily focused on the anatomical distribution of proteins but might include information about subcellular or extracellular localization.

Table 10. Accessing information about high-confidence, low-confidence, and/or predicted *Drosophila* PPIs.

Online resource	Content and navigation tips	URL
Alliance	Table of interactions in the "Molecular Interactions" section of a gene page	https://www.alliancegenome.org/
BioGRID	Table of interactions from a "Search BioGRID" search with a CG number or other identifier and " <i>Drosophila melanogaster</i> " indicated as the species	https://thebiogrid.org/
COMPLEAT	At the "Browse/Download" menu tab, a gene name and species search will retrieve complex membership	https://www.flyrnai.org/compleat/
Complex Portal	List of complexes that include the protein, following a search with the CG number or other identifier	https://www.ebi.ac.uk/complexportal/home
DIP	List of interactions from a "Node" search using a CG number or other identifier in a "Node Identifier"	http://dip.doe-mbi.ucla.edu
DroID	Table of interactions from a DroID Search followed by clicking "Display Interactions"	http://droidb.org
FlyBase	Network display and table of PPIs in the "Interactions" section of a gene report, including type of assay and curated literature source for each PPI	https://flybase.org/
FlyMine	Table of interactions and network display in the "Interactions" section of a gene page	https://www.flymine.org
G2F	Count of interactions that links to a table of interactions, including database source for each PPI	https://www.gene2function.org/search/
GeneMANIA	Network diagram of interactions with links to sources (choose <i>Drosophila</i> on the drop-down list left of search)	https://genemania.org
IntAct	Network diagram and table of interactions from the results of a "Quick Search"	https://www.ebi.ac.uk/intact/home
MIST	Network view and table of interactors from a search with CG number or other identifier; search options include "fly" (comprehensive search) and "DroRI" (high-confidence PPI search)	https://fgrtools.hms.harvard.edu/MIST/
Monarch	Table of interactions, expandable to view of details such as type of evidence and source	https://monarchinitiative.org/
NCBI Gene	Table of interactions in "Interactions" section of a gene page, including database source and type of evidence ("description") for each PPI	https://www.ncbi.nlm.nih.gov/gene/
STRING	Network view and table of interactions from a search with a CG number or other identifier as "Protein name" and " <i>Drosophila melanogaster</i> " as species	https://string-db.org
UniProt	"Interactions" subsection of a protein page includes "Subunit" information if the protein is a subunit of a complex; table of binary interactions from IntAct; links to meta-databases of PPI information	https://www.uniprot.org/

reports or G2F results tables) and a focus on human orthologs (e.g. at MARRVEL). Orthologs are displayed in the "Homolog" section of Monarch gene pages and are also indicated in the "Ortholog phenotype" and "Ortholog Disease" sections of these pages; the primary source of ortholog relationships for *Drosophila* genes at Monarch appears to be a single source, the Protein Analysis THrough Evolutionary Relationships (PANTHER) resource (Thomas et al. 2022). UniProt groups both paralog and ortholog predictions in a "Similar Proteins" section that allows users to view and filter results, such as by percent amino acid identity.

Structure-based ortholog predictions

Some proteins have low or modest similarity at the primary amino acid level but have more striking similarity at the level of 3D structure, which is suggestive of shared function. Through advances in artificial intelligence (AI) and its application to protein structure predictions, it is now possible to compare one predicted protein structure with others using Foldseek (van Kempen et al. 2023). At Foldseek, a UniProt ID corresponding to an AlphaFold structure prediction for an uncharacterized *Drosophila* protein can be used as a query to look for similar proteins in other model species and/or humans.

Table 11. Accessing *Drosophila* paralog and ortholog predictions.

Resource	Navigation path	URL
Predicted paralogs		
Paralog Explorer	Paralog Explorer>Select Species ID: Fly	https://www.flyrnai.org/tools/paralogs/web/
DIOPT	DIOPT>Search Type: Paralogs	https://www.flyrnai.org/diopt
FlyBase	FlyBase>Gene Page>Paralogs (from DIOPT)	https://flybase.org/
Predicted orthologs in humans and common model organisms		
Alliance	Alliance>Gene Page>Orthology (subset of DIOPT)	https://www.alliancegenome.org/
DIOPT	DIOPT>Search Type: Orthologs (option to view alignment)	https://www.flyrnai.org/diopt
FlyBase	FlyBase>Gene Page>Orthologs (from DIOPT)	https://flybase.org/
UniProt	UniProt>Protein Page>Similar Proteins	https://www.uniprot.org
Predicted orthologs in other species		
FlyBase	FlyBase>Gene Page>Orthologs>Other Organism Orthologs (from OrthoDB; many nonmodel species)	https://flybase.org/
FlyMine	FlyMine>Gene Page>Homology	https://www.flymine.org
OrthoDB	OrthoDB>“Get Gene” search option	https://www.orthodb.org/
DIOPT	DIOPT>Search type: Orthologs and select species “ <i>Drosophila</i> ” and “ <i>Anopheles</i> ” (<i>Anopheles</i>)	https://www.flyrnai.org/diopt
GuideXpress	GuideXpress>Search type:Ortholog search, fly to mosquito (<i>Anopheles</i> , <i>Aedes</i> , <i>Culex</i>)	https://www.flyrnai.org/tools/fly2mosquito/web/
UniProt	UniProt>Protein Page>“Similar Proteins,” with an option to filter based on percent identity	https://www.uniprot.org
VectorBase	VectorBase>Gene Page>“Orthology and Synteny,” with an option to view an alignment (disease vector mosquitos, tsetse fly, other invertebrate vectors)	https://vectorbase.org/vectorbase/app/
Structure-based searches		
Foldseek	A PDB or AlphaFold accession ID can be used as an input structure to identify similar predicted structures	https://search.foldseek.com/search

Table 12. Accessing information about human disease associations for human orthologs of *Drosophila* genes.

Resource	Navigation path	URL
Alliance	Home>Gene page>“Disease Associations”	https://www.alliancegenome.org
FlyBase	Home>Gene report>“Human Disease Associations”	https://flybase.org/
G2F	Home>Search by gene (Fly)>“Human disease count”	https://www.gene2function.org/search/
MARRVEL	Home>Search (Model organism gene)>“MARRVEL it”	https://marrvel.org/
ModelMatcher	Home>Search by gene (fruit fly)	https://www.modelmatcher.net/
Monarch	Home>Search>Gene page>“Ortholog Disease”	https://monarchinitiative.org

Viewing protein alignments

Because many paralog and ortholog pairs are based on computational predictions, it is important to examine these relationships closely. In some cases, the similarity between the 2 proteins is limited to only a specific protein domain; in other cases, the similarity extends across the full length of the protein and/or includes a similar organization of multiple annotated protein domains. At the DIOPT portal, search results include links to pages that display an alignment of the 2 amino acid sequences and annotated domains present in each protein in the pair, allowing for a detailed look at amino acid identity and overall similarity. Some other resources link to alignments at DIOPT or provide their own alignments of proteins in a pair. NCBI BLAST can also be used to align and compare 2 amino acid sequences (e.g. the “align two or more sequences” option in the “blastp” suite).

Identifying orthologs in nonmodel species

For some studies, prioritization and/or planning of further experiments can be dependent or at least partially reliant on the identification of orthologs in nonmodel species, e.g. in other flies in the *Drosophila* genus or in mosquitos. DIOPT, and by extension some resources based on DIOPT, include a limited but growing number of nonmodel species, including mosquitos. In addition, FlyBase gene reports not only display orthologs for species covered by DIOPT but also display ortholog predictions from OrthoDB (Kuznetsov et al. 2023) for other species in the *Drosophila* genus and many other species (Table 11). Another place to view nonmodel species ortholog relationships is VectorBase (Giraldo-Calderon et al. 2022), which provides access to ortholog mapping from orthoMCL (Li et al. 2003) between *D. melanogaster* and several disease vector insect species (at section 7, “Orthology and Synteny,” on gene pages for *D. melanogaster* genes, as can be accessed in a general search at VectorBase with a CG number or other identifier; Table 11).

Step 3, part 2: how do I mine information about paralogs or orthologs?

When a paralog exists within the *Drosophila* genome, information about the gene and its corresponding protein can be mined following workflow Steps 1 and 2 (Fig. 2). Moreover, when one or more orthologs exist in another model organism species or the human genome, a similar path to identifying available information can again be followed. At the Alliance, gene pages for other species are organized as they are for *Drosophila*. Likewise, at VectorBase (Giraldo-Calderon et al. 2022), the results of an ortholog search are linked to similar gene pages for orthologous genes in disease vector species. At G2F, MARRVEL, and Monarch, summary information and external links to information about model species and human orthologs are displayed alongside ortholog search results and/or *Drosophila* gene pages. In addition, DIOPT results tables provide links to results at G2F. Thus, at G2F, MARRVEL, and Monarch, users can quickly view gene names, GO annotations,

phenotype data, and other information for orthologs of a *Drosophila* gene in other common genetic model organisms.

Step 3, part 3: if the gene has human orthologs, how do I find disease-related information?

When a human ortholog or orthologs are identified, a likely next question is to ask if the human ortholog or orthologs are associated with human diseases. A subset of meta-databases and

Table 13. Questions that motivate common experiments and corresponding fly stock reagents.

Question	Relevant <i>in vivo</i> reagents
What is the whole-animal loss-of-function phenotype?	<ul style="list-style-type: none"> • Loss-of-function mutant allele (e.g. null allele) • UAS-RNAi + ubiquitous-GAL4 • UAS-CRISPR system + ubiquitous-GAL4
What is the stage- and/or tissue-specific loss-of-function phenotype?	<ul style="list-style-type: none"> • UAS-RNAi + stage- and/or tissue-specific GAL4 (\pm GAL80ts) • UAS-CRISPR system + stage- and/or tissue-specific GAL4 (\pm GAL80ts)
In what stages, tissues, or cell types is the gene expressed?	<ul style="list-style-type: none"> • “Gold” type MiMIC insertion with T2AGAL4 + UAS-fluorescent reporter • CRISPR knock-in T2AGAL4 allele + UAS-fluorescent reporter
In what stages, tissues, or cell types is the protein expressed?	<ul style="list-style-type: none"> • CRISPR knock-in NanoTag allele + NanoTag Chromobody • CRISPR knock-in NanoTag or other epitope tag allele + Antibody • Protein-specific antibody
Is function cell autonomous or noncell autonomous?	<ul style="list-style-type: none"> • Mutant allele on an FRT chromosome + heat shock-FLP

some specialized resources connect fly genes to disease-relevant information (Table 12). FlyBase associates genes with diseases in 2 ways. First, gene reports include “human disease association” sections that associate mutant alleles with standardized terms known as Disease Ontology (DO) terms (Schriml *et al.* 2019). Second, FlyBase has disease-centric pages that include the lists of relevant *Drosophila* genes and links to human disease-centric databases such as Online Inheritance in Man (OMIM; Amberger *et al.* 2015). G2F summarizes the number of disease terms associated with human orthologs, which links to the terms themselves; links to information about drugs or other compounds known to target the human protein, as annotated by DrugBank (Wishart *et al.* 2018); and links to the corresponding page at MARRVEL. The MARRVEL resource itself is organized around human orthologs and provides access to information about disease associations and variants based on multiple sources, namely, OMIM (Amberger *et al.* 2019), ClinVar (Landrum *et al.* 2020), Geno2MP (<https://geno2mp.gs.washington.edu>), and DECIPHER (Foreman *et al.* 2023). Monarch is organized around phenotypes and likewise provides information relevant to human diseases. *Drosophila* researchers and others can also search and/or register at ModelMatcher (Harnish *et al.* 2022) with the goal of making new connections to persons interested in the human ortholog of a *Drosophila* gene, e.g. clinicians investigating a potential disease-associated variant of the human gene.

Step 4: are there relevant fly stocks or other physical reagents available?

Once a relatively uncharacterized gene has been prioritized for further experimental study, a next step will be to identify existing relevant fly stocks and other physical reagents. Table 13 summarizes questions that might motivate a search for *in vivo* resources and the types of fly stock reagents that can help address

Table 14. Accessing fly stocks and other physical reagents.

Resource	Navigation path	URL
Fly stocks—loss- or reduction-of-function (not necessarily validated)		
BDSC	Home>Browse Stocks>RNAi	https://bdsc.indiana.edu/index.html
DRSC/TRiP gRNA database	Home>Search by “FBgn, gene symbol, or CG annotation” (look for “TRiP-KO” stocks)	https://www.flymai.org/tools/gna_tracker/web/
FlyBase	Home>Gene Report>Alleles, Insertions, Transgenic Constructs, and Aberrations>Classical and Insertion Alleles	https://flybase.org/
FlyBase	Home>Gene Report>Alleles, Insertions, Transgenic Constructs, and Aberrations>Transgenic Constructs	https://flybase.org/
Gene Disruption Project (GDP)	Home>Search with a gene name	https://flypush.research.bcm.edu/pscreen/
RSVP Plus	Home>Search by gene symbol, all drivers>Detail page (click any ID in the “Detail page” column in the search results table)	https://www.flymai.org/cgi-bin/RSVP_search.pl
UP-TORR	Home>at Input Options-Gene, check “TRiP” “NIG” and “VDRC”, enter gene, search	https://www.flymai.org/up-torr/
Fly stocks—overexpression or upregulation (not necessarily validated)		
BDSC	BDSC>Browse Stocks>UAS>UAS lines (non-RNAi)>All	https://bdsc.indiana.edu/index.html
DRSC/TRiP gRNA database	Home>Search by “FBgn, gene symbol, or CG annotation” (look for “TRiP-OE” stocks)	https://www.flymai.org/tools/gna_tracker/web/
FlyBase	Home>Gene Report>Alleles, Insertions, Transgenic Constructs, and Aberrations>Transgenic Constructs (look for “TOE” or “UAS-gene”)	https://flybase.org/
ORF clones, cDNAs, and plasmid vectors		
Addgene	Home>Search (e.g. with “ <i>Drosophila</i> ” as search term)	https://www.addgene.org/
DGRC	Home>Search all>click “Dros genes” (for ORFs) or browse at “clones” or “vectors” lists	https://dgrc.bio.indiana.edu/Home
DNASU	Home>Search (e.g. with “ <i>Drosophila</i> ” as search term, or click on “collections” or “browse by species”)	https://dnasu.org/DNASU/Home.do

Table 15. Accessing precomputed reagent designs for CRISPR and qPCR.

Resource	URL
sgRNA designs ^a	
DRSC Find CRISPR (version 3)	https://www.flyrnai.org/crispr3/web/
CRISPRscan	https://www.crisprscan.org/
flyCRISPR Optimal Target Finder	http://targetfinder.flycrispr.neuro.brown.edu/
FlyBase Wiki list of gRNA design resources	https://wiki.flybase.org/wiki/FlyBase:CRISPR#CRISPR_gRNA_Resources
qPCR primer designs	
FlyPrimerBank	https://www.flyrnai.org/FlyPrimerBank
GTPrimer	https://gecftools.epfl.ch/getprime

^a Resources noted at the FlyBase Wiki as based on FlyBase genome annotation release 6 are listed.

the questions. This includes access to UAS-RNAi fly stocks; UAS-CRISPR system fly stocks, e.g. single guide RNA (sgRNA) fly stocks for CRISPR knockout or CRISPR activation (CRISPRa); and other relevant fly stocks for loss- or gain-of-function studies. Open reading frame (ORF) clones, cDNAs, empty vectors, and other plasmids can be identified in searches at FlyBase, NCBI, and the Berkeley *Drosophila* Genome Project (BDGP; <https://www.fruitfly.org/>) and can be requested directly from repositories including Addgene, the *Drosophila* Genomics Resource Center (DGRC), and DNASU (Table 14). The versatile GDP collection of fly stocks can be queried at the GDP Search Page (<https://flypush.research.bcm.edu/pscreen/>); this page also links to an interface that can be used to search the subset of GDP fly stocks that are MiMIC insertions useful for recombination-mediated cassette exchange (RMCE; Venken et al. 2011; Nagarkar-Jaiswal, DeLuca et al. 2015; Nagarkar-Jaiswal, Lee et al. 2015). At the “browse stocks” page at the Bloomington *Drosophila* Stock Center (BDSC; <https://bdsc.indiana.edu/stocks/index.html>), it is possible to both get a sense of the types of fly stocks available and access full lists of stocks of a given type (RNAi, CRISPR, deletion, etc.). We also want to reiterate that previous reviews of resources available to the *Drosophila* research community include (Mohr et al. 2014) and (Mohr et al. 2021). When the best option is to create a new fly stock or other type of reagent, researchers can turn to available sets of precomputed designs, e.g. for CRISPR sgRNAs or qPCR primer pairs (Table 15). A main consideration prior to using a precomputed resource of reagent designs should be to determine which FlyBase, NCBI, or other genome annotation was used by the developers of the design tool. Last, we reiterate that the FlyBase instance of JBrowse provides a visual annotation, organized by genomic region, of fly stock resources, precomputed reagent designs, and more, as noted at the FlyBase JBrowse wiki page (https://wiki.flybase.org/wiki/FlyBase:JBrowse_Tracks).

Concluding comments

Information common to subsets of genes or proteins

We have primarily focused on meta-databases and resources relevant to all *Drosophila* genes. For some specific types of proteins, such as transcription factors or kinases, mining information in specialized resources can be relevant and useful. Examples include mining of cis-regulatory modules or transcription factor

Table 16. Curated lists of resources relevant to *Drosophila* gene information mining and research.

Resource	Navigation	URL
Alliance	Alliance “Data and Tools” menu tab	https://www.alliancegenome.org/
DRSC	DRSC Online Tools Overview Page	https://fgr.hms.harvard.edu/tools
EBI	EBI “Services” list of resources and tools	https://www.ebi.ac.uk/services/data-resources-and-tools
FlyBase	FlyBase “external resources” links	https://flybase.org/
FlyBase Wiki	FlyBase Wiki, “external resources” section	https://wiki.flybase.org/wiki/FlyBase:External_Resources
FlyBase Wiki	FlyBase Wiki, “new to flies” section	https://wiki.flybase.org/wiki/FlyBase:New_to_Flies
NCBI	NCBI home page	https://www.ncbi.nlm.nih.gov/

binding sites at RedFly (<http://redfly.ccr.buffalo.edu/>; Keranen et al. 2022) or mining of kinase-substrate predictions at iProteinDB (<https://www.flyrnai.org/tools/iproteindb/web/>; Hu et al. 2019). Table 16 provides URLs for webpages that list multiple online resources useful to *Drosophila* researchers. These include the lists of available online resources within a group, namely, summary pages at the DRSC, EBI, and NCBI. They also include lists of online resources from many groups, as curated by the Alliance or by FlyBase. In addition, gene- or protein-focused pages at the meta-databases listed in Table 1 typically include external links relevant to specific types of genes or proteins. Mining additional databases can lead to discovery of additional information and predictions and to identification of additional relevant reagents.

Navigating across databases

There are a number of circumstances in which one might want to navigate across multiple databases, including databases such as OMIM (Amberger et al. 2019) or other human disease-focused platforms. Using ortholog mapping tools to generate a list of human gene symbols, NCBI Entrez Gene IDs or other gene IDs is one way to prepare to effectively navigate among different online resources. In addition, we note that FlyBase supports searches with human disease terms, which can be used to view a list of corresponding fly genes (e.g. look for “Human Diseases” at the “Quick Search” tab, and limit search results to “genes” at the “Filter by data class” menu). Moreover, the knowledgebases listed in Table 1, as well as DIOPT and many other online resources mentioned here, include external links to other resources, e.g. NCBI, facilitating direct navigation across resources. Additionally, we refer researchers to the FlyBase ID conversion tool for help synchronizing and disambiguating the different types of gene and protein IDs used by different resources (<https://flybase.org/convert/id>).

Giving a *Drosophila* gene a name

As you complete a study of a previously uncharacterized *Drosophila* gene with only a CG number identifier, you are likely to start thinking about assigning a text-based name to the gene. Early fly geneticists named mutant fly strains based on what they observed: e.g. white eyes or notched wings. In addition, they used lower case when the mutant allele identified behaved as a recessive allele and upper case for dominant alleles. When the corresponding genes were later assigned names, the names

and capitalization given to the mutant fly strains stuck, giving us gene names like *white* and *Notch*. The naming convention has remained largely the same over time, supplemented by database-friendly systematic identifiers like CG numbers and FBgn identifiers. *Drosophila* researchers typically name fly genes based on some aspect of a mutant phenotype associated with disruption of the gene. Exactly how the name relates to a phenotype, however, is not always obvious. Moreover, although most *Drosophila* gene names derive from English language words, some gene names originate from other languages, or are the names of real or fictional persons. Assigning a name to a *Drosophila* gene can be a lot of fun. Nevertheless, some serious thought should go into it. Guidelines are provided on the FlyBase wiki (<https://wiki.flybase.org/wiki/FlyBase:Nomenclature>), and researchers are encouraged to discuss potential new *Drosophila* gene names with leaders or staff at FlyBase.

Paying it forward

The meta-databases listed in Table 1 rely on researchers to share data and associated metadata. Ideally, we should all be aware of and adhering to community standards related to data sharing, including the FAIR principles, i.e. the idea that data should be findable, accessible, interoperable, and reusable. Sharing data and associated metadata should not be an afterthought. For small- and large-scale projects alike, it is wise to plan data management and metadata annotation in experimental design stages and follow through on those plans. Good records in appropriate formats are critical to curation and reuse now. Good records are also critical to ensuring high-quality outputs as automated tools such as text-mining and AI are applied to experimental datasets, now and in the future. Upon publication of a study that reports new findings for *Drosophila* genes, researchers are encouraged to use the FlyBase Fast-Track Your Paper tool to give a head start to curation and take other steps as needed for specific projects. Special consideration might be needed for large-scale datasets. Deposition of large-scale data and associated metadata to databases such as NCBI Sequence Read Archive (SRA), DroID, or RSVP Plus can help ensure community access. Overall, sharing data and associated metadata helps ensure that new research plans are built on the most relevant and accurate available information, contributing to the quality and efficiency of research studies. Last, we suggest that large-scale data generation and other community resource-focused efforts should continue to be encouraged and supported, e.g. as more precise 'omics technologies become available and as the research literature continues to expand. Future progress depends on access to more complete and accurate datasets, as well as on continued curation of the literature and continued improvement of online resources.

Acknowledgments

We thank Brian Calvi, Adam Carte, Victoria Jenkins, Oguz Kanca, Gillian Millburn, and Jonathan Zirin for helpful comments. We recognize that there are additional valuable resources we did not mention and apologize for omissions.

Funding

The authors are associated with the *Drosophila* Research & Screening Center-Biomedical Technology Research Resource (DRSC-BTRR), which is funded by the U.S. National Institutes of Health (NIH) National Institute of General Medical Sciences

(NIGMS) P41 GM132087. A-RK was supported by Postdoctoral Fellowship Program (Nurturing Next-generation Researchers) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A3A14039622). NP is an investigator of the Howard Hughes Medical Institute.

Conflicts of interest statement

The author(s) declare no conflict of interest.

Literature cited

- Alliance of Genome Resources Consortium. 2020. Alliance of genome resources portal: unified model organism research platform. *Nucleic Acids Res.* 48(D1):D650–D658. doi:10.1093/nar/gkz813.
- Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.Org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43(D1):D789–D798. doi:10.1093/nar/gku1205.
- Amberger JS, Bocchini CA, Scott AF, Hamosh A. 2019. OMIM.Org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* 47(D1):D1038–D1043. doi:10.1093/nar/gky1151.
- Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, Marygold SJ, FlyBase C. 2016. Flybase: establishing a gene group resource for *Drosophila melanogaster*. *Nucleic Acids Res.* 44(D1):D786–D792. doi:10.1093/nar/gkv1046.
- Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science.* 373(6557):871–876. doi:10.1126/science.abj8754.
- Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. 2009. QuickGO: a web-based tool for gene ontology searching. *Bioinformatics.* 25(22):3045–3046. doi:10.1093/bioinformatics/btp536.
- Bittrich S, Bhikadiya C, Bi C, Chao H, Duarte JM, Dutta S, Fayazi M, Henry J, Khokhriakov I, Lowe R, et al. 2023. RCSB protein data bank: efficient searching and simultaneous access to one million computed structure models alongside the PDB structures enabled by architectural advances. *J Mol Biol.* 435(14):167994. doi:10.1016/j.jmb.2023.167994.
- Brameier M, Krings A, MacCallum RM. 2007. Nucpred—predicting nuclear localization of proteins. *Bioinformatics.* 23(9):1159–1160. doi:10.1093/bioinformatics/btm066.
- Brand AH, Perrimon N. 1993. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development.* 118(2):401–415. doi:10.1242/dev.118.2.401.
- Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elisk CG, Lewis SE, Stein L, et al. 2016. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* 17(1):66. doi:10.1186/s13059-016-0924-1.
- Casas-Vila N, Bluhm A, Sayols S, Dinges N, Dejung M, Altenhein T, Kappei D, Altenhein B, Roignant JY, Butter F. 2017. The developmental proteome of *Drosophila melanogaster*. *Genome Res.* 27(7):1273–1285. doi:10.1101/gr.213694.116.
- Duckert P, Brunak S, Blom N. 2004. Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel.* 17(1):107–112. doi:10.1093/protein/gzh013.
- Dunst S, Kazimiers T, von Zadow F, Jambor H, Sagner A, Brankatschk B, Mahmoud A, Spann S, Tomancak P, Eaton S, et al. 2015. Endogenously tagged rab proteins: a resource to study membrane

- trafficking in *Drosophila*. *Dev Cell*. 33(3):351–365. doi:[10.1016/j.devcel.2015.03.022](https://doi.org/10.1016/j.devcel.2015.03.022).
- Foreman J, Perrett D, Mazaika E, Hunt SE, Ware JS, Firth HV. 2023. DECIPHER: improving genetic diagnosis through dynamic integration of genomic and clinical data. *Annu Rev Genomics Hum Genet*. 24:151–175. doi:[10.1146/annurev-genom-102822-100509](https://doi.org/10.1146/annurev-genom-102822-100509).
- Franz M, Rodriguez H, Lopes C, Zuberi K, Montojo J, Bader GD, Morris Q. 2018. GeneMANIA update 2018. *Nucleic Acids Res*. 46(W1):W60–W64. doi:[10.1093/nar/gky311](https://doi.org/10.1093/nar/gky311).
- Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science*. 302(5651):1727–1736. doi:[10.1126/science.1090289](https://doi.org/10.1126/science.1090289).
- Giraldo-Calderon GI, Harb OS, Kelly SA, Rund SS, Roos DS, McDowell MA. 2022. Vectorbase.org updates: bioinformatic resources for invertebrate vectors of human pathogens and related organisms. *Curr Opin Insect Sci*. 50:100860. doi:[10.1016/j.cois.2021.11.008](https://doi.org/10.1016/j.cois.2021.11.008).
- Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, Dos Santos G, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T, et al. 2022. Flybase: a guided tour of highlighted features. *Genetics*. 220(4):iyac035. doi:[10.1093/genetics/iyac035](https://doi.org/10.1093/genetics/iyac035).
- Greenspan RJ. 2004. Fly Pushing: The Theory and Practice of *Drosophila* Genetics. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.
- Gurharsha KG, Rual JF, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O, et al. 2011. A protein complex network of *Drosophila melanogaster*. *Cell*. 147(3):690–703. doi:[10.1016/j.cell.2011.08.047](https://doi.org/10.1016/j.cell.2011.08.047).
- Hales KG, Korey KA, Larracuente AM, Roberts DM. 2015. Genetics on the fly: a primer on the *Drosophila* model system. *Genetics*. 201(3):815–842. doi:[10.1534/genetics.115.183392](https://doi.org/10.1534/genetics.115.183392).
- Hammonds AS, Bristow CA, Fisher WW, Weiszmman R, Wu S, Hartenstein V, Kellis M, Yu B, Frise E, Celniker SE. 2013. Spatial expression of transcription factors in *Drosophila* embryonic organ development. *Genome Biol*. 14(12):R140. doi:[10.1186/gb-2013-14-12-r140](https://doi.org/10.1186/gb-2013-14-12-r140).
- Harnish JM, Li L, Rogic S, Poirier-Morency G, Kim SY, Boycott KM, Wangler MF, Bellen HJ, Hieter P, Pavlidis P, et al. 2022. Modelmatcher: a scientist-centric online platform to facilitate collaborations between stakeholders of rare and undiagnosed disease research. *Hum Mutat*. 43(6):743–759. doi:[10.1002/humu.24364](https://doi.org/10.1002/humu.24364).
- Hu Y, Chung V, Comjean A, Rodiger J, Nipun F, Perrimon N, Mohr SE. 2020. Biolitmine: advanced mining of biomedical and biological literature about human genes and genes from major model organisms. *G3 (Bethesda)*. 10(12):4531–4539. doi:[10.1534/g3.120.401775](https://doi.org/10.1534/g3.120.401775).
- Hu Y, Comjean A, Mohr SE, FlyBase C, Perrimon N. 2017. Gene2Function: an integrated online resource for gene function discovery. *G3 (Bethesda)*. 7(8):2855–2858. doi:[10.1534/g3.117.043885](https://doi.org/10.1534/g3.117.043885).
- Hu Y, Comjean A, Perkins LA, Perrimon N, Mohr SE. 2015. GLAD: an online database of gene list annotation for *Drosophila*. *J Genomics*. 3:75–81. doi:[10.7150/jgen.12863](https://doi.org/10.7150/jgen.12863).
- Hu Y, Comjean A, Perrimon N, Mohr SE. 2017. The *Drosophila* gene expression tool (DGET) for expression analyses. *BMC Bioinformatics*. 18(1):98. doi:[10.1186/s12859-017-1509-z](https://doi.org/10.1186/s12859-017-1509-z).
- Hu Y, Comjean A, Rodiger J, Liu Y, Gao Y, Chung V, Zirin J, Perrimon N, Mohr SE. 2021. FlyRNAi.org-the database of the *Drosophila* RNAi screening center and transgenic RNAi project: 2021 update. *Nucleic Acids Res*. 49(D1):D908–D915. doi:[10.1093/nar/gkaa936](https://doi.org/10.1093/nar/gkaa936).
- Hu Y, Ewen-Campen B, Comjean A, Rodiger J, Mohr SE, Perrimon N. 2022. Paralog explorer: a resource for mining information about paralogs in common research organisms. *Comput Struct Biotechnol J*. 20:6570–6577. doi:[10.1016/j.csbj.2022.11.041](https://doi.org/10.1016/j.csbj.2022.11.041).
- Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. 2011. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*. 12(1):357. doi:[10.1186/1471-2105-12-357](https://doi.org/10.1186/1471-2105-12-357).
- Hu Y, Roesel C, Flockhart I, Perkins L, Perrimon N, Mohr SE. 2013. UP-TORR: online tool for accurate and up-to-date annotation of RNAi reagents. *Genetics*. 195(1):37–45. doi:[10.1534/genetics.113.151340](https://doi.org/10.1534/genetics.113.151340).
- Hu Y, Sopko R, Chung V, Foos M, Studer RA, Landry SD, Liu D, Rabinow L, Gnäd F, Beltrao P, et al. 2019. iProteinDB: an integrative database of *Drosophila* post-translational modifications. *G3 (Bethesda)*. 9(1):1–11. doi:[10.1534/g3.118.200637](https://doi.org/10.1534/g3.118.200637).
- Hu Y, Vinayagam A, Nand A, Comjean A, Chung V, Hao T, Mohr SE, Perrimon N. 2018. Molecular interaction search tool (MIST): an integrated resource for mining gene and protein interaction data. *Nucleic Acids Res*. 46(D1):D567–D574. doi:[10.1093/nar/gkx1116](https://doi.org/10.1093/nar/gkx1116).
- Hu Qian S, Shi M-W, Wang D-Y, Fear JM, Chen L, Tu YX, Liu HS, Zhang Y, Zhang SJ, Yu S-S, et al. 2023. Integrating massive RNA-Seq data to elucidate transcriptome dynamics in *Drosophila melanogaster*. *Brief Bioinform*. 24(4):bbad177. doi:[10.1093/bib/bbad177](https://doi.org/10.1093/bib/bbad177).
- Jambor H, Surendranath V, Kalinka AT, Meistrick P, Saalfeld S, Tomancak P. 2015. Systematic imaging reveals features and changing localization of mRNAs in *Drosophila* development. *Elife*. 4:e05003. doi:[10.7554/eLife.05003](https://doi.org/10.7554/eLife.05003).
- Jenkins VK, Larkin A, Thurmond J, FlyBase C. 2022. Using FlyBase: a database of *Drosophila* genes and genetics. *Methods Mol Biol*. 2540:1–34. doi:[10.1007/978-1-0716-2541-5_1](https://doi.org/10.1007/978-1-0716-2541-5_1).
- Jeppe H, Konstantinos DT, Mads Damgaard P, José Juan Almagro A, Paolo M, Henrik N, Anders K, Ole W. 2022. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv*. <https://doi.org/10.1101/2022.04.08.487609>. preprint: not peer reviewed.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*. 596(7873):583–589. doi:[10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2).
- Kalderimis A, Lyne R, Butano D, Contrino S, Lyne M, Heimbach J, Hu F, Smith R, Stepan R, Sullivan J, et al. 2014. InterMine: extensive web services for modern biology. *Nucleic Acids Res*. 42(W1):W468–W472. doi:[10.1093/nar/gku301](https://doi.org/10.1093/nar/gku301).
- Kall L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res*. 35(Web Server):W429–W432. doi:[10.1093/nar/gkm256](https://doi.org/10.1093/nar/gkm256).
- Kanca O, Zirin J, Garcia-Marques J, Knight SM, Yang-Zhou D, Amador G, Chung H, Zuo Z, Ma L, He Y, et al. 2019. An efficient CRISPR-based strategy to insert small and large fragments of DNA using short homology arms. *Elife*. 8:e51539. doi:[10.7554/eLife.51539](https://doi.org/10.7554/eLife.51539).
- Kanca O, Zirin J, Hu Y, Tepe B, Dutta D, Lin WW, Ma L, Ge M, Zuo Z, Liu LP, et al. 2022. An expanded toolkit for *Drosophila* gene tagging using synthesized homology donor constructs for CRISPR-mediated homologous recombination. *Elife*. 11:e76077. doi:[10.7554/eLife.76077](https://doi.org/10.7554/eLife.76077).
- Keränen SVE, Villahoz-Baleta A, Bruno AE, Halfon MS. 2022. REDfly: an integrated knowledgebase for insect regulatory genomics. *Insects*. 13(7):618. doi:[10.3390/insects13070618](https://doi.org/10.3390/insects13070618).
- Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuerhahn M, Hinz U, et al. 2012. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*. 40(D1):D841–D846. doi:[10.1093/nar/gkr1088](https://doi.org/10.1093/nar/gkr1088).

- Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, et al. 2023. Pubchem 2023 update. *Nucleic Acids Res.* 51(D1):D1373–D1380. doi:[10.1093/nar/gkac956](https://doi.org/10.1093/nar/gkac956).
- Kumar S, Konikoff C, Sanderford M, Liu L, Newfeld S, Ye J, Kulathinal RJ. 2017. Flyexpress 7: an integrated discovery platform to study co-expressed genes using in situ hybridization images in *Drosophila*. *G3 (Bethesda)*. 7(8):2791–2797. doi:[10.1534/g3.117.040345](https://doi.org/10.1534/g3.117.040345).
- Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM. 2023. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* 51(D1):D445–D451. doi:[10.1093/nar/gkac998](https://doi.org/10.1093/nar/gkac998).
- Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, Hoffman D, Jang W, Kaur K, Liu C, et al. 2020. Clinvar: improvements to accessing data. *Nucleic Acids Res.* 48(D1):D835–D844. doi:[10.1093/nar/gkz972](https://doi.org/10.1093/nar/gkz972).
- Larkin A, Marygold SJ, Antonazzo G, Attrill H, Dos Santos G, Garapati PV, Goodman JL, Gramates LS, Millburn G, Strelets VB, et al. 2021. Flybase: updates to the *Drosophila melanogaster* knowledge base. *Nucleic Acids Res.* 49(D1):D899–D907. doi:[10.1093/nar/gkaa1026](https://doi.org/10.1093/nar/gkaa1026).
- Leader DP, Krause SA, Pandit A, Davies SA, Dow JAT. 2018. Flyatlas 2: a new version of the *Drosophila melanogaster* expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. *Nucleic Acids Res.* 46(D1):D809–D815. doi:[10.1093/nar/gkx976](https://doi.org/10.1093/nar/gkx976).
- Lecuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes TR, Tomancak P, Krause HM. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*. 131(1):174–187. doi:[10.1016/j.cell.2007.08.003](https://doi.org/10.1016/j.cell.2007.08.003).
- Lee PT, Zirin J, Kanca O, Lin WW, Schulze KL, Li-Kroeger D, Tao R, Devereaux C, Hu Y, Chung V, et al. 2018. A gene-specific T2A-GAL4 library for *Drosophila*. *Elife*. 7:e35574. doi:[10.7554/eLife.35574](https://doi.org/10.7554/eLife.35574).
- Li H, Janssens J, De Waegeneer M, Kolluru SS, Davie K, Gardeux V, Saelens W, David FPA, Brbic M, Spanier K, et al. 2022. Fly cell atlas: a single-nucleus transcriptomic atlas of the adult fruit fly. *Science*. 375(6584):eabk2432. doi:[10.1126/science.abk2432](https://doi.org/10.1126/science.abk2432).
- Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13(9):2178–2189. doi:[10.1101/gr.1224503](https://doi.org/10.1101/gr.1224503).
- Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 379(6637):1123–1130. doi:[10.1126/science.ade2574](https://doi.org/10.1126/science.ade2574).
- Lussi YC, Magrane M, Martin MJ, Orchard S, UniProt Consortium. 2023. Searching and navigating UniProt databases. *Curr Protoc.* 3(3):e700. doi:[10.1002/cpz1.700](https://doi.org/10.1002/cpz1.700).
- Lyne R, Smith R, Rutherford K, Wakeling M, Varley A, Guilleri F, Janssens H, Ji W, McLaren P, North P, et al. 2007. Flymine: an integrated database for *Drosophila* and anopheles genomics. *Genome Biol.* 8(7):R129. doi:[10.1186/gb-2007-8-7-r129](https://doi.org/10.1186/gb-2007-8-7-r129).
- Marygold SJ, FlyBase C. 2023. Exploring FlyBase data using QuickSearch. *Curr Protoc.* 3(4):e731. doi:[10.1002/cpz1.731](https://doi.org/10.1002/cpz1.731).
- Meldal BH, Forner-Martinez O, Costanzo MC, Dana J, Demeter J, Dumousseau M, Dwight SS, Gaulton A, Licata L, Melidoni AN, et al. 2015. The complex portal—an encyclopaedia of macromolecular complexes. *Nucleic Acids Res.* 43(D1):D479–D484. doi:[10.1093/nar/gku975](https://doi.org/10.1093/nar/gku975).
- Meldal BHM, Perfetto L, Combe C, Lubiana T, Ferreira Cavalcante JV, Bye AJH, Waagmeester A, Del-Toro N, Shrivastava A, Barrera E, et al. 2022. Complex portal 2022: new curation frontiers. *Nucleic Acids Res.* 50(D1):D578–D586. doi:[10.1093/nar/gkab991](https://doi.org/10.1093/nar/gkab991).
- Millburn GH, Crosby MA, Gramates LS, Tweedie S, FlyBase C. 2016. Flybase portals to human disease research using *Drosophila* models. *Dis Model Mech.* 9(3):245–252. doi:[10.1242/dmm.023317](https://doi.org/10.1242/dmm.023317).
- Mirdita M, Schutze K, Moriawaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. Colabfold: making protein folding accessible to all. *Nat Methods*. 19(6):679–682. doi:[10.1038/s41592-022-01488-1](https://doi.org/10.1038/s41592-022-01488-1).
- Mohr SE, Hu Y, Kim K, Housden BE, Perrimon N. 2014. Resources for functional genomics studies in *Drosophila melanogaster*. *Genetics*. 197(1):1–18. doi:[10.1534/genetics.113.154344](https://doi.org/10.1534/genetics.113.154344).
- Mohr SE, Tattikota SG, Xu J, Zirin J, Hu Y, Perrimon N. 2021. Methods and tools for spatial mapping of single-cell RNAseq clusters in *Drosophila*. *Genetics*. 217(4):4. doi:[10.1093/genetics/iyab019](https://doi.org/10.1093/genetics/iyab019).
- Murali T, Pacifico S, Yu J, Guest S, Roberts GG III, Finley RL Jr. 2011. DroiD 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res.* 39(suppl_1):D736–D743. doi:[10.1093/nar/gkq1092](https://doi.org/10.1093/nar/gkq1092).
- Nagarkar-Jaiswal S, DeLuca SZ, Lee PT, Lin WW, Pan H, Zuo Z, Lv J, Spradling AC, Bellen HJ. 2015. A genetic toolkit for tagging intronic MiMIC containing genes. *Elife*. 4:e08469. doi:[10.7554/eLife.08469](https://doi.org/10.7554/eLife.08469).
- Nagarkar-Jaiswal S, Lee PT, Campbell ME, Chen K, Anguiano-Zarate S, Gutierrez MC, Busby T, Lin WW, He Y, Schulze KL, et al. 2015. A library of MiMICs allows tagging of genes and reversible, spatial and temporal knockdown of proteins in *Drosophila*. *Elife*. 4:e0538. doi:[10.7554/eLife.05338](https://doi.org/10.7554/eLife.05338).
- Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, et al. 2019. The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47(D1):D529–D541. doi:[10.1093/nar/gky1079](https://doi.org/10.1093/nar/gky1079).
- Perkins LA, Holderbaum L, Tao R, Hu Y, Sopko R, McCall K, Yang-Zhou D, Flockhart I, Binari R, Shim HS, et al. 2015. The transgenic RNAi project at Harvard Medical School: resources and validation. *Genetics*. 201(3):843–852. doi:[10.1534/genetics.115.180208](https://doi.org/10.1534/genetics.115.180208).
- Rey AJ, Attrill H, Marygold SJ, FlyBase C. 2018. Using FlyBase to find functionally related *Drosophila* genes. *Methods Mol Biol.* 1757:493–512. doi:[10.1007/978-1-4939-7737-6_16](https://doi.org/10.1007/978-1-4939-7737-6_16).
- Rocha JJ, Jayaram SA, Stevens TJ, Muschalik N, Shah RD, Emran S, Robles C, Freeman M, Munro S. 2023. Functional unknowns: systematic screening of conserved genes of unknown function. *PLoS Biol.* 21(8):e3002222. doi:[10.1371/journal.pbio.3002222](https://doi.org/10.1371/journal.pbio.3002222).
- modEncode Consortium, S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*. 330(6012):1787–1797. doi:[10.1126/science.1198374](https://doi.org/10.1126/science.1198374).
- Schmidt EE, Pelz O, Buhlmann S, Kerr G, Horn T, Boutros M. 2013. GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. *Nucleic Acids Res.* 41(D1):D1021–D1026. doi:[10.1093/nar/gks1170](https://doi.org/10.1093/nar/gks1170).
- Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, et al. 2019. Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* 47(D1):D955–D962. doi:[10.1093/nar/gky1032](https://doi.org/10.1093/nar/gky1032).
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13(11):2498–2504. doi:[10.1101/gr.1239303](https://doi.org/10.1101/gr.1239303).
- Shefchek KA, Harris NL, Gargano M, Matentzoglou N, Unni D, Brush M, Keith D, Conlin T, Vasilevsky N, Zhang XA, et al. 2020. The Monarch initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* 48(D1):D704–D715. doi:[10.1093/nar/gkz997](https://doi.org/10.1093/nar/gkz997).
- Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, et al. 2023. The STRING database in 2023: protein-protein association networks

- and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 51(D1):D638–D646. doi:[10.1093/nar/gkac1000](https://doi.org/10.1093/nar/gkac1000).
- Tang HW, Spirohn K, Hu Y, Hao T, Kovacs IA, Gao Y, Binari R, Yang-Zhou D, Wan KH, Bader JS, et al. 2023. Next-generation large-scale binary protein interaction network for *Drosophila melanogaster*. *Nat Commun.* 14(1):2162. doi:[10.1038/s41467-023-37876-0](https://doi.org/10.1038/s41467-023-37876-0).
- Teufel F, Almagro Armenteros JJ, Johansen AR, Gislason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H. 2022. Signalp 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol.* 40(7):1023–1025. doi:[10.1038/s41587-021-01156-3](https://doi.org/10.1038/s41587-021-01156-3).
- Thakur M, Bateman A, Brooksbank C, Freeberg M, Harrison M, Hartley M, Keane T, Kleywegt G, Leach A, Levchenko M, et al. 2023. EMBL's European Bioinformatics Institute (EMBL-EBI) in 2022. *Nucleic Acids Res.* 51(D1):D9–D17. doi:[10.1093/nar/gkac1098](https://doi.org/10.1093/nar/gkac1098).
- Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. 2022. PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.* 31(1):8–22. doi:[10.1002/pro.4218](https://doi.org/10.1002/pro.4218).
- Thummuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O. 2022. Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res.* 50(W1):W228–W234. doi:[10.1093/nar/gkac278](https://doi.org/10.1093/nar/gkac278).
- Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn G, Antonazzo G, Trovisco V, et al. 2019. Flybase 2.0: the next generation. *Nucleic Acids Res.* 47(D1):D759–D765. doi:[10.1093/nar/gky1003](https://doi.org/10.1093/nar/gky1003).
- Tomancak P, Beaton A, Weizmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, et al. 2002. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* 3(12):RESEARCH0088. doi:[10.1186/gb-2002-3-12-research0088](https://doi.org/10.1186/gb-2002-3-12-research0088).
- Tomancak P, Berman BP, Beaton A, Weizmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM. 2007. Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* 8(7):R145. doi:[10.1186/gb-2007-8-7-r145](https://doi.org/10.1186/gb-2007-8-7-r145).
- UniProt Consortium. 2023. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* 51(D1):D523–D531. doi:[10.1093/nar/gkac1052](https://doi.org/10.1093/nar/gkac1052).
- van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Soding J, Steinegger M. 2023. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol.* [Online ahead of print] doi:[10.1038/s41587-023-01773-0](https://doi.org/10.1038/s41587-023-01773-0).
- Venken KJ, Schulze KL, Haelterman NA, Pan H, He Y, Evans-Holm M, Carlson JW, Levis RW, Spradling AC, Hoskins RA, et al. 2011. MiMIC: a highly versatile transposon insertion resource for engineering *Drosophila melanogaster* genes. *Nat Methods.* 8(9):737–743. doi:[10.1038/nmeth.1662](https://doi.org/10.1038/nmeth.1662).
- Vinayagam A, Hu Y, Kulkarni M, Roesel C, Sopko R, Mohr SE, Perrimon N. 2013. Protein complex-based analysis framework for high-throughput data sets. *Sci Signal.* 6(264):rs5. doi:[10.1126/scisignal.2003629](https://doi.org/10.1126/scisignal.2003629).
- Viswanatha R, Li Z, Hu Y, Perrimon N. 2018. Pooled genome-wide CRISPR screening for basal and context-specific fitness gene essentiality in *Drosophila* cells. *Elife.* 7:e36333. doi:[10.7554/eLife.36333](https://doi.org/10.7554/eLife.36333).
- Viswanatha R, Marneli E, Rodiger J, Merckaert P, Feitosa-Suntheimer F, Colpitts TM, Mohr SE, Hu Y, Perrimon N. 2021. Bioinformatic and cell-based tools for pooled CRISPR knockout screening in mosquitos. *Nat Commun.* 12(1):6825. doi:[10.1038/s41467-021-27129-3](https://doi.org/10.1038/s41467-021-27129-3).
- Wang J, Al-Ouran R, Hu Y, Kim SY, Wan YW, Wangler MF, Yamamoto S, Chao HT, Comjean A, Mohr SE, et al. 2017. MARRVEL: integration of human and model organism genetic resources to facilitate functional annotation of the human genome. *Am J Hum Genet.* 100(6):843–853. doi:[10.1016/j.ajhg.2017.04.010](https://doi.org/10.1016/j.ajhg.2017.04.010).
- Wang Y, Cheng T, Bryant SH. 2017. Pubchem BioAssay: a decade's development toward open high-throughput screening data sharing. *SLAS Discov.* 22(6):655–666. doi:[10.1177/2472555216685069](https://doi.org/10.1177/2472555216685069).
- Wang J, Liu Z, Bellen HJ, Yamamoto S. 2019. Navigating MARRVEL, a web-based tool that integrates human genomics and model organism genetics information. *J Vis Exp.* 150:e59542. doi:[10.3791/59542](https://doi.org/10.3791/59542).
- Wang J, Mao D, Fazal F, Kim SY, Yamamoto S, Bellen H, Liu Z. 2019. Using MARRVEL v1.2 for bioinformatics analysis of human genes and variant pathogenicity. *Curr Protoc Bioinformatics.* 67(1):e85. doi:[10.1002/cpbi.85](https://doi.org/10.1002/cpbi.85).
- Wilk R, Hu J, Blotsky D, Krause HM. 2016. Diverse and pervasive subcellular distributions for both coding and long noncoding RNAs. *Genes Dev.* 30(5):594–609. doi:[10.1101/gad.276931.115](https://doi.org/10.1101/gad.276931.115).
- Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, et al. 2018. Drugbank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46(D1):D1074–D1082. doi:[10.1093/nar/gkx1037](https://doi.org/10.1093/nar/gkx1037).
- Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. 2000. DIP: the database of interacting proteins. *Nucleic Acids Res.* 28(1):289–291. doi:[10.1093/nar/28.1.289](https://doi.org/10.1093/nar/28.1.289).
- Yu J, Pacifico S, Liu G, Finley RL Jr. 2008. DroID: the *Drosophila* interactions database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics.* 9(1):461. doi:[10.1186/1471-2164-9-461](https://doi.org/10.1186/1471-2164-9-461).
- Zaru R, Orchard S, UniProt C. 2023. UniProt tools: bLAST, align, peptide search, and ID mapping. *Curr Protoc.* 3(3):e697. doi:[10.1002/cpz1.697](https://doi.org/10.1002/cpz1.697).

Editor: H. Bellen