

GTCGGCAATCCYTAAGATAGCCAAATATTATTATGTTTCAGATACTCAC  
AGGCAATCCAACTGCAGATCCCACTGGAGTCTTTTGAATCAGTGAATTT  
TAAAGCTTAAAGTAAAGTAAAGTAAAGTAAAGTAAAGTAAAGTAAAGT  
ATTCCTCCGGCAAGCGGACTTCTTGGGATTCGAACTGATCTGAAAGGA  
AATAATAAAATCAACACAGTGCACCAACAGCCGGGGCATCTTCATAGA



**FlyBase** EMBL-EBI



# Exploiting single-cell RNA-sequencing data in FlyBase and the Single-Cell Expression Atlas

Damien Goutte-Gattat / Nancy George

FlyBase-Cambridge / EMBL-EBI

# 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



Twitter post header: [Redacted] · Jul 28



Legit question to all single cell RNA-seq people. So after you published your massive data set of your tissue or organ or organism of choice & we all admired the uMAPs what are the rest of us non RNA-seq people supposed to do w/ it? You did all this awesome work, now what?

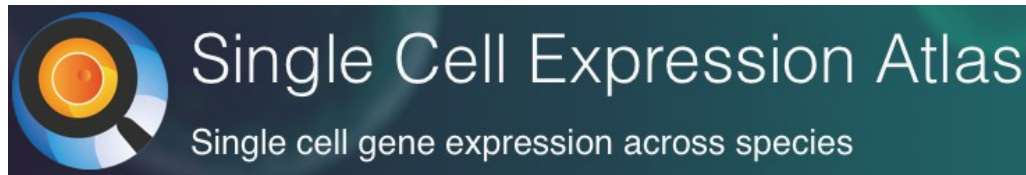
77

78

692



# FlyBase – SCEA collaboration on scRNAseq data



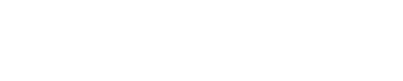
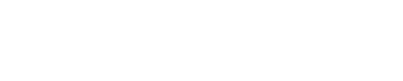
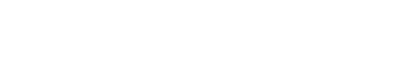
- Dataset validation
- Normalised data processing
- Data storage
- Visual dataset exploration

TACACAATCAGTTAGTTTTCCACCGACAGTCCGCAGAAACCATTTCGACGGC  
GTCGGCAATCCGTAAGATAGCCAAATATTATTATTGTTTCAGATACTCACT  
AGC...CAGCAGTGCAGATCC...TTCAGTGT...CAAATCAGTGAATTC  
...TAAAGCTTCAG...ATTC...G...A  
...ATC...ATCG...AA...G...A...A...A...A  
ATT...CCGGCAAAGCGGACTTTTTCGGAAATGAATGAAAATAAAAAAA  
AATAATAAAAAACAACACAGTGCACACACAGCCGGGGCATCTTCATAGAT  
AACTTCTGCCTGCATTGGTATATGTACTTATCACATAGACATATATATA

- Dataset discovery
- Validation of cell type annotations
- Data summarisation



# The SCEA part



# Import of Fly data from external archives

PubMed ID	Curator	Other peoples comments here	Eligibility	GEO/ENA accession	Citation	Technology	inferred cell types
32396065	AF		partially eligible	GSE146596	Tattikota et al., 2020	inDrops v3 or 10xv2	yes
32900993	AF		eligible	GSE141273	Cho et al., 2020	Drop-seq	yes
32162708	AF		in SCEA already		Cattenoz et al., 2020	10xv2	yes
32487456			eligible - no links to raw data; emailed 30.04.21		Fu et al., 2020	10xv2	provided by email 30
31919193			SRA only links to read2 file - email to check	GSE131971	Staldina et al., 2020	10xv2	
32339165	NG working		eligible - short r2 (91 instead of 98)	GSE146040	Jevitt et al., 2020	10xv2	email sent 18.06.21
33159074	NG working		eligible	GSE136162	Rust et al., 2020	10xv2	email sent 18.06.21
31363221	NG working		eligible	GSE127832	Bageritz et al., 2020	10xv2; Drop-seq	
32815271			eligible	GSE138626	Zappia et al., 2020	Drop-seq	
32815271			eligible	E-MTAB-9444	Zappia et al., 2020	Drop-seq	
31455604			potentially eligible	GSE133204	Deng et al., 2019	10x version? - email to confirm	
31455604			potentially eligible	GSE130566	Deng et al., 2019	10x version?	
31746739			eligible - read1 90bp	GSE134722	Avalos et al., 2019	10xv3	
29909982			eligible - short r2 (88 instead of 98)	GSE107451	Davie et al., 2018	10xv2	
29909982			eligible - short r2 (88 instead of 98)	GSE107451	Davie et al., 2018	Smart-seq2	
31851941			potentially eligible	GSE132274	Guo et al., 2019	10x version? - email to confirm	
31915294			not eligible	GSE120537	Hung et al., 2020	inDrop	
28860209	AF		eligible	GSE95025	Karaiskos et al., 2011	Drop-seq	
30479347			F-MTAB-7194 and F-MTAB-7195	GSE115476	Aries et al., 2018	10xv2	email to request

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



EMBL-EBI



# 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

Nancy George  
 Inferred cell type request template  
 To: Fisher, Malcolm

Sent - ebi.ac.uk 21 June 2021 at 15:04



We would also be grateful if you could assist us further in improving the visualisation of your dataset by including the inferred cell type annotations for your cells if possible. This would simply be a table containing the cell-barcode ID and cell-library mapping along with the inferred cell type for each cell. For an example see below.. We would be happy to accept this as a basic tab or comma separated text file.

cell ID/barcode	Library ID	inferred cell type	cluster
ATCCGACCA	ERR123456	mesophyl	1

I'd be happy to answer any of your questions about inferred cell type. I would be happy to hear any feedback to improve our service.

Dr. Nancy George  
 Bioinformatics  
 Functional Genomics Group  
 EMBL - European Bioinformatics Institute  
 Wellcome Trust Genome Campus, Cambridge UK

Term Type	Term Value	Ontology Class Label	Mapping Confidence	Ontology Class ID	Source
[NO TYPE]	Tm29	transmedullary neuron Tm29	Good	FBbt_00049916	FBbt
[NO TYPE]	Tm3	transmedullary neuron Tm3a	Good	FBbt_00003791	FBbt
[NO TYPE]	distal medullary amacrine neuron	distal medullary amacrine neuron	Good	FBbt_00003769	FBbt
[NO TYPE]	Dm2	Dm2	Good	FBbt_00003769	FBbt
[NO TYPE]	centrifugal neuron C3	centrifugal neuron C3	Good	FBbt_00003744	FBbt
[NO TYPE]	centrifugal neuron C2	centrifugal neuron C2	Good	FBbt_00003743	FBbt

cell barcode	barcode with lane removed	barcode	set	rep	seq	inferred genotype	logged	timestamp	inferred class	subtype	authors cell type	authors cell type - ontology labels
AAACCCAAAGAGCCAA	W1118_24h_A_AAACCCAAAGAGCCAA	W1118_24h_A_4_AAACCCAAAGAGCCAA	W1118	A	4	W1118	24h	neuron	N84	N84	N84	N84
AAACCCAAAGAGTCCG	W1118_24h_A_AAACCCAAAGAGTCCG	W1118_24h_A_3_AAACCCAAAGAGTCCG	W1118	A	3	W1118	24h	neuron	N58	N58	N58	N58
AAACCCAAAGATACAC	DGRP_All_B_8_AAACCCAAAGATACAC	DGRP_All_B_8_AAACCCAAAGATACAC	DGRP	B	8	line_391	36h	neuron	Tm29	Tm29	Tm29	Tm29
AAACCCAAAGACACTG	W1118_24h_A_AAACCCAAAGACACTG	W1118_24h_A_4_AAACCCAAAGACACTG	W1118	A	4	W1118	24h	neuron	N75	N75	N75	N75
AAACCCAAAGACTTTG	DGRP_All_A_6_AAACCCAAAGACTTTG	DGRP_All_A_6_AAACCCAAAGACTTTG	DGRP	A	6	line_508	36h	neuron	Tm3	Tm3	Tm3	Tm3
AAACCCAAAGAAAGT	DGRP_All_B_5_AAACCCAAAGAAAGT	DGRP_All_B_5_AAACCCAAAGAAAGT	DGRP	B	5	line_40	36h	neuron	N39	N39	N39	N39
AAACCCAAAGAGCACTG	W1118_48h_A_AAACCCAAAGAGCACTG	W1118_48h_A_4_AAACCCAAAGAGCACTG	W1118	A	4	W1118	48h	neuron	N155	N155	N155	N155
AAACCCAAAGAGCCGAT	W1118_48h_B_3_AAACCCAAAGAGCCGAT	W1118_48h_B_3_AAACCCAAAGAGCCGAT	W1118	B	3	W1118	48h	neuron	M9	M9	M9	M9
AAACCCAAAGATACAGT	W1118_48h_A_AAACCCAAAGATACAGT	W1118_48h_A_3_AAACCCAAAGATACAGT	W1118	A	3	W1118	48h	neuron	Dm2	Dm2	distal medullary amacrine neuron Dm2	
AAACCCAAAGATACAGT	W1118_48h_B_1_AAACCCAAAGATACAGT	W1118_48h_B_1_AAACCCAAAGATACAGT	W1118	B	1	W1118	48h	neuron	Dm3	Dm3b	Dm3b	Dm3b
AAACCCAAAGATGTTCC	W1118_48h_B_1_AAACCCAAAGATGTTCC	W1118_48h_B_1_AAACCCAAAGATGTTCC	W1118	B	1	W1118	48h	neuron	N143	N143	N143	N143
AAACCCAAAGATTGCA	W1118_Adu1_A_AAACCCAAAGATTGCA	W1118_Adu1_A_1_AAACCCAAAGATTGCA	W1118	A	1	W1118	96h	neuron	N99	N99	N99	N99
AAACCCAAAGATTGGA	DGRP_All_A_7_AAACCCAAAGATTGGA	DGRP_All_A_7_AAACCCAAAGATTGGA	DGRP	A	7	line_505	72h	neuron	C3	C3	centrifugal neuron	Ocentrifugal neuron C3
AAACCCAAAGCAATAGT	W1118_48h_B_1_AAACCCAAAGCAATAGT	W1118_48h_B_1_AAACCCAAAGCAATAGT	W1118	B	1	W1118	48h	neuron	M9	M9	M9	M9
AAACCCAAAGATCCG	DGRP_All_A_2_AAACCCAAAGATCCG	DGRP_All_A_2_AAACCCAAAGATCCG	DGRP	A	2	line_189	24h	neuron	R1.6	R1.6	R1.6	R1.6
AAACCCAAAGCCGACT	DGRP_All_B_3_AAACCCAAAGCCGACT	DGRP_All_B_3_AAACCCAAAGCCGACT	DGRP	B	3	line_320	36h	neuron	T4.15	T5c	T5c	T5c
AAACCCAAAGCCTCAT	DGRP_All_A_5_AAACCCAAAGCCTCAT	DGRP_All_A_5_AAACCCAAAGCCTCAT	DGRP	A	5	line_21	96h	neuron	N79	N79	N79	N79
AAACCCAAAGCGTGAAG	DGRP_All_B_4_AAACCCAAAGCGTGAAG	DGRP_All_B_4_AAACCCAAAGCGTGAAG	DGRP	B	4	line_177	84h	neuron	C2	C2	centrifugal neuron	Ocentrifugal neuron C2
AAACCCAAAGCGAACC	DGRP_All_B_4_AAACCCAAAGCGAACC	DGRP_All_B_4_AAACCCAAAGCGAACC	DGRP	B	4	line_505	48h	neuron	T2	T2	T2	T2
AAACCCAAAGCGAAGT	DGRP_All_B_4_AAACCCAAAGCGAAGT	DGRP_All_B_4_AAACCCAAAGCGAAGT	DGRP	B	4	line_307	48h	neuron	N161	N161	N161	N161
AAACCCAAAGCGACTG	DGRP_All_B_3_AAACCCAAAGCGACTG	DGRP_All_B_3_AAACCCAAAGCGACTG	DGRP	B	3	line_348	48h	neuron	T4.15	T4b	T4b	T4b
AAACCCAAAGCGATGGT	DGRP_All_A_8_AAACCCAAAGCGATGGT	DGRP_All_A_8_AAACCCAAAGCGATGGT	DGRP	A	8	line_307	96h	neuron	Tm9	Tm9b	Tm9b	Tm9b
AAACCCAAAGCGCTG	DGRP_All_A_3_AAACCCAAAGCGCTG	DGRP_All_A_3_AAACCCAAAGCGCTG	DGRP	A	3	line_748	96h	neuron	R7.8	R7.8	R7.8	R7.8
AAACCCAAAGCGGTAT	DGRP_All_B_1_AAACCCAAAGCGGTAT	DGRP_All_B_1_AAACCCAAAGCGGTAT	DGRP	B	1	line_391	36h	neuron	N76	N76	N76	N76
AAACCCAAAGCGTATGG	W1118_48h_B_2_AAACCCAAAGCGTATGG	W1118_48h_B_2_AAACCCAAAGCGTATGG	W1118	B	2	W1118	48h	neuron	Dm9	Dm9	Dm9	Dm9

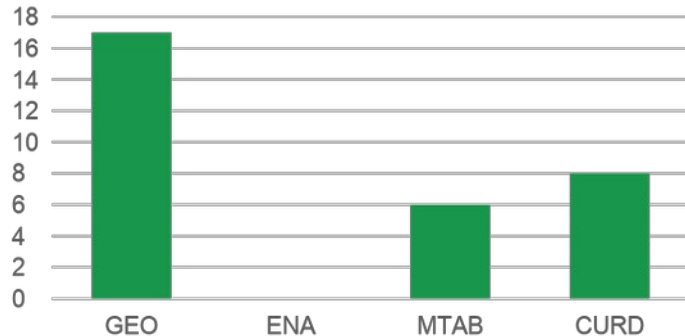


EMBL-EBI

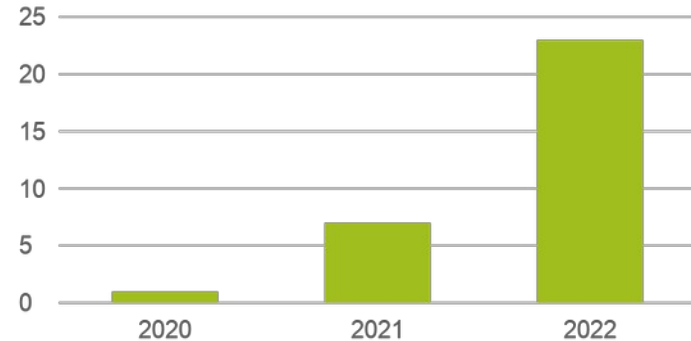
# Datasets in Single Cell Expression Atlas – an overview

- 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 or year shows the increase in Drosophila datasets available
- Total of 28 datasets with 13 including inferred cell types

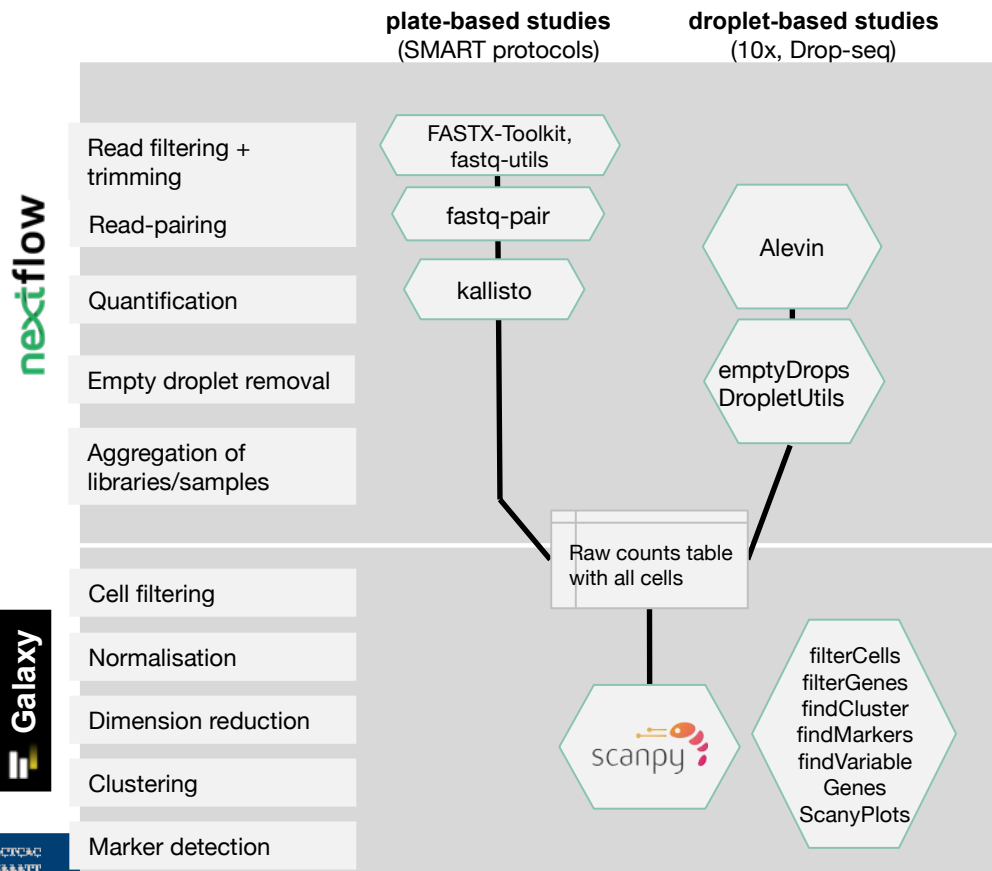
Number of Drosophila datasets



Data ingestion by Year

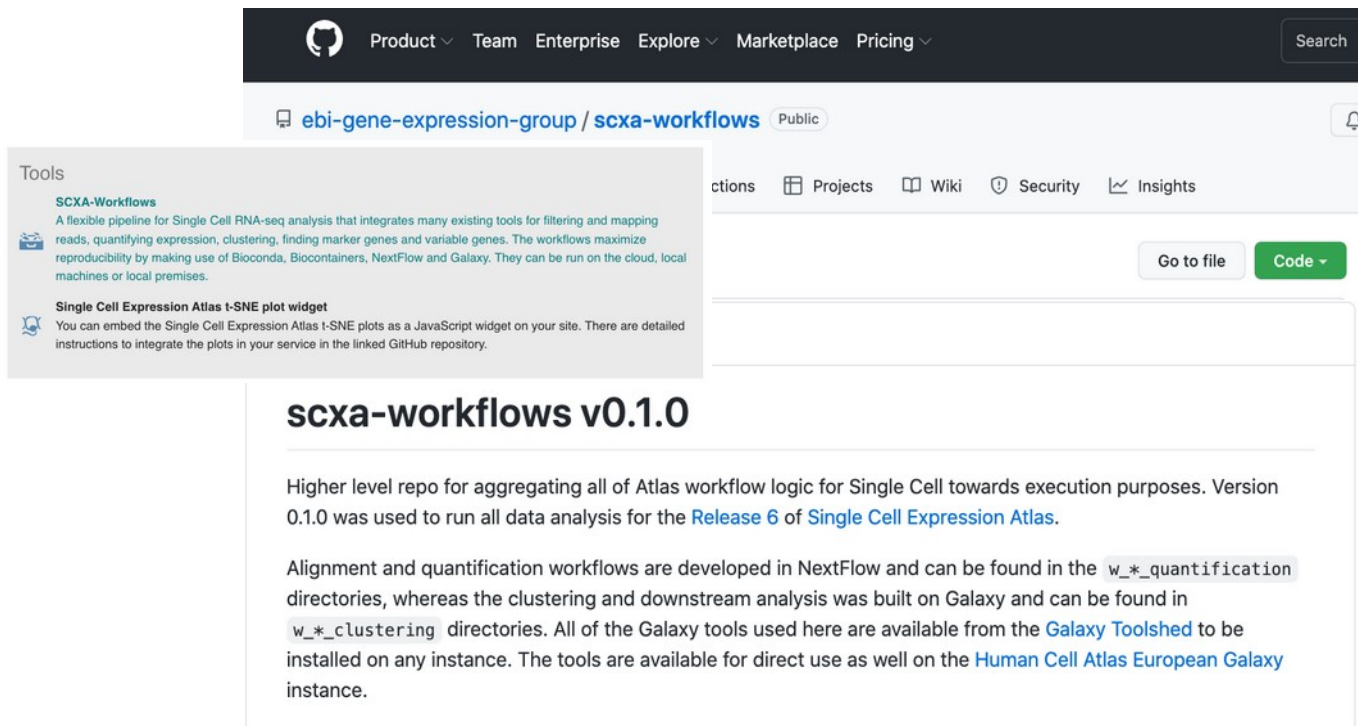


# Standardised analysis pipeline





# Single Cell Analysis Workflows – Galaxy



The screenshot shows a GitHub repository page for 'ebi-gene-expression-group / scxa-workflows'. The repository is public. A 'Tools' sidebar on the left lists two tools: 'SCXA-Workflows' and 'Single Cell Expression Atlas t-SNE plot widget'. The main content area displays the repository name and version 'scxa-workflows v0.1.0'. Below the title, there is a description: '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.' Further down, it details the workflow structure: 'Alignment and quantification workflows are developed in NextFlow and can be found in the w\_\*\_quantification directories, whereas the clustering and downstream analysis was built on Galaxy and can be found in w\_\*\_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.'

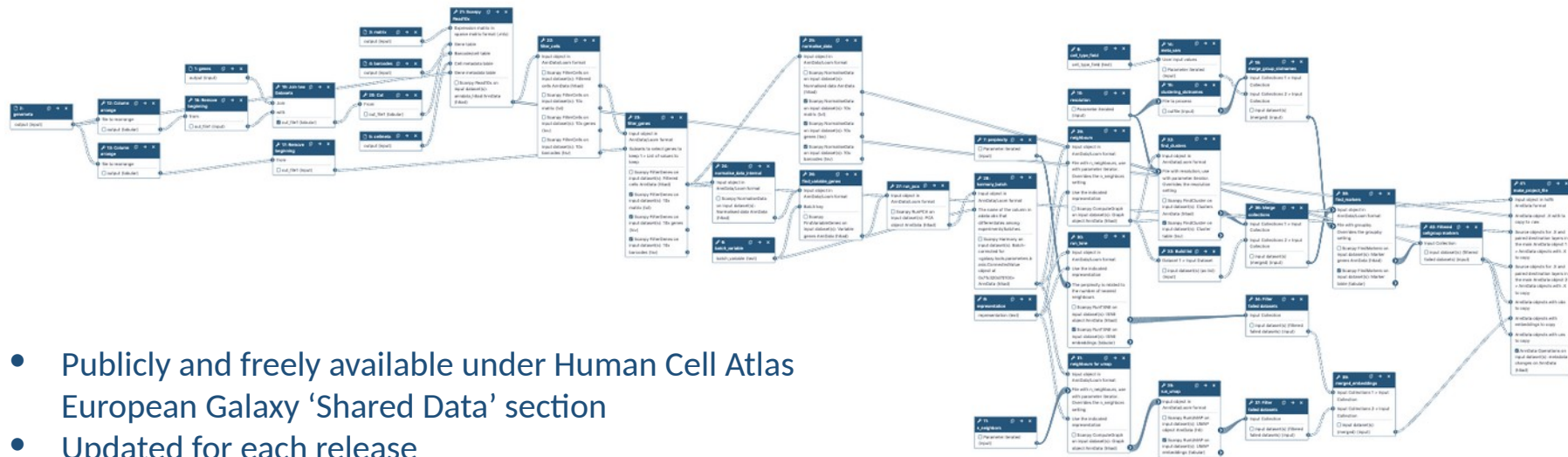
<https://github.com/ebi-gene-expression-group/scxa-workflows/tree/0.1.0>



EMBL-EBI



# 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



EMBL-EBI



# Data visualisation in Single Cell Expression Atlas

## Fly Cell Atlas: single-cell transcriptomes of the entire adult Drosophila - 10x dataset

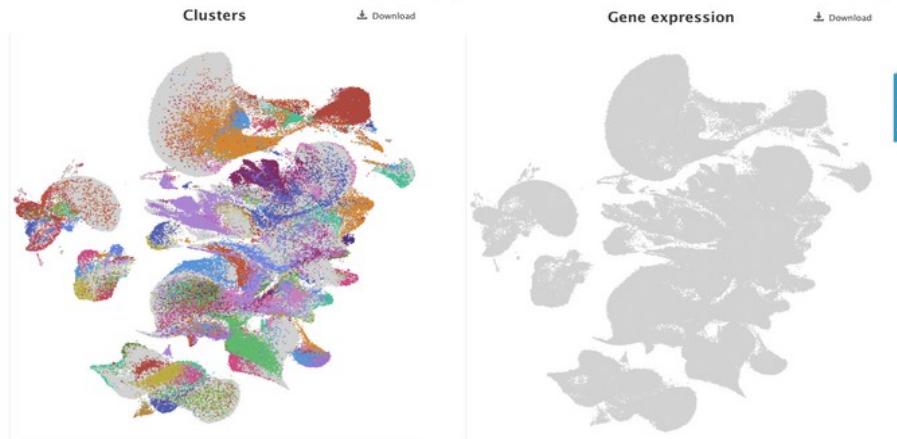
Single-cell RNA-Seq mRNA baseline

Number of cells: 527,530

Organism: *Drosophila melanogaster*

Publication:

• Li H, Janssens J, De Waegeneer M, Kolluru SS, Davie K et al. (2021) *Fly Cell Atlas: a single-cell transcriptomic atlas of the adult fruit fly*



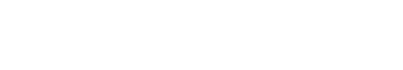
Load date	Q	*drosophila melar	Q	Title	Q	Experimental facts	Number of cells	Download
04-08-2022		Drosophila melanogaster		Adult Phenotypes and Gene Expression in the Brain at Single Cell Resolution After Developmental Alcohol Exposure in Drosophila		<ul style="list-style-type: none"><li>sex</li><li>compound</li><li>dose</li><li>inferred cell type - authors labels</li><li>inferred cell type - ontology labels</li></ul>	88165	<input type="checkbox"/>
04-08-2022		Drosophila melanogaster		Age-related changes in polycomb gene regulation disrupt lineage fidelity in intestinal stem cells		<ul style="list-style-type: none"><li>age</li><li>inferred cell type - ontology labels</li></ul>	24565	<input type="checkbox"/>
04-08-2022		Drosophila melanogaster		The Drosophila Brain on Cocaine at Single Cell Resolution		<ul style="list-style-type: none"><li>sex</li><li>compound</li><li>dose</li><li>inferred cell type - authors labels</li><li>inferred cell type - ontology labels</li></ul>	84603	<input type="checkbox"/>



EMBL-EBI



# The FlyBase part



# Cell type annotations

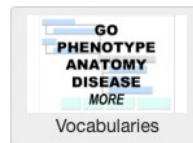
What are “cell type annotations”?

- 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. “plasmacyte-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)
  - ...



General Information			
Term	astrocyte-like glial cell	ID (Ontology)	FBbt:00100505 (Fly Anatomy)
Definition	Neuropil associated glial cell of the central nervous system that has a dendritic morphology and elaborates inside the associated synaptic neuropil (Awasaki et al., 2008). Their nuclei are found at synaptic neuropil surfaces and they extend branched filiform or lamelliform processes that pervade the neuropil (Hartenstein, 2011).[ FlyBase:FBfr0206543 FlyBase:FBfr0207750 FlyBase:FBfr0214261 FlyBase:FBfr0225669 ]		
Also Known As	"ALG" ; "astrocyte" ; "astrocyte-like neuropil associated glial cell" (for all, see Synonyms field below)		

- 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

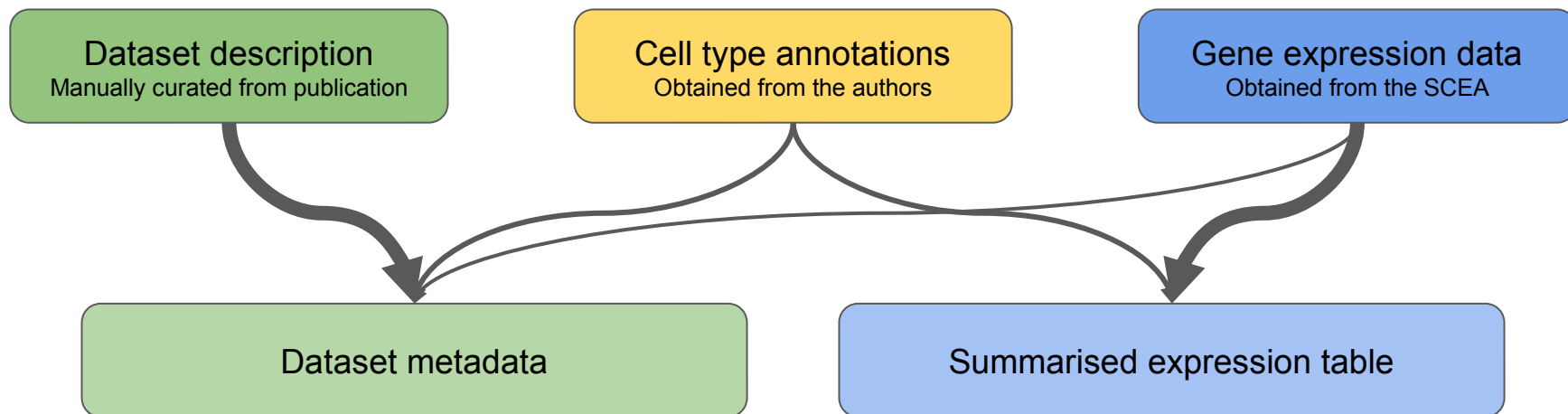
FlyBase should help drosophilists to:

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



# “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)

	Cell #1	Cell #2	...	Cell #15999
<b>FBgn0000001</b>	2697.2354	2022.9265	...	674.3088
<b>FBgn0000002</b>	1348.6177	8766.0151	...	483.5590
...	...	...	...	...
<b>FBgn0009999</b>	2901.354	1934.2361	...	967.1187



# “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
  - = sum(CPMs) / number of cells in the cluster with a non-zero count

To be stored in the FlyBase DB:

Gene ID	Dataset ID	Cell type	Extent of expression	Average expression
FBgn0000017	FBlc0012345	plasmatocyte	0.198	484
FBgn0000017	FBlc0054321	epithelial cell	0.781	527

# What's visible on the website: Dataset reports

Name	scRNAseq_2022_FCA	Species	<i>D. melanogaster</i>
Project type	transcriptome	FlyBase ID	FBic0003845
Parent Project		Data Provider	The Fly Cell Atlas Consortium
Title	Fly Cell Atlas: Single-nucleus RNA-seq study of the entire adult fly		
Accessions	E-MTAB-10519    E-MTAB-10628		
<b>Overview</b>			
Description	A characterization of the diverse populations of cells in the entire body, head, and 15 dissected tissues of the adult <i>Drosophila melanogaster</i> .		(LJ et al., 2022)
Project attributes	tissue type study		(LJ et al., 2022)
Biosample type			
Assay type			
Reagent type			
Result type			
Key genes			
GO term(s)			
SO term(s)			
<b>Details</b>			
Sample preparation			
Protocol			
Mode of Assay			
Data analysis	Reads were processed using the Cell Ranger 3.1.0 software. They were aligned to the <i>Drosophila melanogaster</i> reference genome 6.31. Counts were corrected for the presence of contaminating ambient RNA using DecontX (R package Ceida 1.4.5). Cells with less than 200 genes, less than 500 counts, or more than 5% of mitochondrial genes were excluded. Following dimensionality reduction and clustering, cell types were identified in annotation jamborees by tissue experts and annotations were gathered on the SCoPe platform.		(LJ et al., 2022)
Comments			
Files			
Additional Information	The EMBL-EBI's Single Cell Expression Atlas provides cell-level annotations, clustering data, raw and normalised read counts, and putative marker genes.		(LJ et al., 2022)
<b>Related Datasets</b>			
Component projects (3) <a href="#">Export to HTML</a>			
<b>Project</b>	<b>Type</b>	<b>Title</b>	
scRNAseq_2022_FCA_FEMALE	transcriptome	Single-nucleus RNA-seq on cells from 5-days old female flies	
scRNAseq_2022_FCA_MALE	transcriptome	Single-nucleus RNA-seq on cells from 5-days old male flies	
scRNAseq_2022_FCA_MIXED	transcriptome	Single-nucleus RNA-seq on cells from 5-days old female and male flies	

General summary about the dataset:

- Type of experiment
- Strain
- Tissues of interest
- Experimental conditions
- ...
- Linkouts to actual data store

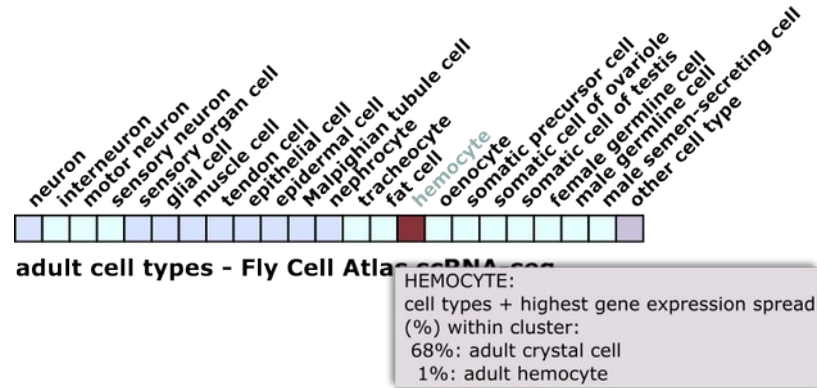
Linked from:

- Reference report
- Cell type CV term report



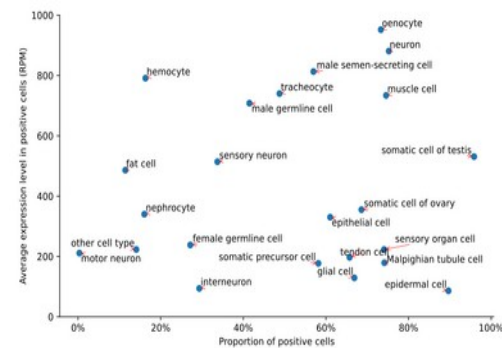
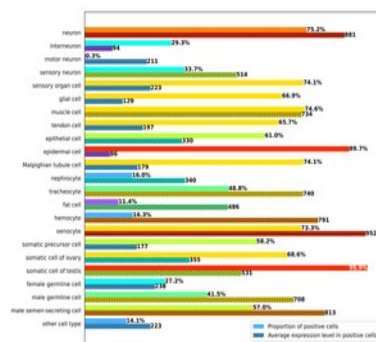
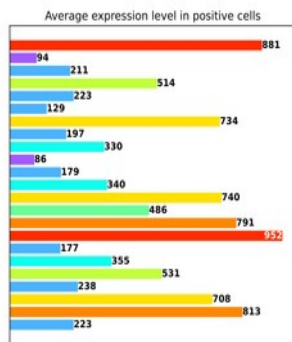
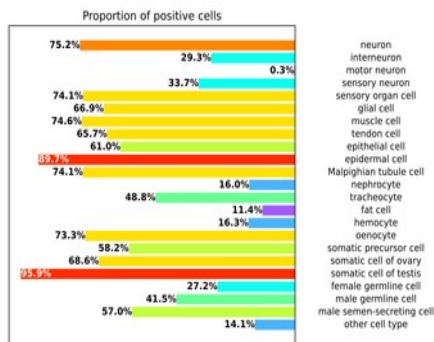
# What's visible on the website: The Cell Types Ribbon

- 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:

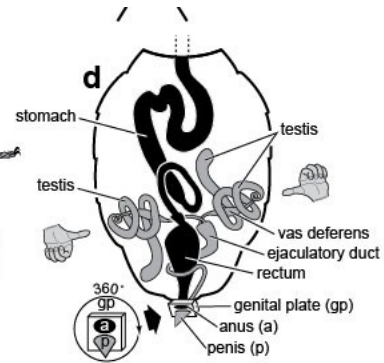
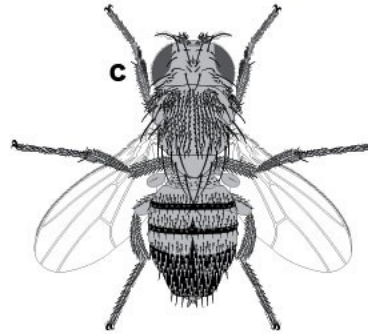


# The anatomograms



# SCEA and FCA – Anatomomograms

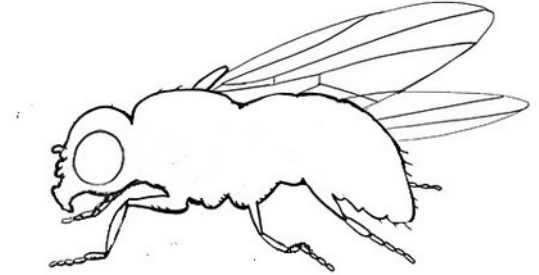
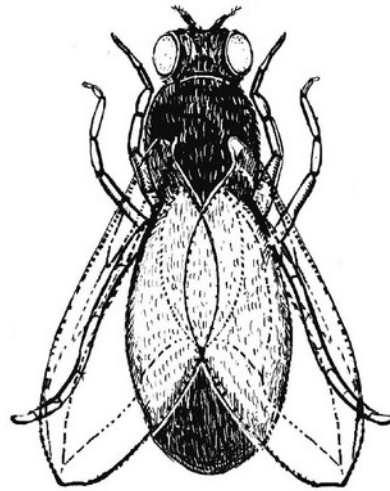
- Anatomomograms planned:
  - Full fly 'overview'
    - ovary
    - testis
    - digestive system:
      - foregut
      - midgut
      - hindgut
    - dorsal vessel (including heart)
    - malpighian tubule
    - optic lobe
    - fat body
    - trachea





# 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'

