



# Analysis of scRNA-Seq from *Drosophila* developmental tissues

**Sílvia Pérez Lluch**

Master in Biostatistics and Bioinformatics

Work area 1

**Enrique Blanco García**

**Javier Luis Cánovas Izquierdo**

June 4th, 2019







Aquesta obra està subjecta a una llicència de [Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya de Creative Commons](https://creativecommons.org/licenses/by-nc-nd/3.0/es/)

## **B) GNU Free Documentation License (GNU FDL)**

Copyright © 2019 SÍLVIA PÉREZ LLUCH.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.

A copy of the license is included in the section entitled "GNU Free Documentation License".

## **C) Copyright**

© (Sílvia Pérez Lluch)

Reservats tots els drets. Està prohibit la reproducció total o parcial d'aquesta obra per qualsevol mitjà o procediment, compresos la impressió, la reprografia, el microfilm, el tractament informàtic o qualsevol altre sistema, així com la distribució d'exemplars mitjançant lloguer i préstec, sense l'autorització escrita de l'autor o dels límits que autoritzi la Llei de Propietat Intel·lectual.

## FINAL REPORT FILE

<b>Title of the Work:</b>	<i>Analysis of scRNA-Seq from Drosophila developmental tissues</i>
<b>Author name:</b>	<i>Silvia Pérez Lluch</i>
<b>Tutor name:</b>	<i>Enrique Blanco García</i>
<b>Area Responsible Teacher:</b>	<i>Javier Luis Cánovas Izquierdo</i>
<b>Date of delivery (mm/aaaa):</b>	<i>06/2019</i>
<b>Program:</b>	<i>Master in Biostatistics and Bioinformatics</i>
<b>Work Area:</b>	<i>1</i>
<b>Language:</b>	<i>English</i>
<b>Key words:</b>	<i>Computational Biology, scRNA-Seq, Drosophila melanogaster</i>
<b>Abstract (in English, 250 words or less):</b>	
<p>Organisms are formed by a huge variety of cell types. Conventional genomics approaches do not allow for the understanding of the diversity of cell populations forming tissues and organs. Recently, and with the aim of scrutinizing the cellular complexity of organisms, single-cell technologies have arisen, challenging the limitations of former methods. In this project, we have analyzed single-cell transcriptomics data generated in different laboratories and using a manifold of technologies. We first explored scRNA-Seq data produced by two independent groups on <i>Drosophila melanogaster</i> brain cells. Data from these reference datasets were generated using four different technologies, allowing for the assessment and correction of putative batch effects caused by the multiple platforms used. By implementing a number of pipelines in bash and R environments, we have processed these datasets from the mapping of raw fastq files to the identification of subpopulations of cells. Cells from the 8 identified clusters expressed several known marker genes involved in neuron and glia differentiation, permitting the full characterization of these populations. The pipeline implemented along this first part of the project was, afterwards, used to analyze scRNA-Seq data generated in our own lab on <i>Drosophila</i> wing imaginal discs. Within our dataset, we distinguished four clusters, although the low expression of known marker genes did not allow for a precise characterization of these populations. Still, we were able to identify several markers showing high variability between clusters, indicating that, indeed, these subpopulations represent different cell types within the wing.</p>	

# Index

<b>1</b>	<b>Introduction.....</b>	<b>4</b>
1.1.	<b>Context and justification of the Master Thesis .....</b>	<b>4</b>
1.1.1.	State of the art.....	4
1.1.2.	Open questions.....	5
1.2.	<b>Objectives of the Master Thesis .....</b>	<b>6</b>
1.3.	<b>Possible approaches and followed method .....</b>	<b>7</b>
1.4.	<b>Thesis planning.....</b>	<b>9</b>
1.4.1.	Resources used for the progression of the project .....	9
1.4.2.	Tasks to be accomplished along the project.....	9
1.4.3.	Task timings and milestones.....	13
1.4.4.	Declaration of risks and contingency plan.....	14
1.5.	<b>Brief summary of obtained products.....</b>	<b>15</b>
<b>2</b>	<b>Analyses on available <i>Drosophila</i> brain scRNA-Seq data.....</b>	<b>17</b>
2.1.	<b>Description of selected reference datasets .....</b>	<b>17</b>
2.2.	<b>Pre-processing of reference datasets .....</b>	<b>18</b>
2.3.	<b>Mapping and quantification of reference datasets.....</b>	<b>19</b>
2.4.	<b>Filtering of low-quality cells and normalization.....</b>	<b>22</b>
2.5.	<b>Visualization of RNA-Seq data by dimensionality reduction methods .....</b>	<b>24</b>
2.5.1.	Identification of highly variable genes.....	24
2.5.2.	Characterization of highly variable genes.....	25
2.5.3.	Visualization of scRNA-Seq data .....	26
2.6.	<b>Batch effect removal .....</b>	<b>28</b>
2.7.	<b>Visualization of known marker genes .....</b>	<b>30</b>
2.8.	<b>Identification of clusters and marker genes .....</b>	<b>32</b>
2.9.	<b>Characterization of identified clusters .....</b>	<b>34</b>
2.10.	<b>Discussion of results .....</b>	<b>37</b>
<b>3</b>	<b>Analysis of MARS-Seq scRNA-Seq data from <i>Drosophila</i> wing imaginal discs....</b>	<b>40</b>
3.1.	<b>Description of the dataset .....</b>	<b>40</b>
3.2.	<b>Pre-processing, mapping, quantification, filtering and normalization of the dataset .</b>	<b>41</b>
3.3.	<b>Visualization by dimensionality reduction methods.....</b>	<b>43</b>
3.3.1.	Identification and characterization of highly variable genes.....	43
3.3.2.	Visualization of scRNA-Seq data .....	43
3.4.	<b>Batch effect removal .....</b>	<b>44</b>
3.5.	<b>Visualization of known marker genes .....</b>	<b>45</b>
3.6.	<b>Identification and characterization of clusters and marker genes.....</b>	<b>46</b>
3.7.	<b>Discussion of the results .....</b>	<b>47</b>
<b>4</b>	<b>Conclusions.....</b>	<b>49</b>
<b>5</b>	<b>Glossary.....</b>	<b>51</b>
<b>6</b>	<b>Bibliography .....</b>	<b>52</b>
<b>7</b>	<b>Appendix .....</b>	<b>55</b>

## Figure index

Figure 1. Pipeline implemented along the progression of the current project ...	13
Figure 2. Gantt chart representing the main tasks of the current project and the proposed timings for each task .....	14
Figure 3. Example of a read from a <i>bam</i> file .....	20
Figure 4. Counts and genes summary statistics on the reference datasets.....	21
Figure 5. Filtering of cells and genes. ....	23
Figure 6. Highly variable genes .....	25
Figure 7. GO Term Enrichment of highly variable genes .....	26
Figure 8. Visualization of scRNA-Seq reference data set by dimensionality reduction methods .....	27
Figure 9. Visualization of scRNA-Seq after CCA correction.....	29
Figure 10. Representation of known neural marker genes .....	31
Figure 11. Representation of known marker genes from non-neural cells.....	32
Figure 12. Visualization of clusters identified by Shared-Nearest-Neighbor .....	33
Figure 13. Expression of highest enriched markers within clusters .....	34
Figure 14. GO Term Enrichment analysis of cluster marker genes .....	35
Figure 15. Characterization of cell clusters .....	36
Figure 16. Mapping and quantification statistics of MARS-Seq scRNA-Seq from fruit fly wing imaginal discs .....	42
Figure 17. Highly variable genes from MARS-Seq.....	43
Figure 18. PCA analysis of wing scRNA-Seq data.....	43
Figure 19. Batch effect correction by CCA .....	44
Figure 20. Visualization of known marker genes.....	45
Figure 21. Clusters and top marker genes of wing scRNA-Seq dataset.....	46
Figure 22. Gene Ontology Term Enrichment of MARS-Seq clusters .....	47

## Table index

Table 1. Description of available reference datasets used along the current project.....	17
Table 2. Number of cells identified per cluster with the different parameters used. .....	33

# 1 Introduction

## 1.1. Context and justification of the Master Thesis

### 1.1.1. State of the art

The relationship between the genotype and the phenotype is one of the eldest and still unsolved questions of Molecular Biology. The RNA is the first representation of the cellular phenotype. The specific set of genes expressed in the cell, the level of their expression and the usage of different transcript isoforms are some of the features that identify a particular cell type. Thus, the precise description of the transcriptome of each cell type is essential to understand the role of the cell within a tissue or an organ.

The development of the Next Generation Sequencing (NGS) technologies, and in particular the implementation of techniques such as the RNA-Seq (massive parallel sequencing of bulk RNA), has allowed for an exponential increase in the sensitivity of the detection of gene expression as well as for the identification of novel genes (Wang et al., 2009). Low expressed genes, unknown transcripts and even transcripts generated from regulatory regions, such as enhancer RNAs (eRNAs), are newly reported thanks to the application of the RNA-Seq technologies to different cell types (Li et al., 2016; Pundhir et al., 2015). In this context, the need of deciphering the specific transcriptome of a particular cell type has become a priority.

During the last decade, RNA-Seq has become the universal tool to uncover the gene expression profile of specific tissues at different developmental time points, but also to identify changes at the transcriptomic level when tissues undergo other processes, such as cancer or degeneration (Wang et al., 2009). However, these experiments have usually been performed in tissues or culture cell lines. Tissues are formed by many distinct cell types, meaning that the transcriptome obtained by bulk RNA-Seq experiments represents an average of the different cell populations conforming the tissue, masking the expression of low represented or rare cell types. On the other hand, culture cell lines, which are

supposed to be homogeneous, have been shown to represent heterogeneous populations with different transcriptomic states (Barron and Li, 2016).

For all these reasons, the need of uncovering the transcriptome at a single-cell level has become a challenge. During the last few years, this ambition has become a reality due to the emergence of single-cell RNA-Seq technologies (scRNA-Seq). By quantifying specific-cell transcriptomes we will understand the heterogeneity existing within a population, uncover differences between cell populations upon disease and even identify specific regulatory networks within each cell type. None of them can be fully understood from bulk RNA-Seq experiments. The need of identifying cell-specific transcription patterns has become so urgent that even huge consortia, such as the Human Cell Atlas (HCA) (Ponting, 2019), have been funded to perform massive scRNA-Seq experiments during the next few years.

#### 1.1.2. Open questions

Although the scRNA-Seq is already a reality, its technology is still in development. scRNA-Seq experiments are, for instance, carried out using many different platforms, such as MARS-Seq, 10X Chromium, Smart-Seq and CEL-Seq, that generate very different outputs. None of these platforms allows, however, for a sensitive and accurate detection of all genes expressed in the cell, meaning that the transcriptome obtained for each single-cell in the experiment represents a subset of the real number of genes that are being expressed (See et al., 2018).

In this context, the new challenge the scientific community is facing is the development of computational tools able to analyze, normalize and integrate all the generated data. On the one side, 10X Genomics technology detection, for instance, is biased towards the 3'/5' ends of transcripts; however, it allows for the removal of redundant reads due to PCR amplifications during the library preparation by the addition of Unique Molecular Identifiers (UMIs). On the other side, Smart-Seq detects full-length transcripts, allowing for the quantification of alternative splicing and intron retention events, for instance, but it cannot discard redundancy due to PCR amplifications, making expression level quantifications less accurate (See et al., 2018). Data obtained from each type of scRNA-Seq assay should be, thus, analyzed in a specific and suitable manner, which already renders into a first challenge, but the results obtained have to be comparable to



the results from the other platforms. This represents a second challenge for the Computational Biologists, as the biases derived from the different experimental protocols generate a batch effect that needs to be corrected before the comparison between samples. New pipelines have been generated aiming to diminish this batch effect and making datasets more comparable (Stuart and Satija, 2019). The third challenge the scientific community is facing is the development of new dimensionality reduction and projection methods to cluster the different cell types and to uncover true relationships between cells and even identify previously unknown cell types. These new methods will permit the integration of the transcriptomic profile of different cells within a samples or from different sources (Kiselev et al., 2019). Finally, the fourth and last challenge that needs to be addressed is the integration of scRNA-Seq data with other single-cell or bulk datasets, such as haplotype, GWAS data, DNA methylation, chromatin accessibility, histone modifications or three-dimensional chromatin structure. This integration will provide for a further understanding of the mechanisms the cell undergoes upon processes such as differentiation and development.

Although intensive research is currently being carried out in this field, the analysis of scRNA-Seq is still an open question, as no standard protocols exist that allow for the normalization and integration of data sets from different sources. However, the fast development of single-cell technologies and the big amount of data generated will enhance the efforts of the research community in order to develop new and more powerful tools to analyze and integrate these data.

## 1.2. Objectives of the Master Thesis

The main objectives of the current proposal are:

1. To fully analyze scRNA-Seq datasets generated from different platforms and in various laboratories, from the mapping to the identification and characterization of cell populations.
2. To benchmark currently available pipelines or libraries generated to perform single/multiple stages of the analysis of scRNA-Seq data.
3. To build our own pipeline with a selection of such tools in order to analyze the scRNA-Seq experiments generated in our laboratory.

The specific objectives are:

1. To collect publicly available scRNA-Seq raw data generated from different platforms, such as 10X, CEL-Seq2 and Smart-Seq2, to define our reference dataset.
2. To map on the fruit fly reference genome and quantify gene expression from the different samples of scRNA-Seq in our reference dataset.
3. To evaluate the impact of several parameters in the analysis (how many reads, how many genes are detected per cell, etc.).
4. To benchmark different available pipelines to integrate the data from the several technologies and remove the batch effect associated to technical issues.
5. To cluster the cells according to their specific transcriptomic signature and identify cell populations and unknown cell types by means of different dimensional reduction methods.
6. To identify marker genes within clusters and perform GO analysis to confirm the biological identity of each group of cells passing the previous steps in our reference dataset.
7. To compare the results obtained with the reference dataset with those published in the original references to evaluate strengths and weaknesses of each method.
8. To gather all tested programs in a single script to allow for faster analysis and parallelization of the process.
9. To use our own pipeline to analyze the scRNA-Seq experiments from *Drosophila melanogaster* wing imaginal disc performed in our lab.

### 1.3. Possible approaches and followed method

The current project is divided into two main areas: first, we aim to learn about the different platforms available to perform single-cell RNA-Seq assays (scRNA-Seq) and to study the output files produced by such platforms; second, we will benchmark the most popular implemented pipelines to perform a thorough analysis of a selection of public scRNA-Seq datasets obtained from different sources.

In our lab, we are interested in the mechanisms involved in cell differentiation and in the determination of cell fate during development. In particular, we are working

in the differentiation of *Drosophila melanogaster* imaginal discs throughout development. Imaginal discs are epithelial sacs in larval stages that give rise to the adult appendages, such as wings, legs and eyes (Ruiz-Losada et al., 2018). In this context, we have recently generated scRNA-Seq data from third instar larvae wing imaginal discs from the fruit fly. Our main objective with the current project is, thus, to find the best tools to analyze our single-cell data and to identify cell populations within the wing disc. To do so, we will, first, evaluate many available pipelines focused on the analysis of scRNA-Seq data and, second, implement a R script gathering all tested tools. As our data belongs to fruit fly cells, we will implement these new tools using also available data from *Drosophila*, mainly from brain tissues. The resulting pipeline will be used to analyze our scRNA-Seq data and compare it to other similar datasets, such as the data obtained also from wing imaginal disc in Aurelio Teleman's lab (Bageritz et al. 2018).

The different datasets have been selected to fulfill several purposes:

1. They all have been performed in *Drosophila* tissues, which is our model organism in the lab. The analysis of these data will help us to identify and select the tools better performing in our dataset of interest.
2. They have been generated using different experimental protocols and laboratories, meaning that they may be subject to a high batch effect. The normalization of all these datasets will allow us to integrate and visualize data from all sources. This expertise will also help us in the integration of our own dataset with similar data from other groups.

The analysis of scRNA-Seq is complex and is still under development, meaning that the scientific community is investing a lot of time and resources to face this new challenge. Taking advantage of all the tools that are being developed, we plan to assess the performance of the most popular ones in several datasets and to gather them in a single pipeline.

The pipelines selected to perform the current project have been selected also according to different criteria:

1. They have been already reported by other groups for analysing scRNA-Seq experiments, proving their suitability for this kind of analysis.

2. Most of them run under R environment, so they are appropriate to be integrated in our final script.
3. Some of them can be used to approach similar questions in different ways, such as the batch effect removal from *RunCCA* and *mnnCorrect*, whereas others are complementary and perform slightly different analyses, such as the removal of cell cycle effect of *ccRemover*.

In some steps of the analysis, such as the mapping of the sequencing reads, only one program is going to be used, in this case the STAR pipeline. This is because these programs already represent a standard in the field and have proven to perform in a very efficient manner. In other steps, such as the removal of the batch effect, many programs that pretend to solve the problem have recently arisen, but the question is still open, so the assess of each's one performance may be essential for the success of the analysis.

The final decision of which pipelines are suitable for being integrated within our R script will depend also on various aspects:

1. The usability of the pipeline, this is, if the pipeline can run under our environment and if it runs in a user-friendly manner.
2. The performance with our reference dataset.

The functionality and the performance of the resulting script will be finally assessed through the analysis of our own dataset from fruit fly wing imaginal disc scRNA-Seq.

#### 1.4. Thesis planning

##### 1.4.1. Resources used for the progression of the project

For the development of the current project we will mainly work under the framework of bash and Rstudio.

To perform the scRNA-Seq analyses a set of pipelines and R libraries have been selected Seq and are susceptible of being included in the R script developed during this project.

##### 1.4.2. Tasks to be accomplished along the project

**Objective 1: Task1.** Collection of the reference datasets of scRNA-Seq from *Drosophila* tissues

Briefly, scRNA-Seq data from full *Drosophila melanogaster* (the fruit fly) brains or selected regions, such as midbrain or specific neurons, will be obtained from public repositories, such as NCBI GEO or EBI ArrayExpress. These datasets have been generated in independent labs and using 4 different platforms and represent a broad example of all data types that can be generated using scRNA-Seq technologies (Davie et al., 2018; Li et al., 2017). For instance, 10X and CEL-Seq2 contain Unique Molecular Identifiers (UMIs) and, thus, are 3' biased but PCR-amplification duplicates can be removed, whereas Smart-Seq2 and AdaptedSmart-Seq2 produce full-length transcripts but do not contain UMI information, meaning that PCR duplicates will not be distinguishable, but more reads covering alternative splicing and different transcript isoform usage will be available.

These particular datasets have been finally selected because they belong to isolated cells from *Drosophila melanogaster* (the fruit fly) tissues, and this will allow us to compare them with scRNA-Seq dataset generated in our lab, performed by MARS-Seq from around 350 isolated cells from third instar larvae wing imaginal discs. The first task of the current project, and with the aim of accomplishing the first objective, will be, thus, obtaining the raw data from the specified repositories.

### **Objective 2: Task 2.** Mapping and quantification of scRNA-Seq data

The second task of the work will be to map and quantify the gene expression per cell. To do so, reads will need to be first preprocessed according to their origin (presence or absence of UMIs, demultiplexed cells, etc.). Mapping will be performed with *STAR* (Dobin et al., 2013), a broadly used program that allows for the mapping of continuous and split reads. Split reads correspond to reads that do not map directly to the genome but to split sequences into the transcriptome, such as splice junctions.

To perform the quantification of gene expression the tool *featureCounts*, embedded into the *subread* package (Liao et al., 2014), will be employed.

### **Objective 3: Task 3.** Quality control of the different datasets

The third task of the work will be to perform a quality control (QC) of the different datasets obtained. The QC will be performed using two different programs: the

QC tool of the *Seurat* package (Butler et al., 2018) and the *scater* program (McCarthy et al., 2017). These packages will inform about the number of genes detected per cell, the number of UMIs, etc. If needed, UMI-tools can be also considered for the removal of PCR duplicates at this step, although it only runs under command line (Smith et al., 2017). This QC step will also allow us to filter out the cells presenting high level of contamination of mitochondrial RNAs, typical of low-quality libraries, or cells with low coverage, for instance. Both *Seurat* and *scater* perform other analyses, such as normalization, so, eventually, we can also perform fulfill other tasks with these tools.

This first part of the project corresponds to the main block of the project according to the working plan and is expected to be accomplished within the first 5 weeks (see Gantt chart in Figure 2).

**Objective 4: Task 4.** Normalization and batch effect removal

The fourth task of the project will be the normalization of the diverse datasets, taking only the cells that have passed the previous step. In this step, batch effects due to the performance of the experiments using different experimental protocols or in different labs will be removed. To do so, different pipelines will be evaluated: *RunCCA*, from the *Seurat* package (Butler et al., 2018), and *mnnCorrect* (Haghverdi et al., 2018). *RunCCA* runs a canonical correlation analysis to identify similar cell types between datasets and cluster them closely in a two-dimensional space. Instead, *mnnCorrect* computes the distance between similar cell types through mutual nearest neighbor and allows for the removal of these distances of the original expression matrices.

After normalization, new expression matrices will be generated. To validate that the normalization has reduced the multiple putative biases, we will use dimensionality reduction techniques, that will also allow us to visualize the distance between different cell types. Three different methods to reduce the dimensions of the scRNA-Seq expression matrices will be used: Principal Component Analysis (PCA) (Venables et al., 2002), t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten et al., 2008) and Uniform Manifold Approximation and Projection for R (UMAPR) (Becht et al., 2018).

**Objective 5: Task 5.** Clustering of the cells according to their expression profile

The fifth task of the project will be to cluster the cells from all experiments according to their gene expression patterns. The expression of particular known markers will allow us to classify different cell populations within our reference dataset as well as, eventually, identify unknown cell types. To perform these analyses, we will use several tools: the Clustering through Imputation and Dimensionality Reduction (*CIDR*) tool allows for the imputation of missing data within the scRNA-Seq matrices and the classification of cell types through dimensionality reduction methods (Lin et al., 2017); the *quickCluster* tool from *scran* (Lun et al., 2016) and *FindClusters* from *Seurat* (Butler et al., 2018) will also be evaluated.

**Objective 6: Task 6.** Identification of marker genes within clusters

The identity of each cluster will be first assessed by the identification of the marker genes representative of each population of cells. This task will be performed by, first, visualizing known marker genes in the dimensionality reduction plots and, second, by automatic detection of highly variable genes between clusters, task that will be assessed with the *FindMarkers* tool of *Seurat* (Butler et al., 2018).

**Objective 6: Task 7.** Gene Ontology term enrichment analysis of different cell clusters

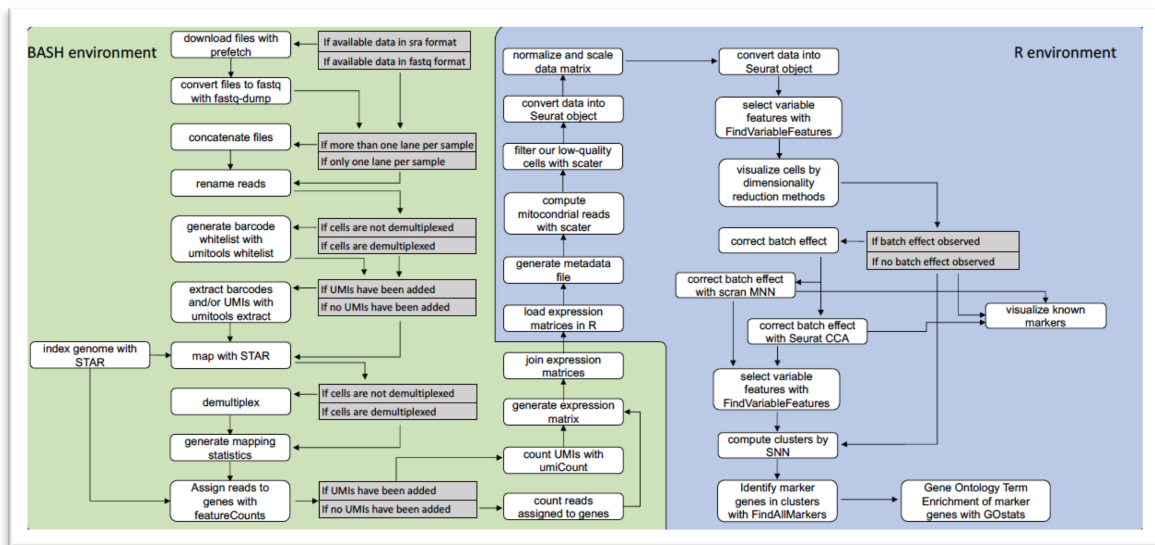
The seventh task of the project will be to perform Gene Ontology term enrichment analyses of the different clusters identified according to the marker genes. To do so, we will use the *GOstats* package (Falcon and Gentleman, 2007), that allows for the identification of Gene Ontology and KEGG pathway terms enriched within each cell cluster.

**Objective 7: Task 8.** Comparison of obtained results with originally published ones

At this step, we will have performed all the basic analyses of our reference dataset. Thus, the next task of the project will be to compare the results obtained with the data originally published in the different referenced papers. This second block is expected to be fulfilled within weeks 5 to 9.

**Objective 8: Task 9.** Implementation of a script to perform scRNA-Seq analysis on our own datasets

The first part of the pipeline will be implemented in bash, as most of the programs that perform preprocessing, mapping and quantification of raw data run in this environment. Instead, the evaluated pipelines in R will be finally gathered in a single R script to allow for a faster analysis of future datasets. Figure 1 depicts a scheme of the pipeline implemented along the progression of the current project (see also appendix, all sections). The implementation of the pipeline will be performed along the progression of the full project.



**Figure 1.** Pipeline implemented along the progression of the current project.

**Objective 9: Task 10.** Analysis of our own scRNA-Seq data from *Drosophila* wing discs using our new pipeline

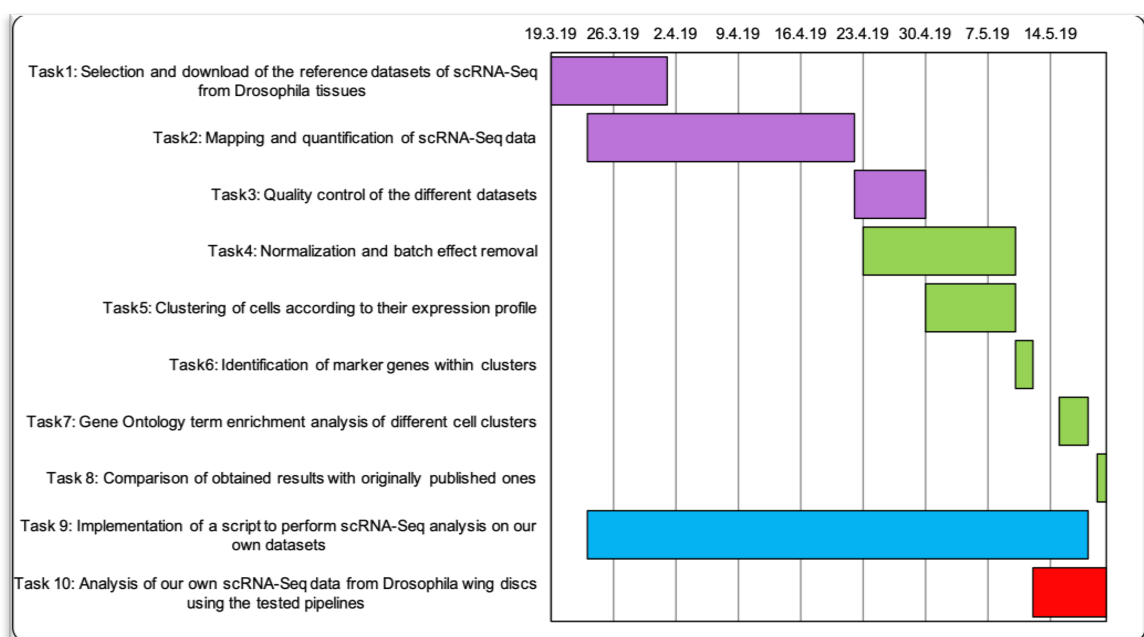
The last task of the project will be the usage of our bash and R pipelines for the analysis of single-cell RNA-Seq data generated in our lab (unpublished data) on isolated cells from third instar larvae imaginal disc of *Drosophila melanogaster*. We will finally compare our data with data from Teleman's lab also in the wing disc (Bageritz BioRxiv 2018). This last part of the project will be accomplished during the last two weeks of the project.

1.4.3. Task timings and milestones

The Figure 2 represents the Gantt chart describing the 10 tasks proposed for the current project as well as the timings proposed for each task. Colors represent the expected milestones. The first milestone, in purple, involves obtaining the data of reference datasets in Table 1 and the basic analyses of mapping,



quantifying and quality controls. The second milestone, in green, consists on the normalization of the data, the clustering of the different cell types and the identification of marker genes, altogether with the characterization of the different clusters through GO term enrichment analyses and the comparison with the original reports. The third milestone, in blue, involves the implementation of the bash and R scripts gathering all tested pipelines. Finally, the fourth milestone, in red, consists on the usage of the selected pipelines in our scripts to analyze the scRNA-Seq data generated in our lab and its comparison to the results obtained from other scRNA-Seq data of wing imaginal disc, from Teleman's lab.



**Figure 2.** Gantt chart representing the main tasks of the current project and the proposed timings for each task.

#### 1.4.4. Declaration of risks and contingency plan

The current proposal is ambitious and implies a certain risk. The major risk we can envision is the inability of running certain scripts. Noteworthy, many of the pipelines described in the previous sections are currently being developed, and it may happen that they do not work under our environmental conditions. As a contingency plan, we may try to, in parallel, run pipelines developed to thoroughly analyze scRNA-Seq data, from the mapping to the clustering of cell types, such as *Seurat*.

The second major risk we can envision is the time and the computational resources needed to process such large amount of data. To solve this problem, we will implement a job array to run scripts that need to be parallelized, for instance, for the mapping of demultiplexed files or for quantification for each independent cell.

#### 1.5. Brief summary of obtained products

During the progression of the current project we will generate different products:

- Work plan. Along the work plan we have indicated the main tasks to be accomplished along the progression of the current project in relationship to the corresponding objectives. Timings and milestones have been also defined (Figure 2).

- Memory. All the work performed along the project will be reported in a written memory, specifying the state of the art in the field, the objectives, the methodology, the results obtained and a general discussion about the integration of our results in a broader context.

- Product: In our particular case, we will also implement a R script gathering all suitable programmes tested along the project. The pipeline code will be included in the appendix of the written memory.

- Virtual presentation: At the end of the master thesis we will also summarize the methodology followed and the results obtained with the publicly available datasets along the project as well as with our own data set.

- Project self-evaluation: Finally, we will present a report including all the information about the difficulties found along the project and what could have been done to improve it.

#### 1.6. Brief summary of the following chapters

In chapter number 2 the analysis performed on available brain samples will be described. This analysis follows the script found in the appendix section *Brain*. A part from the plots presented in the main report, alternative plots and more details and analysis can be found in the appendix. Besides, alternative clustering

pipelines used to analyze the brain reference dataset can be found in appendix sections *CIDR* and *scran*.

Chapter 3 will be devoted to the analyses of MARS-Seq scRNA-Seq generated in our laboratory. The script used to generate this data, including further details and analyses, will be found in appendix section *MARS*. Both chapters 2 and 3 will include a last discussion section, where obtained results will be discussed and compared in the context of the state of the art on the field and the recent publication publications on similar datasets.

For analyses performed in chapter 2 and 3, the series of bash scripts implemented to preprocess, map and quantify the dataset will also be available in the appendix section *Bash\_scripts*.

Chapter 4 will correspond to the conclusions of the project. Final chapters of the report will include the glossary, the references section and the appendix.

## 2 Analyses on available *Drosophila* brain scRNA-Seq data

### 2.1. Description of selected reference datasets

With the final objective of implementing a suite of pipelines to analyze our own scRNA-Seq experiments on *Drosophila melanogaster* third instar larvae wing imaginal discs, we have selected a set of recently published high-quality scRNA-Seq reference datasets from fruit fly brain (Table 1).

GEO accession	platform	UMIs	Tissue	genotype	Age	# cells	Publication
GSE107451	10X Chromium	Yes	Brain	w <sup>1118</sup>	adult 3 days	6,257	Davie et al. 2018.
GSE107451	10X Chromium	Yes	Brain	DGRP-551	adult 3 days	3,157	Davie et al. 2018.
GSE107451	10X Chromium	Yes	Brain	DGRP-551	adult 50 days	3,098	Davie et al. 2018.
GSE107451	Smart-Seq2	No	dorsal fan-shaped body neurons	R23E10>GFP	Adult	45	Davie et al. 2018.
GSE107451	adapted Smart-Seq2	No	dorsal fan-shaped body neurons	R23E10>GFP	Adult	34	Davie et al. 2018.
GSE107451	CEL-Seq2	Yes	dorsal fan-shaped body neurons	R23E10>GFP	Adult	22	Davie et al. 2018.
GSE100058	Smart-Seq2	No	projection neurons	GH146>GFP	Pupa	200	Li et al. 2017.

**Table 1.** Description of available reference datasets used along the current project.

The first selected dataset has been obtained from Li and collaborators paper on 1046 GH146>mCD8GFP labeled projection neurons of fruit fly pupal brains, generated by using an adaptation of Smart-Seq2 protocol (Li et al., 2017). Cells sequenced with this technology do not contain UMI information, and they have been loaded into the GEO database in a demultiplexed manner (cell by cell). This dataset is rather exhaustive in terms of sequencing coverage (more than 1 million reads per cell, according to the authors). Due to the large size of the raw data files, the computational processing of the whole dataset is extremely time-consuming. Because of this reason, and as the objective of the current project is to compare different scRNA-Seq platforms and pipelines, not the exhaustive analysis of the reference datasets, we have randomly selected the first 200 cells loaded into the database for our analysis.

The selected second dataset has been generated by Davie and collaborators, also on different populations of *Drosophila* brain cells (Davie et al., 2018). In this case, up to four different technologies have been utilized to produce the scRNA-Seq datasets (Table 1). First, data were generated on around 29,000 DGRP-551 and 28,000 *w<sup>1118</sup>* brain cells by 10X Chromium technology. These two fly strains correspond to wild type animals in order to identify possible batch effects. To generate these data, brains from the two strains were obtained at different times after metamorphosis, to identify changes at transcriptomic level during aging. Due to the large size of this dataset, we have selected data from the two wild type strains at two different time points: 3 and 50 days after metamorphosis, with the aim of identifying differences in cell populations due to aging (removal of specific cell types, etc.). 10X technologies allow for the addition of UMIs, and data has been loaded into the GEO database in a multiplexed manner. Next, and with the goal of identifying rare cell types, CEL-Seq2 and two different protocols of Smart-Seq2 (Smart-Seq2 and AdaptedSmart-Seq2 from now onwards) have been applied to obtain RNA-Seq data on 22, 45 and 34 R23E10>GFP marked cells, respectively. This marker is specific of a subset of dorsal fan-shaped body neurons (dFB), a very rare cell type in fly brains (only 9 cells per hemisphere). From these three methods, only CEL-Seq2 cells contain UMIs. According to the authors, all cells were sequenced at a coverage of around 50,000 reads per cell, this is, 20 times less than in the previous dataset.

Selected datasets were obtained from GEO database (Barrett et al., 2005) in SRA format. Files were converted to raw fastq files by using the *fastq-dump* tool from the *sra-toolkit* (<http://ncbi.github.io/sra-tools/>).

## 2.2. Pre-processing of reference datasets

Prior to the mapping step, raw data has been pre-processed in different ways depending on the type of data provided by the authors.

Datasets generated with Smart-Seq2 and AdaptedSmart-Seq2 are already demultiplexed, and as no UMIs are included, they can be mapped directly without any pre-processing step.

For datasets containing UMIs (10X and CEL-Seq2), two fastq files are provided: one containing the sequencing read and the other one informing about the cell

barcode (BC) and the UMI attached to each molecule. The BC will allow for afterwards demultiplexing of cells, and UMIs will inform about PCR duplicates generated during the library preparation. To mark each sequencing read with the corresponding BC and UMI, a whitelist containing all BCs used in each run has to be generated. Whitelists are generated using the *whitelist* tool from *umitools* (Smith et al., 2017). BCs from 10X and CEL-Seq2 were defined by using the `--extract-method=regex` option, which allows for the identification of BC/UMI patterns with regular expressions. By using the *regex* option, for instance, BCs and UMIs can be identified by using the pattern `--bc-pattern="(P<cell_1>.{12})(P<umi_1>.{6})"`. This example depicts a read where the first 12 nucleotides represent the BC and the following 6 nucleotides represent the UMI. This is, actually, the BC pattern used in CEL-Seq2 samples, whereas 10X cells are marked by BC of 16 and UMIs of 10 nucleotides. To generate the whitelist, the number of BCs present in the sample can be automatically inferred from the data or, alternatively, can be given to the program as long as an estimate of the number of cells processed within the experiment is available. We have taken advantage of this second option, and the *whitelist* tool has identified and corrected the BCs according to the number provided in each case.

In a second step, the whitelist is used to extract the BCs and the UMIs from the reads in which they are contained to be incorporated into the sequencing read identifier. This step is performed by the tool *extract* from *umitools*. In this way, after mapping and quantifying the reads, we will be able to demultiplex the cells and identify duplicates.

### 2.3. Mapping and quantification of reference datasets

Mapping scRNA-Seq datasets has been performed with *STAR* (Dobin et al., 2013), a Gold Standard to map RNA-Seq reads. Prior to mapping, *STAR* has to be used to index the reference genome. In our case, we have downloaded latest release (95) of *Drosophila melanogaster* genome and annotation from Ensembl (Cunningham et al., 2019).

Mapping with *STAR* can be run under many different options. Here we have applied the same parameters as in Li et. al paper (Li et al., 2017), this is `--outFilterScoreMinOverLread 0.4` (alignment will be output only if its score,

normalized by read length, is higher than this value) -  
 outFilterMatchNminOverLread 0.4 (alignment will be output only if the number of  
 matched bases, normalized by read length, is higher than this value) -  
 outFilterMismatchNmax 999 (alignment will be output only if it has fewer  
 mismatches than this value) -outFilterMismatchNoverLmax 0.04 (alignment will  
 be output only if its ratio of mismatches to \*mapped\* length is less than this  
 value)". The mapping output file is a *bam* file (<http://samtools.github.io/hts-specs/>), consisting on the read identifier and the coordinates of the read within  
 the reference genome (figure 3).

```
SRR5685313.17906.1      83   2R   17564146   255   8S142M   =
      17564146   -142
      AGAGACAGCGGACGAACCGAGAATTTATGGATGTATAAACAAAATACGAGAAACCGTCTTCACC
TAAAACGCCTGCATATGTGTGTATGAATATCGTGTATTTTGCTAATCCTGTAGCTCTTTATTGGAGAATT
TAGCAAAGCCCCATCC  HGF@ACC/CGGGFEGGC/HFHFFF?EBBGGF
```

**Figure 3.** Example of a read from a *bam* file.

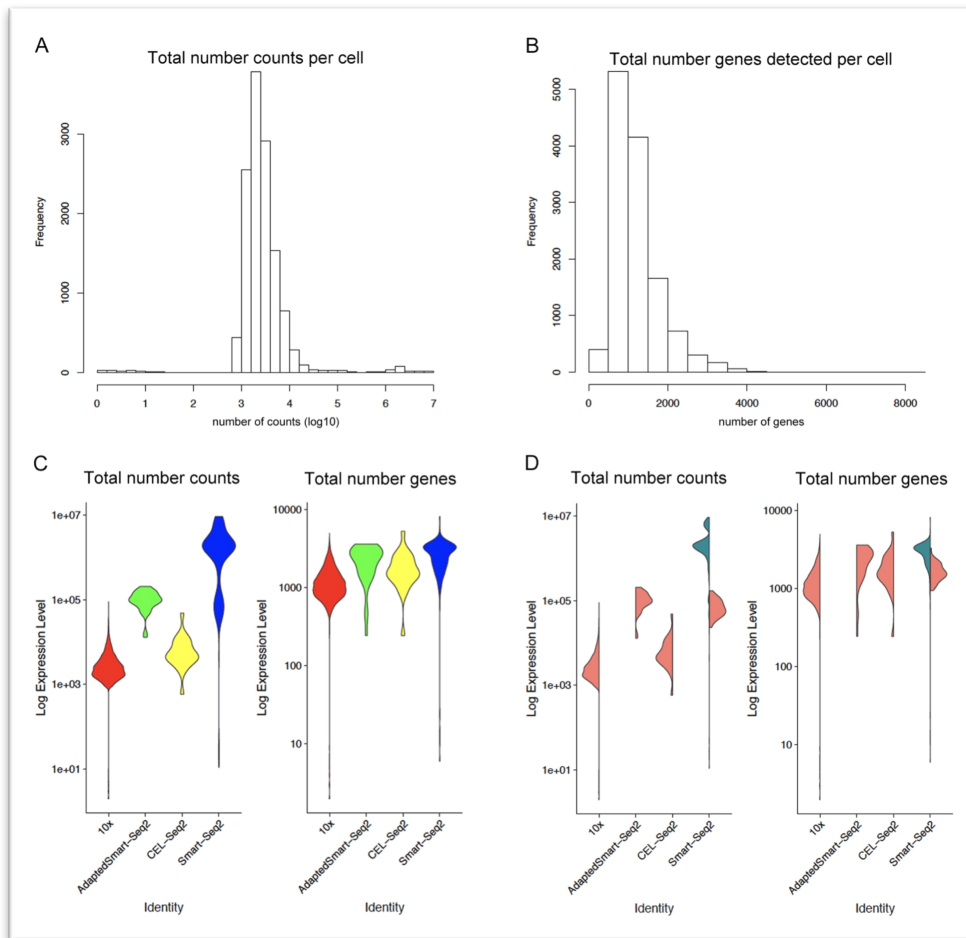
To assign mapped reads to genes, we used *featureCounts*, from the *subread* package, providing also the annotation of the reference genome in GTF format. Thus, this tool assigns each mapped read to a gene and attaches this information to the mapping file, tagging the read as assigned or not and, if so, attaching the name of the corresponding gene. When running *featureCounts*, another file, called *gene\_assigned*, is generated, in which the name of each feature (e. g. gene) is stored with the total number of reads assigned to it.

So far, cells containing and no containing UMIs have been processed in parallel. However, the objective of selecting diverse platforms is, in part, to assess the differences derived from the removal of PCR duplicates, usually thought to generate a bias in gene expression quantification, through their identification by the addition of UMIs. Thus, to quantify how many reads have been assigned at each gene per cell, we have followed different strategies according to the presence or not of UMIs.

For technologies not containing UMIs, such as Smart-Seq2 and AdaptedSmart-Seq2, counts assigned to reads and stored in the *gene\_assigned* file are

recovered and saved in an expression matrix where all genes of the genome (17,739 in our annotation) are represented.

For UMI-containing genes, *umiCount* tool from *umitools* was used to identify and isolate non-duplicated reads. A *counts* file is generated with all genes, all cell BCs and the number of deduplicated reads associated to each gene per each BC. At this step, BCs can be demultiplexed and expression matrices for each different cell can be generated with the number of counts per gene.



**Figure 4.** Counts and genes summary statistics on the reference datasets. **A.** Frequency of total number of counts per cell (log scaled). **B.** Frequency of total number of detected genes per cell. **C.** Left panel, total number of counts per cell represented by platform. Right panel, total number of genes per cell represented by platform. **D.** Left panel, total number of counts per cell represented by platform and split by paper. Right panel, total number of genes per cell represented by platform and split by paper.



We finally gathered all expression matrices in a single one, generating a big expression matrix of 17,739 rows (genes) and 12,799 columns (cells).

Figure 4 depicts summary statistics on total number of counts and number of genes identified per cell. As seen in figure 4A, most of cells present between 1,000 and 10,000 mapped reads, although some cells reach more than 1 million reads (e.g. cells from Li's Smart-Seq2). Figure 4B shows that, although the range of mapped reads is very high, in most of cases, only between 500 and 2,000 genes are identified per cell.

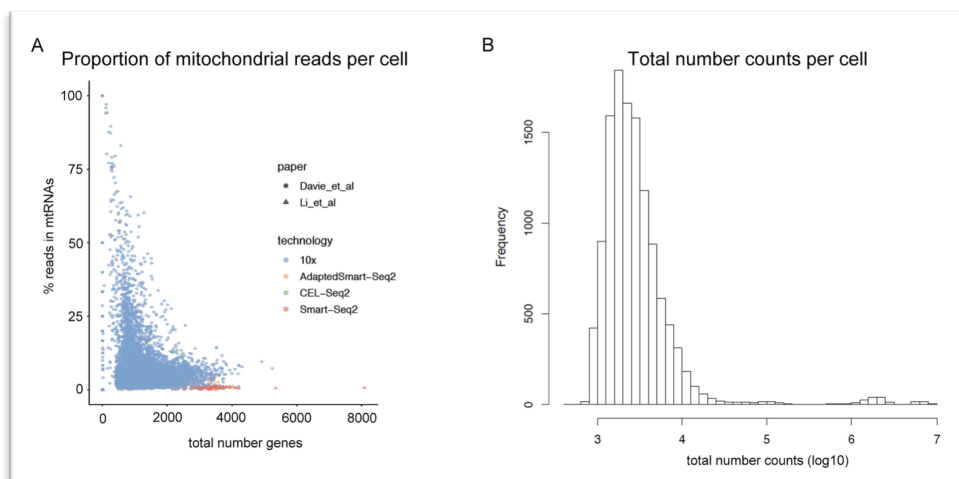
To distinguish which platform detects more features per cell, we have plotted total number of counts and genes per cell grouping the cells per technology, figure 4C. Smart-Seq2 cells present the highest number of total reads, likely due to the higher sequencing coverage of the 200 selected cells from Li et al. As expected, 10X and CEL-Seq2 cells show a much lower number of counts, because sequencing coverage was much lower and, besides, only non-duplicated reads are represented. Smart-Seq2 is also the platform that detects more genes per cell, although closely followed by AdaptedSmart-Seq2 and CEL-Seq2, with a much lower coverage. Finally, in figure 4D, we have plotted again total number of counts and genes but splitting the datasets by publication. Smart-Seq2 is the only technology that has been used in both works, in the 200 cells from Li et al. and in 44 cells in Davie et al. The plots confirm that Li's cells (in green), present higher number of reads than Davie's Smart-Seq2 cells (in pink). However, the number of detected genes per cell is not that much different between the two datasets.

#### 2.4. Filtering of low-quality cells and normalization

As seen in figure 4, there are many cells that present very few reads. This is mainly due to the sequencing of low-quality libraries. Another marker of low-quality cells is the proportion of mapped reads falling into mitochondrial genes (mtRNAs). These low-quality cells have to be removed before going on with the analysis, as they could bias the final results.

To do so, first we have generated a single-cell experiment with *scater*, an R package specifically designed to perform quality control and normalisation of

single-cell RNA-Seq assays (McCarthy et al., 2017). A *scater* single-cell experiment is a large data frame containing the expression matrix and all metadata associated to the dataset. In our case, the metadata informs about the technology used to perform the scRNA-Seq, the genotype and the age of the animal and the paper they belong to. We have first removed cells with low number of reads, this is, with more than 3 median absolute deviations below the median log-library size or with less than 200 detected genes. Next, we have eliminated genes not detected in at least 10 cells. Finally, we have removed cells where more than 20% of reads fell into mtRNAs. We have obtained a list of 38 mtRNAs from Ensembl annotation and have annotated them as spike-ins in the *scater* experiment with the *isSpike* tool. This spike-in control gets integrated into the single-cell experiment. In figure 5A, the proportion of reads assigned to mitochondrial genes per cell is represented as a scatter plot. Interestingly, in cells where more features are identified, less reads fall over mitochondrial genes, whereas in cells where less than 1,000 genes are detected most of reads are captured by the mtRNAs, confirming the hypothesis that high-quality libraries are void of mitochondrial reads.



**Figure 5.** Filtering of cells and genes. **A.** Proportion of reads in mitochondrial genes by sequencing platform. Most of cells containing a high proportion of reads in mtRNAs belong to 10X technology. **B.** Number of counts per cell after cell filtering.

After filtering, we have ended up with 12,086 cells, most of them showing more than 1,000 reads (figure 5B). From the 713 filtered cells, 605 have been removed by their proportion of mitochondrial genes, most of them obtained by 10X

Chromium technology. Finally, from the 17,739 initial genes, only 10,747 are kept after filtering and will contribute to the subsequent analyses.

Last step prior to data visualization and analysis is the normalization. Data from the *scater* experiment is retrieved and a single-cell experiment from *Seurat* R package is generated. It was developed by Satija lab and it is specifically designed to perform all possible steps on single-cell RNA-Seq analyses (Butler et al., 2018). Thus, counts were log normalized and scaled by using *Seurat* tools, and normalized/scaled data was stored within the *Seurat* object.

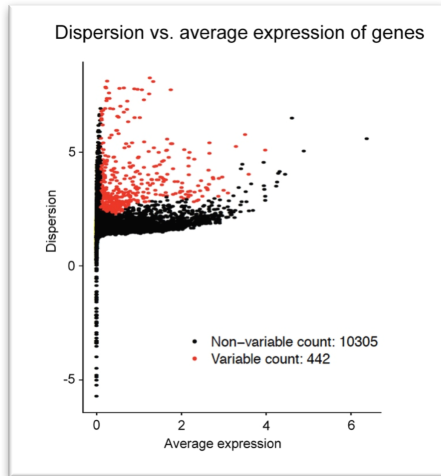
## 2.5. Visualization of RNA-Seq data by dimensionality reduction methods

Dimensionality reduction methods are a way of reducing the variance explained by thousands of variables, such as all genes in the genome, in a lower number of features, facilitating their visualization in a two or a three-dimensional plot. Common reduction methods used to visualize scRNA-Seq data are Principal Component Analysis (PCA) (Jolliffe, 2002), t-distributed Stochastic Neighbor Embedding (tSNE) (Hinton, 2008) and Uniform Manifold Approximation and Projection (UMAP) (Becht et al., 2018). PCA is based on the identification of linear combinations of the original values, through the computation of their eigenvectors and eigenvalues. Distances between samples in a PCA plot represent real distances existing among them. tSNE and UMAP are machine learning algorithms that work under PCA. Thus, they first compute also the principal components, but they perform a non-linear analysis to identify the relationship between the neighboring points. With this approximation, both tSNE and UMAP allow for a representation of the data in a clustered manner, meaning that closest samples are represented closer and forming clusters in the plot. However, distances between samples and, especially between clusters, may not be real, as they depend on the stochasticity of the hyperparameters selected.

### 2.5.1. Identification of highly variable genes

With the objective of better understanding the cell populations on the reference datasets, we have started by computing and visualizing the scRNA-Seq with all three methods. PCA, tSNE and UMAP can be run with all cells from the dataset, but they may be biased by the mild contribution of non-variable genes in the dataset. Thus, and similarly to the approach usually followed to analyze bulk

RNA-Seq data, we have first identified the higher variable genes between samples. *Seurat* has a particular tool that performs this calculation on normalized/scaled data (*FindVariableFeatures*). There are several methods to



compute variable features in *Seurat*, but we have selected the *mean.var.plot*, which calculates dispersion while controls for the relationship between variability and mean expression of features (figure 6).

By using this method, we have identified 442 variable genes.

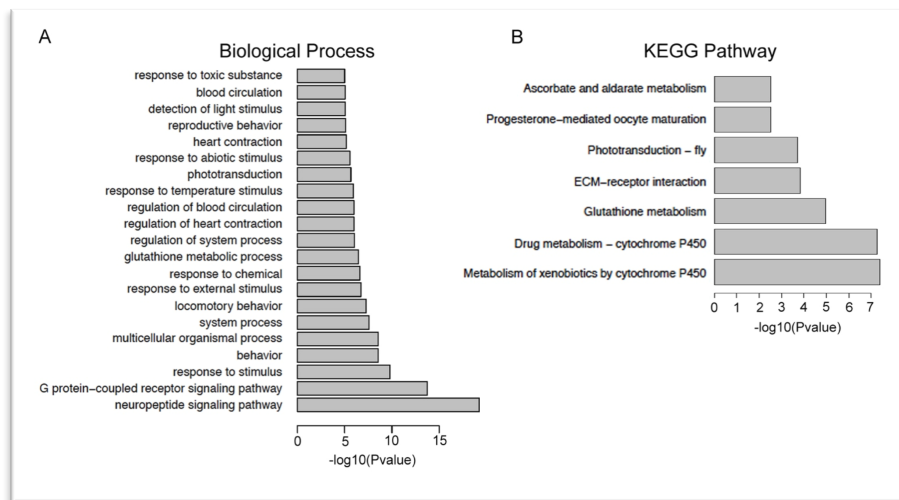
**Figure 6.** Highly variable genes. Dispersion is represented by average expression per gene. Variable genes are depicted in red.

### 2.5.2. Characterization of highly variable genes

We next aimed to characterize these set of variable genes. We hypothesized that, being variable genes in brain cells, they should be related to neural specific functions. To check our hypothesis, we performed a Gene Ontology (GO) Term Enrichment analysis of this dataset. Gene Ontology is a public database where functions related to each gene have been identified and annotated (Ashburner et al., 2000; The Gene Ontology, 2019). It is presented in a tree-based structure, meaning that some categories gather several more specific ones (for instance, metabolism would gather catabolism and anabolism, and catabolism would gather glycolysis, lipolysis and so on). A GO Term Enrichment analysis pretends to identify a set of significantly enriched categories within a subset of genes in comparison to a bigger subset (the Universe). There are many classifications from Gene Ontology that can be interrogated (Biological Process, Cellular Component, Molecular Function, KEGG Pathways, etc). In our case, we decided to identify enriched terms in Biological Process and KEGG Pathways, as they are usually the most meaningful ones to elucidate biological questions.

To perform the GO Term Enrichment analyses we used the *GOstats* (Falcon and Gentleman, 2007) and *KEGG.db* (Carlson et al., 2016) R packages, loading also the annotation of *Drosophila* with all gene identifiers and all categories associated to each gene (*org.Dm.eg.db*).

As Universe of genes, we selected all genes represented in our final expression matrix, this is, genes that are at least in 10 cells and have passed all thresholds in the filtering step (10,746 genes). *GOstats* allows for the identification of over or under-represented categories in the dataset. We selected only categories over-represented and, in this case, with a P value lower than  $10^{-5}$ , as many categories were enriched in this dataset (figure 7A).



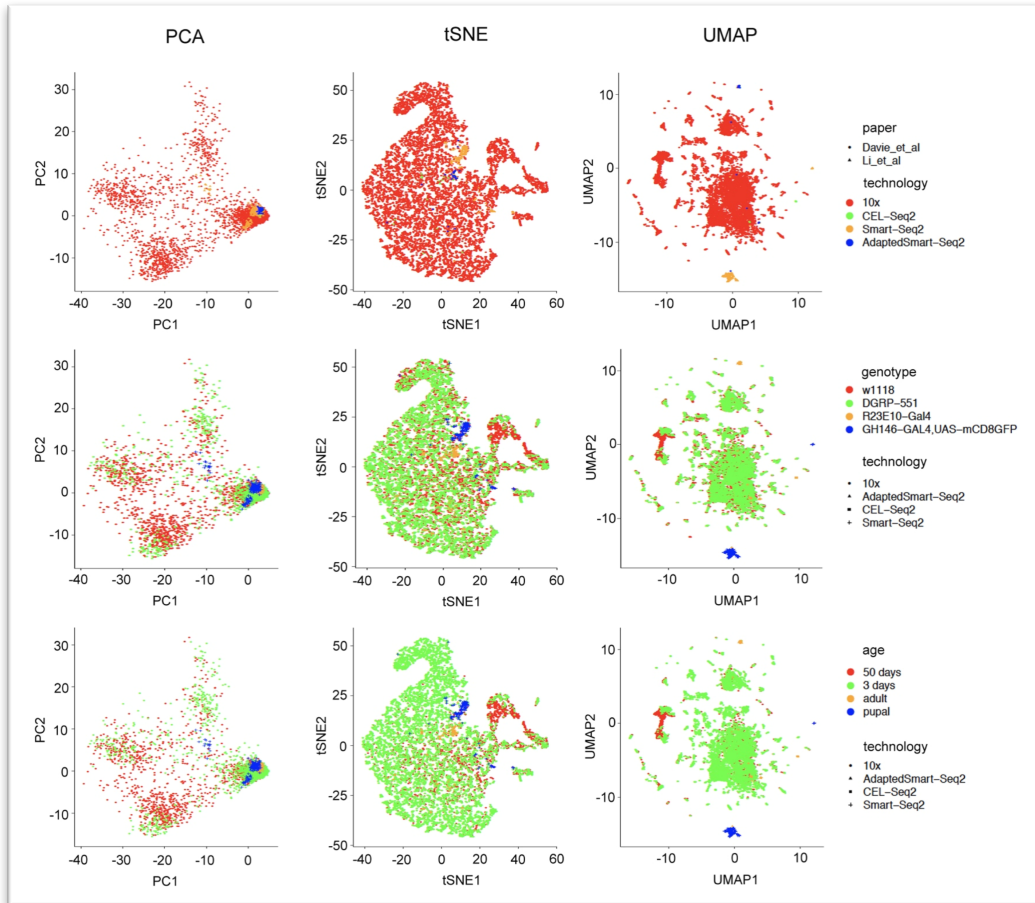
**Figure 7.** GO Term Enrichment of highly variable genes. A. Biological Process terms over-represented in highly variable genes. B. KEGG pathway terms over-represented in highly variable genes.

As expected, highly variable genes in the reference dataset are related to neural activity, such as neuropeptide signaling pathway, response to stimulus, behavior and phototransduction. Many enriched categories belong to the same “tree branches”, meaning that the same set of genes contribute to the enrichment of similar categories of the tree (e.g. regulation of heart contraction, regulation of blood circulation and blood circulation). In figure 7B, enriched KEGG pathways are represented (with a P value lower than 0.01). Among them, we can find pathways related to hormone biosynthesis and phototransduction, both related to neural specific activities.

### 2.5.3. Visualization of scRNA-Seq data

As mentioned before, we are going to compare how dimensionality reduction methods discriminate different cell populations. To do so, we are going to plot PCAs, tSNEs and UMAPs on all cells. Principal components will be computed

taking into account the variable features identified in the previous section, as this will enhance the differences between cell populations, consuming significantly less memory and computing resources. We are going to visualize all cells according to the different variables stored in the metadata of the single-cell experiment.



**Figure 8.** Visualization of scRNA-Seq reference data set by dimensionality reduction methods. Each column depicts one methodology, and in each row a different variable has been grouped by colors or by dot shape. Left panels, cells are represented in a PCA plot. Middle panels, cells are represented in a t-SNE plot. Upper panels, cells are colored by technology. Right panels, cells are represented in an UMAP plot. Middle panels, cells are colored by genotype. Lower panels, cells are colored by age.

As seen in figure 8, results are very different depending on the dimensionality reduction method used. PCA distinguishes four main populations of cells within the cloud of dots, whereas more subpopulations are observable in tSNEs and UMAPs. When coloring cells by technology (upper panels), both PCA and tSNE integrate the four technologies as a single one. In this sense, while all cells from one particular platform tend to homogeneously cluster together (see Smart-Seq2

–in orange– and AdaptedSmart-Seq2 –in blue–, for instance), such minor classes are usually found embedded within the major class of cells, belonging to the 10X technology. In contrast, UMAP shows some clusters of cells that correspond to the diverse platforms used to generate the scRNA-Seq data (upper right panel). Noteworthy, these subpopulations coincide not only with the several technologies but also with the different fly genotypes (middle panels). This could mean that they actually represent true rare subpopulations of cells, which are difficult to identify when performing scRNA-Seq on full tissues, like the experiments with 10X, where thousands of cells are processed in parallel, but would emerge after selecting specific cell types.

When comparing the genotypes by different colors (middle panels), the two wild type strains used in 10X technology overlap almost perfectly in all three dimensionality reduction plots, with the exception of a small subpopulation of cells that is only present in  $w^{1118}$  animals. Very interestingly, this subpopulation of cells is also highlighted when animals age is plotted (lower panels), indicating that it is actually a population of cells only present in  $w^{1118}$  elder animals. This population of cells may not be actually related to the fly strain but to the age of the animals, as 50 day-old animals were only sequenced in  $w^{1118}$  strain.

## 2.6. Batch effect removal

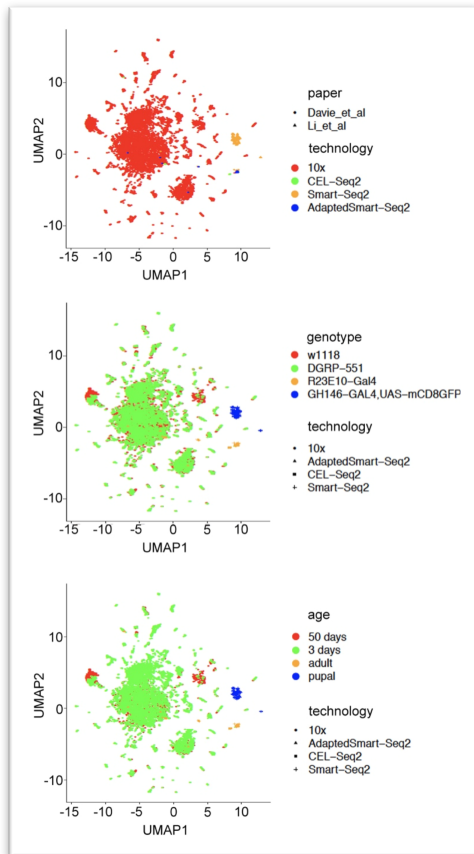
Although it is very likely that the segregation of the multiple scRNA-Seq techniques in separated clusters is due to the actual biological variability between the different genotypes, we aimed to discard any possible confounding factor, such as batch effect, that, due to technical issues, could be masking the real variability between the cell populations.

Thus, we performed a batch effect removal analysis by canonical correlation analysis (CCA). CCA identifies the dimensions in which the different batches show higher correlation and represents the samples in these dimensions (See et al., 2018). To do so, both batches have been analyzed independently. From the count matrices, filtering, normalization, identification of variable features, scaling and analysis of principal components have been performed for each dataset separately.



In our case, we assume that the origin of the putative batch effect may come from the usage of four different technologies to obtain the scRNA-Seq reads. As the analysis must be performed pair-wise, we have decided to divide the dataset into two batches according to the presence or the absence of UMIs (10X/CEL-Seq2 and Smart-Seq2/AdaptedSmart-Seq2, respectively).

In consequence, we have generated matrices with all cells for each batch that were independently analyzed (see appendix, section Brain). We obtained 1,041 variable genes for batch 1 (Smart-Seq2/AdaptedSmart-Seq2) and 446 for batch 2 (10X/CEL-Seq2). We next applied CCA correction from *Seurat* package (standardization and normalization default options) and reanalyzed the data as before (identification of variable features, scaling and calculation of principal components and t-SNE and UMAP dimensions) (figure 9).



**Figure 9.** Visualization of scRNA-Seq after CCA correction. Upper panel, cells are colored by technology. Middle panel, cells are colored by genotype. Lower panel, cells are colored by age.

When visualizing the remaining 12,086 cells by experimental technology (figure 9 upper panel), we still observe a group of cells from Smart-Seq2 that clearly cluster far apart from the others. Indeed, a second cluster with few Smart-Seq2 and AdaptedSmart-Seq2 is also observed in the plot, meaning that they likely represent different subsets of cells not represented in the other datasets. This is actually confirmed when checking the plots of the different genotypes (figure 9 middle panel).

Cells coming from Smart-Seq2 and AdaptedSmart-Seq2 from the *R23E10-GAL4* flies cluster altogether, whereas Smart-Seq2 cells from *GH146-GAL4* cluster further away. Interestingly, after CCA correction, the subpopulation of cells from *w<sup>1118</sup>* 50 days-old that appeared in the previous plots (see figure 8, lower panels) is not as evident now, although two small clusters of cells still appear corresponding to this genotype and age.



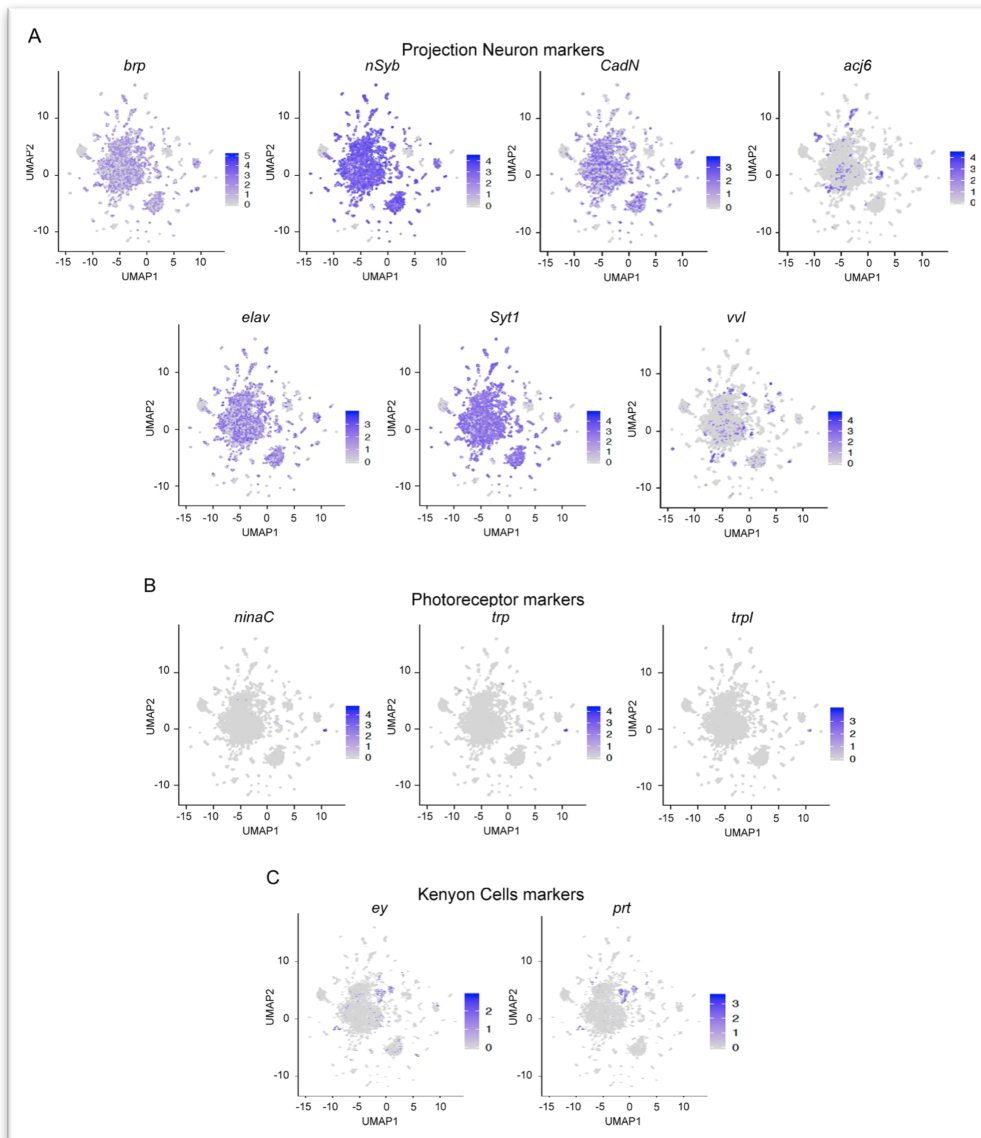
Thus, although we think that these analyses demonstrate there was not a huge batch effect due to the various platforms used on the different papers, the fact that, now, we see a better distinction between the two Smart-Seq2 datasets has prompted us to select this second option to continue with the further data analyses.

## 2.7. Visualization of known marker genes

Once all normalization and correction steps have been performed, we will visualize the level of expression of known marker genes to identify subpopulations of cells within the reference datasets.

Known marker genes were extracted from the reference papers (Davie et al., 2018; Li et al., 2017) and their expression was plotted in the UMAP dimensionality reduction plots. Some examples of markers of subpopulations of neurons are represented in figure 10 and other cell types non-neuronal are depicted in figure 11 (for more examples see the appendix section *Brain*).

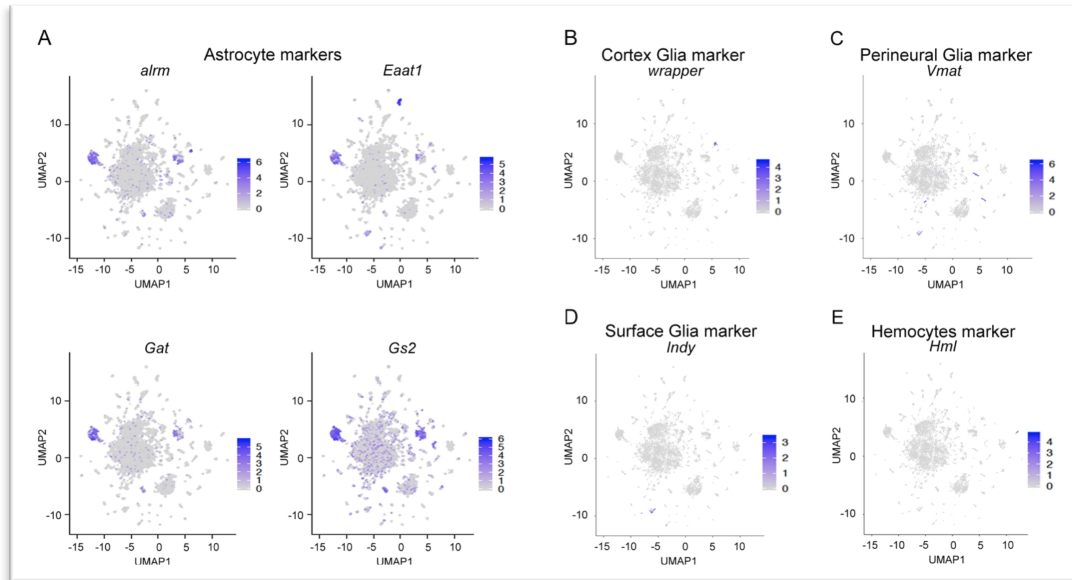
Whereas, according to the different Projection Neuron markers tested (figure 10A), this cell type is spread all around the UMAP plot, photoreceptors (figure 10B) and Kenyon Cells (figure 10C) are localized in few small and well-defined clusters. Photoreceptors are the neurons that reside in the ommatidia, which is the functional unit that forms the compound eye in the fly. Each ommatidium consists in 20 cells, and only 8 of them are photoreceptors, consistent with the low abundance of this cell type in the dataset (Katz and Minke, 2009). All cells in the Photoreceptor population belong to 10X technology isolated in Davie et. al, which is the largest population (compare to figure 9, upper panel), confirming that rare populations will be detectable only by analyzing several thousands of cells. Interestingly, markers from Projection Neurons and Kenyon Cells are mutually exclusive. Kenyon Cells are the neurons that reside in the mushroom bodies, and are related to olfactory memory (Widmann et al., 2018). Such class of cells are also very scarce and and, although both markers coincide in the same populations, they do not cluster in a single population of cells in the UMAP plot.



**Figure 10.** Representation of known neural marker genes. **A.** Projection Neuron marker genes. **B.** Photoreceptor marker genes. **C.** Kenyon Cells marker genes.

The astrocytes are one of the most abundant types of glial cells in *Drosophila*, being responsible for the coverage of the neuropil (Freeman, 2015). In the UMAP plots, astrocytes are distributed into two main clusters of cells and few minor additional populations (figure 11A). The other three types of glial cells (figure 11B-D) are much less abundant, and they represent very small mutually exclusive clusters. Finally, hemocytes are immune cells responsible for the phagocytosis of apoptotic cells (Gold and Bruckner, 2015) and constitute a very small population of cells in our dataset (figure 11E).

All in all, these results suggest that the analyses performed here allow for the identification of well-known populations of cells from the *Drosophila* brain, and that the cells corresponding to these populations cluster close together.



**Figure 11.** Representation of known marker genes from non-neural cells. **A.** Astrocyte marker genes. **B.** Cortex Glia marker genes. **C.** Perineural Glia marker genes. **D.** Surface Glia marker genes. **E.** Hemocytes marker genes.

## 2.8. Identification of clusters and marker genes

The next objective of our project was to identify the different subpopulations of cells conforming our diverse datasets in an unsupervised manner.

We evaluated several packages to identify cell clusters: *quickCluster* from *scrn*, *CIDR* (Lin et al., 2017) and *FindNeighbors/FindClusters* from *Seurat*. However, *Seurat* was the tool proven to perform better in our data. Therefore, *Seurat* is the tool selected to perform this task (see appendix sections *Scran* and *CIDR* for further details of the other two packages).

To identify cell clusters within our reference datasets, we first computed the shared-nearest-neighbors (*SNN*) of the cells. *SNN* algorithm allows for the unsupervised clustering of the samples given a number of “k” nearest-neighbors.

In order to identify clusters, we tried two different options, “k = 20” and “k = 40” nearest-neighbors. In both cases, the analysis is performed by interrogating the variable genes detected after CCA correction. To run *FindClusters* we used the default parameters, this is, “n.start = 10” (number of random starts), “n.iter = 10”

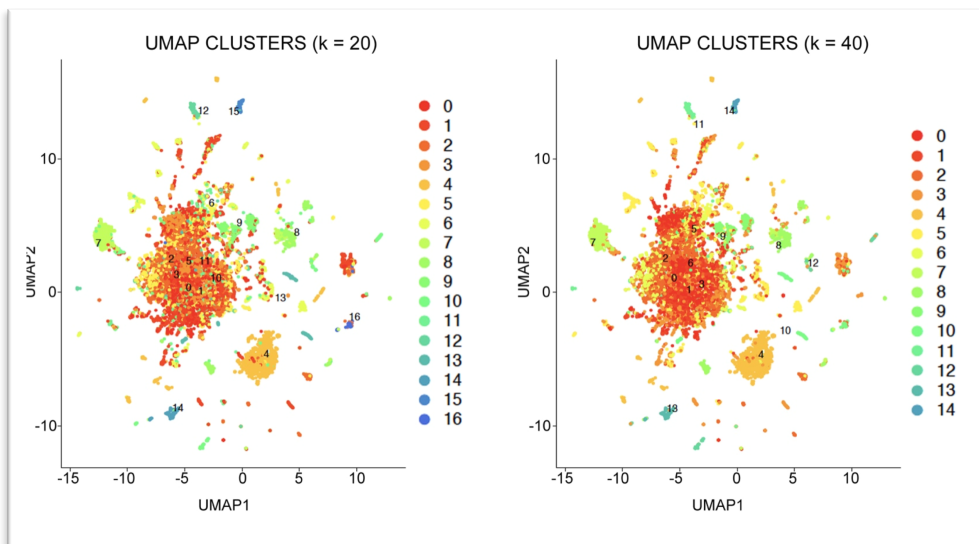
(maximal number of random iterations per start) and “resolution = 1.0” (lower and higher numbers generate a larger and smaller number of communities, respectively).

By running SNN with “k = 20” we identified up to 17 clusters, while “k = 40” identified only 15 (table 2).

k = 20	cluster ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	# cells	1766	1633	1413	1366	1221	1174	792	618	449	384	363	301	173	169	103	84	77
k = 20	cluster ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14		
	# cells	1989	1636	1516	1401	1223	1101	1044	593	482	389	222	180	129	100	81		

**Table 2.** Number of cells identified per cluster with the different parameters used.

To check the performance of the clustering, we next visualized both sets of clusters in the UMAP-dimensionality plots (figure 12).



**Figure 12.** Visualization of clusters identified by Shared-Nearest-Neighbor with “k = 20”, left panel, and “k = 40”, right panel.

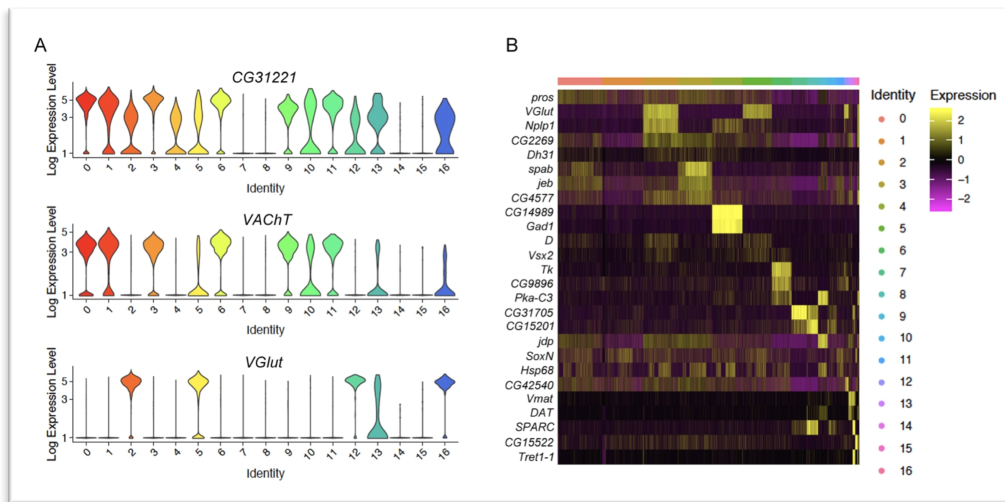
Identification of clusters by SNN seems to perform, in general, quite satisfactorily. The central cloud of cells seems to be formed by 5 to 7 clusters, depending on the number of neighbors selected, while surrounding populations of cells are usually formed by one single cluster

Although the UMAP representation suggests that more clusters should be identified according to the distribution of cells in the plot, a priori we cannot know how many clusters to expect, and we cannot know whether the identified ones are correct or this number represents an over-representation of the real number of cell populations.

## 2.9. Characterization of identified clusters

We finally aimed to assess the identity of the clusters identified by SNN. To do so, we took advantage of the known neural and non-neural markers (see figures 9 and 10) and we also performed Gene Ontology Term Enrichment analysis of the 17 clusters obtained with  $k = 20$ .

First, we selected the highest enriched marker genes from each cluster with the tool *FindAllMarkers* from *Seurat*. This tool identifies the marker genes that contribute more to the identity of the cluster (figure 12).

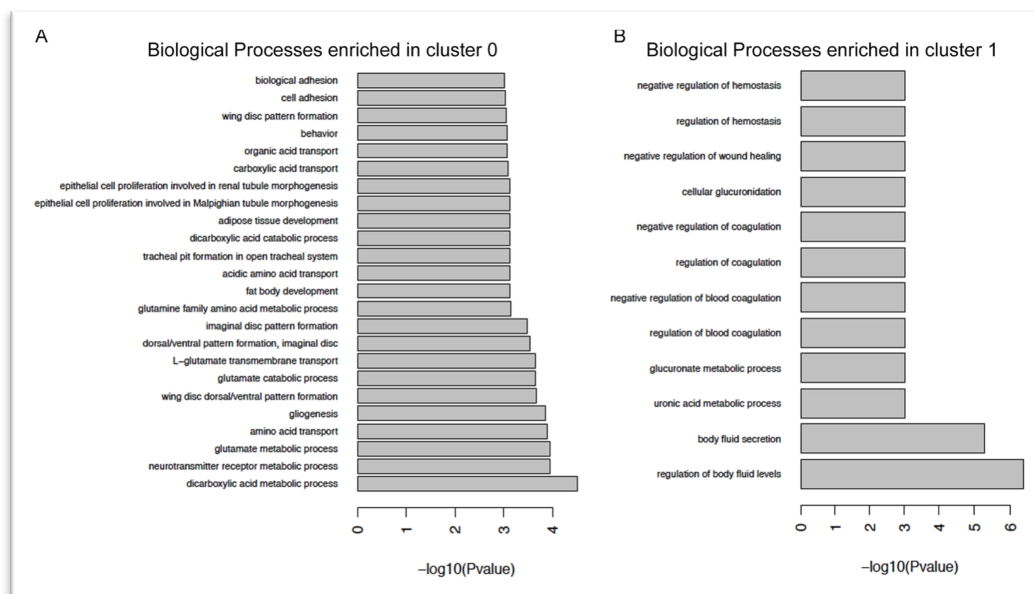


**Figure 13.** Expression of highest enriched markers within clusters. A. Expression of three genes defined as markers according to their log(foldChange). B. Heatmap of expression of top 5 % genes with highest fold change of expression between clusters.

In figure 13A, we plot the expression of three of these markers, as an example, in the 17 clusters identified. The first marker represented (*CG31221*) is an uncharacterized gene, but it shows a very particular expression pattern, from completely silenced in clusters 7, 8, 14 and 15, to very highly expressed in clusters 0, 1, 3, 6, 9 and 11. The expression pattern of the next marker, *VAcHT*, overlaps in actually very similar to the previous one. *VAcHT* is the *Vesicular acetylcholine transporter* gene, and it is responsible for the transport of acetylcholine into synaptic vesicles (FlyBase, (Thurmond et al., 2019)). Very interestingly, the last marker in the plot is *VGlut*, whose expression is completely complementary to *VAcHT*. *VGlut* is actually a transporter protein that resides in glutamatergic and dopaminergic nerve terminals (FlyBase, (Thurmond et al., 2019)), confirming that both cell types are correctly discriminated in our clusters.

From all marker genes, we next selected the top 5 % according to their expression fold change across clusters (figure 13B). Interestingly, many of these highly variable marker genes are actually very well characterized in the literature, and most of them have essential roles during neural development. We can see how some genes are very cluster-specific, such as *Gad1* and *CG14989*, whereas others show a broader expression pattern across cell types, like *pros*. This is consistent with the role of these proteins: *pros* promotes neural differentiation and is present both in neural progenitors and in many differentiated cells, whereas *Gad1* is an enzyme involved in GABA neurotransmitter synthesis, being expressed only in a subset of neurons (FlyBase, (Thurmond et al., 2019)).

To finally characterize the identity of the clusters we performed a GO Term Enrichment analysis on the marker genes of each cluster. Biological Processes and KEGG Pathways were assessed in this analysis (figure 14).



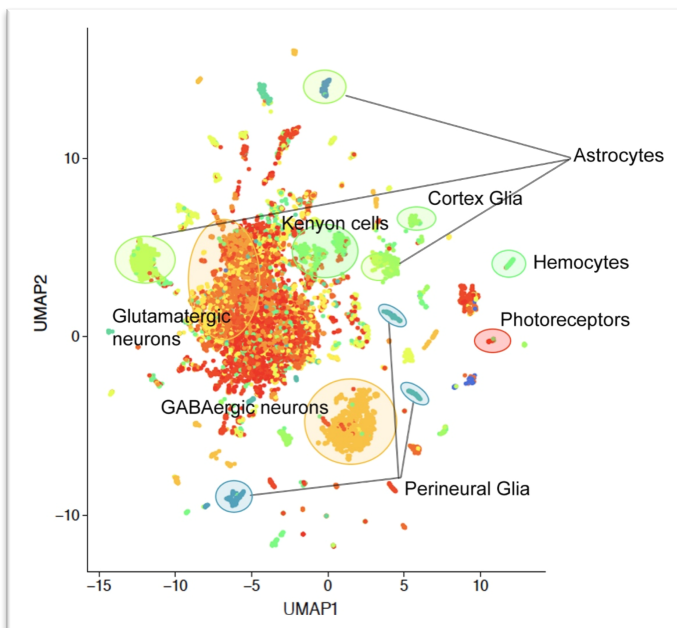
**Figure 14.** GO Term Enrichment analysis of cluster marker genes. A. Biological Processes enriched in cluster 0. B. Biological Processes enriched in cluster 1.

Unfortunately, the GO Term Enrichment analysis did not allow us to determine the identity of many of the clusters. In figure 14, we show, as example, the Biological Processes enriched for clusters 0 (A) and 1 (B). According to the GO Term Enrichment, cluster 0 could be identified as glutamatergic neurons, as some of the processes are related to glutamate metabolism. However, we have seen that *VGlut* is actually higher expressed in cluster 2 (figure 13B). One possibility is

that, actually, these two clusters represent a single one that has been separated according to the expression of other genes. In fact, UMAP from clusters in figure 12 suggested that clusters 0 to 4 were very close, and so the PCA, meaning that the transcriptomic signature of these populations is very similar.

GO Term Enrichment of cluster 1 (figure 14B) shows, instead, categories related to blood coagulation and hemostasis. Both terms, to our knowledge, cannot be assigned to any neural cell type in particular.

All in all, gathering all the information extracted from the cell clusters, we have been able to fully characterize 8 of them, mainly according to the pattern of expression of known cell markers and through the identification of top variable marker genes (figure 15). *VGlut* is expressed, for example, in clusters 2 and 5 and, thus, these were the clusters marked as Glutamatergic Neurons, whereas *Gad1* was expressed exclusively in cluster 4, and we determined that this big cluster of cells corresponded to GABAergic neurons.



**Figure 15.** Characterization of cell clusters. Cluster identity has been determined by the expression of cluster-specific marker genes.

Noteworthy, the clusters identified by SNN do not coincide perfectly with the expression pattern of the markers. For instance, astrocytes markers indicate that there are, at least, three populations of cells that correspond to this cell type

whereas, according to *Seurat*, these three populations belong to, at least, three different clusters (7, 8 and 15). One possible explanation to this observation is that the genes used as markers are actually expressed in several cell subtypes and, in combination with other markers, they represent different subpopulations of astrocytes. *Seurat* would capture the difference between these subpopulations

but it would not have enough resolution to distinguish them with the tested markers.

For many of the marker genes there is not enough information available to determine the identity of the cluster and, thus, we have not been able to fully characterize the identity of each cluster.

Finally, when checking the correspondence of the clusters to the scRNA-Seq platforms used to generate the data (figure 9, upper panel), we see that cluster 16 is, indeed, formed exclusively by AdaptedSmart-Seq2 cells, although some cells processed with this technology are also scattered around the plot. Cells belonging to Smart-Seq2 platform gather in a large cluster of cells and two additional small clusters (figure 9, upper panel). All these cells belong to several different *Seurat* clusters, especially the cells from Li's Smart-Seq2, meaning that, although they are represented close in the UMAP plot, SNN is able to identify them as different cell types. This observation argues against the performance of the batch effect removal by CCA, though.

## 2.10. Discussion of results

To conclude the analysis of *Drosophila* brain scRNA-Seq data, we would like to compare the results obtained along our project with the reference works (Davie et al., 2018; Li et al., 2017) and discuss the main differences.

Along the project, we have analyzed, in parallel, scRNA-Seq data processed from 4 different platforms and two independent publications. The data generated in these publications were very extensive, so, to reduce computing time and to avoid memory issues, we subset the number of samples obtained from each paper. Thus, we ended up with 200 Smart-Seq2 cells from Li's Projection Neurons and with 45 Smart-Seq2, 34 AdaptedSmart-Seq2, 22 CEL-Seq2 and more than 12,000 10X Chromium from Davie's central nervous system.

The first challenge we faced up was the processing of the data, as 10X and CEL-Seq2 samples contain UMIs, whereas the others do not. Datasets were, then, processed independently (see pipeline in figure 1) to quantify either UMI counts or total number of counts per feature (gene).

Our next goal was to remove any putative batch effect due to the different platforms and the different labs in which data were generated. However, the batch



effect observed here was not very significant and it could actually happen that other possible confounding factors affected the distribution of cells in the dimensionality reduction plots, such as the genotype and the age of the animals. Still, we performed CCA correction to try to reduce the possible contribution of the several platforms used to generate the data to the distribution of cells; however, no much differences were observed after CCA correction. It is very likely that the large number of 10X cells present in the analysis is driving all variability between samples in the full analysis. This is, actually, the power of this technology, by which you can easily obtain thousands of high-quality cells in a fast and affordable manner. It is, by far, the technology that is performing better in our analysis.

We finally wanted to identify and characterize cell clusters in our reference dataset. By using *Seurat* package, we have obtained 17 clusters, much less than the number of clusters obtained in the reference publications (35 in Li et al. and 87 to 151 clusters, depending on the resolution, in Davie et al.). This is likely due to the fact that many of the cells show few differences and they cannot be distinguished in the clustering as a whole. However, by subsetting each individual cluster afterwards, we would be able to distinguish more cell types within each one. Indeed, in Li et al., the authors initially identify 12 Projection Neuron clusters, and it is after analyzing the data with a newly developed machine-learning iterative algorithm that they distinguish up to 35 clusters of cells. Davie and colleagues identify between 87 and 151 clusters from their population of cells. It is likely due to the large dataset generated in the paper, as well as to the diverse time points (animal age) used to produce the data.

Very interestingly, with the analyses performed, we were able to identify clusters of cells expressing many of the known marker genes reported in the two papers, indicating that, although the number of processed cells was much lower, we were able to recover most of the known cell populations present in the fruit fly brain.

Although the main objective of the current project was not to perform and exhaustive analysis of the reference datasets but implementing a personalized pipeline to apply it afterwards to our own data, we can envision some improvements for further analyses of these datasets.

First, we suggest to analyze all datasets without removing the duplicates of UMI-containing samples. We would expect the batch effect to be even lower than the observed one and Smart-Seq2 samples from Li et al. to cluster altogether with the 10X cells.

Second, we would implement more clustering methods; although the ones tested for this project did not improve the performance of the Seurat clustering, there are many more methods that could be optimized for this dataset, as the machine learning algorithm developed in Li's paper (Li et al., 2017).

Third, we would represent also a higher number of known *Drosophila* brain markers in the dimensionality reduction plots. In addition, we suggest to explore deeper the expression pattern of the marker genes identified by *Seurat* to try to characterize other clusters in the reference dataset.

Last, we would *subcluster* the clusters already identified to foster the small differences of cells showing a similar transcriptional profile, enhancing the characterization of smaller populations of cells.

# 3 Analysis of MARS-Seq scRNA-Seq data from *Drosophila* wing imaginal discs

## 3.1. Description of the dataset

The last part of the current project involves the analysis of a scRNA-Seq dataset performed by Marina Ruiz-Romero in our group.

The single-cell dataset generated in our lab was performed with cells isolated from two wing imaginal discs from *Drosophila melanogaster*. Imaginal discs are the structures in the larvae that give rise to the diverse appendices in the adult animal and, in this case, to the adult wings (for a review see (Ruiz-Losada et al., 2018)). The adult wing, actually, evaginates during metamorphosis from the central part of the wing disc, called wing pouch. The rest of the wing imaginal disc give rise to other structures in the adult, such as muscle and part of the thorax. Thus, to isolate cells giving rise only to the adult wing, the wing pouch was marked by expressing the fluorescent protein GFP under the control of the *nubbin* driver (*nub>GFP*). Afterwards, third instar larvae were manually dissected and cells were disaggregated by enzymatic digestion of the extracellular matrix.

Cells from two independently disaggregated wing discs were analyzed by flow cytometry and GFP positive cells were sorted onto one 384-well plate, this is, 192 cells per disc. To remove the putative batch effect produced by the independent processing of the two tissues, libraries for single-cell sequencing were prepared in two batches. In each batch, 50% of cells from each disc were included. Libraries were prepared by using the MARS-Seq technique. This technology, that has not been analyzed before in this project, is one of the first methodologies that was implemented to perform scRNA-Seq experiments (Jaitin et al., 2014) and it is actually very similar to the original CEL-Seq protocol, in the sense that amplification of barcoded RNA is performed by *in vitro* transcription, but libraries are prepared within the well-plate.

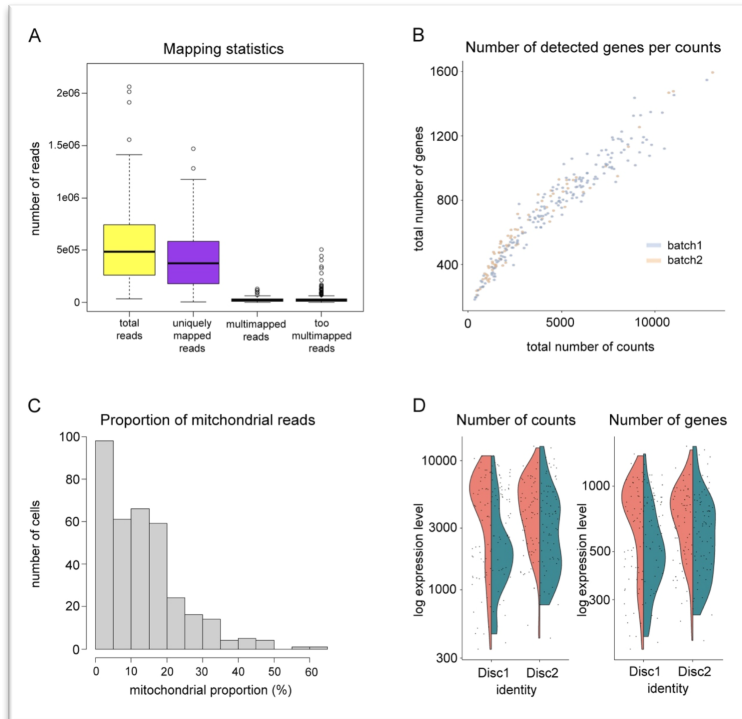
### 3.2. Pre-processing, mapping, quantification, filtering and normalization of the dataset

Our data contain UMI information and, in consequence, it will be processed as 10X and CEL-Seq2 datasets. However, in our case, fastq files had been already demultiplexed by the sequencing facility and two files were provided per cell: read1, containing the read sequence, and read2, containing the cell barcode and the UMI. Besides, two runs of sequencing were performed to increase the sequencing coverage and improve the gene detection per cell.

Thus, the first step of the preprocessing consisted in the concatenation of the two fastq files generated in each sequencing run per cell. We next generated a whitelist with the cell barcodes by using the whitelist options `--extract-method=regex --bc-pattern="(P<cell_1>.{6})(P<umi_1>.{8})"`. These options indicate the *whitelist* tool that cell barcodes correspond to the first 6 nucleotides of the read, whereas the remaining 8 nucleotides represent the UMI. Finally, cell barcodes were extracted from read2 and incorporated into read1 sequence identifier. However, to perform the extraction of the barcodes, we were forced to change the extract method to `--extract-method=string --bc-pattern=CCCCCNNNNNNNN`, as the “regex” method failed when running.

Mapping of scRNA-Seq reads was performed by using the same options as described for the reference datasets, and mapping statistics was computed per cell (figure 16).

According to the statistics file, from the 384 cells, only 353 generated mapping statistics, meaning that the missing 31 cells did not contain any read, either because the library did not work or because no cell fell into the well when sorting. Median number of reads for the remaining cells is 500,000, although some cells reach the 2 million reads. Uniquely mapped reads represent an 80% of the total number of reads, whereas the proportion of multimapped and too multimapped reads (this is, more than 5 mapping positions in the genome and, thus, discarded) is, in general, very low (figure 16A).



**Figure 16.** Mapping and quantification statistics of MARS-Seq scRNA-Seq from fruit fly wing imaginal discs. **A.** Boxplot representing the total number of reads, the uniquely mapping and the multimapping. **B.** Number of detected genes regarding number of counts per cell. Cells are represented according to the batch they were processed in. **C.** Proportion of mitochondrial reads per cell. **D.** Number of counts (UMIs) and genes detected per cell. Cells have

been split by disc (disc1, disc2) and by batch (pink for batch1 and green for batch2).

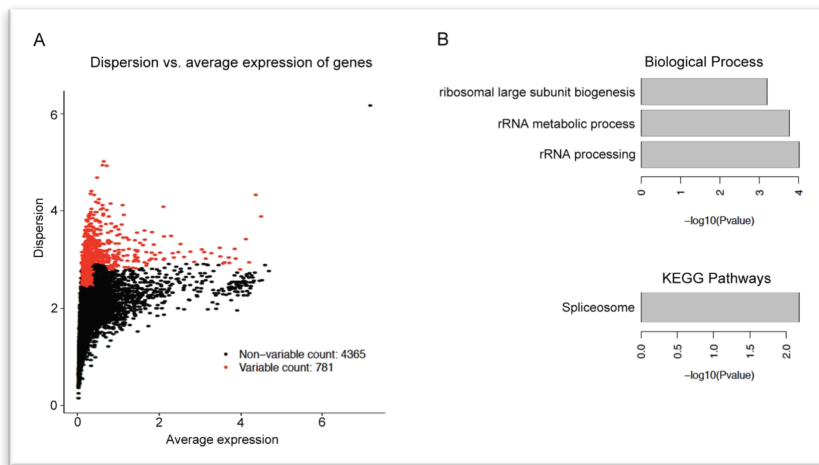
Mapped reads were, afterwards assigned to *Drosophila* genome, PCR duplicates were discarded and UMIs in each cell were quantified. As seen in figure 16B, we have not reached a plateau in gene detection, indicating that increasing the sequencing coverage could also increase gene detection per cell.

Low-quality cells were next filtered out according to the proportion of reads falling into mitochondrial genes (figure 16C). Cells presenting more than 20% reads in mitochondrial genes, cells which are more than 3 MADs below the median log-library size as well as cells for which less than 50 genes were detected were filtered out. Genes not detected in at least 5 cells were also discarded. After filtering, 262 cells and 5,146 genes were kept for further analyses. From these 262 cells, 125 belong to disc1, and 137 to disc2. However, 178 cells were kept in batch1, while only 84 cells from batch2 passed the thresholds. This observation suggests a putative batch effect due to the independent processing of each batch. The lower number of counts and genes detected in the different batches is also observable in violin plots in figure 16D.

Expression matrix was finally log normalized and scaled for further analyses.

### 3.3. Visualization by dimensionality reduction methods

#### 3.3.1. Identification and characterization of highly variable genes

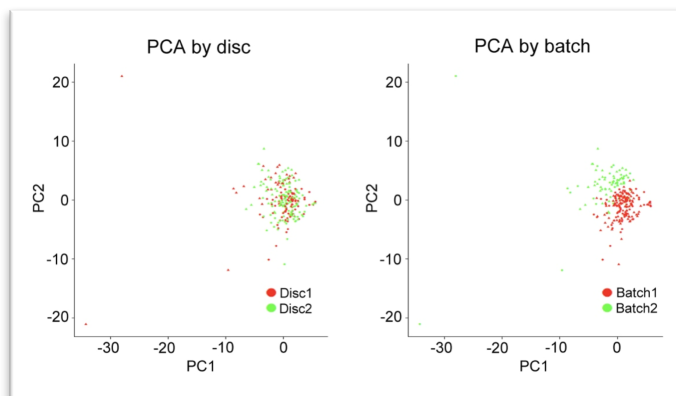


**Figure 17.** Highly variable genes from MARS-Seq. **A.** Dispersion is represented by average expression per gene. Variable genes are depicted in red. **B.** Biological Processes (upper panel) and KEGG Pathways (lower panel)

enriched in variable genes.

To visualize the MARS-Seq wing cells by dimensionality reduction methods, we first computed the highly variable genes from the 5,146 genes in the expression matrix. From these genes, 781 were identified as variable (figure 17A). GO Term Enrichment analysis of this subset of genes only showed categories related to RNA processing and splicing (figure 17B).

#### 3.3.2. Visualization of scRNA-Seq data



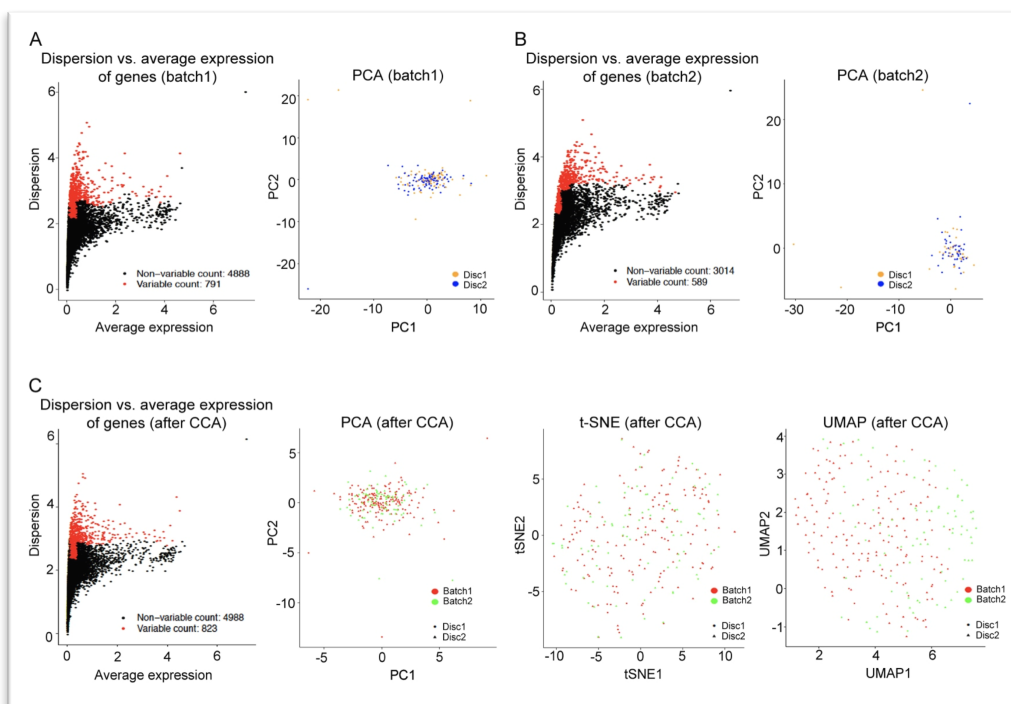
**Figure 18.** PCA analysis of wing scRNA-Seq data. Left panel, cells are colored by disc. Right panel, cells are colored by library batch.

We next computed principal components taking into account the 781 variable genes in the dataset. PCA plots show no batch effect due to the independent disaggregation of the larval wings (figure 18, left panel); however, and as expected, there is a strong batch effect promoted from the library preparation batches that drives the first two principal components in the PCA plot. The same batch effect is observed in t-SNE and UMAP dimensionality reduction plots (see appendix, MARS section). In this case, the batch effect cannot be ignored, and batch effect correction is essential.

### 3.4. Batch effect removal

Batch effect removal in MARS-Seq dataset was assessed by two different methods: mutual-nearest-neighbor (MNN) from *scran* (*FastMNN*) and canonical correlation analysis (CCA) from *Seurat* (*RunCCA*). In this section, however, we are going to detail the latest, as it performed better and was the selected tool for our further analyses (see appendix MARS-Seq section for *FastMNN* results).

Thus, as in the previous chapter, we reloaded each expression matrix according to the library batch each cell was processed in. In the previous sections, we saw that a higher number of cells was discarded from batch2, likely due to an issue in the library preparation of this subset of cells. So, we performed an independent analysis for the two datasets, being the batch1 formed by 178 and batch2 by 84 cells (figure 19).



**Figure 19.** Batch effect correction by CCA. **A.** Variable genes and PCA of cells from library batch1. **B.** Variable genes and PCA of cells from library batch2. **C.** Variable genes and dimensionality reduction plots of cells after CCA correction (from left to right, PCA, t-SNE and UMAP).

After processing, we identified 791 variable genes within batch1 and 589 variable genes within batch2 (figure 19A and B). When performing PCA analysis for each batch, cells from disc1 and disc2 seem to gather without showing any particular

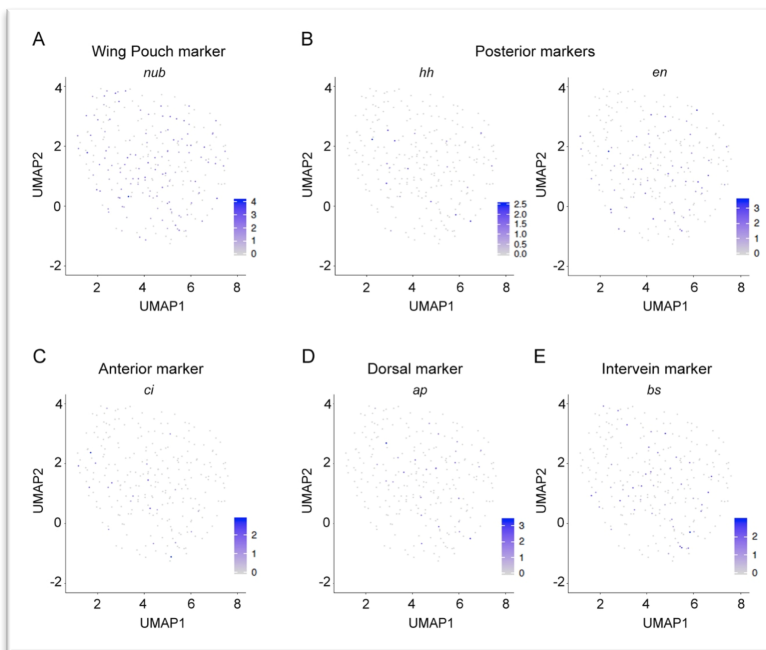
clustering, confirming that no batch effect was generated when isolating the cells from the tissues.

After CCA correction, 823 genes were selected as variable (figure 19C). The CCA correction has certainly removed most of the batch effect detected across libraries, as seen in PCA and t-SNE plots. UMAP plot still shows segregation of some cells according to their batch, but it has been greatly reduced (compare to UMAP in appendix, section MARS-Seq).

### 3.5. Visualization of known marker genes

Although no evident subpopulations of cells were observed in the dimensionality reduction plots, we next aimed to identify cell types and clusters in the dataset.

First approach was, then, to visualize known wing marker genes in UMAP dimensionality reduction plots (figure 20).



**Figure 20.** Visualization of known marker genes. **A.** Wing pouch marker gene. **B.** Posterior compartment marker genes. **C.** Anterior compartment marker gene. **D.** Dorsal compartment marker gene. **E.** Intervenein marker gene.

The driver selected to isolate the wing pouch cells from the rest of cells of the wing imaginal disc was

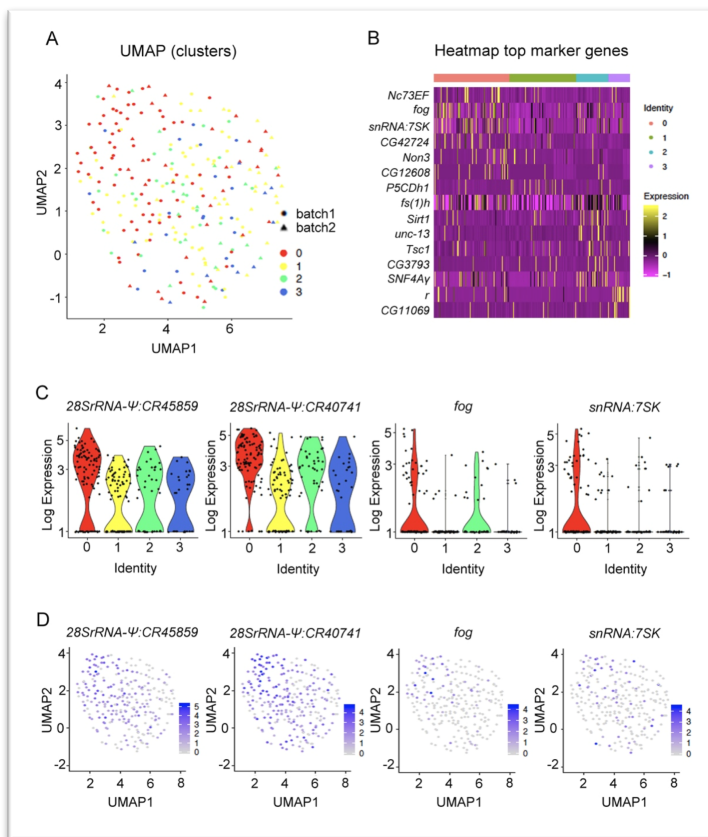
*nubbin*, and thus, we expected this gene to be expressed in almost all cells of the dataset (figure 20A). Unfortunately, the expression level is either very low or absent in many cells, indicating that either we have not reached enough sequencing coverage to detect the expression of ubiquitous genes or that MARS-Seq technology does not perform as expected for this kind of samples. The same happens for the other marker genes tested (figure 20B-E). Anterior and posterior markers, for instance, should be complementary, representing each one around



50% of the cells, but the expression level of all of them is very low and no specific pattern is observed.

### 3.6. Identification and characterization of clusters and marker genes

To finally determine the identity of the cells in our dataset, we asked *Seurat* to identify clusters and the marker genes representing these clusters within our experiment (figure 21). In this way, we identified 4 clusters (figure 21A, for PCA and t-SNE reduction plots see appendix, section MARS-Seq). Cluster 0 was formed by 101 cells, cluster 1 by 89 cells, cluster 2 by 43 cells and cluster 3 by 29 cells. A part from cluster 0, the other 3 clusters were not very well discriminated in the UMAP plot, being all cells mixed in a central cloud of dots. Again, this seems to be caused by the lack of detection of many marker genes in all cells belonging to a cluster, as seen in figure 21B, where expression of top representative marker genes of each cluster is depicted. No evident expression pattern is observed in the heatmap, indicating that the clustering is driven by the expression of few common genes.



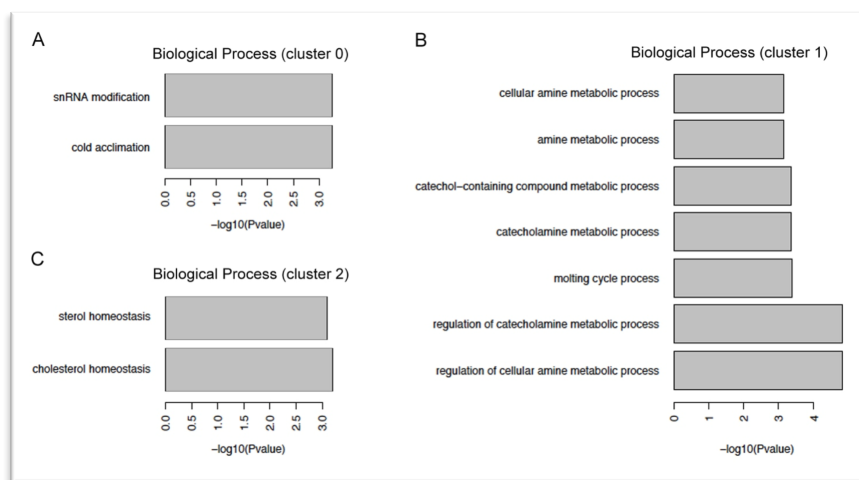
**Figure 21.** Clusters and top marker genes of wing scRNA-Seq dataset. **A.** UMAP plot of clusters identified by *Seurat*. **B.** Heatmap of expression of top marker genes found in each cluster. **C.** Expression of the 4 marker genes showing highest variability across clusters. **D.** UMAP plot of expression of the same marker genes as in C.

From the marker genes with highest variability across clusters, we selected the top 4 and depicted their expression as violin plots (figure 21C) and in each cell in UMAP dimensionality

reduction plots (figure 21D). The top two marker genes were, actually, pseudogenes (*CR45859* and *CR40741*) (Thurmond et al., 2019), and they were

higher expressed in cluster 0 than in the other populations. The third top marker gene, *fog*, is a gene involved in gastrulation and axon guidance and it is expressed in cluster 0 and 2. The fourth marker gene, *snRNA:7SK*, is a small nuclear RNA and it is almost exclusively expressed in cluster 0. All in all, none of top marker genes seems to be strongly related to wing development or morphogenesis.

With the aim of further characterizing the clusters, we finally performed a GO Term Enrichment of the representative genes of each cluster (figure 22).



**Figure 22.** Gene Ontology Term Enrichment of MARS-Seq clusters. **A.** Biological process categories enriched in cluster 0. **B.** Biological process categories enriched in cluster 1. **C.** Biological process categories enriched in cluster 2.

From the 4 clusters, only 3 showed categories enriched for Biological Process classification. The enriched categories were related to metabolic (cluster 1) and hormone (cluster 2) processes. Cluster 0 only showed two categories: snRNA modification and cold acclimation, with no apparent relationship.

### 3.7. Discussion of the results

In the last part of the project, we have analyzed scRNA-Seq data generated by Marina Ruiz-Romero in our laboratory. The sample was obtained from developing tissues from *Drosophila melanogaster*, in particular, from third instar larvae wing imaginal discs. Two wing imaginal discs were manually isolated and sorted into 384-well plates. Libraries were performed according to MARS-Seq protocol, developed in Ido Amit's lab as a high-throughput improvement of the CEL-Seq protocol (Jaitin et al., 2014). Two batches of libraries were generated, each one

containing 50% of cells of each isolated wing. Samples were finally pooled after library preparation and two runs of sequencing were performed, reaching a coverage of 500,000 reads per cell, on average. However, despite the high coverage, the number of detected genes per cell is not very high. Further analysis of the samples uncovered an issue in one of the library batches, for which more than 50% of cells were removed due to a low quality of the library. Upon cells removal, the batch effect still was driving cell variability, as seen in the dimensionality reduction plots. Thus, batch effect correction was essential to perform further analyses on these data. After CCA batch effect removal, cells clustered altogether independently of the disc and the library batch.

Nevertheless, the quality of the libraries was not high enough to identify different cell populations within the dataset. This is evident when representing the expression of known wing marker genes. There is not clear expression pattern of any of them in any subpopulation of cells. The same happens with marker genes identified by *Seurat*. All this indicates that, either the sequencing coverage is not high enough, or quality of libraries is very limited. Given that sequencing coverage in MARS-Seq samples reached the 0.5M reads/cell, ten times more than the sequencing performed in 10X Chromium samples from Davie et al. (Davie et al., 2018), we speculate that the technology used to generate the data was not the most appropriate for our samples.

Another factor that could have influenced our analyses is the low diversity of the cells in the dataset. As stated above, only cells from the wing pouch were isolated to perform the scRNA-Seq experiments. Although the wing pouch is formed by many different cell types (anterior/posterior and dorso/ventral compartments, boundaries, veins and interveins) (Ruiz-Losada et al., 2018), the transcriptional signature of these cell types is very similar, as all of them represent epithelial cells in a similar differentiation state. The lack of variability between cell types would not allow for a clear discrimination between them. Indeed, in a recent work from Boutros and Teleman labs (Josephine Bageritz, 2018) on full third instar larvae wing imaginal discs, cells from different parts of the wing disc cluster separately, but cells from the wing pouch cluster altogether and no subpopulation is observed within this region, suggesting that, certainly, the variability between these cells is not high enough to discriminate any particular cell type.

# 4 Conclusions

1. The reference datasets analyzed from Li et al. and Davie et al. publications, generated on brain isolated cells using 4 different technologies show strong differences at the level of sequencing coverage. At level of gene detection, the differences are highly reduced.
2. These referent datasets also show a mild batch effect generated by the usage of four different technologies. CCA correction reduces satisfactorily this bias.
3. Dimensionality reduction methods distinguish several populations of cells within the reference datasets. From PCA, t-SNE and UMAP, UMAP is the method that discriminates more clusters of cells, allowing for a better visualization of the data.
4. Expression of known marker genes has allowed for the identification of several populations of known brain cell types in *Drosophila*, such as photoreceptors and astrocytes.
5. By using unsupervised methods, we have been able to identify up to 17 clusters of cells. Highly variable marker genes representing these clusters have also been identified, and their expression shows specific patterns along the different clusters of cells.
6. From the 17 identified clusters, we have been able to fully characterize 8 clusters, mainly due to the expression of known and automatically identified marker genes. Gene Ontology Term Enrichment analysis of cluster marker genes has not provided significant information on most of them.
7. From the 4 methodologies employed to produce the scRNA-Seq datasets, 10X Chromium seems to be the one performing better, given the low sequencing coverage reached and the high number of detected genes per cell.
8. From the several pipelines evaluated during the first part of the project, *Seurat* is the most complete and user-friendly one, performing almost all possible tasks in the analysis. However, other pipelines have been used to perform specific analyses, such as the low-quality library filtering performed by *scater*. The

pipelines selected from the previous study have been used to analyze our scRNA-Seq data generated in *Drosophila* wing imaginal discs.

9. Cells sequenced using MARS-Seq technology in our lab present a high sequencing coverage (500,000 reads per cell, on average). However, *plateau* of gene detection per number of counts has not been reached.

10. This dataset shows a strong batch effect due to the two library preparation batches. CCA correction also removes almost all batch effect in this dataset. No batch effect is detected for the independent handling of the two wing imaginal discs.

11. After batch effect correction, highly variable genes between cells are related to RNA processing categories.

12. Known marker genes visualized in UMAP dimensionality reduction plots show very low levels of expression and no clear expression patterns in the population of cells.

13. Unsupervised cluster detection has identified 4 different clusters, not well-discriminated within the dimensionality reduction plots.

14. Top highly-variable marker genes in each cluster show very sparse expression within each population of cells.

15. The low level of expression of marker genes and the low variability between cells in the dataset has not allowed to fully characterize any of the clusters identified.

# 5 Glossary

BC, barcode

CCA, canonical correspondence analysis

CEL-Seq, cell RNA-Seq

CIDR, Clustering through Imputation and Dimensionality Reduction

DNA, deoxyribonucleic acid

EBI, European Bioinformatics Institute

eRNA, enhancer RNA

GEO, Gene Expression Omnibus

GWAS, genome-wide association study

HCA, Human Cell Atlas

MNN, Mutual Nearest-Neighbors

NCBI, National Center for Biotechnology Information

NGS, Next Generation Sequencing

PCA, Principal Component Analysis

PCR, Polymerase Chain Reaction

QC, Quality Control

RNA, ribonucleic acid

RNA-Seq, high-throughput sequencing of RNA

SNN, Shared Nearest-Neighbors

scRNA-Seq, single-cell RNA-Seq

t-SNE, t-distributed Stochastic Neighbor Embedding

UMAP, Uniform Manifold Approximation and Projection

UMI, Unique Molecular Identifier

## 6 Bibliography

- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-29.
- Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res* 33, D562-566.
- Barron, M., and Li, J. (2016). Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. *Sci Rep* 6, 33892.
- Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*.
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36, 411-420.
- Carlson, M.R., Pages, H., Arora, S., Obenchain, V., and Morgan, M. (2016). Genomic Annotation Resources in R/Bioconductor. *Methods Mol Biol* 1418, 67-90.
- Croset, V., Treiber, C.D., and Waddell, S. (2018). Cellular diversity in the *Drosophila* midbrain revealed by single-cell transcriptomics. *Elife* 7.
- Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M.R., Armean, I.M., Bennett, R., Bhai, J., Billis, K., Boddu, S., *et al.* (2019). Ensembl 2019. *Nucleic Acids Res* 47, D745-D751.
- Davie, K., Janssens, J., Koldere, D., De Waegeneer, M., Pech, U., Kreft, L., Aibar, S., Makhzami, S., Christiaens, V., Bravo Gonzalez-Blas, C., *et al.* (2018). A Single-Cell Transcriptome Atlas of the Aging *Drosophila* Brain. *Cell* 174, 982-998 e920.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21.
- Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics* 23, 257-258.
- Freeman, M.R. (2015). *Drosophila* Central Nervous System Glia. *Cold Spring Harb Perspect Biol* 7.
- Gold, K.S., and Bruckner, K. (2015). Macrophages and cellular immunity in *Drosophila melanogaster*. *Semin Immunol* 27, 357-368.
- Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 36, 421-427.

- Hinton, L.J.P.v.d.M.a.G.E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9, 2579-2605.
- Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., *et al.* (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343, 776-779.
- Jolliffe, I.T. (2002). *Principal Component Analysis*. (books.google.com).
- Josephine Bageritz, P.W., Erica Valentini, Svenja Leible, Michael Boutros and Aurelio A. Teleman (2018). Gene expression atlas of a developing tissue by single cell expression correlation analysis. *BioRxiv*.
- Katz, B., and Minke, B. (2009). Drosophila photoreceptors and signaling mechanisms. *Front Cell Neurosci* 3, 2.
- Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet*.
- Li, H., Horns, F., Wu, B., Xie, Q., Li, J., Li, T., Luginbuhl, D.J., Quake, S.R., and Luo, L. (2017). Classifying Drosophila Olfactory Projection Neuron Subtypes by Single-Cell RNA Sequencing. *Cell* 171, 1206-1220 e1222.
- Li, W., Liu, C.C., Kang, S., Li, J.R., Tseng, Y.T., and Zhou, X.J. (2016). Pushing the annotation of cellular activities to a higher resolution: Predicting functions at the isoform level. *Methods* 93, 110-118.
- Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923-930.
- Lin, P., Troup, M., and Ho, J.W. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* 18, 59.
- Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* 5, 2122.
- McCarthy, D.J., Campbell, K.R., Lun, A.T., and Wills, Q.F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33, 1179-1186.
- Ponting, C.P. (2019). The Human Cell Atlas: making 'cell space' for disease. *Dis Model Mech* 12.
- Pundhir, S., Poirazi, P., and Gorodkin, J. (2015). Emerging applications of read profiles towards the functional annotation of the genome. *Front Genet* 6, 188.
- Ruiz-Losada, M., Blom-Dahl, D., Cordoba, S., and Estella, C. (2018). Specification and Patterning of Drosophila Appendages. *J Dev Biol* 6.
- See, P., Lum, J., Chen, J., and Ginhoux, F. (2018). A Single-Cell Sequencing Guide for Immunologists. *Front Immunol* 9, 2425.
- Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* 27, 491-499.
- Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. *Nat Rev Genet*.



The Gene Ontology, C. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* 47, D330-D338.

Thurmond, J., Goodman, J.L., Strelets, V.B., Attrill, H., Gramates, L.S., Marygold, S.J., Matthews, B.B., Millburn, G., Antonazzo, G., Trovisco, V., *et al.* (2019). FlyBase 2.0: the next generation. *Nucleic Acids Res* 47, D759-D765.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10, 57-63.

Widmann, A., Eichler, K., Selcho, M., Thum, A.S., and Pauls, D. (2018). Odor-taste learning in *Drosophila* larvae. *J Insect Physiol* 106, 47-54.

# 7 Appendix

To facilitate the handling of the documents and the visualization of the code and the plots generated along the project, appendix sections corresponding to R pipelines (*Brain*, *CIDR*, *scran* and *MARS*) will be provided as *html* files attached to this document.

Bash scripts are detailed below.

## *Bash\_scripts*

### **01\_Concatenate\_files**

```
#!/bin/bash

#$ -N concatenate
#$ -cwd
#$ -j y # Merge error/output logs
#$ -o logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=6G,h_rt=2:00:00 # ask for memory and time as needed
#$ -t 1-768 # substitute n by the number of files you have

# Input files
input=$( cat /sperez/masterBioinfo/01_datasets/MARS_Seq/input.txt | /bin/sed -
n ${SGE_TASK_ID}p ) &&

cp /sequences/fly/single-cell/fastq/2016-11-05/$input.gz fastq/ &&
cp /sequences/fly/single-cell/fastq/2017-02-02/$input.gz . &&
gzip -d fastq/$input.gz &&
gzip -d $input.gz &&
cat $input fastq/$input > fastq/$input.cat.fq &&
rm fastq/$input &&
rm $input &&

exit
```

### **02\_Rename\_reads**

```
#!/bin/bash

#$ -N rename
#$ -cwd
#$ -j y # Merge error/output logs
```

```

#$ -o logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=6G,h_rt=2:00:00 # ask for memory and time as needed
#$ -t 1-768 # substitute n by the number of files you have

# Input files

input=$( cat / sperez/masterBioinfo/01_datasets/MARS_Seq/input.txt | /bin/sed -
n ${SGE_TASK_ID}p ) &&

awk                               'BEGIN{FS=":"}{print                $1}'
sperez/masterBioinfo/01_datasets/MARS_Seq/fastq/$input.cat.fq          >
renamed_fastq/$input &&

exit

```

### 03\_Generate\_whitelist

```

#!/bin/bash

#$ -N whitelist
#$ -cwd
#$ -j y # Merge error/output logs
#$ -o logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=6G,h_rt=2:00:00 # ask for memory and time as needed
#$ -t 1-2 # substitute n by the number of files you have

# Input files
input=$( cat /sperez/masterBioinfo/01_datasets/MARS_Seq/input_2.txt | /bin/sed
-n ${SGE_TASK_ID}p ) &&

mkdir whitelist/$input &&
cd whitelist/$input &&

~/local/bin/umi_tools                whitelist                --stdin
/sperez/masterBioinfo/01_datasets/MARS_Seq/renamed_fastq/"$input"_2.fq --
extract-method=regex --bc-pattern="(P<cell_1>.{6})(P<umi_
1>.{8})"                --log2stderr                >
/masterBioinfo/01_datasets/MARS_Seq/whitelist/$input/"$input"_whitelist.txt &&

cd ../.. &&

exit

```

### 04\_Extract\_barcodes

```

#!/bin/bash

#$ -N extract_umi

```

```

#$ -cwd
#$ -j y # Merge error/output logs
#$ -o logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=6G,h_rt=2:00:00 # ask for memory and time as needed
#$ -t 1-384 # substitute n by the number of files you have

# Input files
input=$( cat
/no_backup_isis/rg/sperez/masterBioinfo/01_datasets/MARS_Seq/input_2.txt |
/bin/sed -n ${SGE_TASK_ID}p ) &&
mkdir extracted_files/$input &&
cd extracted_files/$input/ &&

~/local/bin/umi_tools extract --extract-method=string --bc-
pattern=CCCCCNNNNNNNNN --stdin
sperez/masterBioinfo/01_datasets/MARS_Seq/renamed_fastq/"$input"_2.fq --
read2-in
sperez/masterBioinfo/01_datasets/MARS_Seq/renamed_fastq/"$input"_1.fq --
stdout
sperez/masterBioinfo/01_datasets/MARS_Seq/extracted_files/$input/"$input"_
1_extracted.fastq --read2-
out=/sperez/masterBioinfo/01_datasets/MARS_Seq/extracted_files/$input/"$inp
ut"_2_extracted.fastq

cd /sperez/masterBioinfo/01_datasets/MARS_Seq/

exit

```

## 05\_Mapping

```

#!/bin/bash

#$ -N map_fastq
#$ -cwd
#$ -j y # Merge error/output logs
#$ -o logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=6G,h_rt=6:00:00 # ask for memory and time as needed
#$ -t 1-384 # substitute n by the number of files you have

# Input files
input=$( cat /sperez/masterBioinfo/01_datasets/MARS_Seq/input_2.txt | /bin/sed
-n ${SGE_TASK_ID}p ) &&
mkdir mapping/$input &&
cd mapping/$input &&

/sperez/masterBioinfo/tools/STAR-2.7.0f/bin/Linux_x86_64/STAR --genomeDir
/sperez/masterBioinfo/tools/genomeDir/ --sjdbGTFfile /sperez/ma
sterBioinfo/tools/subread-1.6.4-Linux-
x86_64/annotation/Drosophila_melanogaster.BDGP6.95.gtf --

```

```

outFilterMultimapNmax 5 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --
outFilterMismatchNover
ReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --
alignMatesGapMax 1000000 --readFilesIn
/sperez/masterBioinfo/01_datasets/MARS_Seq/extracted_files/$input/"$input
"_2_extracted.fastq --outSAMtype BAM SortedByCoordinate &&

cd / rg/sperez/masterBioinfo/01_datasets/MARS_Seq/ &&

exit

```

## 06\_Mapping\_statistics

```

#!/bin/bash

#$ -N map_statistics
#$ -cwd
#$ -j y # Merge error/output logs
#$ -o ../logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=1G,h_rt=0:05:00 # ask for memory and time as needed
#$ -t 1-384 # substitute n by the number of files you have

# Input files
input=$( ls / sperez/masterBioinfo/01_datasets/MARS_Seq/mapping/ | /bin/sed -
n ${SGE_TASK_ID}p ) &&
cd mapping/$input/ &&

head -6 Log.final.out | tail -1 | awk 'BEGIN{FS = "|"}{print "total_reads", $2}' >
total_number_reads.txt
head -9 Log.final.out | tail -1 | awk 'BEGIN{FS = "|"}{print
"uniquely_mapped_reads", $2}' > uniquely_mapped_reads.txt
head -10 Log.final.out | tail -1 | awk 'BEGIN{FS = "|"}{print $2}' | awk 'BEGIN{FS
= "%"}{print "prop_uniquely_mapped_reads", $1}' >
proportion_uniquely_mapped_reads.txt
head -24 Log.final.out | tail -1 | awk 'BEGIN{FS = "|"}{print
"multimapping_mapped_reads", $2}' > multimapped_reads.txt
head -25 Log.final.out | tail -1 | awk 'BEGIN{FS = "|"}{print $2}' | awk 'BEGIN{FS
= "%"}{print "prop_multimapped_reads", $1}' >
proportion_multimapped_reads.txt
head -26 Log.final.out | tail -1 | awk 'BEGIN{FS = "|"}{print
"too_multimapped_mapped_reads", $2}' > tooMultimapped_reads.txt
head -27 Log.final.out | tail -1 | awk 'BEGIN{FS = "|"}{print $2}' | awk 'BEGIN{FS
= "%"}{print "prop_tooMultimapped_reads", $1}' >
proportion_tooMultimapped_reads.txt
head -30 Log.final.out | tail -1 | awk 'BEGIN{FS = "|"}{print $2}' | awk 'BEGIN{FS
= "%"}{print "prop_tooShort_reads", $1}' > proportion_tooShort_reads.txt

cat          total_number_reads.txt          uniquely_mapped_reads.txt
proportion_uniquely_mapped_reads.txt          multimapped_reads.txt

```

```
proportion_multimapped_reads.txt          tooMultimapped_reads.txt
proportion_tooMultimapped_reads.txt      proportion_tooShort_reads.txt    >
/sperez/masterBioinfo/01_datasets/MARS_Seq/statistics/"$input"_mapping_statistics.txt
```

```
rm total*
rm uniquely*
rm proportion*
rm multimap*
rm too*
```

```
cd /sperez/masterBioinfo/01_datasets/MARS_Seq/
```

```
exit
```

## 07\_Join\_statistics

```
#!/bin/bash
```

```
#$ -N joinStatistics
#$ -cwd
#$ -j y # Merge error/output logs
#$ -o logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=6G,h_rt=2:00:00 # ask for memory and time as needed
```

```
# Run
```

```
for i in $(cat samples.txt);
do
    join $i tmp.tsv > tmp2.tsv
    mv tmp2.tsv tmp.tsv
done
```

```
exit
```

## 08\_Assign\_features

```
#!/bin/bash
```

```
#$ -N featureCounts
#$ -cwd
#$ -j y # Merge error/output logs
#$ -o logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=6G,h_rt=6:00:00 # ask for memory and time as needed
#$ -t 1-384 # substitute n by the number of files you have
```

```
# Input files
```

```
input=$( ls / sperez/masterBioinfo/01_datasets/MARS_Seq/mapping/ | /bin/sed -
n ${SGE_TASK_ID}p ) &&
mkdir featureCounts/$input &&
cd featureCounts/$input &&
```

```
/sperez/masterBioinfo/tools/subread-1.6.4-Linux-x86_64/bin/featureCounts -a
/sperez/masterBioinfo/tools/subread-1.6.4-Linux-
x86_64/annotation/Drosophila_me
lanogaster.BDGP6.95.gtf -o gene_assigned -R BAM
/sperez/masterBioinfo/01_datasets/MARS_Seq/mapping/$input/Aligned.sortedB
yCoord.out.bam -T 4 &&
```

```
samtools sort Aligned.sortedByCoord.out.bam.featureCounts.bam
assigned_sorted.bam &&
samtools index assigned_sorted.bam.bam &&
```

```
cd /sperez/masterBioinfo/01_datasets/MARS_Seq/
```

```
exit
```

## 09\_UMI\_quantification

```
#!/bin/bash
```

```
#$ -N umi_toolsCount
#$ -cwd
#$ -j y # Merge error/output logs
#$ -o logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=6G,h_rt=6:00:00 # ask for memory and time as needed
#$ -t 1-384 # substitute n by the number of files you have
```

```
# Input files
```

```
input=$( ls /sperez/masterBioinfo/01_datasets/MARS_Seq/mapping/ | /bin/sed -
n ${SGE_TASK_ID}p ) &&
mkdir count/$input &&
cd count/$input &&
```

```
~/local/bin/umi_tools count --per-gene --gene-tag=XT --per-cell -l
/sperez/masterBioinfo/01_datasets/MARS_Seq/featureCounts/$input/assigned_
sorted.bam.bam -S "$input"_count
s.tsv.gz &&
```

```
cd /sperez/masterBioinfo/01_datasets/MARS_Seq/ &&
```

```
exit
```

## 10\_Generate\_expressionMatrix

```

#!/bin/bash

#$ -N generate_matrices
#$ -cwd
#$ -j y # Merge error/output logs
#$ -o logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=6G,h_rt=2:00:00 # ask for memory and time as needed
#$ -t 1-384 # substitute n by the number of files you have

# Input files
input=$( ls /sperez/masterBioinfo/01_datasets/MARS_Seq/mapping/ | /bin/sed -
n ${SGE_TASK_ID}p ) &&
cd matrices/ &&

# Run
zcat
/sperez/masterBioinfo/01_datasets/MARS_Seq/count/$input/"$input"_counts.tsv
.gz | grep -v "gene" | awk '{print $1, $3}' > "$input"_counts.tsv &&
cat /sperez/masterBioinfo/tools/genes2.txt | join - -a1 -a2 -e "0" -o "0,2.2"
"$input"_counts.tsv > "$input"_counts_allGenes.tsv &&
rm "$input"_counts.tsv &&
cd /sperez/masterBioinfo/01_datasets/MARS_Seq/ &&

exit

```

## 11\_Join\_matrices

```

#!/bin/bash

#$ -N joinMatrices
#$ -cwd
#$ -j y # Merge error/output logs
#$ -o logs/$JOB_NAME.$TASK_ID.log # folder for the logs
#$ -l virtual_free=6G,h_rt=2:00:00 # ask for memory and time as needed

# Run

for i in $(cat samples.txt);
do
    join $i tmp.tsv > tmp2.tsv
    mv tmp2.tsv tmp.tsv
done

exit

```