Summarising scRNAseq expression data in FlyBase

Damien Goutte-Gattat [1]

[1] Dept. of Physiology, Development and Neuroscience, University of Cambridge, Cambridge CB2 1TN, UK

Abstract

Single-cell RNA sequencing has proved an invaluable tool in biomedical research. The ability to survey the transcriptome of individual cells offers many opportunities and has already paved the way to many discoveries in both basic and clinical research. With the increasing amount of single-cell transcriptomic data available, including whole-organism single-cell transcriptomic atlases, biological databases face a challenge to integrate these data and make them easily accessible to their users. In this Short Paper, we describe how the fruit fly-specific database FlyBase is making use of single-cell RNA sequencing datasets to let fly researchers quickly know in which cell types genes are known to be expressed.

Keywords

Gene expression, single-cell RNA sequencing, cell types, biocuration

1. Introduction

First introduced in 2009, single-cell RNA sequencing (scRNAseq) has become a powerful tool to investigate cell states and functions [1]. Since 2017, many studies have leveraged the technique in the fruit fly *Drosophila melanogaster*, leading to various new insights [2], and the number of scRNAseq datasets from flies is only expected to grow quickly in the coming years. Recently, a consortium of 40 fly laboratories, the Fly Cell Atlas project, completed and reported the first single-cell transcriptomic atlas of the entire adult fruit fly, sequencing more than half a million cells of 250 different types across 17 tissues [3].

FlyBase is the Model Organism Database (MOD) for all data related to *Drosophila melanogaster* [4]. It provides access to a wide range of scientific information either manually curated from the published literature or from high-throughput research projects. Expression data are a particularly important subset, and we strive to give access to as many resources as possible to allow fly researchers to find out in which tissues and at which developmental stages their gene of interest is known to be expressed. To that end, we already make use of several high-throughput projects such as modENCODE [5], FlyAtlas [6], and FlyAtlas 2 [7]. We now want to exploit available scRNAseq datasets to offer our users a new window on per-cell type gene expression.

2. Aims

We want to help FlyBase users to: (a) discover the available *Drosophila* scRNAseq datasets; (b) get some information (metadata) about these datasets; (c) get a quick overview of the expression data from those datasets.

In this paper, we will focus on (c), and explain how we leverage the expression data provided by scRNAseq experiments to obtain and present a *per-cell type summary* of the expression data, so that users can gen an “immediate” answer (straight from the “Gene Report” page) to the following questions: What are the cell types in which a given gene is expressed? What is the proportion of cells of a given type in which the gene is expressed? What is the average expression level of the gene across all cells of that type?

An important design decision is that no scRNAseq dataset will ever be stored within FlyBase itself. Instead, we will only store some metadata as well as a simplified version of the gene expression data – only what is needed to cover the questions listed above. Users wanting to inspect the full data will be directed to a pre-existing external data store where the data will already be available.

ICBO 2022, September 25–28, 2022, Ann Arbor, MI, USA
EMAIL: dpg44@cam.ac.uk (A. 1)
ORCID: 0000-0002-6095-8718 (A. 1)

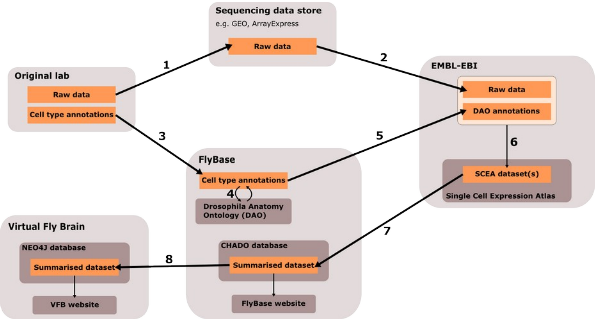


Figure 1: Flow of scRNAseq data into FlyBase and VFB. Authors of a scRNAseq paper upload their raw sequencing data to a public data store such as the NCBI's GEO or the EMBL-EBI's ArrayExpress (1). The raw data are fetched by EMBL-EBI data curators (2). FlyBase curators request the cell type annotations from the the authors (3), convert the original cell type labels to terms from the Drosophila Anatomy Ontology (4), and provide the converted annotations to the EMBL-EBI curators (5). The raw data and their annotations are analysed according to a standard pipeline and the results are published on the EMBL-EBI's Single Cell Expression Atlas website (6). FlyBase curators fetch the analysed data and produce a *summarised version* (7), which is used to feed the FlyBase website. Virtual Fly Brain curators fetch the summarised data from FlyBase and convert them into a graph representation (8), which is used to feed the VFB website.

3. Data pipeline

3.1. Data source

All the scRNAseq data used in FlyBase are obtained through the Single Cell Expression Atlas (SCEA), the EMBL-EBI resource for gene expression data at the single cell level [8]. We thus have a single data provider that is independent of all the individual research projects. The SCEA data curators process the original raw data, as deposited by the original authors in common data stores such as the Gene Expression Omnibus or Array Express (Figure 1, steps 1–2), according to a standard pipeline with common parameters that allows for cross-dataset comparisons. As an added benefit, the SCEA provides the processed data in a common format, independently of the actual scRNAseq method originally used (e.g., Smart-seq2 or 10× sequencing).

As part of a collaboration between FlyBase and the EMBL-EBI, FlyBase curators assist the EMBL-EBI curators in two ways: a) FlyBase curators alerts the EMBL-EBI curators of any upcoming fly scRNAseq dataset of interest, based on the monitoring of fly literature that is routinely done at FlyBase; b) FlyBase curators obtain the cell type annotations from the datasets' authors, convert them to proper ontology terms (see next section), and provide them to the EMBL-EBI curators (Figure 1, steps 3–5).

3.2. Converting to ontology terms

Among other fly resources, FlyBase maintains the Drosophila Anatomy Ontology (DAO), the reference ontology for *Drosophila melanogaster* anatomical structures and cell types [9]. Upon obtaining the cell type annotations for a given dataset from the upstream authors, FlyBase curators make sure that the annotations use cell type names that correspond to existing terms in the DAO. There are several reasons why this step is needed:

1. the authors may have used a "common" cell type name, not knowing that this cell type is formally known under another name in the DAO (an example is "astrocyte", for which the actual term in the DAO is "astrocyte-like glial cell");
2. an annotation may refer to a cell state rather than, or in addition to, a cell type (an example is "plasmatocyte-prolif", which was used in a dataset to annotate a cluster of proliferating plasmatocytes; the appropriate ontology term in that case is "plasmatocyte", leaving the cell state aside);
3. an annotation may incorrectly refer to an organ or a tissue rather than to the cell types that make up this organ or tissue (an example was a cluster annotated as "dorsal vessel", where the correct cell type name should have "cardial cell");
4. an annotation may refer to several cell types (an example was "OPN innervating DA1, VA1 or DC3 glomerulus"; the correct term in such case is the closest cell type that encompasses all cell types in the cluster, in this instance "olfactory projection neuron").

Lastly, there is one case where the original annotation cannot be converted to an ontology term: when the cell type is either not clearly identified (e.g. "btl-GAL4 positive cell, likely to be ovary cell"), or identified only for what it is not (e.g. “non-hemocyte”). Such clusters will be excluded from the expression data summarisation described below.

Of note, this “conversion” of the original annotations is not destructive: the original annotations from the upstream authors are at all time preserved. FlyBase curators merely add a new set of ontology-compliant annotations alongside the original annotations. Both sets of annotations are ultimately available on the Single Cell Expression Atlas.

3.3. Ontology updates

In the process of converting the original annotations to DAO-compliant annotations, the need to update the DAO itself may arise, for mostly two reasons: a) a dataset reports a new cell type, which must be added to the ontology (an example was “adipohemocyte”, a subtype of plasmatocytes found in a scRNAseq analysis of the lymph gland [10]); b) the ontology is found to lack some “grouping terms”, useful to offer a better qualification of some clusters (an example was the addition of “endocrine cell of the ring gland”, to encompass the various types of endocrine cells found in the ring gland). FlyBase data curators then work alongside FlyBase ontology editors to update the DAO as needed and the newly added terms are used to annotate the scRNAseq clusters.

In some cases, this process of updating the DAO to better annotate a dataset may actually start before the dataset is even published, through a direct collaboration between FlyBase and the dataset’s authors. This was notably the case of the Fly Cell Atlas dataset [3].

3.4. Summarising the expression data

Once the raw sequencing data have been processed by the SCEA standard pipeline, the most important output is the *Normalised Expression Matrix*: for any single dataset, this is a large matrix of *n* cells by *m* genes, with *n* the total number of cells in the dataset and *m* the total number of identified genes; each matrix cell contains the normalised reads count for a particular gene in a particular cell. This matrix is accompanied by the *Experiment Design Table*, which contains annotations for each single cell (notably the cell type annotations, both as submitted by the original authors and as corrected by the FlyBase curators as described above).

From those two files, we proceed to what we call the *per-cell type summarisation* of the gene expression data (Figure 1, step 7). For each gene and for each identified cell type, we extract three measures: a) the *cluster size*, that is the number of cells annotated with that cell type; b) the *spread* or *extent of expression*, the proportion of cells in the cluster in which the gene is expressed (the number of cells in the cluster with a non-zero reads count for that gene, divided by the cluster size); and c) the *average expression* of the gene in positive cells (the total number of reads across all cells in the cluster, divided by the number of cells with a non-zero count). Those measures are then stored in the Chado database at the heart of FlyBase, from where they can be exploited on the FlyBase website.

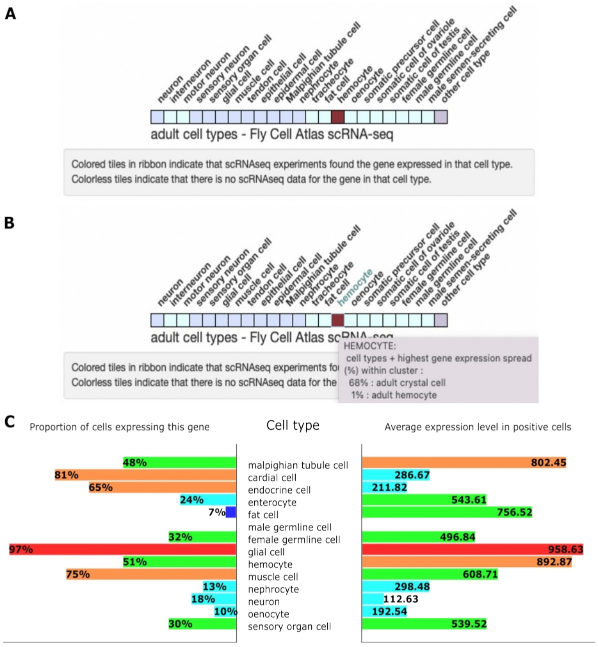
The summarised data are also slated to be used by Virtual Fly Brain [11]. To that effect, a subset of the data, corresponding to the cell types present in the adult brain only and to genes with an extent of expression of more than 0.9 (genes that are expressed in more than 90% of cells of a given type), is loaded from the Chado database to the Neo4j graph database (Figure 1, step 8) that powers the Virtual Fly Brain website.

4. User-visible outcomes

The summarised expression data will be exploited on the FlyBase website in several incremental steps. The first step was realised for the 2022_03 release of FlyBase in June 2022. In this release, a *cell type ribbon* was added to the Gene Report page (Figure 2A). The ribbon lists 22 high-level cell types, each of those being coloured according to the extent of expression of the current gene in cells of that type. Because several more precise cell types are agglomerated into a single high-level cell type, a list of all the precise subtypes, with their corresponding extent of expression, is given in a tooltip box shown when the mouse pointer is left hovering over a ribbon cell (Figure 2B). To avoid cluttering that tooltip, cell types where the extent of expression is less than 1% are filtered out.

At present, the cell type ribbon is only fed with the data from the Fly Cell Atlas project [3]. In a second step and as more datasets will be accumulated into FlyBase, we will update the ribbon to display consolidated data from all the available scRNAseq datasets.

Lastly, in a future release we will add a new type of graph to the “high-throughput expression



Colored tiles in ribbon indicate that scRNAseq experiments found the gene expressed in that cell type.
Colorless tiles indicate that there is no scRNAseq data for the gene in that cell type.

Colored tiles in ribbon indicate that scRNAseq experiments fo
Colorless tiles indicate that there is no scRNAseq data for the

HEMOCYTE:
cell types + highest gene expression spread (%) within cluster :
68% : adult crystal cell
1% : adult hemocyte

Figure 2: Graphical display of summarised expression data on FlyBase. **(A)** The cell type ribbon. Each cell in the ribbon corresponds to one of the main *Drosophila* cell types. Each cell is coloured depending on the fraction of cells of that type in which the current gene is expressed according to the Fly Cell Atlas dataset. **(B)** When the mouse pointer is hovering over one cell, a tooltip shows the list of cell subtypes in which the current gene is expressed, along with the proportion of cells expressing the gene in each subtype. **(C)** Mockup of a more complete graphical representation of summarised expression data that is planned for a future release of FlyBase. For each cell type, this graph will display both the proportion of cells of that type in which the gene is expressed (left) and the average expression level in all cells that do express the gene (right).

data” section of the Gene Report page. The exact modalities of this new graph are yet to be determined, but il will display both the extent of expression and the average expression level for a given gene across high-level cell types (Figure 2C).

5. Acknowledgements

This work was supported by joint grants BB/T014008/1 (FlyBase) and BB/T014563/1 (EMBL-EBI) from the UK Biotechnology and Biological Sciences Research Council (BBSRC) and the US National Science Foundation Directorate of Biological Sciences (NSF/BIO).

6. References

[1] Haque, A., Engel, J., Teichmann, S.A., Lönnberg, T., A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 2017, 9, 1–12.

[2] Li, H., Single-cell RNA sequencing in *Drosophila* : Technologies and applications. *WIREs Dev Biol* 2021, 10.

[3] Li, H., Janssens, J., De Waegeneer, M., Kolluru, S.S., et al., Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science* 2022, 375, eabk2432.

[4] Larkin, A., Marygold, S.J., Antonazzo, G., Attrill, H., et al., FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res* 2020, 49, D899–D907.

[5] Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., et al., The developmental transcriptome of Drosophila melanogaster. *Nature* 2011, 471, 473–479.

[6] Chintapalli, V.R., Wang, J., Dow, J.A.T., Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nat Genet* 2007, 39, 715–720.

[7] Leader, D.P., Krause, S.A., Pandit, A., Davies, S.A., Dow, J.A.T., FlyAtlas 2: a new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. *Nucleic Acids Research* 2018, 46, D809–D815.

[8] Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A.M.-P., et al., Expression Atlas update: from tissues to single cells. *Nucleic Acids Research* 2019, gkz947.

[9] Costa, M., Reeve, S., Grumbling, G., Osumi-Sutherland, D., The Drosophila anatomy ontology. *Journal of Biomedical Semantics* 2013, 4, 32.

[10] Cho, B., Yoon, S.-H., Lee, D., Koranteng, F., et al., Single-cell transcriptome maps of myeloid blood cell lineages in Drosophila. *Nat Commun* 2020, 11, 4483.

[11] Milyaev, N., Osumi-Sutherland, D., Reeve, S., Burton, N., et al., The Virtual Fly Brain browser and query interface. *Bioinformatics* 2012, 28, 411–415.