GLAD: an Online Database of Gene List Annotation for Drosophila

Yanhui Hu1, Aram Comjean1, Lizabeth A. Perkins1, Norbert Perrimon1,2, Stephanie E. Mohr1

1. Drosophila RNAi Screening Center, Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA
2. Howard Hughes Medical Institute, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

Corresponding author: Yanhui Hu, yanhui_hu@hms.harvard.edu

Abstract

We present a resource of high quality lists of functionally related Drosophila genes, e.g. based on protein domains (kinases, transcription factors, etc.) or cellular function (e.g. autophagy, signal transduction). To establish these lists, we relied on different inputs, including curation from databases or the literature and mapping from other species. Moreover, as an added curation and quality control step, we asked experts in relevant fields to review many of the lists. The resource is available online for scientists to search and view, and is editable based on community input. Annotation of gene groups is an ongoing effort and scientific need will typically drive decisions regarding which gene lists to pursue. We anticipate that the number of lists will increase over time; that the composition of some lists will grow and/or change over time as new information becomes available; and that the lists will benefit the scientific community, e.g. at experimental design and data analysis stages. Based on this, we present an easily updatable online database, available at www.flyrnai.org/glad, at which gene group lists can be viewed, searched and downloaded.

Key words: GLAD, Drosophila, genes

Introduction

The Drosophila genome was first published in 2000 [1, 2] and so far five major updates to the genome assembly have been released (versions 2-6) [3, 4]. Based on recent FlyBase release (FB2015_02, May 4, 2015), the Drosophila genome is thought to contain 17,622 annotated genes, of which 13,903 are protein-coding genes. There are many advantages to interrogate the full Drosophila genome, such as in genetic or RNAi screens. However, often it is more appropriate and/or more feasible to screen a sub-set of genes, such as due to limitations on time, availability (e.g. of reagents) and/or costs. Choosing an appropriate sub-set of genes is largely guided by scientific interest. In some cases, researchers build lists for functional studies based on other ‘omics data, such as transcriptomics or proteomics data, prior to a functional genomics study. In other cases, genes are grouped based on common features such as biochemical functions (e.g. kinases) or biological processes. In either case, the quality and completeness of the library will impact results, as genes inadvertently left off a list will not be included in the study, and the presence of genes that do not belong on the list will needlessly use up resources and/or affect the analysis of the results.

The availability of high quality annotated groups of related genes (hereafter, “gene groups”) allow scientists to quickly focus on relevant genes, such as in the context of functional genomics screens in tissue culture cells or in vivo. However, mechanisms for distribution and update of gene groups have been limited. At the time the Drosophila genome was pub-
lished, efforts were made to compile several gene groups (see for example [5-8]). However, given the number of changes to gene annotations since then, both in terms of defining genes and understanding their functions, updating and adding to existing lists becomes important. Over the past years, the Drosophila RNAi Screening Center (DRSC) at Harvard Medical School (HMS) has put together several gene groups based on the needs of specific screening projects, as well as to support organization of reagent collections at the Transgenic RNAi Project (TRiP) at HMS. We have recognized over time a) that gene groups are of value to the community for applications additional to RNAi screens; b) that the gene groups benefit from careful curation and review by experts; and c) that the lists change over time, such that they would benefit from being available in an easily updateable database rather than as static lists. We expect gene groups to be of value to researchers at the study design stage, where the lists can help guide decisions regarding what genes are interrogated in a given screen or other assay, and at data analysis stages, such as by providing a supplement to existing groups used in gene set enrichment analyses.

Results & Discussion

Compilation and annotation strategy for gene groups

As mentioned above, several gene groups were annotated in conjunction with release of the Drosophila genome in 2000 (see for example [5-8]). More recently, FlyBase has begun to associate a number of genes with gene groups. The first release of the FlyBase gene groups (FB2015_02, released May 4, 2015) includes 178 gene groups, with the number of genes in a group ranging from 1 to 168. As we had high-throughput functional genomics screening in mind, the approaches we took to defining, building and annotating gene groups draw on knowledge not available in 2000 and are complementary to the approaches taken by FlyBase. In general, our focus is on larger sets of genes and, given the goals of large-scale functional genomics, we tend to cast a broader net, applying less stringent cut-offs for inclusion in a gene group. So far, we have annotated 23 major gene groups with 29 sub-groups. For example, kinases are annotated as belonging to one of two subgroups: protein kinases and non-protein kinases, and the transcription factors (TFs), related proteins and other DNA-binding proteins are organized into four groups, DNA-binding with transcription factor activity; transcriptional co-factors; chromatin regulation; and possible TFs, which we assign to proteins predicted to be TF based only on low confidence data (Table 1). The number of genes in major gene groups ranges from 53 to 3,683. Currently, GO annotation is the major resource for identification of groups of genes relevant to a particular molecular function, biological process or sub-cellular localization. Several individual laboratories have also built databases in particular areas (e.g. GlycoFly [9] and FlyTF [10]). In addition, a large amount of information exists in free text format in the literature. Although our strategy differed for each group, depending on available resources, in general we built the lists using one or more of the following approaches: a) mining of organized and digitalized information from existing annotation resources and databases including generic gene and protein annotation, e.g. gene ontology and UniProt, as well as specialized resources, e.g. transporterDB [11] or FlyTF [10, 12]; b) mining of information in free text format from the literature; c) mining lists from relevant publications on Drosophila or other species; d) direct curation or review by experts (Table 1). The strategy used to build a Drosophila kinases gene group is outlined in Fig. 1. To help guide studies or analyses that use the gene groups, when possible we have assigned confidence scores that help separate high- and low-confidence associations of a gene with a given group. See Methods and Table 1 for additional details regarding annotation.

Features of the user interface

The gene group resource, which we call GLAD for gene list annotation for Drosophila, is available online at www.flyrnai.org/glad. Users can choose a gene group of interest from a drop-down menu. At the results page (Fig. 2), information about how the list was built is indicated, along with detailed information regarding members of the group. The table of genes includes FlyBase gene identifiers, gene symbols, sub-group annotations and if available, a confidence score. Tables can be downloaded for off-line analysis. At the user interface, a link to UP-TORR [13] is provided so users can quickly identify corresponding cell-based or in vivo RNAi reagents from public resources. A form provides an opportunity for the research community at large to suggest changes or additions to a given list (see below).

Feedback will improve the quality of the resource

Although we have made a concerted effort to evaluate all available resources and used best available methods for building each list, there remains room for improvement, in particular as new knowledge places new genes in a given group. To facilitate
community updates, we welcome and encourage researchers to use the form at each gene group list to provide feedback and/or alert us to relevant publications. We will evaluate feedback and modify gene groups accordingly. Annotation of new gene groups is also an ongoing effort, and scientific interest will typically drive decisions regarding which gene groups to build next. We welcome feedback from the community regarding which groups not already covered by GLAD or FlyBase gene groups should be added. With community input as well as continued curation by bioinformatics experts, we anticipate that the GLAD resource will improve and expand over time, further increasing its value to the community. Examples of studies that used these gene groups include a primary cell-based screen of autophagy-related factors [14] and an in vivo screen of transcription factors [15].

**Table 1:** Summary of approaches to gene group compilation, annotation and expert review.

<table>
<thead>
<tr>
<th>Gene group</th>
<th>Sub-group</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autophagy-related*</td>
<td>Mapped from literature [17]</td>
<td></td>
</tr>
<tr>
<td>Chaperone and heat shock proteins</td>
<td>GO/UniProt annotation supplemented by human chaperon/HSP list from Dr. Susan Lindquist lab</td>
<td></td>
</tr>
<tr>
<td>Cytoskeletal</td>
<td>Interactive Fly, UniProt and publication [6], reviewed by Dr. Norbert Perrimon</td>
<td></td>
</tr>
<tr>
<td>Glycoproteins</td>
<td>GlycoFly database [9]</td>
<td></td>
</tr>
<tr>
<td>GPCRs*</td>
<td>GO annotation, in collaboration with Dr. Mathias Beller</td>
<td></td>
</tr>
<tr>
<td>Kinases*</td>
<td>Nomenclature/GO/domain annotation supplemented with human KP list [23] and publications [22] [7], reviewed by Dr. Richelle Sopko</td>
<td></td>
</tr>
<tr>
<td>Phosphatases*</td>
<td>Protein Phosphatase</td>
<td></td>
</tr>
<tr>
<td>Major signaling pathways</td>
<td>flyReactome, reviewed by Dr. Herve Agaisse</td>
<td></td>
</tr>
<tr>
<td>Autophagy-related factors*</td>
<td>GO annotation and FlyTF.org [10]</td>
<td></td>
</tr>
<tr>
<td>Transcription factors, related proteins and other DNA-binding proteins*</td>
<td>DNA-binding with transcription factor activity</td>
<td></td>
</tr>
<tr>
<td>Trans-membrane proteins*</td>
<td>DRSC in collaboration with NYU supplemented by TMHMM prediction [33]</td>
<td></td>
</tr>
<tr>
<td>Transporters</td>
<td>TransporterDB [11]</td>
<td></td>
</tr>
<tr>
<td>Ubiquitin-related*</td>
<td>DRSC in collaboration with NYU</td>
<td></td>
</tr>
</tbody>
</table>

* Indicates availability of a corresponding DRSC RNAi library for cell-based screens.
Methods

Compilation of gene groups

Specific resources used to annotate the gene groups are shown in Table 1. Five groups listed below exemplify the range of approaches we took. 1) Most of the major signal transduction pathways were assembled manually by Dr. N. Perrimon. 2) The main source for the autophagy-related factors list was orthologs [16] of factors identified in a mammalian proteomics study [17]. 3) The mitochondrial gene group was built by combining gene annotation, relevant databases (MitoDrome, MitoMiner), mapping of mammalian orthologs and experimental proteomics data [18-20]. 4) The transcription factors and other DNA-binding factors list was built based on GO annotation and domain annotation. More specifically, the genes annotated with relevant GO terms such as “sequence-specific DNA binding transcription factor activity” and/or genes annotated with relevant domain at InterPro database. The list was then supplemented with the genes annotated at FlyTF database [10, 12]. The sub-categories were assigned based on GO term and high confidence was assigned to TFs associated with experimental evidence at FlyTF database. 5) As outlined in Fig. 1, the kinase list was initially assembled based on genes annotated as having “kinase activity,” supplemented by genes annotated at InterPro [21] as containing a “kinase” domain, supplemented with information from two publications [7, 22], and further supplemented following mapping of Drosophila orthologs of human genes included on a list of kinases [23]. The compiled list was reviewed by Dr. Richelle Sopko with specific expertise and interest in the field. Family assignment was applied according to Manning et al. [22].

User interface implementation

The GLAD website was created at the DRSC and hosted on web servers provided and maintained by the Harvard Medical School (HMS) Research Computing Group. The website was created in PHP using silex as a web framework. The PHP web code pulls the lists from tables from the mysql database of flyrnai and prepares data for display. HTML/javascript and jquery was used to create sortable output tables.

Figure 1: Strategy for assembly of a Drosophila kinases gene group. As outlined, we incorporated information from several sources. See main text and Table 1 for relevant URLs and reference citations.
Figure 2: User interface for the GLAD gene group resource (www.flyrnai.org/glad). Features of the user interface include search or download of specific lists, one-click transfer of the list to UP-TORR for identification of corresponding RNAi reagents, and the option to provide feedback regarding a list (e.g. suggest new genes or relevant publications).
Acknowledgements

We would like to thank Drs. Herve Agaisse (Yale School of Medicine), Jeff Axelrod (Stanford School of Medicine), Mathias Beller (Heinrich Heine Universität Düsseldorf), Phillip Karpowicz (University of Winsor), John J. Kim (University of Michigan), Henry Krause (University of Toronto), Susan Lindquist (MIT), Ralph Neumuller (Harvard Medical School), Richelle Sopko (Harvard Medical School), Bing Ye (University of Michigan) and Jonathan Zirin (Harvard Medical School) for taking the time to review specific gene groups. We also thank Dr. Steven Marygold at FlyBase for helpful discussion. The DRSC and TRip are supported by NIH NIGMS R01 GM067761, R01 GM084947 and R24 RR032668. SEM is additionally supported in part by the Dana Farber/Harvard Cancer Center, which is supported in part by NCI Cancer Center Support Grant # NIH 5 P30 CA06516. NP is an Investigator with the Howard Hughes Medical Institute.

Competing Interests

The authors have declared that no competing interest exists.

References

31. Ben-Shlomo I, Yu Hsu S, Rauch R, Kowalski HW, Hsuheh AJ. Signaling receptors: a genomic and evolutionary perspective of
