Exploiting single-cell RNA sequencing data in FlyBase

Damien Goutte-Gattat,[1,3] Nancy George,[2] Irene Papatheodorou[2] and Nick Brown[1],
[1]FlyBase Group, Department of Physiology, Development and Neuroscience, University of Cambridge
[2]Gene Expression Team, European Bioinformatics Institute (EMBL-EBI)
([3]E-mail: dpg44@cam.ac.uk)

Single-cell RNA sequencing (scRNAseq) 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. For the fruit fly alone, nearly a hundred of scRNAseq datasets have already been published since the first reported use of the technique in fly laboratories in 2017 – a number that is only expected to grow quickly in the coming years. This increasing amount of single-cell transcriptomic data available, including whole-organism single-cell transcriptomic atlases, creates a challenge for biological databases to integrate these data and make them easily accessible to the research community. Here we present the approach adopted by FlyBase.

Data Processing

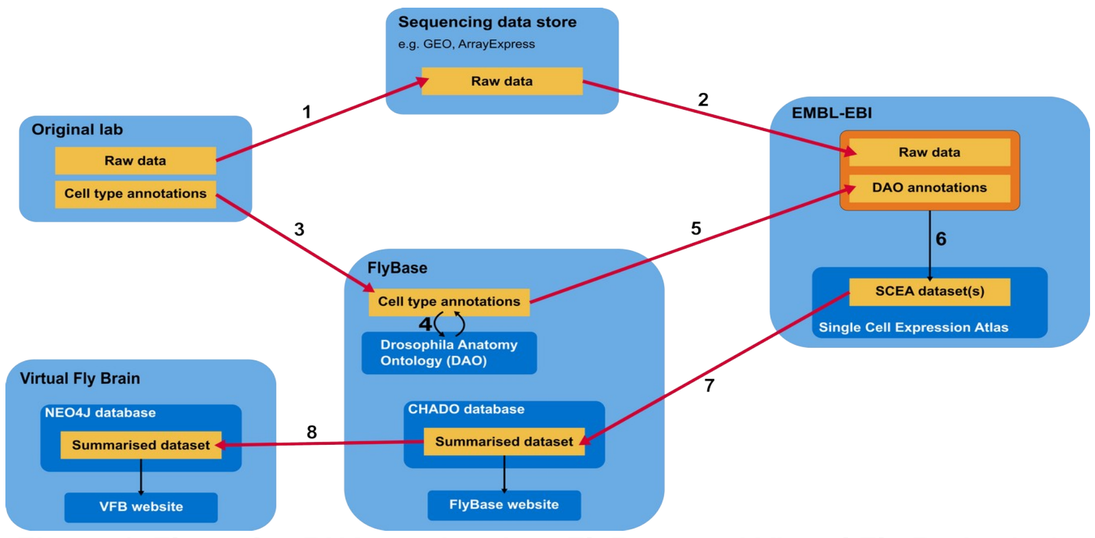


Figure 1. Flow of scRNAseq data into FlyBase and Virtual Fly Brain. Authors of a scRNAseq paper upload their raw sequencing data to a public data store such as the NCBI's Gene Expression Omnibus (GEO) ot the EMBL-EBI's ArrayExpress (1). The raw data are fetched by EMBL-EBI data curators (2). FlyBase curators request the cell type annotations from the authors (3), convert the original cell type labels to terms from the Drosophila Anatomy Ontology (DAO, see Figure 2), 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, see Figure 3), which is used to feed the FlyBase website (Figures 4 and 5). Virtual Fly Brain (VFB) curators fetch the summarised data from FlyBase and convert then into a graph representation (8), which is used to feed the VFB website.

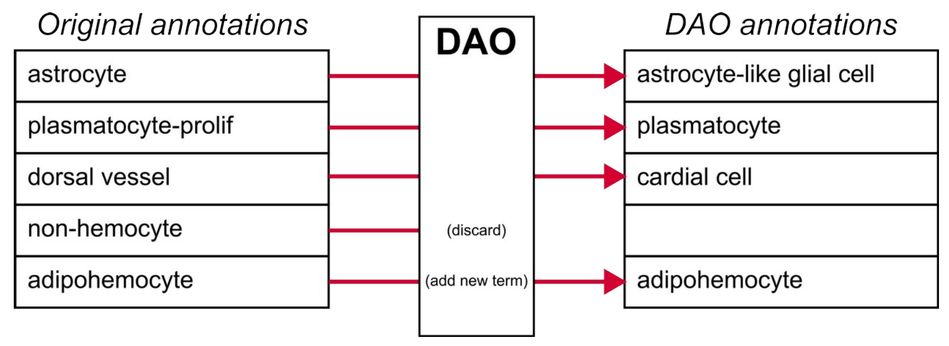


Figure 2. Converting cell type annotations. To ensure consistency of annotations across datasets, cell type annotations provided by upstream authors are converted to standard ontological terms from the Drosophila Anatomy Ontology (DAO). As part of this process, clusters of insufficiently identified cell types are excluded of the downstream data summarisation; conversely, newly identified cell types are added to the ontology.

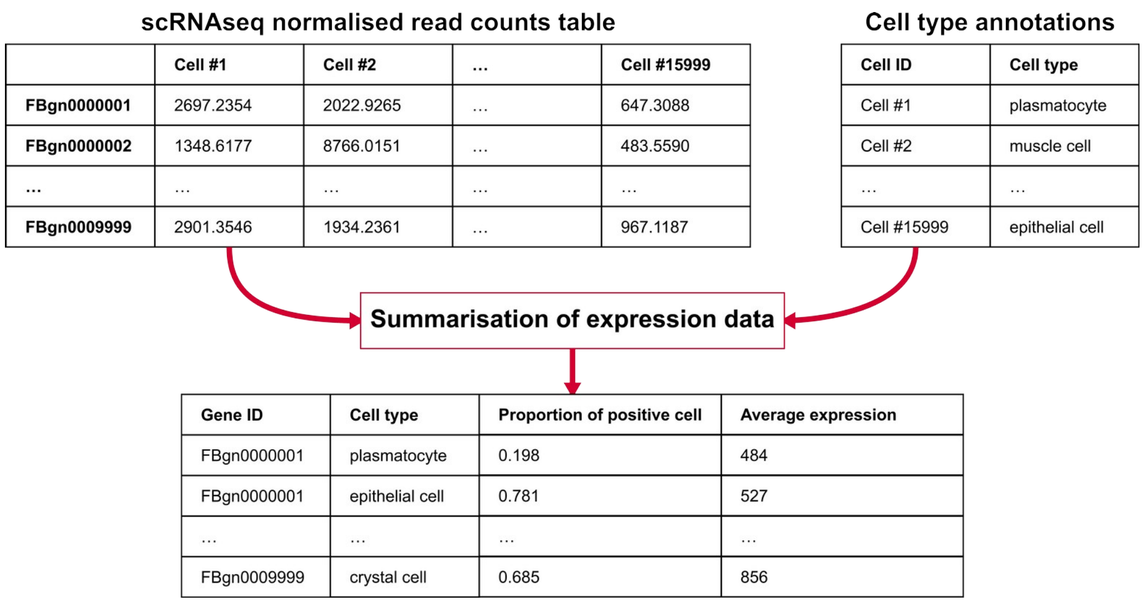


Figure 3. Summarising gene expression data. scRNAseq expression data can be pictured as a very large table where each column represents a single cell and each row represents a gene (top left table); each value in that table is the normalised read count for one gene in one cell, expressed in counts per millions of mapped reads (CPMs). We combine this table with the cell type annotations that associate a cell type to each individual cell (top right table) and summarise the whole by extracting, for each couple {gene, cell type}: (1) the proportion of cells of that type in which the gene is detected at all, and (2) the average CPM in cells of that type that do express the gene. The resulting summarised expression data (bottom table) is loaded into the database and will feed the displays shown on the FlyBase website.

User-visible Output

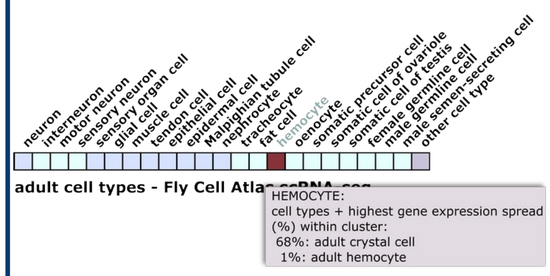
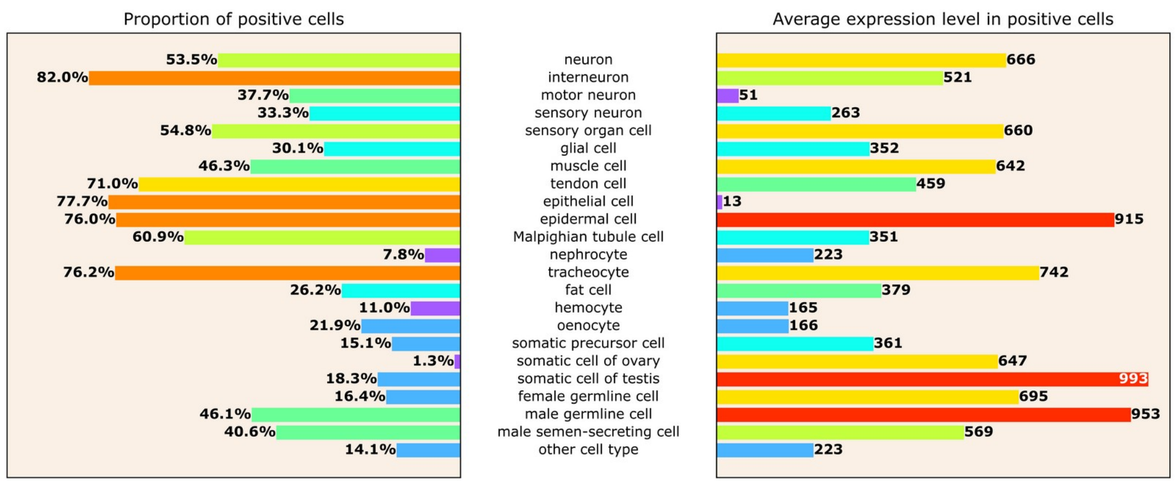


Figure 4. The cell type ribbon. Each tile in the ribbon corresponds to one of the main *Drosophila* cell type. Each tile is coloured depending on the fraction of cells of the corresponding type in which the current geme is expressed, according to the Fly Cell Atlas dataset.

Figure 5. Future graphical display. 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 panel) and the average expression level in all cells that do express the gene (right panel).



FlyBase is the Model Organism Database (MOD) for the fruit fly *Drosophila melanogaster*. It aims to provide access to all the data generated by the fly research community, through curation of the published literature and exploitation of high-throughput research projects.

The Single Cell Expression Atlas (SCEA) is the European Bioinformatics Institute's resource for single-cell RNA sequencing. It compiles publicly available raw data from a wide range of single-cell RNA sequencing studies across 12 different model organisms.