
Damien Goutte-Gattat / Nancy George
Single-cell RNA sequencing

- Introduced in 2009 on mouse cells
- First used on *Drosophila melanogaster* in 2017
- ~100 fly scRNAseq papers since then (as of July 2022)
  - including the 2022 *Fly Cell Atlas*
  - current pace is ~3–6 papers every month
- Generates huge amount of data... not necessarily easy to exploit
FlyBase – SCEA collaboration on scRNAseq data

- Dataset validation
- Normalised data processing
- Data storage
- Visual dataset exploration

- Dataset discovery
- Validation of cell type annotations
- Data summarisation
The SCEA part
## Import of Fly data from external archives

<table>
<thead>
<tr>
<th>PubMed ID</th>
<th>Curator</th>
<th>Other people comments here</th>
<th>Eligibility</th>
<th>GEO/ENA accession</th>
<th>Citation</th>
<th>Technology</th>
<th>inferred cell types</th>
</tr>
</thead>
<tbody>
<tr>
<td>3235685</td>
<td>AF</td>
<td>potentially eligible</td>
<td>GSE146586</td>
<td>Teitelköt et al., 2020</td>
<td>inDrop v3 or 10xv2</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>3200093</td>
<td>AF</td>
<td>eligible</td>
<td>GSE141273</td>
<td>Cho et al., 2020</td>
<td>Drop-seq</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>3205286</td>
<td>AF</td>
<td>in SCEA already</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32457885</td>
<td></td>
<td>eligible - no links to raw data, emailed 30.04.21</td>
<td>GSE128901</td>
<td>Fu et al., 2020</td>
<td>10xv2</td>
<td>provided by email 30</td>
<td></td>
</tr>
<tr>
<td>30191907</td>
<td></td>
<td>eligible - no need to read file - email to check</td>
<td>GSE131971</td>
<td>Bieda et al., 2020</td>
<td>10xv2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32358140</td>
<td>NG</td>
<td>eligible - short-2 (91 instead of 90)</td>
<td>GSE140340</td>
<td>Juuti et al., 2020</td>
<td>10xv2</td>
<td>email sent 13.06.21</td>
<td></td>
</tr>
<tr>
<td>32185714</td>
<td>NG</td>
<td>eligible</td>
<td>GSE131162</td>
<td>Rust et al., 2020</td>
<td>10xv2</td>
<td>email sent 13.06.21</td>
<td></td>
</tr>
<tr>
<td>32453221</td>
<td>NG</td>
<td>eligible</td>
<td>GSE127832</td>
<td>Bagotz et al., 2020</td>
<td>10xv2, Drop-seq</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32415721</td>
<td></td>
<td>eligible</td>
<td>GSE138626</td>
<td>Zapka et al., 2020</td>
<td>Drop-seq</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32356271</td>
<td></td>
<td>eligible</td>
<td>GSE-ATAB-9444</td>
<td>Zapka et al., 2020</td>
<td>Drop-seq</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31455504</td>
<td></td>
<td>potentially eligible</td>
<td>GSE132324</td>
<td>Deng et al., 2019</td>
<td>10x version?</td>
<td>email to confirm</td>
<td></td>
</tr>
<tr>
<td>31455504</td>
<td></td>
<td>potentially eligible</td>
<td>GSE130568</td>
<td>Deng et al., 2019</td>
<td>10x version?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31740339</td>
<td></td>
<td>eligible - read 17xv20</td>
<td>GSE134722</td>
<td>Juslin et al., 2019</td>
<td>10xv3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29005862</td>
<td></td>
<td>eligible - short-2 (88 instead of 90)</td>
<td>GSE107451</td>
<td>Durie et al., 2018</td>
<td>10xv2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29005862</td>
<td></td>
<td>eligible - short-2 (88 instead of 90)</td>
<td>GSE107451</td>
<td>Durie et al., 2018</td>
<td>10xv2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31851941</td>
<td></td>
<td>potentially eligible</td>
<td>GSE132274</td>
<td>Guo et al., 2019</td>
<td>10x version?</td>
<td>email to confirm</td>
<td></td>
</tr>
<tr>
<td>31915894</td>
<td></td>
<td>not eligible</td>
<td>GSE120537</td>
<td>Hung et al., 2020</td>
<td>inDrop</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26812289</td>
<td></td>
<td>eligible</td>
<td>GSE36025</td>
<td>Kaukinen et al., 2018</td>
<td>Drop-seq</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31455504</td>
<td></td>
<td>potentially eligible</td>
<td>GSE133224</td>
<td>Anwar et al., 2018</td>
<td>10xv2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Considerations:
- **Dataset is public**
- **Raw data is eligible**
- **Allowed technology type**
- **Sufficient metadata**
- **Request inferred cell types**

### Key:
- Completed analysis in SCEA
- Rejected – private or raw data not available
- Pass review and awaiting curation
- Curation review requires more information – awaiting author reply
Incorporating inferred cell types

- Standardised email and inferred cell type format request
- Initial curation of inferred cell type by EBI curator followed by FCA curation review and introduction of new FBbt terms where needed
- Cell barcode and Library run mapping allows visualization of inferred cell type in the Single Cell Expression Atlas knowledgebase
Datasets in Single Cell Expression Atlas – an overview

- Data come from various sources – GEO; ENA, ArrayExpress and subsets of existing datasets
- Might be worth investigating ENA for missed datasets
- Automatic pipeline from FCA identifies Drosophila single cell dataset papers for curation
- Data ingestion year by year shows the increase in Drosophila datasets available
- Total of 28 datasets with 13 including inferred cell types
Standardised analysis pipeline

plate-based studies (SMART protocols)
- FASTX-Toolkit, fastq-utils
- fastq-pair
- kallisto
- emptyDrops DropletUtils
- Raw counts table with all cells
- filterCells
- findMarkers
- findVariable
- Genes
- ScanyPlots

droplet-based studies (10x, Drop-seq)
- Alevin

nextflow
- Read filtering + trimming
- Read-pairing
- Quantification
- Empty droplet removal
- Aggregation of libraries/samples
- Cell filtering
- Normalisation
- Dimension reduction
- Clustering
- Marker detection
Single Cell Analysis Workflows – Galaxy

**scxa-workflows v0.1.0**

Higher level repo for aggregating all of Atlas workflow logic for Single Cell towards execution purposes. Version 0.1.0 was used to run all data analysis for the Release 6 of Single Cell Expression Atlas.

Alignment and quantification workflows are developed in NextFlow and can be found in the `{w_e}_quantification` directories, whereas the clustering and downstream analysis was built on Galaxy and can be found in `{w_e}_clustering` directories. All of the Galaxy tools used here are available from the Galaxy Toolshed to be installed on any instance. The tools are available for direct use as well on the Human Cell Atlas European Galaxy instance.

https://github.com/ebi-gene-expression-group/scxa-workflows/tree/0.1.0
Standardised analysis pipeline for SCEA – Galaxy

- Publicly and freely available under Human Cell Atlas European Galaxy ‘Shared Data’ section
- Updated for each release
- Workflows per release can be imported and edited by users
- Intuitive and easy-to-use
- t-SNE and UMAP visualization via the UCSC browser
- Can analyse either published or personal data using the workflows
Data visualisation in Single Cell Expression Atlas
The FlyBase part
Cell type annotations

What are “cell type annotations”?

- Association { Single cell ID → identified cell type }
- Needed to answer the question “Which cell type(s) is this gene expressed in?”
- Typically not deposited on data repositories alongside the raw sequencing data
  - Sometimes provided as supplementary data: Association { cluster # → identified cell type }
  - Not enough if we don’t also have the association { cell ID → cluster # }
- Typically not using a controlled vocabulary → great variability in the original annotations
  - Use of variable “common” names e.g. “astrocyte” / “astrocyte-like glial cell”
  - Referring to organ/tissues e.g. “dorsal vessel” / “cardial cell”
  - Referring to cell states e.g. “plasmatocyte-prolif”
  - Uncertain identification e.g. “btl-GAL4 positive, likely to be ovary cell”
Validating / “translating” cell type annotations

- FlyBase’s Controlled Vocabularies (CVs)
  - Drosophila Anatomy Ontology (DAO / FBbt)
  - Drosophila Phenotype Ontology (DPO)
  - Drosophila Development Ontology (FBdv)
  - ... 

- Translation of original annotations into terms from the DAO
  - Enriching the DAO in the process if needed

- Translated annotations are sent back to the EBI curators

- Both original and translated annotations are ultimately provided to SCEA users
  - *authors labels | ontology labels*
Post-SCEA processing of scRNAseq data at FlyBase
FlyBase’s aims for scRNAseq data

FlyBase should help drosophilists to:

1. **discover** the available fly scRNAseq datasets
   - What datasets are provided by a given paper?
   - What are all the datasets containing data for a given gene? obtained using a given transgene?
   - What are all the datasets containing data for a given cell type?

2. **get some information** about these datasets
   - How was a dataset generated?
   - Where can the actual data be found?

3. **get a quick overview of the expression data** from these datasets
   - 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 level of expression of the gene across all cells of that type?
Post-SCEA processing

Dataset description
Manually curated from publication

Cell type annotations
Obtained from the authors

Gene expression data
Obtained from the SCEA

Dataset metadata

Summarised expression table
“Summarisation” of expression data

Gene ID × Cell ID matrixes provided by the SCEA:

- “Raw counts” (number of reads aligned to a gene)
- “Normalised counts” (counts per millions of mapped reads, CPMs)

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Cell #1</th>
<th>Cell #2</th>
<th>…</th>
<th>Cell #15999</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBgn00000001</td>
<td>2697.2354</td>
<td>2022.9265</td>
<td>…</td>
<td>674.3088</td>
</tr>
<tr>
<td>FBgn00000002</td>
<td>1348.6177</td>
<td>8766.0151</td>
<td>…</td>
<td>483.5590</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>FBgn0009999</td>
<td>2901.354</td>
<td>1934.2361</td>
<td>…</td>
<td>967.1187</td>
</tr>
</tbody>
</table>
“Summarisation” of expression data

For each couple { Gene, Cell type [as ontological term] }, extract:

- the extent of expression
  - the proportion of cells of that type in which that gene is detected at all
  - = number of cells in the cluster with a non-zero count / total number of cells in the cluster

- the average expression
  - the average CPM in cells of that type that do express that gene
  - = \( \frac{\text{sum}(\text{CPMs})}{\text{number of cells in the cluster with a non-zero count}} \)

To be stored in the FlyBase DB:

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Dataset ID</th>
<th>Cell type</th>
<th>Extent of expression</th>
<th>Average expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBgn0000017</td>
<td>FBlc0012345</td>
<td>plasmatocyte</td>
<td>0.198</td>
<td>484</td>
</tr>
<tr>
<td>FBgn0000017</td>
<td>FBlc0054321</td>
<td>epithelial cell</td>
<td>0.781</td>
<td>527</td>
</tr>
</tbody>
</table>
### What’s visible on the website: Dataset reports

<table>
<thead>
<tr>
<th>Name</th>
<th>FlyBase ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>sciRNAseq_2020_FCA</td>
<td>FBc0003840</td>
</tr>
</tbody>
</table>

#### General summary about the dataset:
- Type of experiment
- Strain
- Tissues of interest
- Experimental conditions
- ... Linkouts to actual data store

#### Linked from:
- Reference report
- Cell type CV term report
What’s visible on the website: The Cell Types Ribbon

- Added to the Gene Report in 2022_03
- Tiles colored based on “extent of expression” of the current gene in the indicated cell types
- Fed from the Fly Cell Atlas dataset (FCA)
What’s next for 2023?

- You decide! FCAG Survey by the end of this year!
- Main proposal: Adding a new graphical display in the “High-Throughput Expression Data” section
  - fed solely from the FCA dataset as the “canonical” dataset
  - proposed designs:
The anatomograms
SCEA and FCA – Anatomograms

- Anatomograms planned:
  - Full fly ‘overview’
    - ovary
    - testis
  - digestive system:
    - foregut
    - midgut
    - hindgut
  - dorsal vessel (including heart)
  - malpighian tubule
  - optic lobe
  - fat body
  - trachea
Anatomograms – Whole fly overview

- Initial 'top level' page for datasets with multiple organs
- An overview of the fly split into two parts
  - Top view
  - Side view with organs
Anatomograms – ovary project

- Ovary anatomogram will be split into (at least) three parts:
  - Ovary
  - Ovariole timeline
  - Germarium ‘zoom in’