POLITECNICO DI TORINO

MASTER's Degree in PHYSICS OF COMPLEX SYSTEMS

MASTER's Degree Thesis

Study of cellular heterogeneity of mouse cerebral cortex, through joint scRNA-seq and scATAC-seq analysis, derived from SNARE-seq technique.

Supervisors

Prof. Stefano DI CARLO

Dr. Roberta BARDINI

Candidate

LORENZO MARTINI

December 2020

Summary

Single-cell RNA sequencing analysis is part of Next Generation Sequencing (NGS) and allows investigating the gene expression profile of thousands of cells simultaneously. Through this experiment, one can study the cellular heterogeneity and try to find new rare cell types. Cellular heterogeneity analysis and cell-type identification are open challenges in this context because the current bioinformatic analyses focus on a machine learning approach to process transcriptomic data, which lacks in the reliability of the results, mainly due to the impossibility of a strong validation. In this work, we propose a computational approach for investigating cellular heterogeneity based on the study of multiple biological information simultaneously. The fundamental idea is that looking at various levels other than gene expression can help to have a broader view of the whole biological mechanism that identifies the cell. In this regard, we found a recent technique called SNARE-sequencing, which, from the same sample of cells, provides not only scRNA-seq but also single-cell ATAC sequencing. ScATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is an epigenetic analysis technique to assess genome-wide chromatin accessibility i.e. allows studying the chromatin state and the accessibility of the genes. The two pieces of information describe different biological cellular mechanisms but are complementary, thus the general hypothesis is that the joint analysis of transcriptomic and epigenetic data can help the cellular heterogeneity study.
We utilized the dataset provided by the SNARE-seq to work. It consists of a collection of 10309 cells from samples of adult mice brain cortex. To process the dataset, we employed two well-known pipelines, Monocle and Seurat. The first step consists of performing the analysis of the gene expression separately. Using firstly Seurat, and then Monocle, we elaborated the data to obtain a classification based on unsupervised clustering machine learning algorithms. This describes the usual analysis performed for a scRNA-seq experiment and represents what we want to improve or at least validate through epigenetic data. Using what the SNARE researchers found during their studies as a reference, we obtained a total of 21 clusters. These are what the algorithm suggests to be different cell types, even if, at this point, there is no information about biological cell type. As a common

proceeding, we performed a differential analysis of the gene expression, meaning finding which genes present different expression patterns between clusters. These genes are what characterize a group of cells and, more precisely, a cell type.
Next, we followed the same processing with the accessibility data through Cicero and Signac, companion packages of the previous ones. This step aims to understand how one can manage epigenetic data and use them to study cellular heterogeneity. While the clustering process appears to be informative and with results similar to the previous one, the differential analysis is not effective. In particular, the function struggles to find features that are differentially accessible and reliably identify one cluster. Besides the similar studies done for the expression data, Cicero provides the possibility to estimate the co-accessibility score of the data and to find cis-regulatory networks (CCAN) that can be important to understand regulatory mechanisms like enhancer-promoter. The co-accessibility describes how peaks relate to each other, meaning it identifies the peaks that are accessible with the same patterns. It will allow focusing on the regulatory network that controls the expression inside the cell. Signac, instead, provides tools for motif analysis and especially ways to integrate scATAC-seq data with scRNA-seq.
Before proceeding to the joint analysis, it was necessary to establish a reference classification of the cells, to compare the unsupervised cluster partitions, and make sure that the algorithms were recognizing cellular heterogeneity and not some other features. We performed an independent classification of each cell through the exploration of the expression of known markers. To do so, we assessed the expression level of sets of genes we derived, firstly, by the gene suggest by the SNARE researchers, then from a study of literature marker through the DropViz platform. The results showed how the final classification makes clear that the unsupervised clustering is identifying cellular heterogeneity, and we have been able to label the clusters with cell types. To evaluate the consistency with the clustering results, we calculated the Normalized Mutual Information, which gives a value between 0 and 1, indicating how close the classifications are.
After the separate analyses, the study focused on the correlation between the results, trying to understand the relations between expression and accessibility of notable genes, like cluster markers. The first approach starts from the overlap of the classifications derived from the separate dataset to find the sensible differences in the cluster partition. In this way, one can label cells with the cluster results of one of the two datasets while visualizing using UMAP dimensional reduction based on the other. This showed how the overall classifications agreed, with some differences in some subdivisions. In particular, epigenetic data appear to not properly divide subtypes within specific groups, like Oligodendrocytes or Inhibitory neurons. It means that accessibility data recognize cellular heterogeneity at a more general level than gene expression.
The second approach has been to create a gene activity matrix from the accessibility

data. There are two approaches to generate the latter. The first is based on the assessment of the accessibility of promoter regions of the genes. The second takes into account also the accessibility of co-accessible peaks through what we previously obtained with Cicero. The gene activity matrix allows studying the overall accessibility of a gene, and therefore link the epigenetic data directly to genes, instead of looking only at peaks. Through this analysis, we have been able to show how gene expression and accessibility are related, but, specifically, we determined that some features characterize the same groups of cells both at the transcriptional and epigenetic levels. The latter is not a trivial statement and proves how the joint analysis can help to validate the results of the clustering process of the gene expression through epigenetic data. This is a starting point for future works that could aim to study cell type markers not only at a transcriptional level but also at an epigenetic one.
In conclusion, multimodal sequencing is promising for cellular heterogeneity studies, in particular, we showed how the joint analysis of scRNA-seq and scATAC-seq can help in this field.

Acknowledgements

First of all, I want to sincerely thanks my supervisors Prof. Stefano Di Carlo and Dr. Roberta Bardini, for offering me this thesis and always be kind and available to help me out whenever I needed it. I am grateful to have been able to work in this field, even if I was unfamiliar with it. In this regard, I want to thanks again Dr. Bardini, that spent her time teaching me all the necessary biological background. Despite the distance due to the warring situation, she always has been supportive and encouraging, pushing me to do my best, helping me through every aspect of this work.

I also want to thank all the Professors that I met during these academic years. They always have been helpful and prepared me for the future.

I want to thank my parents, who have supported me throughout all these years and have encouraged me to be the best version of myself. And finally, to my friends, for being the great people they are, who helped me during hard times and have always been the best friends I could wish for.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Next-generation sequencing (NGS), or massively parallel sequencing, is a term to indicate various sequencing technologies, which have revolutionized genomic research, thanks to their increase in throughput and accuracy. These technologies allow the sequencing even of the entire genome of an organism within a single day, with high accuracy. Moreover, in this last decade, there has been the rise of single-cell experiments, which implement the NGS technologies at a cell resolution, meaning, for example, one can obtain from a tissue sample sequences reads from each cell. The NGS techs are not only used to sequence DNA but also RNA. Single-cell RNA sequencing is becoming the central point of interest of many biological and bioinformatic studies, which aims to investigate the cellular transcriptomic profiles, to improve the understanding of cellular regulation, cell differentiation, but also cancer characterization and neuronal diseases[1]. Cellular heterogeneity analysis and cell-type identification are open challenges in this context because the current bioinformatic analyses focus on a machine learning approach to process transcriptomic data, which lacks in reliability of the results, mainly due to the impossibility of a strong validation. In this work, we propose a computational approach for investigating cellular heterogeneity based on the study of multiple biological information simultaneously. The fundamental idea is that looking at various levels other than gene expression can help to have a broader view of the whole biological regulatory state that identifies the cell. The epigenetics of the cell could bring a different view on the gene expression and regulation. Thus the general hypothesis is that the joint analysis of transcriptomic and epigenetic data can help the cellular heterogeneity study. To validate this key assumption, we worked on the first SNARE-seq dataset, which provides a set of cells with both the sequencing analysis. With that, the experimental design consisted of both separate studies of the datasets, but also the comparison and the joint analysis of the two, to understand if they come to the same results and therefore validate each other. In particular, we wanted to know if differentially expressed genes were

also differentially accessible in the same group of cells, validating, therefore, the classification.

The single-cell analyses are the starting point and a central element of this thesis. In particular, the work revolved around the bioinformatic analysis of the data produced by single-cell experiments. The informatic part helps to elaborate on these large datasets, and analyze the actual differences between cells' transcriptomic profiles, to study the cellular heterogeneity through machine learning algorithms. The work started questioning how to interpret the results of the elaborated data. In particular, there was a question about the ability to recognize cellular types reliably, given the results of the unsupervised clustering algorithms. In this regard, we investigated the field of multimodal single-cell experiments. The idea was to study the cellular heterogeneity, looking not only at the transcriptomic profile but also at some other cellular information. In particular, we employed epigenetic data. The underlying idea was to investigate the cellular heterogeneity on different biological levels in the hope of incrementing the biological information, and therefore improving the heterogeneity analysis. The main goal of this thesis has been to understand if the joint analysis of transcriptomic and epigenetic data can help the cellular heterogeneity study. To do so, it has been firstly processed the data separately to observe the unrelated results.In this way, the results are what one obtains during a typical scRNA-seq analysis, that is what we want to improve or validate through the multimodal study. The separate analysis of scATAC-seq data instead aims to look at what epigenetic data alone can say about the dataset. Then the idea has been to compare them directly on a qualitative level, meaning if the division of the cells performed through unsupervised clustering on the two datasets, separated them in the same way or at least similarly. This helped to assess whether the different data types recognize the same heterogeneity, and therefore validate on a general level the clustering process. Afterward, we looked for shared features to try to cross-validate the clustering classification. Since the epigenetic data work on peak and not through genes, we implemented something called gene activity matrix that allows studying directly the overall accessibility of a gene. Thanks to that, it has been possible to assess if groups of cells were identified by both the differential expression of a gene and also by its differential accessibility. In this way, one can validate a cluster through the accessibility of its differentially expressed genes. The results showed that the scATAC data can validate the gene expression results, increasing reliability, and posing a more solid base for future and more specific cell-type analysis.
But before starting with the proper core of the thesis is significant to establish a background. Therefore in this introductory chapter, it will be discussed the experimental context, the biological theory, the dataset research, and the study of state of the art. It will follow a summary of the chapter to better navigate through

the work.

1.1 Biological background

Everyday biology researchers come across complex biological organisms, which are composed of an incredible ensemble of highly specialized tissues. Biology teaches us that different tissues are formed of very different cells that perform a great variety of functions. These cells are divide into types, like neurons or blood cells, and present distinct structures, so it is easy to divide them apart. However, within these general cell types, the cells appear to be highly differentiated to perform specific functions inside the tissue. But unfortunately, to identify and classify these different sub-types is challenging because they do not present sensible phenotypical variations. The cellular heterogeneity studies aim indeed to find ways to discover new cell types and better understand the ones already known. But what makes cells different?
Given an organism, all the cells inside it possess the same genome, meaning that they share the DNA sequences. The difference is which parts of it are effectively used by the cell. Said in a better way, part of the long DNA sequence is composed of the so-called genes. Genes are particular sequences of nucleotides that encode for the synthesis of genomic products like proteins, but also structural RNA. The flow of information from DNA to the products follow the important Central Dogma of Biology. Given the gene sequence stored in the long DNA molecule in the nucleus, it is transcripted, and a molecule of messenger RNA (mRNA) is synthesized. The mRNA is then taken out of the nucleus and is translated into proteins through the ribosomes. The proteins are the final product and are important because they carry out the functions of the cells. So to summarize, what differentiates the cells is their functions, which are carried out by the proteins that derive from the mRNA molecules which are transcriptions of the genes inside the DNA. Therefore a way to study cellular heterogeneity is to investigate the content of all mRNA molecules inside cells.This means to study the gene expression since different types will express (i.e. transcribe into mRNA) different sets of genes and also in varying quantities. However, this type of analysis it is not trivial since gene expression is a highly regulated process with a various complex biological mechanism. First of all, at the DNA level, there are epigenetic processes that control the accessibility of DNA regions. Other than that, there is an incredible network of regulatory proteins that can enhance or block the transcription of genes. Moreover, the RNA molecules outsides the nucleus are not translated into proteins in 1:1 manners, but from one RNA molecule the cell can synthesize several copies of the protein. So in conclusion the gene expression is a key factor in cellular heterogeneity. For this reason, access the information about the gene expression profile is fundamental to

advance in this field. Therefore the technological advancements in the RNA-seq have been crucial to the development of this biology branch. For this reason, it is useful to review the technical aspect of the scRNA-seq.

1.2 ScRNA sequencing

The RNA sequencing indicates the techniques with which the mRNA molecules are collected and sequenced. The first step is the preparation of the sample. For single-cell investigations, the chosen tissue is separated into its constituent cells allowing, therefore, the high resolution at the cellular level distinctive of the technique. This can be done following different methodologies, including mechanical and micromanipulation through pipette or nanotube. But lately, it has gained popularity the microfluidic technology that enables high-throughput single-cell profiling of even tens of thousands of cells, with high capture efficiency and a reduced cost. With this method, cells are separated through microfluid droplet manipulation and matched with microbeads, which are then employed to identify cells uniquely[2]. After the separation, cells must be first of all lysed, allowing the capture of the RNA fragments that were present inside them. The mRNA sequences must then reverse-transcribed into first-strands of cDNA (complementary DNA), a step which is necessary since DNA molecules are much more biologically stable and resistant to degradation. The resulting cDNA undergoes the preparation for sequencing, meaning fragmentation and barcoding, so one can trace back the cell the fragments came from, and subsequently, they are amplified through PCR or in vitro transcription, and finally, specialized adapters are ligated to the ends of each piece. The barcoding is done through the so-called unique molecular identifiers (UMI) short sequences that tag the fragments and help to reduced errors and biases due to amplification[2]. The sequencing process is carried out by specific platforms like Illumina[3]. The result is a large dataset of short reads that needs to be mapped to its appropriate location on a reference genome, a process called sequence alignment. The file can be left in this raw version, or it can be processed to obtain a matrix format, where genes constitute rows, cells constitute columns, and values within the matrix are read counts representing the expression of a particular gene in a cell. This final product is the starting point for the bioinformatic analysis.
The latter review is a brief explanation of the experiment that can be performed using several commercially-available protocols, like Chromium from 10x Genomic[4], which allows obtaining large scRNA-seq datasets. The latter is one of the most used and provides also the processed file in a format that can be directly and easily input in the pipelines.
The central point of this study, however, is the bioinformatic analysis of the output[5]. This includes all the data elaborations to interpret the biological

information resulting from the experiments. The usual points are:

- Low dimensionality visualization: the gene expression matrix can be seen as a set of M points in an N-dimensional space, where M is the number of cells (i.e. number of rows) and N is the number of genes (i.e. number of columns). One can visualize these points in a 2-dimensional representation, through the use of appropriate algorithms, such as t-SNE and UMAP.
- Clustering: the clustering process aims to separate the cells into groups based on differential gen expression. This should represent the partition of the cellular heterogeneity.
- Differential expression: after the clustering, one can find which are genes differentially expressed between clusters and then define the markers that characterize each of them.

During this work two very well-known pipelines, Monocle 3 and Seurat, were mainly used, offering a wide range of useful functions, to perform such analyses. They are not the only ones available but are the most complete and widely adopted in this field. There is no one better than the other since they allow for different investigations of the data, meaning it is helpful to work with both and try also to identify possible differences.

1.3 Datasets

Even if it may seem superfluous, the analysis of the available Dataset landscape is quite important because from it one can choose which direction to take based on the quantity and quality of the accessible data. Therefore it was really useful to group various Datasets and categorize them according to certain criteria:

- Organism: ScRNA experiments are performed on a wide range of more or less complex organisms. The choice must be made thinking about the fact that it is better to consider a well researched and known organism, so one can compare possible results with the literature already present.
- Tissue: this is relevant especially for more complex organisms like vertebrates where different tissues mean different cell types. One can find also some large datasets which include different tissues but when researching cellular heterogeneity is better to focus only on one.
- Age/time point: it mainly differentiates between adult organisms i.e. fully developed, and embryonic stages. The latter is useful to study the pluripotent cells and their evolution using the pseudotime analysis.

- Temporal points: linked to the previous point, it indicates the multiple temporal points at which samples were obtained. They can be connected to embryonic stages or to times after artificial manipulation of some cell culture.
- Number of cells: self-explanatory, it is helpful to choose datasets with a great number of cells.
- File type: each dataset is available in different formats, mainly divided into RAW and processed formats. The first ones are the raw outputs of the experiments (that needs pre-elaboratation before using them in the pipelines), while the second ones are different formats of expression matrices derived by the RAW data.
- Project/study: it is helpful to connect the data to the projects or studies which implemented them, so one can compare with their results and also understand the limitation of their analysis.

As far as search is concerned one can follow mainly two strategies.
The first is to use databases and platforms that provide various datasets. One example is the platform PanglaoDB[6] a database that makes available datasets already organized by species (mainly Homo Sapiens and Mus Musculus), tissue, and the number of cells. One problem with the latter is the dimensions, in terms of the number of cells of the data listed, in the sense that only a small number of these exceed a thousand cells and therefore are suboptimal for broad cellular heterogeneity analysis. Another helpful tool is the webpage of the Brain Atlas project of the Allen Institute[7]. This project aims to investigate the taxonomy of mouse and human cell type at a deeper level, using not only transcriptomic data but also morphological and electrophysiological ones and therefore obtain a more extensive view of the cellular heterogeneity. In an ideal scenario, one would like to have both the transcriptomic and the morphological data of the same cell in such a way to pre-assign a cellular type on a phenotypical base and afterward analyze the gene expression, but this is still not possible and there are only separate datasets. Anyway, the Allen Institute provides a good number of different datasets which include samples from several neuroanatomical areas. Generally speaking, the brain samples data are interesting because they contain a great variety of cell types that may differ in functionality based on position or area of the brain. Moreover, they are used in the study of some neurodegenerative diseases through the study of the gene expression profile of particular cells. The Allen Institute works closely with the NIH's BRAIN Initiative Cell Census Network (BICCN) whose goal is to "generate comprehensive 3D common reference brain cell atlases that will integrate molecular, anatomical, and functional data for describing cell types in mouse, human, and non-human primate brains"[8]. The latter also provides several transcriptomics, epigenomics datasets, morphology, and connectivity data of mouse, human, and

primate brain, that one could also easily use in addition to the previous ones. Therefore, in general, is useful to look for projects and researches that focus on a precise area and provides several data not only on gene expression but also other analysis that ideally could then help in some cross-validation.
The second strategy is to look for published articles about the topic and look at the availability of the data used in that study; except for rare cases (especially in the case of brand-new papers), these are always obtainable. In the most common case, the article will provide under the heading "Data availability" the number or the link to the Gene Expression Omnibus database (GEO)[9]. The latter is part of the National Center for Biotechnology Information (NCBI)[10] which provides access to biomedical and genomic documents, libraries, data, projects, and researches. The GEO section is a public functional genomics data repository that archives next-generation sequencing data submitted by the research community. One can browse directly the repository, but it is not recommended due to a lack of convenient filters enabling searches based on the parameters previously listed. As mentioned the best thing to do is, while one is studying the state of art, take the accession numbers of various papers and find directly through it on the GEO database. On the page of a dataset, one can find useful information from the experiment type to a summary of the work done and its overall design. The most important part is the data section; here one can find the sections:

- Samples: usually there is not only one experiment, but there are several that differ for example for the time point of the sample or different conditions.
- SRA run: the link to the Sequence Read Archive (SRA) where one can download the unprocessed data i.e. the sequence reads, output of the experiment.
- Supplementary file: all the processed files given by the researchers, like expression matrices or metadata files, in different formats with the explanation of how they were been obtained. Because which supplementary files are available it is up to who published the data, one can find a great variety especially about the formats and the type.

The raw data are always available but they can not be used directly, they need to be processed. However, this step, even if one can do it using suitable packages, can be long and time-consuming, so it is better to use the already processed data.
Now let's spend some words on what we found. As previously mentioned a good portion of found datasets arrive from the Allen Brain Atlas; they derive from different portions of the human and mouse brain and are sampled from different time points, that makes it a good collection to use for different analysis. It is worth mentioning the fact that the experiments from which the data derive, follow a new approach that involves only the nuclei of the sample cells, instead of the whole cell. This is part of new techniques that aim to improve the single cell sequence

analysis and will be discussed in more detail later. Other datasets were collected from the publications of the already mentioned pipelines Monocle and Seurat. These are available on the respective sites. For the first one, the publications concern mainly the pseudotime analysis and trajectories construction, and use as the primary dataset a sample of primary human myoblast "as a model system of cell differentiation to investigate whether ordering cells by progress revealed new regulators of the process"[11]. This is a collection of hundreds of cells taken at different moments of a serum-induced differentiation, and it is the base dataset for all the pseudotime publications. In addition to this, they also employed datasets of mouse lung epithelial cells at 4 different embryonal stages and one adult stage. It is also worth mentioning the latest paper from the Monocle platform that includes a huge dataset of several mouse embryos staged during the so-called organogenesis, when the three germ layers formed from gastrulation differentiate in the different organs of the mouse, providing in this way a 'mouse organogenesis cell atlas' (MOCA) allowing the study of this developmental processes at a transcriptional level[12]. These datasets focus on the pseudotemporal analysis and could be used for similar studies.
To conclude, one can obtain a huge variety of datasets with different characteristics. The choice depends on the typology of study one wants to perform.

1.4 State of the Art

Before starting with any type of study it is really important to understand the related problems and limitations, the current state of research, and the possible innovation in the field. Therefore it is important to do a careful study of the State of the Art, focusing on the newest advancements.

One open question is related to the reliability of the cellular classification performed by the algorithms. In particular how one can ensure that the clustering process is truly recognizing cellular types and not something else, and how can one cross-validate this categorization. The principal issue is the lack of datasets that provides not only the transcriptional data but also a classification based on other analyses like for example morphological ones[13]. This would incredibly help the study because with a base classification one could ensure the quality of the clustering algorithms, at least of the major cell types, and consequentially focus on the identification of rare and unknown cellular types. However such pre-classified datasets are not available mainly due to technical difficulties. While the technologies for the gene expression experiments, as previously mentioned, allows working with several thousands of cells, morphological analysis can not keep up with such high-throughput experiments due to the necessity of user inspection. Nevertheless, there are alternatives, one can use datasets with classification based on known markers.

However, this brings out other issues because those markers are considered at the protein level, and could not match the gene expression. Unfortunately, as already explained in section 1.1, the connection between transcription and translation is not a direct 1 to 1 correlation, meaning that one can not simply think every mRNA molecule is translated into one protein, and therefore it is complicated to establish a relation between the abundance of protein molecules and the effective gene expression levels. This is the reason why classification based on proteins is not completely reliable because it is based on the preknowledge of some markers (i.e. genes/proteins that from previous literature that can be used to identify some cell types) of certain cell types, and the presence of those markers is enough to give an identity without further inspection, even if as we said the gene expression-protein relation is not trivial, and could lead to misleading classification.
Summarizing one would like to have datasets with a classification based on other characteristics, which however is not always possible, and even when there is one it may not always be reliable, so one has to find other solutions to cross-validate or at least define a ground point to measure the quality of the analysis.
The field of all the Next Generation Sequencing (NGS) which includes all the sequencing at single-cell resolution, is incredibly dynamic and keeps improving and changing every day. Therefore it is useful to find and understand which are the very latest innovations, and figure out if there are possible solutions to the current problems or different approaches not taken into account yet. Fortunately, the Satija Lab[14], the portal which provides the Surat pipeline, works closely with the different aspects of the NGS and offers every year a little conference called "Single Cell Genomics Day" that recaps all the latest and most interesting innovations and researches. It focuses mainly on the newest and most promising researches of the year, most of which are not published yet. Below there is a review of all the studies discussed, with a commentary on what caught the attention and fueled this thesis work.
Let's start with "Multiplexed human genetic studies". One might want to understand human genetic variation and to do so has to use samples from various individuals, but how one can recognize from which individual the cells come from. The first approach is to look for a single nucleotide variant to distinguish the original genotype and use it as a second barcode. The main disadvantage is the need to previously genotyping all the individuals to identify the variants and this could be costly. New approaches overcome this problem, like "Vireo"[15] that enables the identification of different genotype without the pre-analysis of them, reducing time and cost of the analysis. Similar to the latter, other examples are "souporcell"[16] and "scSplit"[17] which autonomously identity candidate variants, and create statistical models to then uniquely identify all the cell genotypes. One can see its application to a fetus sample where these pipelines successfully recognize the mother cells as different genotypes.

The next one is a "Statistical model for scRNA-seq data" and analyze the presence of zeros in the gene expression matrix. Due to the nature of the experiment, the matrix is extremely sparse with a lot of elements that are zeros, and the study "Droplet scRNA-seq is not zero-inflated"[18] shows how the number of zeros for gene closely fit a negative binomial distribution, that allows to better understand and statistically analyze the expression matrix, for example implementing tailored PCA (as in the case of the GLM-PCA) that take in consideration this aspect, minimizing the technical bias.
There are also innovations in the experimental sensitivity, meaning new techniques that improve the detection of mRNA molecules, but it is strictly related to the technical side, so will not be explained further.
Another interesting study is the one related to "Cross-species alignment of single-cell data"[19], that through the comparison of scRNA-seq experiments of related species try to infer new cell types and understand how certain cell changes between different species. One related work[20], explore the different neuron population of mouse and marmoset and is able to identify an interneuron subtype only present in the marmoset cortex.
Interesting is the research on "Predicting cellular interactions" that try to infer from the gene expression how the cells interact with each other through the analysis of the expression of the genes that encode for ligand/receptor proteins. The "cellphoneDB"[21] is a related project that identifies cases where both receptor and ligand are cell-type-specific and use a statistical framework supported with ligand/target known links to predict interactions and cellular communication network.
The last research that has been taken into consideration is the one which then has been taken as a start for this work. Before talking about it, it is helpful to explain one thing; it has been profoundly discussed the scRNA-seq, but the latter it is only one of the NGS techniques, that is the one that targets the RNA molecules, however, one can also sequence the DNA and the whole genome, or analyze the epigenetic of the cell. Epigenetics is the study of heritable phenotype changes that do involve alterations in the DNA sequence, or in other words the changes in the chromatin state that influence the gene accessibility and therefore have a role in the gene activity and expression. Hence is fair to say that one could use the information in one methodology to improve the analysis and interpretation of the other, taking into consideration that the two processes work on very different time scales. This is the focal point of the last work presented at the "Single Cell Genomics Day" which proposes a massively parallel, simultaneous, multi-omics sequence analysis under the name of SNARE-seq[22]. The latter is a high-throughput experiment that enables the joint capture of DNA and RNA molecules with shared barcodes so that one can obtain double information from the same cell. In detail, the SNARE-seq is a single-cell high-throughput experiment, combination of scRNA-seq and scATAC-seq.

The scATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is an NGS technique whose tack is to assess the accessibility of the chromatin of all the genome. It is a faster and more sensitive epigenetic analysis than similar methods like DNAse-seq. The process makes use of hyperactive mutant Tn5 Transposase which insert sequencing adapters into open regions of the chromatin probing all the genome. The DNA fragments that are tagged are then purified and amplified through PCR, from which one identifies the features which will form the count matrix, that in this case are called peaks. The process that defines the features is called peak calling and does not simply call all the fragment regions, but take into account all cells, align all the fragments to the genome, and identifies the regions with a high number of fragments through pipelines like MACS2[23]. Therefore the final result is a matrix similar to the gene expression matrix, where the columns are the cells and the rows are the peaks which are usually written in the form chN:pos1-pos2 where chN indicates the N-th chromosome and pos1-pos2 are the starting and ending base pair position inside the chromosome. This matrix can be treated similarly to the gene expression one, through appropriate pipelines like Cicero and Signac, which are, respectively, portions of Monocle and Seurat.
The multimodal technique here presented has great potential, because allows having both gene expression measures and gene accessibility from the same set of cells, so one can directly see how the analysis of the two information could give different results. One could study the cellular heterogeneity obtained from the two and analyze how they are related, in particular, how the accessibility of a gene influences its expression.
In conclusion, after studying the state of art and the issues with the validation of classification, it has been decided to focus on the SNARE-seq a new multimodal technique that provides a dataset of cells with both scRNA-seq and scATAC-seq data thus giving a wider sight on the biological processes the occur inside the cells. This work will aim to separately analyze the data, overlap the results, compare them with a classification based on literature markers, and try to infer the relations of expression and accessibility. It will be used the dataset provided by the paper related to the SNARE-seq, which consists of 10309 of cells from adult mouse cortex samples.

1.5 Chapters summary

After laying the foundations with this introduction is time to describe the work done.

In chapter 2 we started with the transcriptomic data analysis. We processed the data with both Monocle and Seurat. This reflects the commons work done on scRNA-seq datasets. In this way, we obtained the first division in clusters and

performed the differential analysis.

In chapter 3 we followed the same workflow to study the scATAC-seq data. In this case, the features are the peaks, small regions of the genome. This part aims to try to cluster the cells independently from the expression information, to then try to understand how the accessibility data influence the cellular heterogeneity alone. However, when we tried a differential analysis of the accessibility, the peaks features seemed not to effectively describe the differences between the unsupervised clusters.

In chapter 4 we performed an independent cell-type classification of the dataset. The goal was to obtain a ground reference to compare the clusterings and to have a first biological ground to then make a hypothesis on cell types. We employed two methodologies. One analyzing literature markers, two with transfer label technique from a previously classified dataset.

In chapter 5 we considered all the previous analyses together. The main goal is to understand if the clustering made with the different datasets are consistent with each other, meaning if the cells are clustered in the same ways despite the different information provided. This first step allows validating the expression results with accessibility data. However, the peaks accessibility is not trivial to study differentially, so we implemented the gene activity matrix. With it, it was possible to analyze the activity (i.e. the overall accessibility) and compared it to the expression of marker genes and differentially expressed genes (DE). This because one wants to find if DE genes for a cluster are also differentially active, and thus validate, on an epigenetic level, the cluster identity.

Now, after the introduction, it is time to start with the explanation of the work.

Chapter 2

Gene Expression Analysis

After the study of the state of art, it is time to start to work with the data. In this chapter, the focus will be on the gene expression analysis, so it will include an explanation of the usual workflow previously mentioned. Here we present the process to obtain a 2-dimensional visualization of the cells clustered together. This represents the usual work done during a typical scRNA-seq analysis, and it is what this work wants to improve on. As already said the two pipelines mainly used are Monocle and Seurat, where there is no one better than the other, since they work differently and have different strengths, like for example, Monocle provides pseudotemporal analysis[24][25] while Seurat has better differential analysis and QC metrics functions. The dataset used comes from mouse cortex samples; in particular, there are available two datasets, one from an adult mouse (2 months) and one from a mouse just after the birth (postnatal day 0, P0). Since for the time being, there is no particular interest in a pseudotemporal analysis, we adopted the adult mouse sample. The latter consists of a collection of 10309 cells, where RNA libraries were prepared for sequencing using standard Dropseq protocol[26]. All the necessary material is available on the GEO database with the accession code GSE126074, where are provided three files, the sparse matrix with expression counts values, and two tab-delimited text files showing barcodes and gene names. These are the starting point of the work and will be the input of the pipelines. The chapter is divided according to the pipeline used, so will start with Seurat and then Monocle. It is worth mentioning that Seurat has been recently updated, improving some of its aspects and, therefore, will be proposed the results with both versions.

2.1 Gene expression analysis using Seurat pipeline

Seurat is a package for R, developed and updated by the Satija lab [27][28], which offers a great variety of functions for the exploration of single-cell RNA-seq data,

in particular, helps to interpret the cellular heterogeneity based on gene expression and offers ways to integrate with other types of single-cell data. It moreover provides a variety of tutorials aimed to learn the basics and the most relevant functions that are used for the expression analysis. One of these tutorials was initially taken into account mainly to understand how to set the values of some function's parameters. Now let's talk about the actual workflow. First of all, one needs to load the data; as previously mentioned the dataset is provided with three files, the matrix content in a sparse format and two separate text files which represent the barcodes, i.e. the cells, that identify the columns of the matrix, and the gene names that, instead, identify the rows. Given the high sparsity of the gene expression matrix, the latter is often given in a sparse format to save memory space, but it is necessary to transform it in a dense format. It is very simple to do so with Seurat, which provides the function "Read10X" that takes as input the three files and returns a unique molecular identified (UMI) count matrix. The elements of the resulting matrix are predominantly zeros and the values represent the number of molecules for each detected gene. With the count matrix, one can create the Seurat object through the function "CreateSeuratObject" which take as an input the latter and some optional parameters, like "min.cells" (Include features detected in at least this many cells) and "min.features" (Include cells where at least this many features are detected). The Seurat Object is the fundamental element of the pipeline and contains both the data and all the analysis results. In this way, one can easily store every different operation outcome did on the data and access them directly through the subparts of the Seurat Object. Once the SO is created, one can start the pre-processing workflow which includes the quality control (QC) and the normalization and scaling of the data. The QC metrics are used to ensure that the data do not have low-quality elements that can negatively affect the analysis[29]. The common metrics are:

- The number of unique genes per cell: cell with few detected genes are usually low-quality, but also a number too high may suggest an incorrect detection.
- The total molecules count per cell: as the previous point, a too high or too low count is not good.
- The percentage of reads of mitochondrial genome: extensive mitochondrial contamination usually indicates a dying cell.

The first two are automatically calculated at the creation of the Seurat Object, while the last one can be easily derived with the function "PercentageFeatureSet". The latter need a pattern to recognize, which in this case is "MT-" that identifies the mitochondrial genes. The dataset provided has been already purified of very low-quality detections, but it is always helpful to check anyways. As one can see from the Figure 2.1, each black point represents a cell, and there are no points

with too high or too low features (usually are considered too high if exceed 2500 features and too low if do not reach 200). The third graph shows how there are not

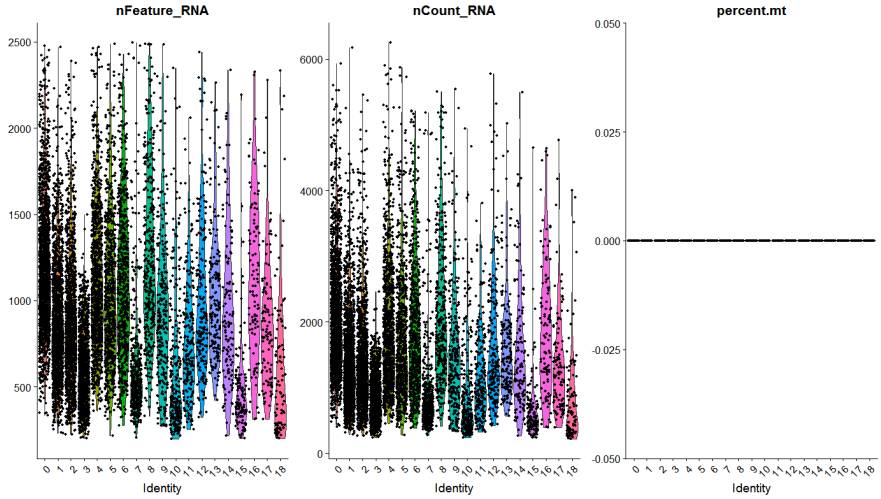

Figure 2.1: Violin plot of QC metrics for the RNA dataset

detected mitochondrial genes, showing the good quality of the dataset. It is whort mentioning that on the horizontal axis the elements called Identity are none other than the batches of the experiment. The next step after the controls on the data quality is to normalize them. Normalization is necessary "to remove cell-specific bias, which can affect downstream applications"[5].The function "NormalizeData" allow performing the normalization, which by default use a "LogNormalize" method, that takes the features counts for each cell, divides them by the total count for that cell, multiplies for the scale factor that is given as an input to the function, and finally log-transforms them. As an input one can set the normalization method and the scale factor that is usually set to 10000 as suggested by the Seurat tutorial. After that, the Seurat vignette explains how focusing on the high variable genes can improve the analysis of biological signals in single-cell datasets[30], therefore it is helpful to calculate and create a subset of features that highly variate between cells. To do so one can directly model the mean-variance relationship, using the "FindVariableFeatures" function[31], which takes as an input the method for the selection ("vst" is the default, but one can use also "mean.var.plot" or "dispersion") and how many features to return, which in this case was set to 2000, that is about a tenth of the total features. These selected features will be later used, especially for the dimensional reduction operation like PCA. Before moving to the dimensional reduction analysis, one has yet to scale the data (remember that before performing PCA one always must scale the data). The "ScaleData" function can do this, and performs two operations, first shift the expression of every gene so that the mean

value across cells is 0, and then changes it so that the variance is equal to 1. This process is important but can take a long time, so instead of rescaling on all the features, one can do so only on the subset derived from the previous step. After scaling the data it is time to perform the dimensional reduction. We performed the PCA where only the subset of variable features defined in the previous passage are taken into account, however, this is not mandatory but can be used any set of features by just defining the features argument of the "RunPCA" function. The latter function computes the first 50 principal components, but not all of them are very informative, and it is better to choose how many of them it is helpful to include. To do so, one can use a heuristic method that is an Elbow plot (with the "ElbowPlot" function), which plots the standard deviation of each PC (Fig.2.2). Once plotted one can see how the higher number of PC corresponds to a lower standard deviation until, between 20 and 30 components, this does not change much more. Taking the standard deviation as a measure of informativeness, it

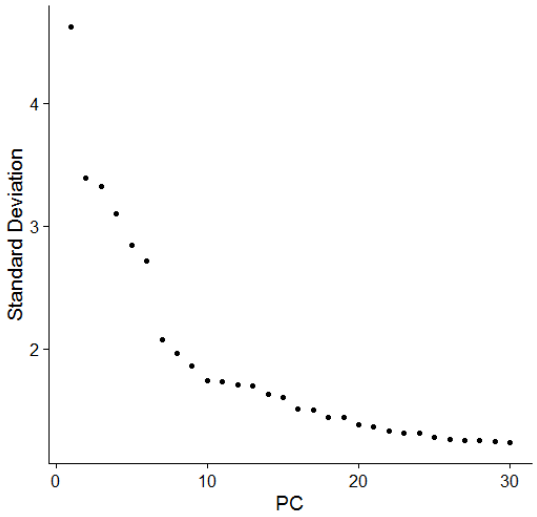

Figure 2.2: Elbow Plot of PCA components

makes sense to consider for the downstream operation, a number between 20 and 30 PCs. For the next steps, concerning clustering and low-dimensional visualization, where the functions require setting the number of dimensions to consider (i.e. the principal components), both the value (20-30) have been used showing how the differences, especially in the visualization, are minimal. If, on the other hand,we would have chosen a lower value, the results would be most affected.
As said, next is the clustering process. The Seurat v3 is based on the graph clustering approach[32][33]. The first step is to use the "FindNeighbors" which construct a K-NearestNeighbors embedding all the cells, using the distances in the PCA space, so that cells with similar expression are linked with edges. After the

creation of the graph, the function partitions the cells to optimize the standard modularity function, using the Louvain algorithm or SLM[34]. The "FindClusters" function performe that, and takes as an input the resolution parameter which influences the granularity of the resulting clustering[35]. The resolution greatly changes the results, in particular, a higher resolution will result in a higher number of clusters. Choosing the optimal value for this parameter is not trivial. The Seurat guidelines suggest that resolution "between 0.4-1.2 typically returns good results for single-cell datasets of around 3K cells. Optimal resolution often increases for larger datasets."[28]. In this case, where the dataset is composed of about 10K cells, it is then better to choose a higher resolution, but instead of considering only the result with one chosen resolution value, it has been instead repeatedly performed the clustering process with different values. The resolutions and the relative number of total clusters obtained are reported in the Table 2.1. From the paper concerning the SNARE-seq, one can see how the researchers identified on the same dataset a total of 21 clusters, so the resolution that gets closer is 1.3. For this reason, we considered the classification made with the latter value

Resolution	0.4	0.5	0.8	1.2	1.3
Number of clusters	12	13	19	20	21

Table 2.1: Number of clusters returning with different resolutions

as a reference, however, the other ones, have not been discarded but will be later compared with Monocle's results and with the classification made in the next chapter. Until now all the calculations have not produced visual outcomes, but one can try to plot the cells as points on a graph. After all, the gene expression matrix can be viewed as M points in N-dimensional space (M cells, N features), but first is necessary to reduce the dimensionality to plot them on a flat graph. To do so one has to run non-linear dimensional reduction such as UMAP or tSNE. The goal of this operation is not just plotting all the cells on a 2-DIM space, but do so in a meaningful way so that cells with similar expression profiles are grouped. In this way, the clusters determined above with the graph-based algorithms will localize near one another. Seurat provides different non-linear reduction methods, and the most used are UMAP (Uniform Manifold Approximation and Projection) and tSNE (t-distributed stochastic neighbor embedding)[36]. Even if both alternatives have good performances, lately UMAP method is increasingly common, due to the fact it is faster and better visualize the differences between clusters for RNA-seq datasets then tSNE[37]. In figure 2.3 one can see the two, with the relative cluster assignment. As previously mentioned we plotted the results with 20 and 30 dimensions, and one can see there are no great differences (the one with 20

dimensions is just flipped). It is useful to report also the case where all the previous functions have taken as input only 10 dimensions, where one can see noticeable changes (Fig. 2.5). There are a bunch of observations that can be done at a qualitative level.

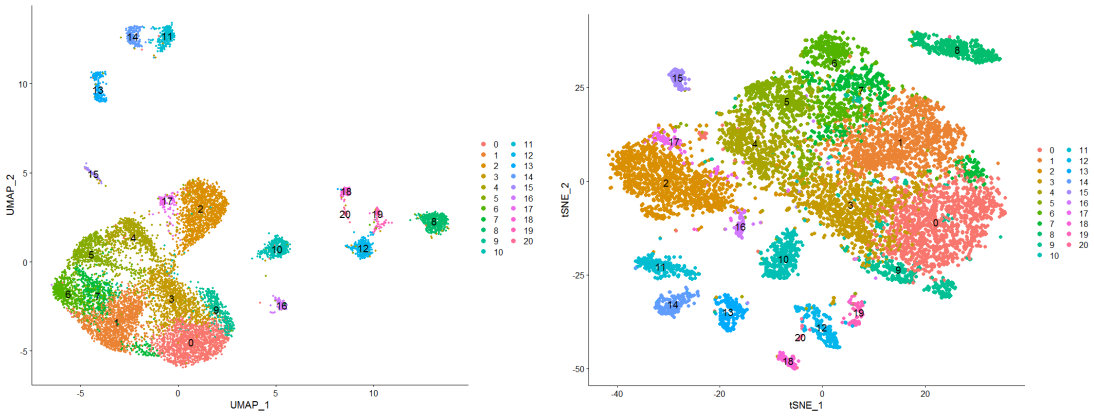

Figure 2.3: Seurat plot with both UMAP (left) and t-SNE (right), considering 30 PCA components

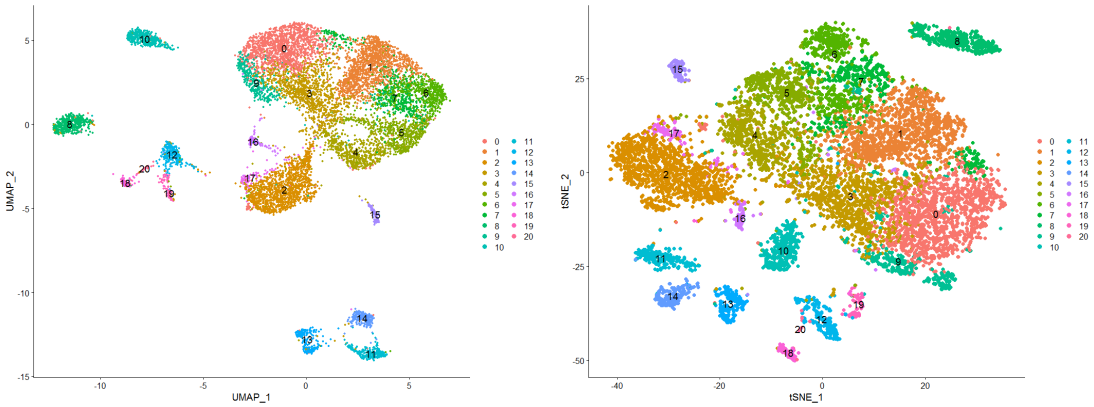

Figure 2.4: Seurat plot with both UMAP (left) and t-SNE (right), considering 20 PCA components

First of all, from the general notion of brain sample composition, it is expected to observe a major population of neurons, plus some smaller populations of non-neuronal cells such as astrocytes or oligodendrocytes. Looking at Fig. 2.3 one can make some hypotheses; the biggest portion of cells are togeher in this large cluster, which in turn is partitioned in different clusters, and other cells are well-separated in smaller groups. So is reasonable to assume that the large one will comprise the

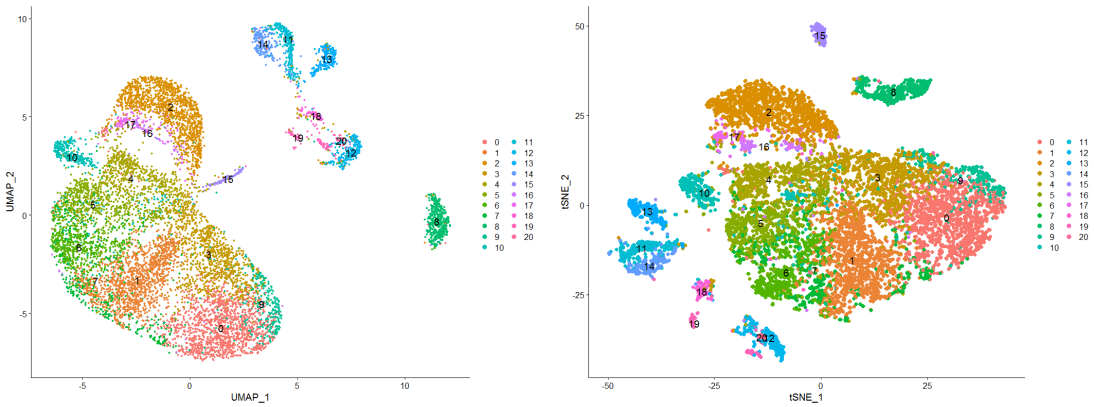

Figure 2.5: Seurat plot with both UMAP (left) and t-SNE (right), considering 10 PCA components

various neuron types, while the smaller ones will be some non-neuronal populations. From the same figure, it can be noticed how the different dimensions applied for the non-linear dimensionality reduction influence the result. Qualitatively, one can see how the cases with 20 and 30 dimensions seem pretty similar, the main difference it is the separation of the smaller groups, while in the case of 10 dimensions, there are less well-separated clusters. This is a consequence of the fact that more principal components (from the PCA step) can bring up relevant sources of heterogeneity, which translates into more diversity when clustering and performing UMAP. Therefore one might be inclined to include more PCs, but it has two limitations. First, too many PCs, when performing calculations, considerably increase the time of the operations, and second, can be a source of overfitting, meaning that the results are too linked to the dataset and are not able to generalize. So it is recommended to choose this parameter after some analysis, like the heuristic one previously mentioned, and, when in doubt, prefer the higher value. Therefore it has been taken as a reference, the results with 30 PCs.
Once the cells are partitioned and visualized, it is useful to understand what makes the clusters different from each other. The clustering algorithm works in such a way that divides the cells based on their expression profile, or in other words, whatever different genes are expressed or not between them. Seurat provides an easy way to perform differential expression test, through the functions "FindMarkers" and "FindAllMarkers". The first receives in input one or two identities generally clusters ID and identifies all markers (positive and negative) of the first argument compared with the second, or with all the other cells if no second argument is passed. The function returns a table with the following columns:

- Gene name.

- p_ val: the unadjusted p-value.
- avg_ logFC: log fold-change of the average expression between clusters taken into consideration. It can assume positive and negative values and indicates how much more the gene is expressed in the first cluster (negative values indicates the opposite).
- pct.1: the percentage of cells where the gene is expressed for the first cluster.
- pct.2: the percentage of cells where the gene is expressed for the second cluster.
- p-val-adj: Adjusted p-value with bonferroni correction.

When looking for good markers of one cluster compared to all the remaining cells, the negative values of avg_logFC are not relevant, so it is better to consider only positive ones by setting the "only.pos" argument of the "FindMarkers" to TRUE. It is worth to mention that looking only at high-valued avg_logFC genes it is not optimal, but it is better to take into account also the pct.1 and pct.2 values. A gene is a good marker when it is expressed in the majority of cells of the cluster considered and in a few other cells not belonging to it, meaning a high pct.1 value and a low pct.2. The "FindAllMarkers" perform the same calculation but automate the process for all the clusters. We run the latter on the dataset and the result is a list of 2770 possible markers for the 21 clusters. From them, it is easy to select the top 2 genes for each cluster for avg_logFC value. The best markers, as said before, are the ones with high avg_logFC value, high pct.1, and low pct.2. Below are reported the feature plot (through "FeaturePlot") and violin plot (through "Vlnplot") of some of them. Looking at them one can do some observations. First of all, consider the plot of the genes Scl1a3 (Fig. 2.6) and Plp1 (Fig. 2.7), which are, respectively, markers of clusters 8 and 12. These are some of the stand-alone groups that were supposed to represent populations of non-neuronal cells. These markers besides having high avg_logFC values respect the other conditions, and one can see from the feature plot how they well represent the origin cluster. This is a consequence of being clusters already well separated from the others, and therefore it is easier finding what makes them so different, or in other words, the optimal markers. Now let's focus on the Rorb gene, which is listed as a marker of cluster 1 (Fig. 2.8). The latter is part of the big agglomerate that we supposed to represent the big population of neurons. Therefore, defining rigorously the subtypes of this population becomes challenging. As one can see, especially from the violin plot, the Rorb gene is expressed in cluster 1 (exactly in 70% of its population), but also in a lot of cells of nearby clusters. Since these clusters have cells with similar expression profiles, it is hard for the algorithm to distinctly separate all the cells, and, therefore, it is challenging to find optimal markers. The last gene is Fam19a1 (Fig. 2.9), which shows how the "FindMarkers" function is not perfect. This gene

is listed as a marker of both the cluster 0 and 10, and one can already understand how this is not optimal. From both feature and violin plot, it is clear how Fam19a1 it is highly expressed in both these clusters, but also a lot of other cells. For this reason, this gene it is not defining the cluster, despite its high avg_logFC value. This also explains the need for a low pct.2. The last plot is the HeatMap of the top 4 genes for each cluster (Fig. 2.10), showing their expression abroad all of them. It is quite clear how the genes considered are highly expressed in its origin cluster, and not so much in the others.

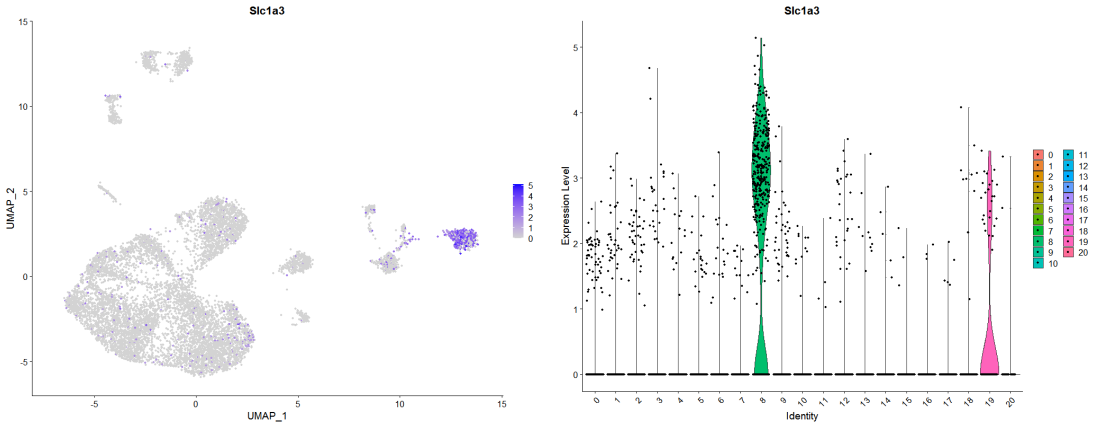

Figure 2.6: Feature plot and violin plot of Slc1a3

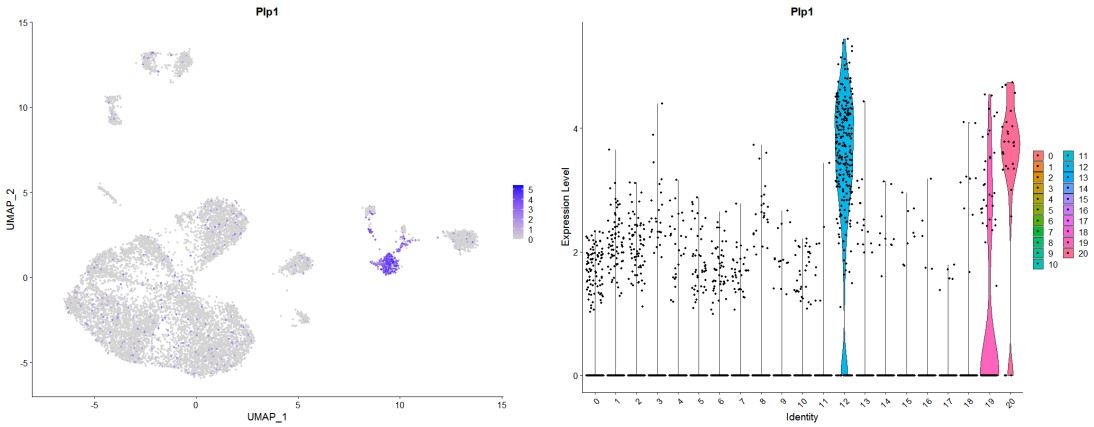

Figure 2.7: Feature plot and violin plot of Plp1

In conclusion, the differential expression analysis can help to find suitable markers for clusters of smaller populations but gives no clear results for similar subtypes, or even fails to find optimal ones. This shows how the correlation between cellular

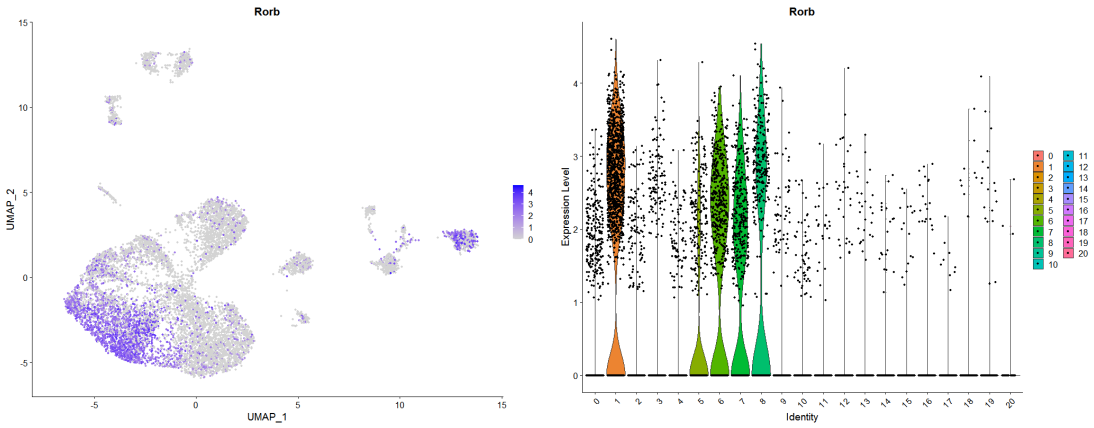

Figure 2.8: Feature plot and violin plot of Rorb

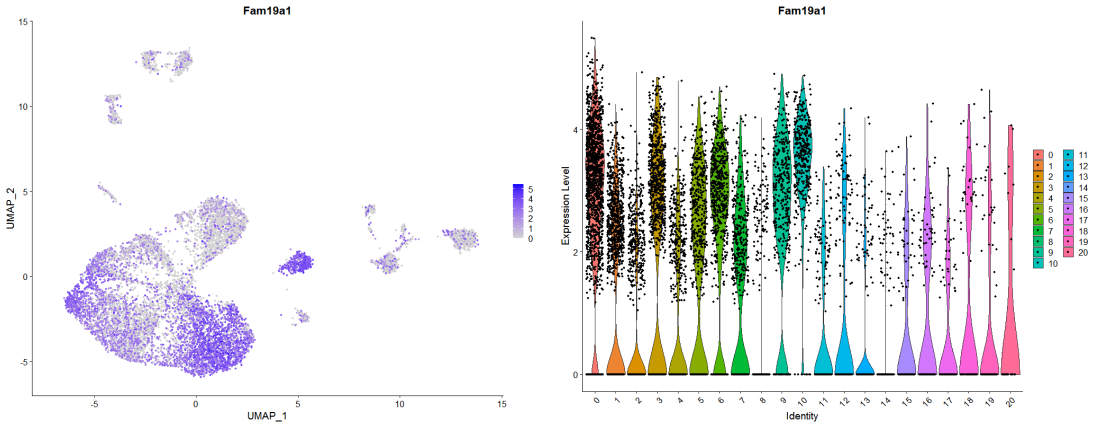

Figure 2.9: Feature plot and violin plot of Fam19a1

heterogeneity and gene expression is not trivial, and finding optimal markers for future classification of other dataset is challenging. It is useful now to perform the same analysis with the Monocle pipeline and see if some differences appear.

2.2 Gene expression analysis using Monocle pipeline

The Monocle[38], similarly to Seurat, is an R package that provides a toolkit for analyzing single-cell gene expression experiments, to study complex biological processes. As always, the first step is loading in the data, but differently from Seurat, Monocle requires to load the three files separately. It is helpful to use the

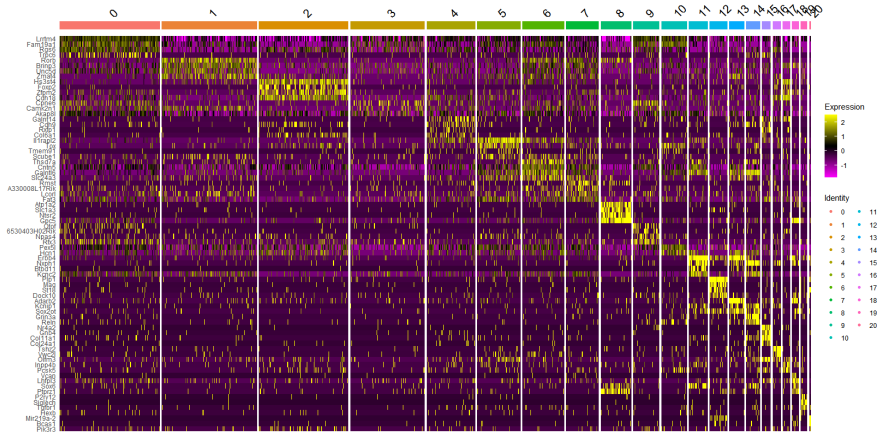

Figure 2.10: Heatmap of the top 4 genes for each cluster

"Matrix" package[39], which lets easily read and load the matrix file in a sparse format, transforming it into the dense format. There is also a need to load the cell and gene files. In this case, both files contain only the cell's barcodes and the gene's names, but Monocle accepts in input also files where besides these, there are also other metadata, such as cell type, culture condition, day captured for the cells, biotype,and gc content for genes. All the optional metadata will be stored in the same object and can be easily called. Like for the Seurat pipeline, Monocle makes use of a fundamental object to store all the data and calculation results. The object belongs to the cell data set class, derived from the class SingleCellExperiment of Bioconductor. One creates the class through the function "new_cell_data_set" which requires the input of the three files already mentioned, creating the starting point of the analysis. Before going on with the proper analysis, it is useful to use the function "detect_genes" which counts how many cells express each gene over a given threshold and, for each cell, counts how many genes are expressed. The two results are added to the cds as metadata for both cells and genes. This will be helpful for later operations. Unlike Seurat, Monocle does not encourage to perform quality control calculations, but it is not a problem because we already made them previously. The next step is to preprocess the data or, in other words, to normalize, scale, and perform dimensionality reduction. The pipeline perform all these operations by one Monocle function "preprocess_cds", which is a substantial difference from Seurat. This can be an advantage, given the faster processing time of Monocle, but also a disadvantage because one loses the possibility to fine-tune all the parameters. About the dimensionality reduction, Monocle provides the common PCA, but also offers the Latent Semantic Indexing (LSI), which transforms the expression matrix into a tf-idf matrix and performs SVD (Singular Value Decomposition). We performed the calculation with both methods but,

since with Seurat there was a focus on PCA, now the reference method will be LSI. Next to the preprocess, it is the visualization of the data. Monocle 3 uses UMAP by default the function "reduce_dimension" which also takes as input the preprocess method used. The "plot_cells" function plots the results. At this point, the graph is still a collection of grey points because there are no assigned cluster identities to the cells. In this regard, Monocle offers the function "cluster_cells" which uses the Louvain/Leiden community detection technique[40][41] and stores cluster assignments in the CDS. Now, once the "plot_cells" is called, it colors the cell based on these identities. In addition to clusters, the function returns the partition assignment, where partitions are well-separated supercluster, found using a statistical kNN method introduced in the PAGA algorithm[42].One has to set also the resolution parameter. Monocle does not explain how to properly chose it, so it has initially followed a vignette where they used the value $1e^{-5}$. With that, the function can identify only 10 clusters, which are the already divided populations. That few clusters are not enough if one wants to study the cellular heterogeneity at a deeper level. So it has tried with $1e^{-4}$ but, again, the result is a 12 cluster classification. We, thus, set the parameter to $1e^{-3}$, where, finally, there is the identification of 21 clusters, in line with the expectations (Fig. 2.11). After one obtained the cluster division is again helpful to find what genes make the

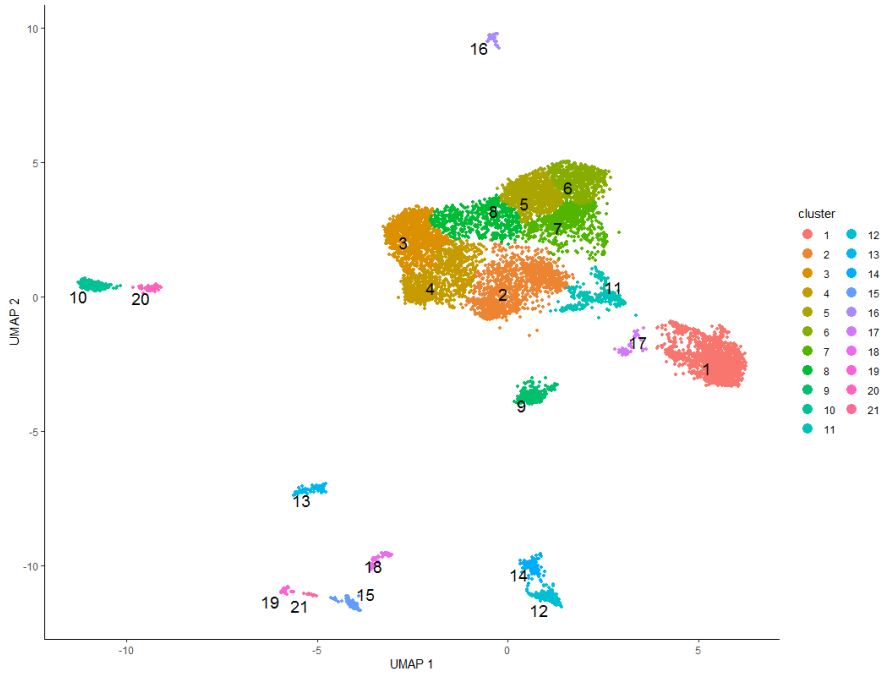

Figure 2.11: Monocle UMAP visualization of the cells, divided in 21 clusters

cluster different from one another, through differential expression analysis. Monocle

provides the function "top_markers" which identifies most specifically expressed genes between clusters, but also between other possible classifications defined in the cell metadata. It returns a table with the following columns:

- Gene_id.
- Cell_group.
- Marker_score: a general value between 0 and 1, which describes reliability as a marker for that cluster.
- Mean_expression: the mean expression value of the gene of the cells inside the cluster.
- Fraction_expressing: the portion of cells that express the gene in the cluster, and it is similar to the pct.1 value.
- Specificity: a value between 0 and 1, which describes how much specific is the marker for the cluster.
- Pseudo_R2.
- Marker_test_p_value: p-value.
- Marker_test_q_value: q-value.

In this case, the function returns 525 possible markers. Again it is useful to look for the top markers for each cluster based on the pseudo R^2 value or the marker score. We selected the top 2 genes, according to both parameters, filtering the markers with fraction_expressing > 0.1 and specificity > 0.15. The filtering mirrors the same selection based on the values pct.1 and pct.2, used for Seurat. Through the "plot_genes_by_group" one can visulalize the expression value of the top 1 marker throughout all the clusters (Fig. 2.12).

Between the results, one can notice similar genes as the Seurat analysis, like Adarb2.The latter has both a high marker score and pseudo R^2, and also a good specificity making it a possible optimal marker for that cluster. In an upcoming chapter, the dataset will be classified through literature markers, and after identifying the corresponding cell type, one can propose it as a possible marker. However, also with Monocle, there is the problem of genes, like Fam19a1, that correspond to multiple clusters, and therefore they can not differentiate between them. An explanation could be that the clustering process had recognized different clusters that in reality are the same cell type, or had mispositioned some cells through them. Again the problem comes from the big population of cells where clusters are not well-separated and rigorously diving them becomes challenging.

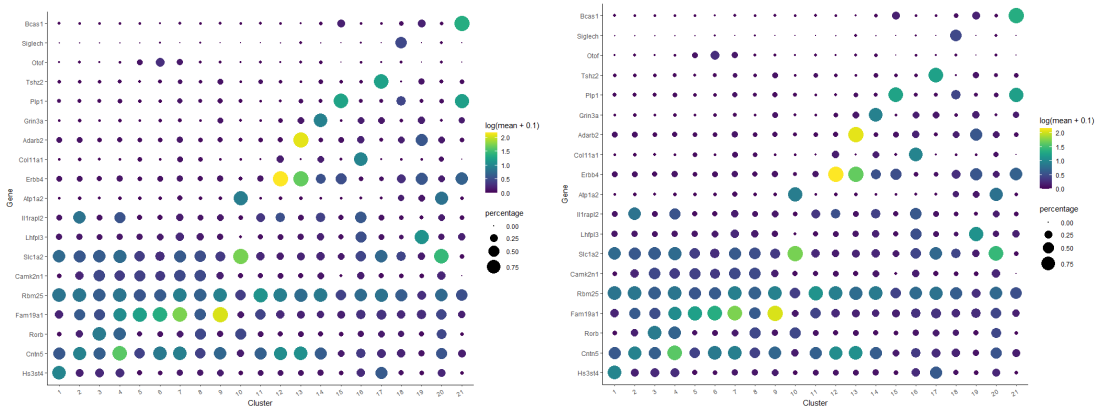

Figure 2.12: Expression of identified markers, based on Marker score (left) and pseudo R^2 (right)

In conclusion, one can see how the results from the two pipelines agree with each other in both the clustering and the differential analysis. It is possible to visualize the two cluster partitions on the other plot. This is possible with the simple operation:

$$cds\_rna@clusters@listData[["UMAP"]][["clusters"]] <- ad.m.rna@active.ident \tag{2.1}$$

This takes the labels of each cell found with the Seurat clustering process (right-hand side) and replaces the label of the Monocle cds (left-hand side) with them. One can also do the opposite with:

$$ad.m.rna@active.ident <- cds\_rna@clusters@listData[["UMAP"]][["clusters"]] \tag{2.2}$$

Fig. 2.13 shows the results. The main difference is again on the big population where the two pipelines both find eight different clusters, but the divison is not the same. The smaller well-separated populations instead agree with each other, apart from the two clusters on the left side of the Monocle visualization (10 and 20) that Seurat identifies as one. There is no a priori better clustering between them, but later analysis could confirm one of the two.

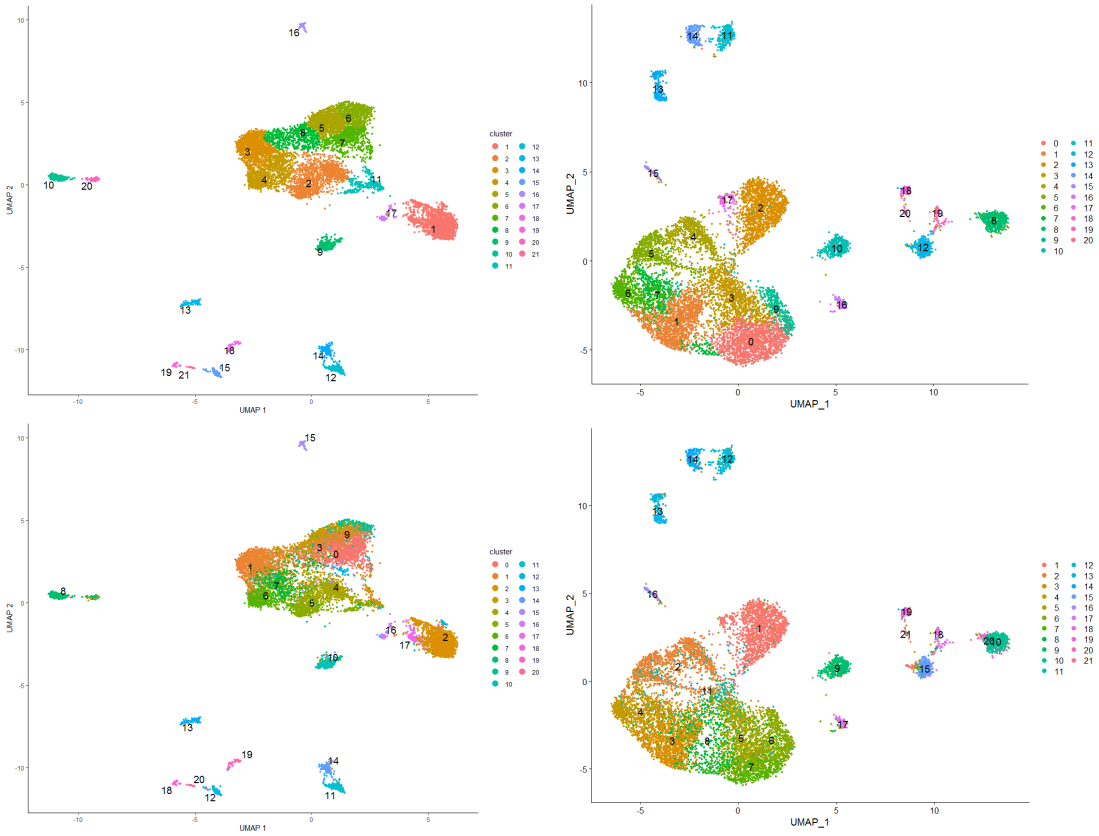

Figure 2.13: Monocle with Seurat labels (bottom left), Seurat with Monocle labels (bottom right)

Chapter 3

ATAC Analysis

The scRNA analysis is just the start of the work. The dataset, as previously said, provides, also, the scATAC-seq data. This part aims to use similar analyses performed for the expression data on the epigenetic data and separately study the results to understand the power of only peaks analysis. ATAC stands for Assay for Transposase-Accessible Chromatin using sequencing and is a technique to probe the DNA to assess the chromatin accessibility of all the genome[43]. The DNA inside the nucleus of a cell is mostly packed into the compact structure that is the chromatin. Not all the DNA is always packed, but some regions are accessible to let the transcription of the genes. The accessibility is a necessary but not sufficient condition for the expression of the gene because even if the gene appears to be accessible, this gives no information on its effective expression. Moreover, the two biological mechanisms are not static, but dynamically change with different time scales[44]. For these reasons, the correlation between these two mechanisms, and even more between the experimental data, is far from trivial. So before trying to understand the links between them, it is helpful to analyze separately the ATAC data[45]. To study this type of data there are appropriate packages, extensions of the now well-known Monocle and Seurat, called Cicero[46] (companion R package of Monocle) and Signac[47] (extension of Seurat). The ATAC analysis consists of similar steps to the RNA one, with the addition of further calculations. Cicero and Signac provide different types of operations, so again we elaborated the dataset with both. Cicero provides co-accessibility and cis-regulatory network calculation, while Signac allows to manipulate multiple epigenetic information simultaneously like motif enrichment analysis and useful fragment visualization.

3.1 ATAC analysis using Cicero pipeline

Cicero is an R package, design to work with Monocle. Its main function is to examine the co-accessibility of the genome from single-cell chromatin accessibility data to predict cis-regulatory networks[46][48]. Cicero makes use of the CDS object, but with some modification to hold this type of data. First of all, instead of the genes, now the features are the peaks. The peaks are small regions of the genome, identified during the ATAC experiment. The actual length of a peak can vary a lot, but it is always much smaller than the general gene length. These features are composed of the chromosome number, the starting position, and the ending one, like chr1_10390134_10391134. Therefore, the matrix has the peaks as rows, the barcodes, the same as for the expression matrix, as columns, and all the reads per cell per peak as elements. Because of the sparsity of ATAC data, it is not expected more than 1 read per cell per peak, and so Cicero suggests to binarize the matrix switching to an "on-off" model, stating that can give clearer results. We performed this binarization step only with Cicero, while we used the original matrix with Signac. After loading data, the CDS goes through the same steps as the previous chapter.

cds_atac <- detect_genes(cds_atac)

cds_atac <- estimate_size_factors(cds_atac)

cds_atac <- preprocess_cds(cds_atac, method = "LSI")

cds_atac <- reduce_dimension(cds_atac, reduction_method = 'UMAP', preprocess_method = "LSI")

When working with the accessibility data, it is strongly recommended to use LSI as the preprocessing method due to better performance rather than PCA. One can, at this point, run the clustering function to obtain a classification based only on the accessibility data. There will be a later comparison with the previous one. The resolution parameter was initially set to $1e^{-3}$ as in the RNA case, but it returned only 16. For this reason, it was increased to $1.6e^{-3}$, giving 19 clusters. Even if it is still less than what we obtained during the RNA analysis, this is the value chosen for the time being. The reason is that one can further increase the resolution, obtaining how many cluster one wants, but does not imply a better classification. The difference in the cluster identified, can be a consequence of the dissimilar type of data, and the study of this difference could be informative. Anyways, in this case, the cells are again grouped in a big population, plus some other smaller well-separated groups, but it is early to make suppositions on parallelisms (Fig. 3.1) Once one obtained the clustering division, it would be helpful to understand what makes the cluster different, similarly to the differential gene expression. However, using the "top_markers" function on the ATAC data does not give a clear result as for the expression analysis. First of all, the computation time is much longer, probably due to the increased number of features to consider. Moreover, the results

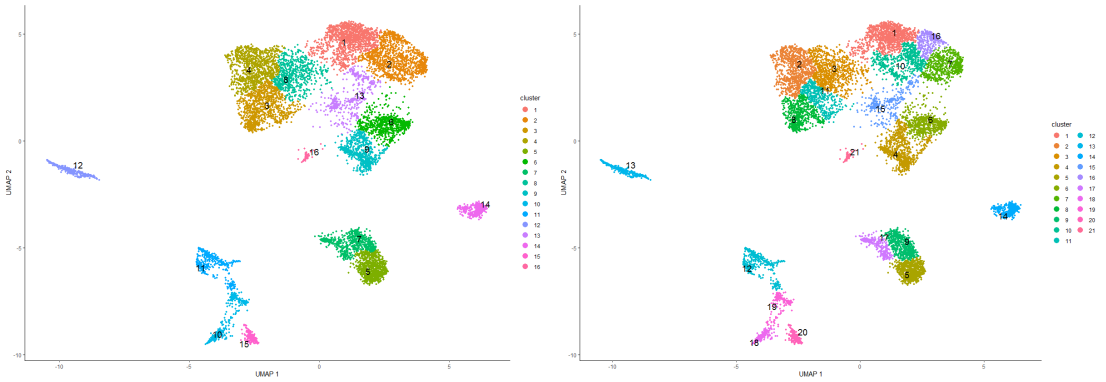

Figure 3.1: Clustering results of ATAC data with resolution $1e^{-3}$ and $1.6e^{-3}$

are not what one would want. All the reported features have both a really low marker score and pseudo R^2 and are not discriminative of the cluster. Twelve clusters have the same peak (chr13_9011697_9011890) found by the function, and one can see from Fig 3.2 how it is accessible through all the clusters. Some other

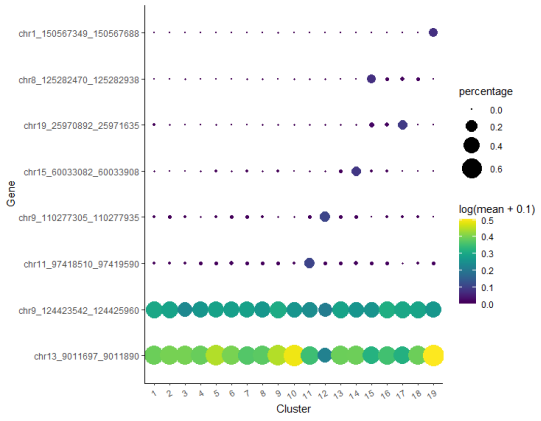

Figure 3.2: Accessibility of peaks from differential analysis

peaks, instead, appear to be accessible only in one particular cluster, and even if they show low marker score value, have good specificity. In general, the differential analysis performed on this type of data does not give optimal results. This is due to the nature of the ATAC data and the binarization of the matrix performed at the start. For this reason, these results are not further investigated, at least for the time being.

Fortunately, the analysis does not stop here, but with Cicero, one can perform additional calculations. The first thing is the calculation of co-accessibility scores between peaks. The score is a value between -1 and 1 between pairs of accessible

peaks within a certain distance set by the user. The co-accessibility defines how two peaks are correlated, meaning if there is a relation such that when one is accessible also the other is, and the other way round. So pairs with high values might identify regions belonging to correlation mechanisms like enhancer-promoter. Because the data are extremely sparse, to estimate the co-accessibility, one needs to modify the original CDS. In particular, it requires aggregating similar cells to obtain denser data. To do so, Cicero provides the "make_cicero_cds" function. The latter uses a k-nearest-neighbors method to create sets of overlapping cells, which are constructed starting from the coordinate of a reduced dimensionality map, in this case, UMAP. The function takes as input the original CDS and the UMAP coordinates and returns a new CDS with only 2914 cells left, and the same number of features.

After we obtained the cicero_cds, we proceed to the calculation. The process is quite complicated and computationally heavy and consists of three steps. Cicero gives the possibility to run the three parts separately or to use the function "run_cicero" which performs all the calculations with default parameters. The separated functions are:

- estimate_distance_parameter: calculates the distance parameter given random windows on the genome.
- generate_cicero_models: using the parameter found at the previous step and through the graphical LASSO, calculates the co-accessibility score of overlapping windows.
- assemble_connections: creates the final list of co-accessibility scores through the output of the previous function.

Calling the functions separately, give more flexibility and control on all the middle parameters, which need to be changed especially depending on the organism considered. The default settings are optimal for human and mouse organisms, so in this case, it is not necessary to run separately. Anyways, before performing the calculation one needs also a genome coordinates file containing the lengths of each of the chromosomes, that must match the same genome coordinates used during the peak calling step of the creation of the ATAC data. The genome reference used for this data is the build GRCm38 (Genome Reference Consortium Mouse Build 38), also called mm10. Cicero already provides in the package the gene annotation of the mouse but the build mm9. This is not a problem, because one can find all the genome references on the already mentioned NCBI portal, in the "Assembly" section. It is worth mentioning that due to the heavy computational work required, the calculation takes a very long time, so it is advisable to make sure that all the inputs and parameters are correct to run it the minimum times. Indeed the result is a table with three columns, the two peaks, and the co-access value, and has a

total of 23640402 elements, that are couples of peaks. Cicero provides a very useful plotting function to visualize interestingly the co-accessibility. Before this, however, the gene annotation file must be loaded. The annotation has all the information for each gene, but in particular,the following are important:

- Start and end: the starting and ending point of the gene.
- Strand: the DNA strand on which it is (+ or -).
- Gene_name.
- Transcript_ID.
- Gene_biotype: the type of the gene like protein-coding or snRNA.
- Chromosome.

For the plotting are necessary only these columns, so it is useful to create a smaller file composed only of these. One needs to also rename the columns to match the requirements of the plotting function. Then one can use the "plot_connections" function, which takes as input the chromosome, the extremes of the region one wants to plot, the annotation file, and a lower cutoff for the co-accessibility score. The resulting figure shows, in the lower part, all the genes and peaks present in the region given as input, and a series of lines connecting the peaks, that identify the connection and their "strength".
In addition to the pairwise accessibility, it is possible also to estimate cis-regulatory networks that are groups of peaks that are highly co-accessible with each other. In a similar way to the clustering process, the function "generate_ccans" uses the Louvain community detection algorithm[40] to find sites that appear to tend to be co-accessible and group them. The function takes as input the connection table resulted from the previous calculation. These results are not completely useful alone but are necessary for the creation of the Cicero gene activity matrix. The latter is a new count matrix where the features are no more the peaks but are again the genes, and the matrix elements do not describe the expression but the accessibility of the gene, based on the accessibility of peaks at promoters and of the ones highly co-accessible to them. In this way, one obtains an object that can help to estimate the accessibility of the genes themselves and not the various peaks that do not accurately match the genes' body. This topic, however, will be the central part in a later chapter with the joint analysis of scRNA and scATAC.

In conclusion, Cicero gives useful tools to estimate the relation between peaks, other than the possibility to process scATAC data in a similar way to scRNA data.

3.2 ATAC analysis using Signac pipeline

After Cicero, we processed the data with Seurat and the additional package Signac. At first, Seurat recommended using the same workflow, already discussed, to work with scATAC data. Therefore we set the parameters of the functions like in the scRNA case. The only difference is that, instead of performing PCA, we performed LSI, running the TF-IDF normalization and SVD separately. Unfortunately, the results were not excellent. Fig.3.3 shows the final clustering, where one can see something different from previous results. First of all, setting the resolution to 1.3, it returns only 14 clusters. One needs to increase the value to 3 to obtain 19 clusters, but it is an arbitrary choice that does not bring reliable information. Moreover, the cells visualized are not well organized in well-separated clusters and form strange and unexpected structures. One could argue that it is just a matter of visualization and they retain the information, but, if one labels the cells with the gene expression classification, one can clearly see how they disagree with each other (Fig. 3.3). This means that there are problems within the workflow that do not allow a correct elaboration of the data.The critical point rise from the intrinsical dynamics of the data. The accessibility of genomic regions does not vary as much as the expression between cells, and, therefore, it is hard to find the optimal variable features to base the following calculations. For this reason, the "FindVariableFeatures" function is unable to correctly identify the features to use in the following steps, in particular during the UMAP reduction, resulting in the unusual visualization of the cells. Fortunately, recently Seurat updated the

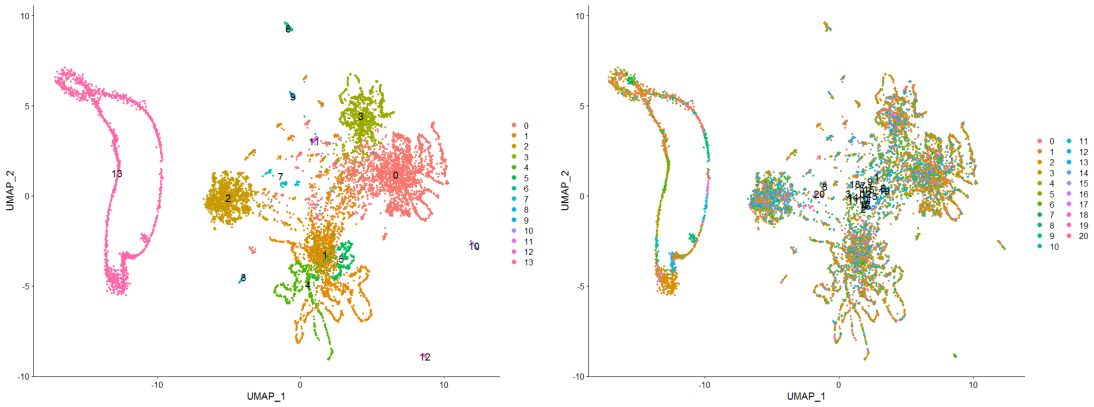

Figure 3.3: scATAC data analysis with the older version of Seurat

Signac package, with new tools specially designed for the epigenetic data. In the following,we reviewed the new Signac workflow, focusing on the differences and new features. First of all, the object to contains the data is still a SeuratObject, but, the chromatin data, are stored through a ChormatinAssay. The latter is a custom

assay which adds several slots for additional data helpful for ATAC-seq analysis:

- Ranges: a "GRanges" object which contains the coordinates ranges of each peak inside the count matrix.
- Motif: a "Motif" object, that can be obtained starting from the count matrix.
- Fragments: A "Fragment" object, a list of all fragments inside all single cells.
- Seqinfo: a "Seqinfo" object which contains information about the genome from which the data were mapped, in particular, the length of chromosomes.
- Annotation: a "GRanges" object for the annotation of the genes.
- Links: a "GRanges" object for links between peaks, like the connections estimated by Cicero.

So the first step is to create this Assay through the function "CreateChromatinAssay". As input, other than the count matrix, it takes a genome reference, in this case, the mm10 genomic build, which must be the same genome used to create the data. It is really useful also to load the fragment file, which contains the unique fragment reads for all cells. This is a very large file that can be slow to work with, but has the advantage to include all the fragments and not only the ones mapped to peaks. The fragment file for the SNARE-seq dataset is not publicly available, but it can be obtained from the raw data, through the help of the Sinto package[49]. Fortunately, the Satijalab has already processed the raw data and provides directly the fragments file. Then the newly created assay is taken as input for the creation of the actual SeuratObject. Signac, unlike Cicero, provides tools for Quality Controls, specifically for chromatin data. First is the nucleosome signal, which calculates approximately the ratio of mononucleosomal to nucleosome-free fragments, which describes the relationship between the length of peaks and the length of DNA wrapped around a single nucleosome that must not be high (not more than 4) for a good quality experiment. The second is the transcriptional start site (TSS) enrichment score[50] which defines the ratio of fragments that are centered on TSS or flanking regions. Low values of enrichment (values less than 2) usually mean poor quality experiments. Fig. 3.4 shows how the results respect the limits.

After that, it is time to normalize and reduce the dimensionality of the data. As previously said, the nature of ATAC data makes it hard to find variable features, so instead, it is better to use the "FindTopFeatures" function to identify the top n% features to consider in the following calculation. To reduce the dimensionality it is again adopted the LSI method[51]. It can often happen that the first component from LSI correlates to sequencing depth instead of biological variation[52], and so it is better to not consider that, for the downstream calculations. To assess this there

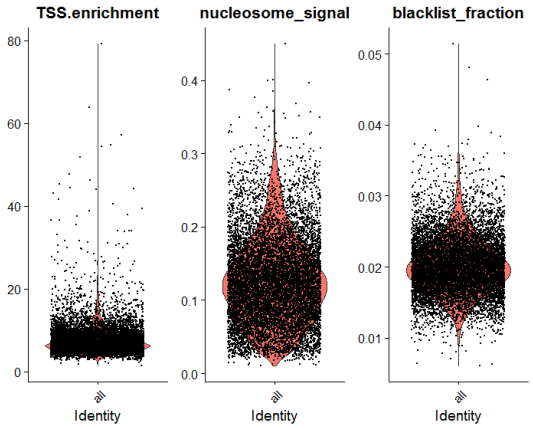

Figure 3.4: QC mterics for epigenetic data

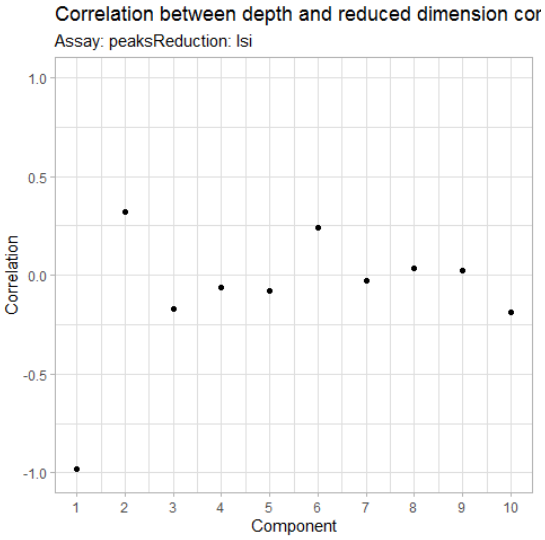

Figure 3.5: Correlation of the LSI components

is the "DepthCor" function, and the plot in Fig. 3.5 shows the results. As one can see the first component has a stronger correlation so we performed the following calculation without it. The next step is, as always, clustering and visualization. Here things do not change much, we implemented UMAP non-linear dimensional reduction, and the clustering method is the same. We set the resolution to 1.3 like in other clustering processes, and 17 clusters resulted. However, two clusters (15 and 16) have a really small population (cells each), so they are not informative. This time, at least, the visualization makes more sense, the cells are not anymore in those strange patterns like before (Fig. 3.6). Again there is a big population of

cells divided into subclusters, with other separated smaller populations, similar to what we previously found.

Even if with Cicero, the differential analysis did not been completely successful,

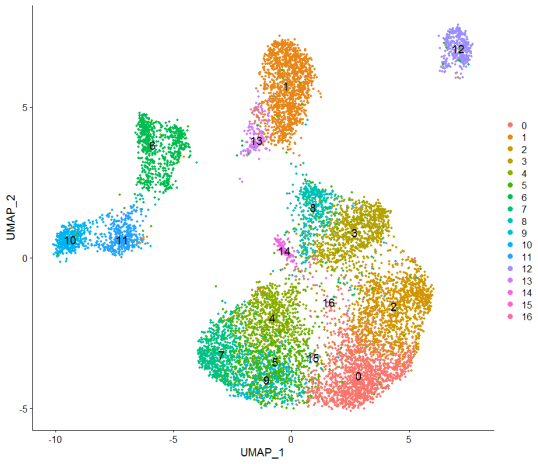

Figure 3.6: Signac final clustering

we performed it also with Signac. The function to use is the same as for the scRNA analysis, which resulted in a total of 4834 differentially accessible peaks. This time around, the features have high values of avg_logFC (which represents how much differentially accessible are for that cluster) but the fraction of cells that access that peak is very low. One can see from some violin plots and feature plots how some peaks are accessible only in a few cells overall, even if they identify some particular clusters. This is an improvement of the previous differential analysis done with Cicero but, still, it is not sufficient to find peaks that can be treated as epigenetic "markers", like the genes markers.

The ATAC analysis presents some criticalities. First of all, besides the new Signac tools, the workflow to process epigenetic data is the same for the gene expression, even though the dynamics of these biological mechanisms are different. For example, the ATAC count matrix has entries that are mainly ones, apart from zeros due to the sparsity of this type of experiment. So the peaks accessibility is treated as an on/off mechanism, rather than having different levels like the gene expression. Whereby the analysis, which tries to find the differences between the cell's accessibility profiles, becomes challenging. Moreover, the features of the dataset are the peaks, which have the problem of not being unique like the genes. While genes, given the reference genome, are the same between experiments, this is not true for peaks, which depend on the peak calling process of the experiment. Therefore it is hard to find results that can be generalized, like with the gene markers. It is better, therefore, to study the relation within peaks (for example,

with the co-accessibility), instead of considering them alone, and even more, relate them with the gene expression. This will be just the work of the last chapter, where we compared the epigenetic results with the gene expression. But first, it is helpful to classify the cells in some way to have a reference, to evaluate the results obtained up to this point.

Chapter 4

Dataset classification

In this chapter, we work to obtain an independent classification of biological cell types of the cells. It will allow validating the fact the unsupervised clustering is recognizing cellular heterogeneity and will give us a biological reference to make hypotheses. Until now, the work focused on the separate analysis of expression and accessibility data. We separated the cell into clusters based on them, and there was a good agreement between the results of the pipelines. But none of the partitions has been identified as known cell types, they are just groups of cells that the unsupervised algorithm found to be similar, and none other information relates to them. During the scRNA, we made some assumptions looking at the 2-dimensional visualization and the unsupervised clustering. The clustering separated the cells into a big population plus some well-separated, smaller groups. On the general knowledge of the cellular composition of the brain, one can therefore assume that the big group is probably composed of neuronal cells (which make up the majority of the brain) while the small ones can be non-neuronal cell types like astrocytes[53][54]. However, without an inspection of the expression profile of the cells, one can not assign cell types to cluster or even be sure that the clusters are truly identifying cell types. For this reason, it is helpful to implement some sort of unrelated classification to understand two things. One, if the algorithm is working properly and the clusters represent the cellular heterogeneity. Two, creating a classification to use as a reference to compare the various results obtained and understand which perform better. In this chapter, we implemented two strategies to do so. First, the classification is based on the expression of literature gene markers[55]. Then we performed a label transfer technique to classify cells based on an unrelated dataset with cell labels already present. They have different advantages and disadvantages, and most importantly none of them is completely correct, but they are only approximations.

4.1 Literature markers analysis

The first approach to give the cells a cell type identity start from the expression of distinct genes considered to be specific for a precise cell and are called markers. They derive from various biological works that propose what they have found studying some particular cells. Here the first problem emerges; some of them originate from protein analyses rather than from RNA expression. While it is true that there is a direct relationship between gene and protein, the same thing cannot be said for the expression of a gene and its translation to protein. Indeed, the mechanisms inside a cell are such that from a single mRNA molecule can be synthesized a large number of proteins. Not to mention also that not all the RNA fragments are translated, but can be degraded before that. For this reason, comparing expression and protein levels can be misleading. Unfortunately, even if there are a lot of known optimal marker genes for different cell types, it is still challenging to obtain a reliable set of markers to classify a general dataset. Even if this method is an approximation and it is not completely optimal and stable, it can at least give an idea of the cellular heterogeneity of the dataset used until now. To do this it has been worked with the older version of Monocle for R version 3.4.3, which provides some useful functions to tag each cell directly through the assessment of their expression profile. The annotation is produced based on a simple set of functions manually provided. These functions are generally simple comparison operation which accepts as input the expression values of the gene for a cell and returns TRUE to tell that the cell matches the imposed criteria. The idea is to construct a series of functions for each cell type one wants to identify, which detect whether one or more genes are expressed over a given level. An example could be to label a cell if the expression of a certain gene X is > 0, meaning the expression is active. Instead of zero, one could set that to a higher value to identify the cells where that gene is highly expressed to not mislabel cell where that is expressed but only in low values. Besides, multiple genes can be compared for one class, through a logical operation like OR, AND, etc. One stores all the functions in a small structure called "CellTypeHierarchy" (CTH). The creation of this object is the first step, through the function "newCellTypeHierarchy". After one needs to load all the gating functions, one for each cell type of interest, employing "addCellType" which takes as input the CTH, the name of the type, and the operation to identify it. Once all of them are ready in the data structure, it is time to classify all the cells in the dataset. The function "ClassifyCells" applies each gating function of the CTH to each cell in the provided CDS to which it adds a new column to the pDATA table with the results. It is important to mention that besides the types manually defined, there are two more labels: "Unknown" and "Ambiguous". The first is assigned when the cell matches no criteria specified in the CTH, while when it satisfies multiple criteria is marked as Ambiguous. Once all cells are labeled it is

easy to count how many cells there are for each type and visualize them on a pie chart so one can have a first impression of the quality of the classification.

We implemented three different classifications, with varying labeling rules based on different sets of markers. The first uses the same markers proposed by the researchers of the SNARE technique for annotating clusters, the second originates from the portal DropViz[54], and the third is a mixture of the two. For all of them, the workflow is:

- Choice of the markers and function for each cell type.
- Run of the classification on the expression dataset.
- Visualization of the labels on the plotted cells.
- Evaluation of the results compared to unsupervised clustering
- Annotation of the clusters based on the most frequently occurring labels.

4.1.1 First Run

We implemented the first run of classification through the assessment of the expression of the same markers used for the SNARE publication[22]. They provide a list of cell types with one or two genes to label them. Because there is no more information about them (like how highly expressed they are in the clusters), the classification functions are very simple. We chose one gene, which identifies one type if it is expressed in a cell i.e. if the gene expression is >0. The Table 4.1 report all the functions, with the corresponding category name. The categorization proposed includes ten varieties of Excitatory Neurons (Ex), four Inhibitory Neurons (In), plus six non-Neuronal cell types. Excitatory Neurons are subdivided based on the spatial layer of the cortex to which they belong, so for example EX-L5 means neurons coming from the fifth layer one of the most internal. The non-Neuronal cells are Astrocytes, Oligodendrocytes, their progenitors (OPC), Microglia, and Claustrum cells. Besides all these classes, the classifying algorithm also returns the Ambiguous and Unknown cells. The number of labeled cells that received a cellular identity is 3460. The remaining 6849 cells are divided into 4512 Ambiguous and 2337 Unknown. One can already see how only a rough third of the total number is actually labeled, with the majority of cells being uncertain. This is due to the classification functions, which look for the absolute presence of a marker gene. The majority of the genes, even if one considers them to be a marker of cell type, can be also expressed in minimal quantity in other ones, so the algorithm can not uniquely give them an identity. This becomes even worst between the Excitatory neurons since they are very similar between layers and tend to express similar genes. It is useful to represent the labels on the plots previously obtained. In Fig.4.1 we

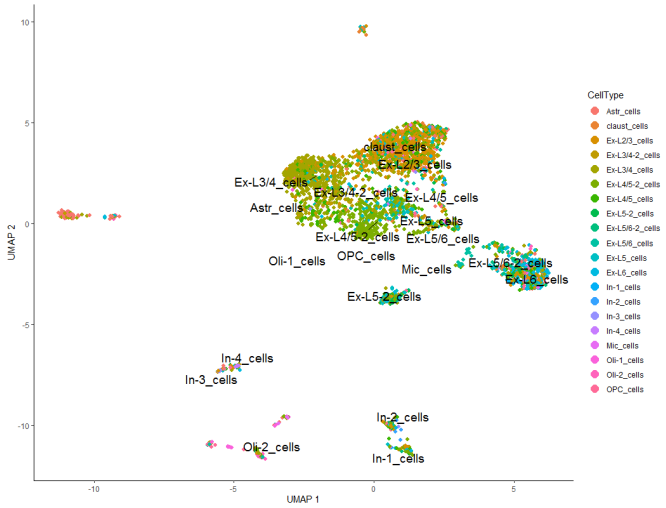

Figure 4.1: Monocle UMAP visualization with the labels from the first marker classification

present all the cells, except for the Ambiguous and Unknown, on the Monocle representation. On the plot, the names of the cell types are positioned on the centroid of all the cells with that identity. One can see how some cell types match with known clusters like In-1, In-2, while others seem not to be well-identified like Oli-1. One can look at the number of labeled cells per cluster. As expected, the big population is composed of various Excitatory Neurons showing, however, how there is not a clear distinction between them. For some of them, there is a clear and unique match with the cell types, like for the first cluster composed mainly of Ex-L6 cells. Some cell types, like Ex L2/3, represent multiple clusters meaning that the unsupervised algorithm could have found subtypes or the resolution it is too high and the differentiation is meaningless. After a qualitative analysis, it is better to perform some statistical metrics calculation to understand at a mathematical level the similarity between two classifications. In this regard, it has been implemented standard clustering comparison measures[56][57], in particular, Rand Index and Mutual Information. The Rand index is a measure of the similarity between two clustering classifications and is related to the accuracy. It has a value between 0 and 1, where 0 means no classification pairs agree while 1 indicates a perfect match. Mutual Information is a measure of the mutual dependence between the two variables and is linked to the concept of the entropy of a random variable. Mutual Information measures the information that two sets share and can be used to evaluate how close are two classifications. But these two have some problems related to the dimensions of the sets, and the possible different number of partitions between the two. So it is better to calculate their modified versions, in particular,

the Adjusted Rand Index (ARI) and the Normalized Mutual Information (NMI). To do so we adopted the R package "aricode" which provides easy functions to calculate these values. So we evaluated, with these metrics, the concordance of the first run of classification with the unsupervised clustering, taking into account only the cells labeled with cell types (not Ambiguous and Unknown),
RI = 0.8429034
ARI = 0.1970836
MI = 0.6585482
NMI = 0.2519598
The NMI is what we mainly taken into account, because it is the best to fix for the different number of partitions between the two classifications. The value of 0.25 for it can seem quite low but it is not a bad result considering the generality of the function employed for the first classification. The main problem is related to the high number of Ambiguous cells due to the gating functions that do not take into account the expression levels, and the few markers considered. So we implemented a second run.

4.1.2 Second Run

The second run focused more on trying to improve the previous one, especially to reduce the Ambiguous classification that inflated the other. To do so, we changed the set of marker genes, the gating functions, and the set of cell types. Instead of using the set of genes provided by SNARE research, it has been studied the literature for constructing a possibly new collection. In particular, we used the platform "DropViz"[54]. This platform provides tools to explore the mouse brain cellular heterogeneity, through the exploration of cell expression profiles of hundreds of thousands of cells. These cells come from nine regions of the adult mouse brain, which are explorable separately. Then for each region are identified several cell types, which one can explores one at a time or can compares them one another. The platform is really helpful because provides both canonical markers for each cell type and markers derived from differential expression. Moreover, one can explore the expression levels of a given gene throughout the cell types, giving information about the possible level to use to identify the types. DropViz has everything one needs for this work. It has data for the adult mouse brain and one can focus only on the cortex, which is the same sample used for this study dataset, and gives reliable markers and a way to explore their expression levels. Therefore we created a new set of markers and relative gating functions, shown in Table 4.2. There are some differences from the previous one. First of all, the number of cell types is only 17, because the platform does not give a clear classification of all the Excitatory Neurons like before, so here there are only 7 of the 10 different cell types related to them. On the other hand, the cell types chosen are better defined.

The other fundamental difference is that this time around, we considered multiple genes for the identification. When this happens, the identification functions become a bit more complex, each gene-level is evaluated separately, and they are linked with some logical operator. For example, for the identification of the Astrocytes, we considered the genes "Slc1a3" and "Apoe", combined with the operator OR, meaning that the cell is labeled when one of the two conditions is met. In an attempt to reduce the inflation of the Ambiguous cells, we chose the expression levels of the markers greater than zero.To avoid the possibility of becoming too strict in the classification, the values are relatively low, but always consistent with the DropViz information. After setting the classification rules, one can procide with the classification. It resulted in the labelling of 4098 cells, 2486 Ambiguous, and 3725 Unknown cells. There is an increase in labeled cells and a decrease of Ambiguous ones, as desired, but at the cost of a lot more Unknown ones. The increase of Unknown cells is due to the new levels set, which makes it unable to identify cells with lower expressions, especially between Excitatory Neurons. However, about the non-Neuronal cells, there is an improvement. All the cell types of this kind have the number of the labeled population increased. This is because more strict rules on the genes guarantee that cells of rarer type, which can exhibit the expression in small quantities of some Excitatory Neurons, are not mislabelled or become Ambiguous. In Fig.4.2 one can see the cells with their identity. Again

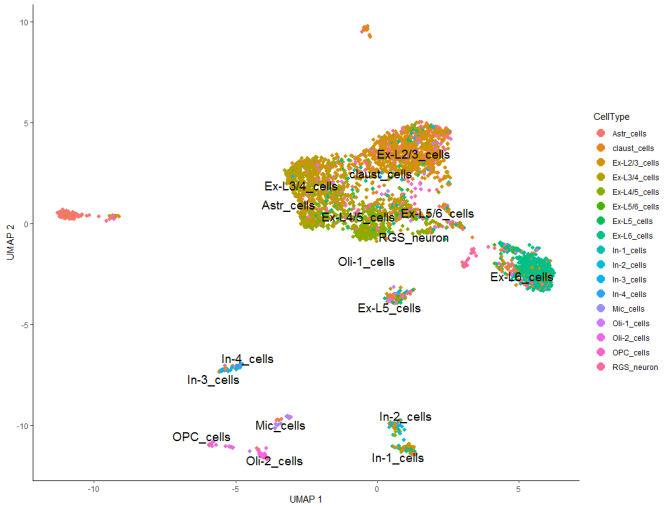

Figure 4.2: Monocle UMAP visualization with the labels from the second marker classification

the cell type tags are placed in the centroids of each group, and some of them seem not well-placed, like Oli-1. If one looks at the populations for each cluster, one can find which cell type better represents that, and results are similar to the previous

classification. This means that despite the variations, the two rounds agree with each other, but also with the unsupervised partition, giving the first hint that the algorithm used is finding cellular heterogeneity. However, before assigning the identities to the clusters, we performed one last round of classification that aims to combine the best of both. Before that, it is useful to evaluate again the results of the labeling against the clusters labels, with the already mentioned metrics. The results are:
RI = 0.852115
ARI = 0.2779283
MI = 0.8705825
NMI = 0.3296452
There is an increase in all the values, probably related to the greater number of labeled cells between the smaller population. The last round of classification aims to reduce even more the Ambiguous cell, mixing the information of the previous two sets.

4.1.3 Third Run

We performed one last classification round. The first one focused more on all the Excitatory Neurons types at the cost of more ambiguity while the second gave more information for the identifications of non-neuronal cell types. Therefore, we attempted to implement an analysis that took into account both. A good classification does not necessarily require that most of the cells are labeled, but the rules to identify them must be the most univocal as possible, meaning the lowest possible ambiguity. In this way, the labeled cells are more reliable and become an optimal reference set. The summary of all the cell types and relative gating functions is reported in Table 4.3. First of all, the types of Excitatory Neurons have been increased to eight types (RGS Neurons are the same as Layer 5/6), a middle ground between the two. The set of chosen marker genes did not change, but the expression levels did. In particular, some genes appear to be markers when they are highly expressed, so the thresholds for some cell types now have higher value that the second run. After setting all the parameters we performed the classification. This resulted in a total of 3918 cells labeled, 1946 Ambiguous and 4445 Unknown (Fig. 4.3). The labeled ones decreased in comparison to the second round, but the same is for the Ambiguous cells as desired. The inconvenience is that was not possible to identify a lot more cells, but as previously explained is better to have fewer but more reliable. With this result, one can finally attempt to give an identity to the unsupervised clusters. It can be accomplished mainly in two simple ways. One, if in a cluster there is a considerable majority of one cell type respect the others, it is reasonable to label the group with that. Two, if one cell type is present in only one particular cluster which is not predominated by others, it also

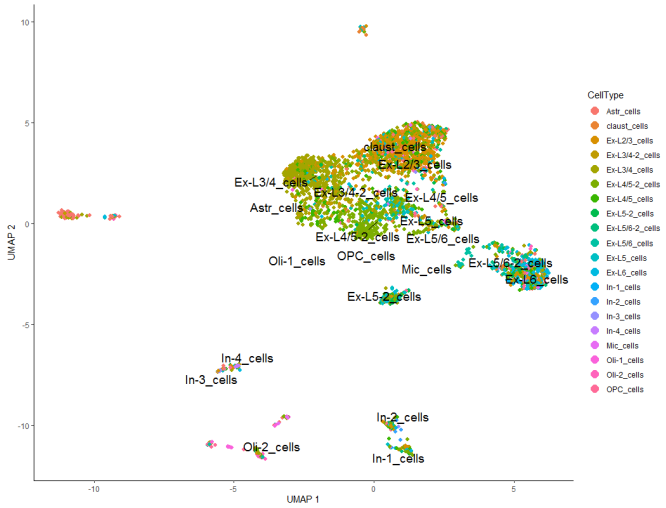

Figure 4.3: Monocle UMAP visualization with the labels from the third marker classification

makes sense to link the two. The second method is usually less reliable because it involves a smaller number of cells. For example, the In-1 cells are included only in cluster 12, even if the latter has multiple types related to it. Therefore, from looking at Table 4.4, conclusions we labeled the clusters, and plotted them in Fig. 4.5

Some observations are useful. Some clusters are easy to associate, like cluster 1,

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Astr_cells	4	4	0	1	0	6	3	0	1	136	0	0	1	0	2	0	0	6	2	10	0
claust_cells	4	3	0	1	1	1	1	2	1	0	0	0	0	3	0	36	0	0	0	0	0
Ex-L2/3_cells	35	64	84	46	297	196	242	128	17	8	22	23	20	17	4	6	4	2	0	6	0
Ex-L3/4_cells	1	18	370	65	14	6	11	72	3	20	5	0	1	0	1	1	0	1	1	5	0
Ex-L4/5-2_cells	16	33	16	107	10	40	25	2	12	0	7	15	12	3	0	0	0	0	0	0	0
Ex-L4/5_cells	14	277	6	51	7	7	10	8	7	1	27	21	1	12	0	10	2	1	1	2	0
Ex-L5/6_cells	7	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ex-L5_cells	0	2	0	2	2	1	0	1	15	1	0	0	0	1	2	0	0	0	0	0	1
Ex-L6_cells	586	31	16	14	12	18	14	11	22	1	9	10	2	5	4	2	5	1	7	3	0
In-1_cells	1	2	0	1	1	0	0	0	1	0	0	15	0	1	0	0	0	0	0	0	0
In-2_cells	1	0	1	0	1	1	1	1	2	1	1	1	0	37	0	0	1	0	1	0	0
In-3_cells	1	1	0	1	1	3	1	2	1	0	0	2	21	3	1	0	0	0	0	0	0
In-4_cells	0	0	2	0	0	1	0	0	0	0	0	0	25	0	1	0	0	0	0	0	0
Mic_cells	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0
Oli-1_cells	0	0	0	0	0	0	1	0	1	2	0	0	0	0	0	0	0	0	0	0	5
Oli-2_cells	0	1	1	1	0	1	4	0	0	0	0	0	0	0	44	0	0	0	0	0	0
OPC_cells	3	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	23	0	3
RGS_neuron	1	1	1	2	0	2	7	3	2	0	2	0	0	0	0	0	29	0	0	0	0

Figure 4.4: Number of classified cells per cluster

where there is an insurmountable majority of Ex-L6 neurons, but unfortunately, it is not the case for all of them. Clusters like 2, 3, and 4 have predominant populations, but also include a not indifferent number of each other cells. The reason is that these cell types are neuron from adjacent layers, so their difference is not well defined, and, therefore, the classification, both from markers and unsupervised clustering, can not be completely precise, but this is not a big problem. It is interesting the case of clusters 5, 6, 7, and 8, which are all identified with Ex-L2/3 Neurons. It means that the unsupervised algorithm has maybe recognized some subtypes of these neurons. This could also explain why during the differential expression analysis, they showed similar marker genes. The cluster 1 includes a small population of Ex-L5/6 cells, meaning that it could be a subtype that was not identfied by the algorithm. Also, as expected, the classification recognizes the different Non-Neuronal types as the smaller groups found in the previous chapters. A bit more problematic are the Inhibitory Neurons, which one can label only through the second method mentioned. It is the case of the already mentioned cluster 12, which has pretty similar populations of Ex-Neurons and In-1, so it would not be clear which label to assign. However, one can notice how the In-1 cells are barely included in other clusters, so it is reasonable to assign it to the mentioned cluster. Regarding again the Inhibitory Neurons, it is noticeable cluster 13 which has two different cell types in it (In-3 and In-4). This means that the clustering algorithm failed to identify the difference in the cellular heterogeneity of that group. Cluster 11 was the only cluster we were unable to consistently label with the classification, since there are no prominent populations or cell types related only to it.
In this way, we achieved a ground reference classification for the dataset, and all the clusters have been identified with cell types. Therefore, from now on, one can start formulating biological hypotheses and not only supposition when performing analysis. As the last thing, we calculated the values of the statistical metrics giving as results:
RI = 0.834823
ARI = 0.3124383
MI = 0.9600847
NMI = 0.3672774
There is an overall but slight increase in the metrics, showing how the two partitions of the cells agree with each other, demonstrating that the unsupervised algorithm is recognizing the cellular heterogeneity.

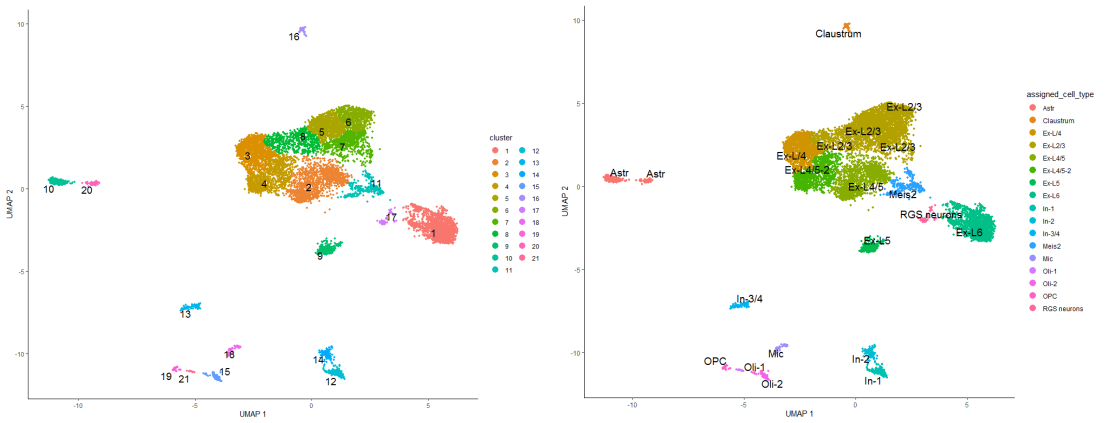

Figure 4.5: Monocle unsupervised clusters labeled after the classification analysis

4.2 Label transfer classification

For completeness, we performed the classification of the cells through a second method. It consists of Integration and Label Transfer[31] of two separate datasets. The method aims to first identify anchors, which represent couples of individual cells coming from the two datasets that are likely to be from the same biological state, that are then used to transfer information from one dataset to another. In this way, one can take one dataset already processed, where the cells have been identified, and use it to classify a distinct dataset. Using this method one obtains complete labeling of all the cells, which are derived from an analysis of all the expression profiles and not only some chosen markers. The drawback is that obviously, one can not be sure of the quality of the classification of the processed dataset, which is still a challenging step, but one can assume it can be better than the analysis made on few marker genes. The Seurat pipeline provides some helpful functions to proceed in this calculation.
First of all, we loaded the scRNA dataset to employ as the reference for the classification of the cells. we chose an Allen dataset of the mouse brain, precisely from the Primary Visual Cortex (Visp)[58].Satija Lab already processed the dataset, and one can directly access its SeuratObject counterpart. The first step is to identify the possible anchors between the two. The function "FindTransferAnchors" performs exactly this step, and takes as input the reference dataset, the query dataset (the one to be labeled), and a reduction method that can be CCA or PCA (recommended for scRNA experiments). The function returns a list of anchors that on can then use to transfer the labels. This is done through the function "TransferData", which takes as input the previous results and returns a list of identities for all the cells. The Seurat Object stores the cell type identities as new metadata (through "AddMetaData"). Since the method is based on the Seurat pipeline, firstly we

plotted the labels on the Seurat plot (Fig.4.6). With this classification, there are

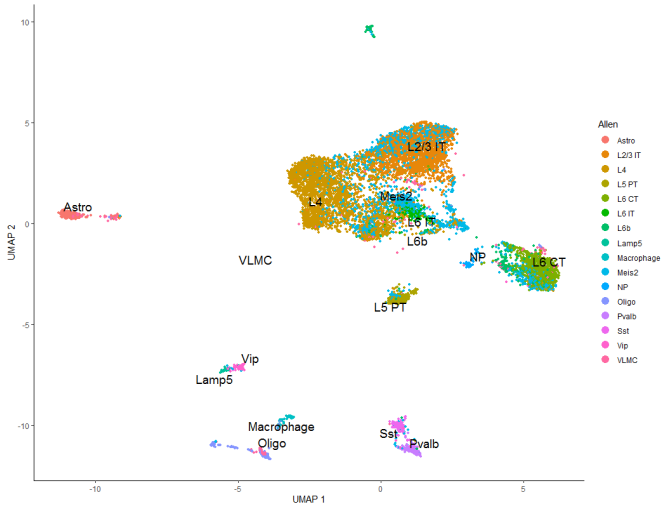

Figure 4.6: Monocle UMAP visualization with the labels from the Allen transfer label method

only 16 identities. Seven of them are Excitatory Neurons, which include the already well-known types based on layers (L4), some more specific cell types (L6IT, L6CT), NP neurons that are related to previously mentioned RGS neurons, and Meis2 that is a general classification of neurons based on the marker Meis2. Instead, for Non-Neuronal cell type, the labels are similar, apart from Oligodendrocytes that are not divided into multiple subtypes apart from VLMC cells (vascular leptomeningeal cells). To be consistent with what previously done, we plotted the predicted labels on the Monocle plot. From it, one can notice that the cluster previously labeled as Claustrum cells, has no separate identity, but contains mainly L6b cells. It makes sense because the claustrum is a small part of the brain near the inner layers, so it can be misrecognized as L6 neurons. Cluster 9 with the previous classification had a small population of L5 neurons but was too small to confidently label the whole group. However, now the L5PT type identifies the same cluster, so it validates the previous hypothesis. For the remaining cells, the two classifications agree with each other, at least for the general clusters' identities. In the same way as before, the predicted labels one can compare to the unsupervised clusters with the metrics:
RI = 0.8439087
ARI = 0.3038634
MI = 1.351484
NMI = 0.5097729
One can notice an increase, especially in Mutual Information. This could be a consequence of the lower number of partition in the predicted labels, and also the

more uniform classification than the previous one.
In conclusion, the transferring label method leads to a complete classification of the dataset based on a pre-processed dataset. It is helpful since, even if its classification could have the same limitation as the marker strategy, it is probably more complete than what we implemented previously. The problem is that it is an indirect analysis since it is the result of integration calculations, that insert multiple points of uncertainty. However, it is useful to try to validate the direct marker analysis with an uncorrelated classification. In the end, what wen obtained are two classifications to evaluate the clustering algorithms and to identify the clusters, so one can perform a joint analysis of expression and accessibility based also on biological knowledge, instead of only unsupervised algorithm results.

Cell type name	Acronym	Marker Genes	Function	Number of labeled cells
Excitatory Neurons Layer 2/3	Ex-L2/3	Rasgrf2	Rasgrf2 > 0	644 cells
Excitatory Neurons Layer 3/4	Ex-L3/4	Rorb	Rorb > 0	754 cells
Excitatory Neurons Layer 3/4b	Ex-L3/4-2	Rmst	Rmst > 0	117 cells
Excitatory Neurons Layer 4/5	Ex-L4/5	Thsd7a	Thsd7a > 0	260 cells
Excitatory Neurons Layer 4/5b	Ex-L4/5-2	Il1rapl2	Il1rapl2 > 0	587 cells
Excitatory Neurons Layer 5	Ex-L5	Galnt14	Galnt14 > 0	168 cells
Excitatory Neurons Layer 5b	Ex-L5-2	Parm1	Parm1 > 0	77 cells
Excitatory Neurons Layer 5/6	Ex-L5/6	Tshz2	Tshz2 > 0	118 cells
Excitatory Neurons Layer 5/6b	Ex-L5/6-2	Sulf1	Sulf1 > 0	80 cells
Excitatory Neurons Layer 6	Ex-L6	Tle4	Tle4 > 0	228 cells
Claustrum cells	claust	Nr4a2	Nr4a2 > 0	44 cells
Inhibitory Neurons 1° type	In-1	Pvalb	Pvalb > 0	10 cells
Inhibitory Neurons 2° type	In-2	Sst	Sst > 0	17 cells
Inhibitory Neurons 3° type	In-3	Npy	Npy > 0	23 cells
Inhibitory Neurons 4° type	In-4	Vip	Vip > 0	13 cells
Astrocytes	Astr	Slc1a3	Slc1a3 > 0	150 cells
Oligodendrocytes 1° type	Oli-1	Itpr2	Itpr2 > 0	50 cells
Oligodendrocytes 2° type	Oli-2	Mal	Mal > 0	47 cells
Oligodendrocytes progenitors	OPC	Vcan	Vcan > 0	53 cells
Microglia	Mic	Apbb1ip	Apbb1ip > 0	20 cells

Table 4.1: Cell types with their markers and results of the first run

Cell type name	Acronym	Marker Genes	Function	Number of labeled cells
Excitatory Neurons Layer 2/3	Ex-L2/3	Rasgrf2	Rasgrf2 > 0.75	1021 cells
Excitatory Neurons Layer 3/4	Ex-L3/4	Rorb	Rorb > 1	587 cells
Excitatory Neurons Layer 4/5	Ex-L4/5	Il1rapl2	Il1rapl2 > 1	503 cells
Excitatory Neurons Layer 5	Ex-L5	Parm1, Bcl11b	Parm1 > 0 AND Bcl11b > 0	16 cells
Excitatory Neurons Layer 5/6	Ex-L5/6	Sulf1, Hs3st2	Sulf1 > 0.25 AND Hs3st2 > 0.3	26 cells
Excitatory Neurons Layer 6	Ex-L6	Foxp2, Syt6	Foxp2 > 0.5 OR Syt6 > 0.5	644 cells
RGS Neurons	RGS	Tshz2	Tshz2 > 0.5	207 cells
Claustrum cells	claust	Nr4a2, Col11a1	Nr4a2 > 0.5 OR Col11a1 > 0.8	259 cells
Inhibitory Neurons 1° type	In-1	Pvalb	Pvalb > 0	24 cells
Inhibitory Neurons 2° type	In-2	Sst	Sst > 0	42 cells
Inhibitory Neurons 3° type	In-3	Npy	Npy > 0	34 cells
Inhibitory Neurons 4° type	In-4	Vip	Vip > 0	27 cells
Astrocytes	Astr	Slc1a3, Apoe	Slc1a3 > 0.5 OR Apoe > 0.5	426 cells
Oligodendrocytes 1° type	Oli-1	Itpr2, Tcf7l2	Itpr2 > 0.5 OR Tcf7l2 > 0.5	91 cells
Oligodendrocytes 2° type	Oli-2	Mal, Mog	Mal > 0.6 OR Mog > 0.6	145 cells
Oligodendrocytes progenitors	OPC	Vcan, Sox6	Vcan > 0.5 AND Sox6 > 0.8	207 cells
Microglia	Mic	Siglech	Siglech > 0.5	28 cells

Table 4.2: Cell types with their markers and results of the second run

Cell type name	Acronym	Marker Genes	Function	Number of labeled cells
Excitatory Neurons Layer 2/3	Ex-L2/3	Rasgrf2	Rasgrf2 > 0	1221 cells
Excitatory Neurons Layer 3/4	Ex-L3/4	Rorb	Rorb > 1	595 cells
Excitatory Neurons Layer 4/5	Ex-L4/5	Il1rapl2	Il1rapl2 > 1	465 cells
Excitatory Neurons Layer 4/5b	Ex-L4/5-2	Cntn5	Cntn5 > 4	298 cells
Excitatory Neurons Layer 5	Ex-L5	Parm1, Bcl11b	Parm1 > 1 OR Bcl6 > 1.5	28 cells
Excitatory Neurons Layer 5/6	Ex-L5/6	Sulf1, Hs3st2	Sulf1 > 1 AND Hs3st2 > 0.3	11 cells
Excitatory Neurons Layer 6	Ex-L6	Foxp2, Syt6	Foxp2 > 0.5 OR Syt6 > 0.5	773 cells
RGS Neurons	RGS	Tshz2	Tshz2 > 2	50 cells
Claustrum cells	claust	Nr4a2, Col11a1	Nr4a2 > 1.5 OR Col11a1 > 2	53 cells
Inhibitory Neurons 1° type	In-1	Pvalb	Pvalb > 0	22 cells
Inhibitory Neurons 2° type	In-2	Sst	Sst > 0	50 cells
Inhibitory Neurons 3° type	In-3	Npy	Npy > 0	38 cells
Inhibitory Neurons 4° type	In-4	Vip	Vip > 0	29 cells
Astrocytes	Astr	Slc1a3, Apoe	Slc1a3 > 1 OR Apoe > 1	176 cells
Oligodendrocytes 1° type	Oli-1	Itpr2, Tcf7l2	Itpr2 > 1 OR Tcf7l2 > 1	9 cells
Oligodendrocytes 2° type	Oli-2	Mal, Mog	Mal > 1 OR Mog > 1	52 cells
Oligodendrocytes progenitors	OPC	Vcan, Sox6	Vcan > 0.5 AND Sox6 > 0.8	32 cells
Microglia	Mic	Siglech	Siglech > 1.5	16 cells

Table 4.3: Cell types with their markers and results of the third run

Chapter 5

Joint Analysis

The last part of this work focus on the joint analysis of the scRNA and scATAC[59][60]. The main goal is to understand if the clustering made with the different datasets are consistent with each other. In particular, we worked with the idea that if a gene identifies the same group of cells both at the transcriptional and epigenetic level, it is probably a sign of consistency, and one can more reliably say that the unsupervised cluster is identifying a cell type.
Until now, the datasets have been studied separately, but, as pointed out, the real strength of SNARE-seq is the double information coming from the same set of cells. For this reason,one can study the correlation of accessibility and expression directly without the need for further hypothesis and integration of two separate datasets. However, it is not that simple, since the biological processes that control the epigenetic mechanisms and transcription events are complex and intricate.We briefly reviewed the biological aspect in the first chapter, but here it is useful to go further on that, especially on the epigenetic part. When one talks about epigenetics is referring to all the process that, without altering the DNA sequence, produce heritable phenotype changes. It includes a great variety of mechanisms like DNA methylation or histone modification, and most of them directly affect gene activity and expression. This because the DNA is structurally packed into a chromatin state, which includes the so-called nucleosome, DNA wrapped around a protein complex called histones. If a region coding for a gene is packed in this structure, it is not accessible and the transcription can not happen. Therefore, there must be a modification in the chromatin state to allow these processes. In eukaryotes organisms, the epigenetic changes are a key factor to cellular differentiation where, during morphogenesis, stem cells become pluripotent cells and then fully differentiated cells. During the differentiation, there are a lot of epigenetic changes, but in an adult fully developed organism, these occur to a lesser extent especially for the normal cell regulation but are more frequently linked to different types of diseases. Hence, unlike gene expression that is constantly regulated and

therefore has a fast dynamic, the accessibility is more settled or at least works on larger time scales than expression changes. One must remember that the single-cell experiments can be viewed as snapshots in time of the cells state, so it is likely to see greater differences between expression profiles rather than between epigenetic profiles, making it even more challenging to properly study the relation between the two. One can understand that the task is not trivial, but this section aims to elaborate, with different approaches, the data, making great use of the property of the SNARE-seq, and trying to answer some questions:

- Is the clustering process partitioning the cells in the same way in both the datasets, or there are sensitive differences?
- Are differently accessible peaks also differently expressed?
- Are there direct relations between marker genes and their accessibility?

But also more in-depth:

- Is the gene activity matrix analysis informative?

Through different tools provided both by Signac and Cicero packages, it has been elaborated the data and results are reported in the following sections.

5.1 Clustering superposition

From chapters 3 and 4,the clustering process divided the 10309 cells of the datasets into multiple clusters. At first, based on the gene expression profiles and then on the accessibility. We performed it with both the already well-known pipelines providing different but qualitatively similar results, but without investigating more in-depth, this is just a superficial consideration. Therefore in this section, we compare all the clustering results, with each other and with the cell type classifications, to understand the differences and find the right resolution values.
The first thing to do is to transfer the labels directly. We already implemented something similar in chapter 2 with the gene expression clusters between the results of Monocle and Seurat (Fig. 2.13). Now the same operation has been performed between the results of expression and accessibility. Remember that during the clustering of ATAC data with Cicero, we adopted two resolutions, which had found a different number of clusters. Fig. 5.1 reports both of them with also the RNA cluster labels.
Before looking at them, however, is helpful to think about what one could expect from this type of analysis, or in other words, how could the clusters of the two datasets be linked:

- Clusters are coupled in a 1:1 manner.

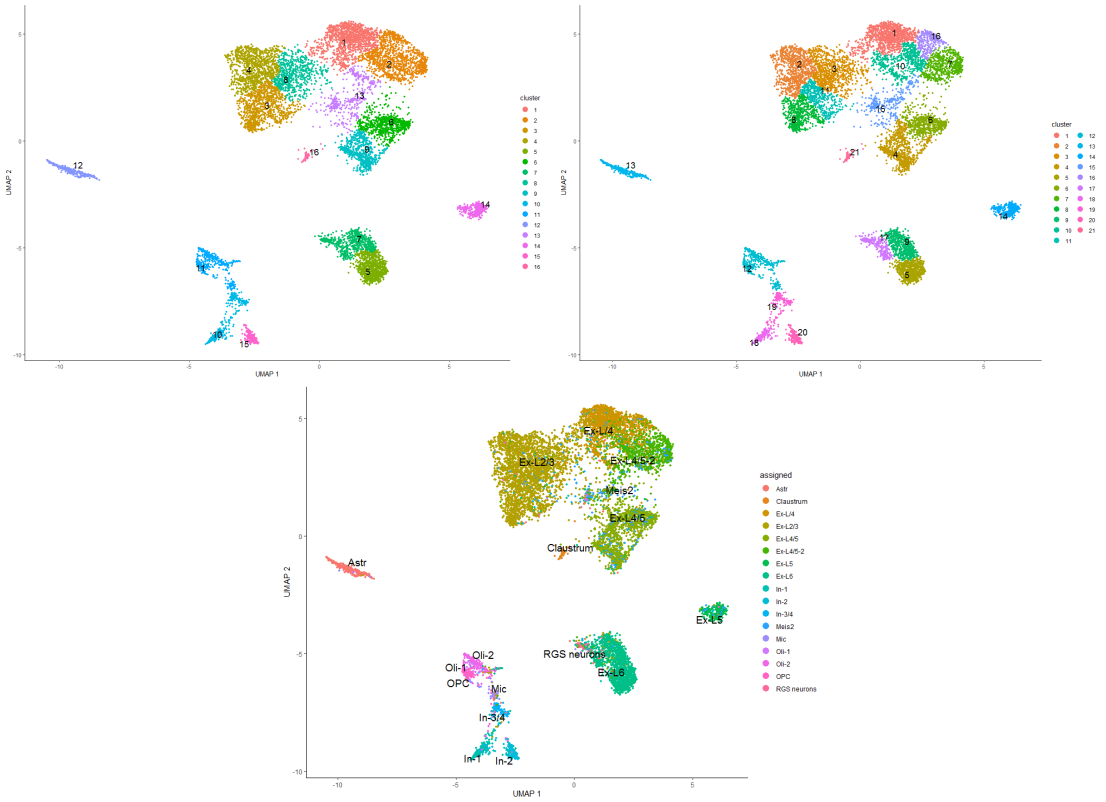

Figure 5.1: Epigenetic data processed with Cicero, with Monocle labels

- Some distinct expression clusters are together for the accessibility data.
- Some clusters in the transcriptomic data are divided in the epigenetic data.

These three cases summarize what can happen in general with a multimodal analysis. It can occur that the two modes are recognizing the same features (first case), or one of the two is finding more sources of heterogeneity (cases two and three). Now, looking at Fig. 5.1, the results are interesting. First of all, the population of Excitatory Neurons found with RNA data superposes well with the big group of cells plotted, even if the division into clusters inside it, it is not well-defined. This is caused by the fact that, especially for the case with higher resolution, there are more clusters identified with the ATAC data. Cell groups like the Claustrum cells have a clear counterpart in the epigenetic data, meaning they are well-defined for both biological aspects. Instead, the family of Oligodendrocytes appears as a unique cluster, meaning that probably their differences rise from the expression profiles only. The most intriguing case is the cluster of EX-L6 cells, which the ATAC data identifies as two separate groups. This is interesting because one can remember that from the marker classification, there was a small population of cells,

not belonging to L6 and not belonging to any other cluster. So the ATAC clustering may have found some heterogeneity, not appreciable with the transcriptomic data only. The problem is that with the lower resolution some differences in smaller clusters are not present (like for Inhibitory Neurons), but the higher one appears to overly partition the Excitatory Neurons. Therefore it is useful to apply the metrics used in the previous chapter to the two clustering resolutions to understand which better agrees. We calculated the Normalized Mutual Information between the two clustering results and the Monocle cluster results, the marker classification, and the transfer label classification. The Table 5.1 reports them. As one can notice the

	Resolution 1	Resolution 1.6
Monocle clustering	0.614105	0.5829951
Marker classification	0.3397086	0.3135536
Transfer label classification	0.4578342	0.424748

Table 5.1: Normalized Mutual Information of ATAC clustering

lower resolution gives better results with all three. So it is fair to assume that as the reference for the ATAC clustering. Nevertheless, the peculiarity of the Ex-L6 cells cluster, which emerged with the higher resolution, will be not discarded.
We qualitatively analyzed the result of Monocle and Cicero and we started the joint analysis of transcriptomic and epigenetic. But during this work, we also profoundly utilized the Seurat pipeline, and therefore it is helpful to review its clustering results and compared them with the previous ones. Again the first thing has been to transfer the labels from Seurat to Signac (Fig.5.2). The results are similar, again the groups of Oligodendrocytes are grouped, also like the Inhibitory Neurons. This confirms that accessibility can identify cell types on a higher level. Unfortunately, there is no more additional information from this visualization, but it is useful to compare the NMI of the various partitions. One can see from the Table 5.2 two things. Cluster partitions from the same biological data have higher concordance (higher NMI values), which one can expect because they derive from the same dataset. Second, the marker classification and the transfer label classification agrees better with the transcriptomic data, also expected since they derive from an expression analysis. However, overall, there is a good agreement between all the cell partitions meaning that the data from the two biological processes are recognizing similar heterogeneities between the cells. But this is only an initial qualitative observation, and therefore is better to study what are the similar features that the different analyses are commonly finding. To do so we implemented the so-called gene activity matrix to assess the accessibility of the genes.

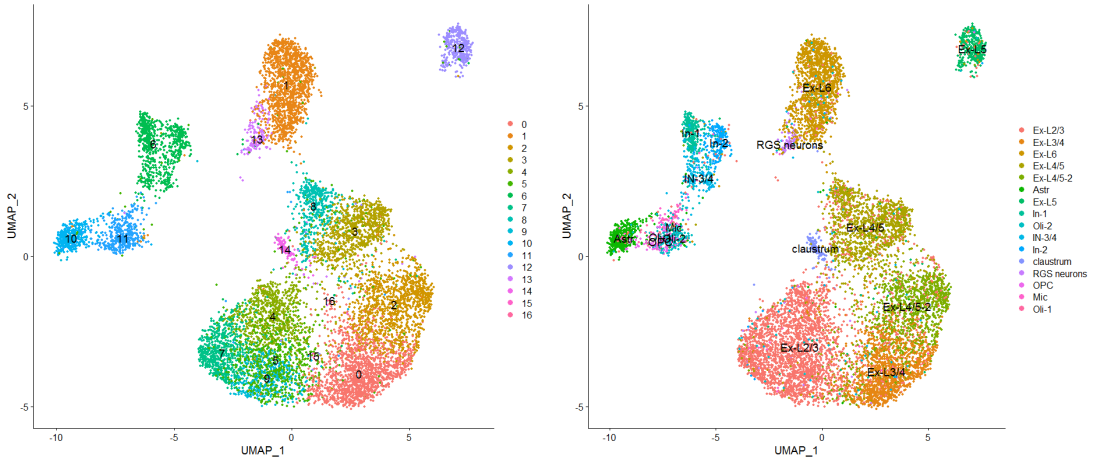

Figure 5.2: Epigenetic data processed with Signac, with Seurat labels

	Seurat RNA	Seurat ATAC
Seurat clustering RNA	//	0.5636672
Monocle clustering RNA	0.676613	0.5846018
Cicero clustering ATAC	0.5898904	0.6890831
Marker classification	0.3788516	0.3396632
Transfer label classification	0.5288887	0.4474821

Table 5.2: Normalized Mutual Information of different classifications.

5.2 Gene Activity Matrix

After this first more qualitative analysis, it is time to attempt to evaluate the correlation of expression and accessibility. The first approach is something cited already in section 3.1. With the Cicero pipeline, we employed its functions to calculate the Cicero connections. The latter are couples of peaks that show co-accessibility, i.e. they have similar patterns of accessibility between cells. This is useful because one must remember that peaks do not uniquely identify genes. First of all, the order of magnitude of their length is much smaller than the genes. Peaks are small regions that can be contained within the coding gene lengths, but not necessarily. Some of them mark DNA regions that are not protein-coding but are relevant as well. For this reason, looking for a relationship between genes and peaks is not trivial. Difficulties arise from the gene regulation process, which includes

a wide range of mechanisms that can increase or decrease the transcription of a specific gene. Some mechanism involves some specific proteins (like Transcriptional Factors or Repressor) which bind to specific regions of DNA and interact with the RNA polymerase complex inhibiting or allowing the start of the transcription. These regions to which the proteins bind can be near the promoter but can also be very far away. For this reason, to evaluate the general accessibility of a gene, one needs to look at three things:

- The accessibility of the promoter region: this is the most important since the promoter is the region where the RNA polymerase binds and starts the transcription.
- The accessibility of peaks inside the gene body: these are less important but can also contribute to the overall accessibility.
- The accessibility of binding regions: as explained above, these are regions that directly regulate the transcription, so their accessibility is relevant.

The last point is the most problematic since the networks of transcriptional factors can be complex, and knowing all their binding regions becomes difficult to analyze. Therefore, instead of looking at all the possible peaks that could be binding sites, we employed the results of the co-accessibility calculations. The underlying hypothesis is that co-accessible peaks define regulatory correlations, in particular, the ones that are co-accessible to the promoter regions are likely to be binding sites to regulatory complexes. With this information, one can try to create the so-called gene activity matrix, which is a matrix, similar to the count matrices for expression, that has the cells as column names, the genes as rows, and the element of the matrix are the gene activity scores. The scores are related to the three aspects mentioned above. With the help of Cicero tools, one can create a gene activity matrix that takes into account all of them, using the co-accessibility evaluated by the same pipeline. The workflow to do so is the following. The first thing to do is to identify the peaks that are promoters and annotate them. To do so, one needs to use an annotation file, which contains all the coordinates and various information for each gene. From it, one takes into consideration the first exon of each transcript, and annotate it as a one base coordinate to indicate the start of the gene sequence. With the function "annotate_cds_by_site", that for each peak indicates the gene if it is its promoter or NA otherwise, one can store the promoter annotations is then stored in the ATAC CDS . After this, there are two steps. First, the function "build_gene_activity_matrix", which takes as input the ATAC CDS and the connection file generated from co-accessibility, and generates an unnormalized gene activity matrix. For a quality check, it is useful to eliminate every row or column with 0 entries. Second is the normalization of the matrix through the function "normalize_gene_activities", which takes as input the list

of the number of accessible sites per cell (can be easily found in the pData of the CDS, called "num_genes_expressed"). The result is a matrix with 15430 rows (genes promoters) and 10309 columns (same number of cells). The first thing that jumps out is that the number of features is less than half of the number of genes in the expression count matrix that was 33160. This means that more than half of the genes did not have peaks on promoters. The cause can be due to experimental errors or biological reasons. The first is related to the peak calling step of the experiment, which could not have been able to correctly identify some peaks, due to too low fragment number. The biological reason is that the experimental "snapshot" of the cells, could have been done in a time of changes in chromatin accessibility, but with some RNA molecules still present. A third possibility is that some of the genes detected in the RNA experiments were present only for few cells and their accessibility signals were not strong enough to call a peak. Anyways, once we obtained this matrix, the idea has been to process it as a count matrix with the Monocle pipeline. The results are interesting. As always,we clustered the data with the unsupervised algorithm and visualized using UMAP dimensional reduction(Fig. 5.3). What is obtained is something different from the plots until now. Almost all

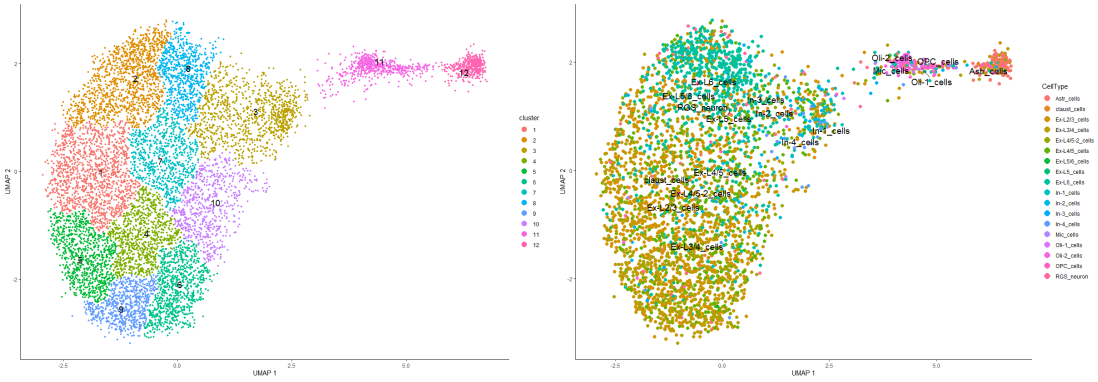

Figure 5.3: Cicero gene activity matix clustering and marker classfification

the cells are plotted in this big group, with only a small population of cells a bit detached. To understand, instead of looking at the clustering, we labeled the cells with the marker classification. In this way, one can understand the composition of the groups. The rightmost cluster identifies the Astrocytes, and the cluster near it is composed of all the cell type of the Oligodendrocytes family. All the remaining neuronal cells are all together but keeping a certain separation within the large group. In particular, the Inhibitor neuron cells are localized all together near the top edge. It is a bit clearer what is happening with the activity matrix. Unlike previous analyses, here the data are differentiating the cells on a more structural level. Instead of recognizing possible subtypes, it is acknowledging the general cell types, meaning it is dividing all the neurons (divided into excitatory, inhibitory,

etc.) from the Oligodendrocytes family and also the Astrocytes. It might appear as a downgrade from what we studied until now, but it is not completely true. We performed the differential analysis as before and the results are interesting. The top markers of the two right clusters appear to be mostly genes that encode for proteins that are Transcriptional Factors and Promoter. In particular, using the platform Uniprot[61] to learn about protein functionality,one can see that the markers of the Oligodendrocityes family (namely Olig1, Olig2, Sox1) are important proteins for the formation and maturation of oligodendrocytes. This is important because it is a cross-validation of the classification made. At first,we identified a small group of cells by the unsupervised algorithm, then we labeled it with the literature markers classification, and now with another type of information, there is the confirmation that they are oligodendrocytes. For the Astrocytes, the top markers do not show particular specificity with that cell type, but the majority of them seem to be related to cell-fate determination, neuronal development, and adult nervous system postnatal development[62]. Unfortunately, a similar analysis for the neuronal cells does not bring much to the table. Instead of performing differential analysis between the unsupervised cluster, we performed it also between the cell types, hoping to find more information. Unfortunately, the results are not better; the ones related to the already mentioned cell types agree, but concerning other cell types did not help more. The functional analysis should be done with more biological knowledge background to find possible relations between these "active" genes and their cell types[63].
The latter results come from the analysis of the gene activity matrix obtained with Cicero, which takes into account the co-accessibility values and, therefore, is applying a complex model. One could argue that maybe the connection calculation is not reliable enough, but it is better to look only at promoter accessibility. Thus, we implemented also this version of the gene activity matrix using Signac. the pipeline provides the function "GeneActivity" which similarly creates the matrix, but only investigating the gene body and promoter regions[64]. This is a simpler model, but not necessarily less valid. The first evident difference is that now the number of features increased (21991 genes). As said also before, the reasons are not clear, maybe, in this case, a simpler model identifies more active genes.
Anyway, again we processed the matrix as previously. We clustered and visualized it, and then we labeled the cells with the marker classification. The results are shown in Fig. 5.4. Again the cells appear to be less well separated, but one can notice that the Inhibitory Neurons are more distinct from other neurons and are also identified with a cluster. However, the clustering algorithm has distinguished a lot of small clusters between the neuron cells, which are probably not so informative. It has also not been able to separate the Astrocytes from the Oligodendrocytes. So, overall, there are some differences from Cicero. It is useful, therefore, to perform a differential analysis, right away between the cell types, to understand if the two gene

activity matrices agree with each other. However, if one tries to look for the same active genes that we found for the Oligodendrocytes, they are not present between the list provided by the Seurat differential analysis. Another way to attempt to detect informative differentially active genes is to perform the differential analysis between the cluster obtained from the epigenetic data. In this way, the cells are partitioned in slightly different ways, and most importantly the total number of cells is higher. For example, all the Inhibitor Neurons are in one cluster and are taken into account all the cells and not only the ones that were labeled with the marker classification. Even if the results do not change that much, it is worth mentioning that between the active genes for the Inhibitor Neurons, we identified the genes GAD1 and GAD2 which are strictly related to this type of cells, meaning, once more, that the gene activity is cross-validating the expression results.

In general, the gene activity matrix obtained with Signac could appear less reliable

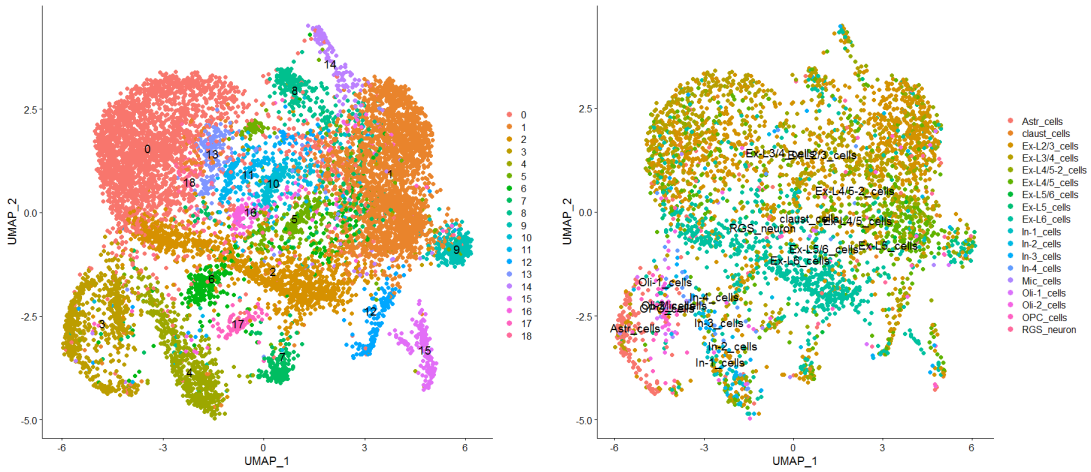

Figure 5.4: Signac gene activity matix clustering and marker classfification

but, in reality, further examinations show interesting results. The following section reports these additional studies, where we tried to combine all the results obtained until now, as a conclusion to the work.

5.3 Comparative analysis of differentially expressed genes, accessible peaks, and active genes

The aim of this last section is to look for the accessibility of marker genes and differentially expressed genes to understand if epigenetic data can validate the expression results.

Summarizing what we have done until now, it all started with the gene expression analysis, which produced a clustering division that was hypothesized to represent cellular heterogeneity. Between these clusters, we identified differentially expressed genes which identify the differences between them based on the data. Next, we studied the epigenetic data, which contain information about the accessibility of genomic regions. Again we performed independently clustering and differential analysis, obtaining information about which peaks were significant for each cluster. After that, we implemented a classification of the cells based on the expression of literature markers. This helped to provide a biological background to the dataset. So in this last chapter,we compared the different clusterings and classifications with each other, showing strong similarities. However, this kind of comparison is a qualitative analysis, and it is better to study it more technically. In particular, it is interesting to understand the correlation between differentially accessible peaks and expressed genes inside clusters. To do so we analyzed the features linked to accessible peaks and their expression, and vice versa. The goal is to see if a difference in accessibility leads to a difference in expression so one can understand if the features that characterized one cluster are the same in both transcriptomic and epigenetic data, and therefore the two information agree on the classification. We started with the observation of the accessible peaks resulted from the differential analysis. To link them to a gene, Signac provides the function "ClosestFeature". It takes a list of genomic regions and employing the annotation file, finds the closest gene to it. Through the list of differentially expressed genes one can search for the obatined genes, hoping to find a match. Unfortunately, this does not happen. This methodology is too simplistic since it assumes that peaks and genes are directly related in a 1:1 manner, but we already explained this is not precise. This proves that epigenetic data must be studied in a connected way instead of as solitary features like expression data. Fortunately, we already described the concept of gene activity matrix, which is the center of this last analysis. The workflow has been like this:

- One starts considering the two gene activity matrices separately, starting with the Signac one.
- We performed the differential analysis of the gene activity, considering five different partitions, Monocle (MG) and Seurat (SG) gene expression clusters, Cicero (CA) and Signac (SA) epigenetic clusters, and the marker classification (Type).
- From the resulting lists of genes, we looked for the presence of a list of genes that includes markers and differentially expressed genes.
- Tables 5.4 and 5.3 report all the results.

The Tables have inside them the number or name of the clusters in which the gene is differentially expressed. The column Agree indicates if the cluster in which the gene is differentially active is in agreement with the cluster that the gene is originally liked to. The X means that no classification detected the gene. Yes means that at least one classification detected it and the results agree with at least half of them. No goes accordingly. The idea is to find some genes that are relevant for the classification at a transcriptional level and an epigenetic level.
Starting from the Signac gene activity matrix, looking at table 5.4, one can see that a good number of genes from the list are differentially active. In particular, the set of markers that we used to identify the Excitatory Neurons appear to be relevant for the differential analysis regardless of the cell partition considered. What does it mean? Take, for example, the gene Rorb, which is the marker of the Ex-L3/4, and was also between the differentially expressed genes of chapter 2. It appears to be differentially active in some clusters that for each classification identify the exact cell type of the marker. So this means no matter the pipeline used, no matter the biological level observed, Rorb identifies the same group of cells, and therefore it is an optimal marker, and one can label those cells confidently with that specific cell type. A different situation is for the markers of the Inhibitor Neurons. They do not appear at any level. This is warring because they seem to not be good markers, even if they were used in the classification. However, two differentially expressed genes, Adarb2 and Erbb4, identify the clusters assigned with the Inhibitor labels and appear to be also differentially active in the same group of cells. Therefore, they are good candidates to be markers of these cell types. The same reasoning can be done for the gene Slc1a2 that appears to be a better marker, at least at an epigenetic level, than Scl1a3, which is strongly related to Astrocytes. Are worth mentioning also the genes Sox6 and Hs3st2. The first is the marker for OPC cells and the function identify it in the correct cluster for the expression clusterings (MG, SG), but when one considers the ATAC clusterings, the analysis identify it as differentially accessible for the Inhibitors Neurons clusters. This could explain the ambiguity between the two families of cell types, found during the marker classification. The second instead is a marker for the Ex-L5/6 cells, but it appears to be differentially active for clusters identified as Ex-L5. The ambiguity arises probably from the fact that cells from close layers are more difficult to distinguish. However, this could be the reason behind the fact that Ex-L5/6 cells were only small number, since the chosen marker appear to be ambiguous at an epigenetic level. The last gene to be mentioned is Fam19a1. It has been discussed also during the differential expression analysis since it appeared to identify multiple clusters. The same uncertainty emerged from the activity investigation, meaning that, even if it is not a good marker for a single cell type, it is strongly differentiated for these various groups of cells both at the transcriptomic and epigenetic level. Thus the consistency between expression and activity helps to validate the first. Looking

instead to the Cicero gene activity matrix, things are different. Here the great majority of the genes on the list do not appear to be differentially active. We found just four of them, reported in table 5.3. Vip gene appear only between the In-4 cells of the marker classification, and no others, so it is not enough to make more hypothesis. The genes Bcas1 and Erbb4, instead, agree with before, in particular, Erbb4 confirms the reliability as a marker for the Inhibitory Neuron cell types (it is also in agreement with other studies[65]). The last one is the Apoe gene, which interestingly enough, was not in previously list. We used it as an Astrocytes marker and appears to be differentially active for exactly the clusters that are labeled as Astrocytes. The Cicero gene activity matrix, as previously explained, derives from the study of the various connection between accessible peaks and it is probably the cause of the discrepancy in the active genes found.
In conclusion to answer the questions posed in the introduction of this chapter:

- Is the clustering process partitioning the cells in the same way in both the datasets, or there are sensitive differences?
  Yes, the unsupervised clustering processes implemented on the datasets agree if each other, meaning that the classification is consistent throughout the different biological levels.
- Are differently accessible peaks also differently expressed?
  No, the accessibility of peaks alone does not show correspondence with the expression of the nearest gene.
- Are there direct relations between marker genes and their accessibility?
  Yes, we found relations between some marker genes and their accessibility inside different clusters.
- Is the gene activity matrix analysis informative?
  Yes, the gene activity matrix appears to be informative, not only to identify accessibility of the known genes but also to find the accessibility of new characterizing genes at a functional level (like Oligo1).

In general, the epigenetic analysis has brought to the confirmation of transcriptomic analysis. In particular, the ATAC data, even if the study of accessibility alone does not add much, showed how for the same groups of cells, the genes that are differentially expressed are also differentially accessible. The latter is not a trivial statement, as the biological mechanisms that regulate gene expression can be complicated and hide direct relations like that. Therefore, it is fair to say that the joint analysis of epigenetic and transcriptomic data helps to improve the study of cellular heterogeneity, for two reasons. First, genes that identify a group of cells at both biological levels make it strongly consistent, confirming that those cells

probably belong to the same biological type. This could help the identification of new cell types that maybe are identified by unsupervised clustering algorithms and can be validated by a gene activity investigation. Second, the gene accessibility analysis could bring to the identification of epigenetic markers. Therefore, one could try to classify cells, like what we done in chapter 4, looking simultaneously at marker expression and marker accessibility.

Gene	MG	SG	CA	SA	Type	Agree
Apoe	10	8	10	X	Astr	Yes
Bcas1	21	20	X	X	X	Yes
Erbb4	12	11	X	X	X	Yes
Vip	X	X	X	x	In-4	Yes

Table 5.3: Differentially active genes for Cicero activity matrix of different classifications.

Gene	MG	SG	CA	SA	Type	Agree
Rasgrf	5/6/7	0	4/7/9	2/4/9	Ex-L2/3	Yes
Rorb	3	1/7	0/2	1	Ex-L3/4	Yes
Il1Rapl2	X	5	4/7	8	X	No
Cntnt5	4	6	2	3	Ex-L4/5b	Yes
Parm1	X	X	X	X	X	X
Bcl6	X	X	X	X	X	X
Sulf1	X	X	X	X	Ex-L5/6	No
Hs3st2	9	10	12	13	X	No
Foxp2	1	2	1	X	Ex-L6	Yes
Syt6	X	X	X	X	X	X
Tshz2	17	16	13	X	RGS	Yes
Nr4a2	16	15	14	16	X	Yes
Col11a1	X	X	X	X	X	X
Pvalb	X	X	X	X	X	X
Sst	X	X	X	X	X	X
Vip	X	X	X	X	X	X
Npy	X	X	X	X	X	X
Slc1a3	X	X	X	X	X	X
Apoe	X	X	X	X	X	X
Itpr2	14/19	14/18	X	11	X	Yes
Tcf7l2	21	20	X	X	Oli-1	Yes
Mal	X	X	X	X	X	X
Mog	X	X	X	X	X	X
Vcan	X	X	X	X	X	X
Sox6	21/19	18/20	6	10/15	In-1, In-2	No
Siglech	X	X	X	X	X	X
Bcas1	21	20	X	X	Oli-1	Yes
Otof	X	X	X	X	X	X
Plp1	X	X	X	X	X	X
Grin3a	X	X	X	X	In-2	X
Adarb2	13	13	14	10	In-3, In-4	Yes
Erbb4	12/13	11/13	6	10	In-1,In-3	Yes
Atp1a2	X	X	X	X	X	X
Lhfpl	X	X	10	X	X	X
Slc1a2	10	8	X	12	Astr	Yes
Camk2n1	X	X	X	X	X	X
Rbm25	X	X	X	X	X	X
Fam16a1	5/6/7/9	0/3/9/10	!2/4/7/9	2/9/13/14	Ex-L2/3	Yes

Table 5.4: Differentially active genes for Signac activity matrix of different classifications.

Chapter 6

Conclusions and future works

With this thesis work, we wanted to understand if the joint analysis of transcriptomic and epigenenomic data could help the cellular heterogeneity study. It started with a common scRNA-seq analysis that is what we want to improve or at least validate with the addition of epigenetic data. We obtained a first division of the cells in clusters based on the transcriptomic data, and the differential analysis produced a list of differentially expressed genes. From here we tried to answer four questions:

- To which results does the epigenetic data analysis alone lead?
- What are the cell types identified by the gene expression clustering?
- Do the transcriptomic results agree with the epigenetic ones?
- Are there features that characterize clusters on both biological levels?

For the first one, we proceeded to study the epigenetic data with the same workflow. From it, one can see how one can also use the accessibility data to cluster the cells in a way that appears to be similar to the RNA data, meaning that it is informative for the cellular heterogeneity studies. However, the differential accessibility analysis seems difficult to perform and does not bring much information. We classified the cells in the dataset with a marker expression investigation that allowed us to identify cell types and label the cluster at a biological level. This showed that the expression data clustering was recognizing particular brain cell types. In particular, we identified eight Excitatory neuron types, four Inhibitory neuron types, and various Glia cells like Astrocytes and Oligodendrocytes. We compared the different subdivisions, showing that epigenetic and transcriptomic data agree with each other, and also their clusters are quite well identifiable with

the cell types just discussed. However, it appears that accessibility data do not completely distinguish between certain subtypes but are only able to recognize more general types, like the Inhibitory neurons and the Oligodendrocytes. Through the implementation of the gene activity matrix we have been able to link the expression and the accessibility of some genes. In this way, we demonstrated how groups of cells are identified by certain genes at both levels, meaning that the epigenetic data can validate the results of the expression data. In this way, we have been able to validate the initial hypothesis, that is the joint analysis of epigenetic and transcriptomic data can help the validation of pipelines' results and, therefore, improve cellular heterogeneity studies.
Moreover, this is just the starting point. We want to suggest some possible continuations of this work. First of all, one can study the activity and the expression of all the features of the dataset. In this way, one could find more genes that are specific for a cell type at both levels, making them good marker candidates for future works. Second, since we saw how the epigenetic information is identifying cell types at a more general level, one could try to reliably identify macro families of cells with accessibility analysis and focus on them with the expression analysis to find possible subtypes. One last possible continuation could be based on the Cicero activity matrix. We noticed how, through it, one can find interesting features strongly related to certain cell types, despite they were not particularly relevant at the expression level. Also, more in general, this could be the starting point for a more deep analysis of the activity of genes as the accessibility of all factors involved in the transcription process. In particular, it seems that the research in this field is sleeping on the possibilities of the Cicero gene activity and its informative power.

Anyway, this field is incredibly active and increasingly promising, thanks to more technological advances for the experiments and different ways to approach the data, and I hope my work might, even minimally, help future researches.

Bibliography

[1] Kandror Elena K Rizvi Abbas H Camara Pablo G et al. «Single-cell Topological RNA-seq Analysis Reveals Insights into Cellular Differentiation and Development.» In: *Nature Biotechnology* 35.6 (2017), pp. 551–60 (cit. on p. 1).

[2] Sarah A. Teichmann Ashraful Haque Jessica Engel et al. «A practical guide to single-cell RNAsequencing for biomedical research and clinical applications.» In: *Genome Medicine* 18 (2017), 9:75 (cit. on p. 4).

[3] *Explore Illumina sequencing technology.* URL: `https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html` (cit. on p. 4).

[4] *Resolving Biology to Advance Human Health.* URL: `https://www.10xgenomics.com/` (cit. on p. 4).

[5] Lee Ji Hyun Hwang Byungjin and Bang Duhee. «Single-cell RNA Sequencing Technologies and Bioinformatics Pipelines.» In: *Experimental Molecular Medicine* 50.8 (2018), pp. 96–14 (cit. on pp. 4, 15).

[6] Johan L M Björkegren Oscar Franzén Li-Ming Gan. «PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data». In: *Databse* 2019 (Apr. 2019), baz046 (cit. on p. 6).

[7] E.S Lein et al. «Genome-wide atlas of gene expression in the adult mouse brain». In: *Nature* 445 (2007), pp. 168–176 (cit. on p. 6).

[8] *Cell Census Network (BICCN).* 2017 (cit. on p. 6).

[9] Lash AE. Edgar R Domrachev M. «Gene Expression Omnibus: NCBI gene expression and hybridization array data repository». In: *Nucleic Acids Res.* 30(1) (2002), pp. 207–10 (cit. on p. 7).

[10] *National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06].* URL: `https://www.ncbi.nlm.nih.gov/` (cit. on p. 7).

[11] Grimsby Jonna Trapnell Cole Cacchiarelli Davide et al. «The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells.» In: *Nature Biotechnology* 32.4 (2014), pp. 381–86 (cit. on p. 8).

[12] Qiu Xiaojie Cao Junyue Spielmann Malte et al. «The Single-cell Transcriptional Landscape of Mammalian Organogenesis.» In: *Nature (London)* 566.7745 (2019), pp. 496–502 (cit. on p. 8).

[13] Berg Jim Gouwens Nathan W Sorensen Staci A et al. «Classification of Electrophysiological and Morphological Neuron Types in the Mouse Visual Cortex.» In: *Nature Neuroscience* 22.7 (2019), pp. 1182–195 (cit. on p. 8).

[14] *Satija Lab.* URL: `https://satijalab.org/seurat/` (cit. on p. 9).

[15] McCarthy Davis J GHuang Yuanhua and Stegle Oliver. «Vireo: Bayesian Demultiplexing of Pooled Single-cell RNA-seq Data without Genotype Reference.» In: *Genome Biology* 20.1 (2019), p. 273 (cit. on p. 9).

[16] Knights Andrew Heaton Haynes Talman Arthur M et al. «Souporcell: Robust Clustering of Single-cell RNA-seq Data by Genotype without Reference Genotypes.» In: *Nature Methods* 17.6 (2020), pp. 615–20 (cit. on p. 9).

[17] Nguyen Quan Xu Jun Falconer Caitlin et al. «SGenotype-free demultiplexing of pooled single-cell RNA-seq.» In: *Genome Biology* 20.1 (2019), pp. 1–290 (cit. on p. 9).

[18] Valentine Svensson. «Droplet ScRNA-seq Is Not Zero-inflated.» In: *Nature Biotechnology* 38.2 (2020), pp. 147–50 (cit. on p. 10).

[19] Yang Ying SDing Hongxu Blair Andrew et al. «Biological Process Activity Transformation of Single Cell Gene Expression for Cross-species Alignment.» In: *Nature Communications* 10.1 (2019), p. 4899 (cit. on p. 10).

[20] Fenna M. Krienen et al. «Innovations present in the primate interneuron repertoire.» In: *Nature (London)* (2020), pp. 262–69 (cit. on p. 10).

[21] Teichmann Sarah A Efremova Mirjana Vento-Tormo Miquel et al. «CellPhoneDB: Inferring Cell-cell Communication from Combined Expression of Multi-subunit Ligand-receptor Complexes.» In: *Nature Protocols* 15.4 (2020), pp. 1484–506 (cit. on p. 10).

[22] Lake Blue B Chen Song and Zhang Kun. «High-throughput Sequencing of the Transcriptome and Chromatin Accessibility in the Same Cell.» In: *Nature Biotechnology* 37.12 (2019), pp. 1452–457 (cit. on pp. 10, 40).

[23] Meyer Clifford A Zhang Yong Liu Tao et al. «Model-based Analysis of ChIP-Seq (MACS).» In: *GenomeBiology.com* 9.9 (2008), R137 (cit. on p. 11).

[24] Tang Ying Qiu Xiaojie Mao Qi et al. «Reversed Graph Embedding Resolves Complex Single-cell Trajectories.» In: *Nature Methods* 14.10 (2017), pp. 979–82 (cit. on p. 13).

[25] Packer Jonathan Qiu Xiaojie Hill Andrew et al. «Single-cell mRNA Quantification and Differential Analysis with Census.» In: *Nature Methods* 14.3 (2017), pp. 309–15 (cit. on p. 13).

[26] *Drop-seq*. URL: `http://mccarrolllab.org/dropseq/` (cit. on p. 13).

[27] Smibert Peter Butler Andrew Hoffman Paul. «Integrating Single-cell Transcriptomic Data across Different Conditions, Technologies, and Species.» In: *Nature Biotechnology* 36.5 (2018), pp. 411–20 (cit. on p. 13).

[28] Satija Rahul Stuart Tim Butler Andrew et al. «Comprehensive Integration of Single-Cell Data.» In: *Cell (Cambridge)* 177.7 (2019), 1888–902.e21 (cit. on pp. 13, 17).

[29] Kim Jong Kyoung Ilicic Tomislav et al. «Classification of Low Quality Cells from Single-cell RNA-seq Data.» In: *Genome Biology* 17.1 (2016), p. 29 (cit. on p. 14).

[30] Kim Jong Kyoung Brennecke Philip Anders Simon et al. «Accounting for Technical Noise in Single-cell RNA-seq Experiments.» In: *Nature Methods* 10.11 (2013), pp. 1093–095 (cit. on p. 15).

[31] Hoffman Paul Stuart Tim Butler Andrew and Satija 29 Rahul. «Comprehensive Integration of Single-Cell Data.» In: *Cell (Cambridge)* 177.7 (2019), 1888–902.e21 (cit. on pp. 15, 47).

[32] Xu Chen and Su Zhengchang. «Identification of Cell Types from Single-cell Transcriptomes Using a Novel Clustering Method.» In: *Bioinformatics (Oxford, England)* 31.12 (2015), pp. 1974–980 (cit. on p. 16).

[33] Simonds Erin F Levine Jacob H et al. «Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells That Correlate with Prognosis.» In: *Cell (Cambridge)* 162.1 (2015), pp. 184–97 (cit. on p. 16).

[34] Guillaume Jean-Loup Blondel Vincent D et al. «Fast Unfolding of Communities in Large Networks.» In: *Journal of Statistical Mechanics* 2008.10 (2008), P10008–12 (cit. on p. 17).

[35] Campello Ricardo JGB Jaskowiak Pablo A and Costa Ivan G. «On the Selection of Appropriate Distances for Gene Expression Data Clustering.» In: *BMC Bioinformatics* 15.S2 (2014), S2 (cit. on p. 17).

[36] Van Der Maaten L.J.P and Hinton G.E. «Visualizing High-Dimensional Data Using T-SNE.» In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–605 (cit. on p. 17).

[37] Healy John Becht Etienne McInnes Leland et al. «Dimensionality Reduction for Visualizing Single-cell Data Using UMAP.» In: *Nature Biotechnology* 37.1 (2018), pp. 38–44 (cit. on p. 17).

[38] Cacchiarelli Davide Trapnell Cole et al. «The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells.» In: *Nature Biotechnology* 32.4 (2014), pp. 381–86 (cit. on p. 22).

[39] *Matrix: Sparse and Dense Matrix Classes and Methods*. URL: `http://Matrix.R-forge.R-project.org/` (cit. on p. 23).

[40] L. Waltman N. J. van Eck V. A. Traag. «From Louvain to Leiden: guaranteeing well-connected communities». In: *Scientific Reports* 9.1 (2019), pp. 1–12 (cit. on pp. 24, 32).

[41] Levine et al. «Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis». In: *Cell (Cambridge)* 162.1 (2015), pp. 184–97 (cit. on p. 24).

[42] F. Alexander Wolf et al. «PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells». In: *Genome Biology* 20.1 (2019), pp. 1–9 (cit. on p. 24).

[43] Xiaoyu Wei Chuanyu Liu Mingyue Wang et al. «An ATAC-seq atlas of chromatin accessibility in mouse tissues.» In: *Scientific Data* 6 (2019), p. 65 (cit. on p. 28).

[44] Kakumanu Akshay Velasco Silvia Ibrahim Mahmoud M et al. «A Multi-step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming from Embryonic Stem Cells.» In: *Cell Stem Cell* 20.2 (2017), 205–17.e8 (cit. on p. 28).

[45] Zhu Qin Yu Wenbao Uzun Yasin. «A Comprehensive Workbench for Single-cell Chromatin Accessibility Sequencing Data.» In: *Genome Biology* 21.1 (2020), p. 94 (cit. on p. 28).

[46] Hannah A. Pliner et al. «Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data». In: *Molecular Cell* 71.5 (2018), 858–871.e8. ISSN: 1097-2765. DOI: `https://doi.org/10.1016/j.molcel.2018.06.044` (cit. on pp. 28, 29).

[47] Tim Stuart, Avi Srivastava, Caleb Lareau, and Rahul Satija. «Multimodal single-cell chromatin analysis with Signac». In: *bioRxiv* (2020). DOI: `10.1101/2020.11.09.373613`. URL: `https://doi.org/10.1101/2020.11.09.373613` (cit. on p. 28).

[48] Aghamirzaie Delasa Cusanovich Darren A Hill Andrew J et al. «A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.» In: *Cell (Cambridge)* 174.5 (2018), 1309–324.e18 (cit. on p. 29).

[49] *Drop-seq*. URL: https://timoast.github.io/sinto/basic_usage.html (cit. on p. 34).

[50] Sloan Cricket A Davis Carrie A Hitz Benjamin C et al. «The Encyclopedia of DNA Elements (ENCODE): Data Portal Update.» In: *Nucleic Acids Research* 46.D1 (2018), pp. D794–801 (cit. on p. 34).

[51] McFaline-Figueroa José L Srivatsan Sanjay R et al. «Massively Multiplex Chemical Transcriptomics at Single-cell Resolution.» In: *Science (American Association for the Advancement of Science)* 367.6473 (2020), pp. 45–51 (cit. on p. 34).

[52] *Analyzing PBMC scATAC-seq*. URL: https://satijalab.org/signac/articles/pbmc_vignette.html (cit. on p. 34).

[53] Codeluppi S. Zeisel A. Munoz-Manchado A. B. et al. «Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-cell RNA-seq.» In: *Science (American Association for the Advancement of Science)* 347.6226 (2015), pp. 1138–142 (cit. on p. 38).

[54] Wysoker Alec Saunders Arpiar Macosko Evan Z et al. «Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain.» In: *Cell (Cambridge)* 174.4 (2018), 1015–030.e16 (cit. on pp. 38, 40, 42).

[55] Rudolf S. N De Bont Eveline S. J. M De Jonge Hendrik J. M Fehrmann et al. «Evidence Based Selection of Housekeeping Genes.» In: *PloS One* 2.9 (2007), E898 (cit. on p. 38).

[56] Santos Jorge M and Embrechts Mark. «On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification.» In: *Artificial Neural Networks* 5769 (2009), pp. 175–84 (cit. on p. 41).

[57] Epps Julien Vinh Nguyen and Bailey James. «Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance». In: *Journal of Machine Learning Research* 11 (2010), pp. 2837–2854 (cit. on p. 41).

[58] *Cell Types Database: RNA-Seq Data*. URL: https://portal.brain-map.org/atlases-and-data/rnaseq (cit. on p. 47).

[59] Ramani Vijay Cao Junyue Cusanovich Darren A et al. «Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells.» In: *Science (American Association for the Advancement of Science)* 361.6409 (2018), pp. 1380–385 (cit. on p. 53).

[60] Schug Jonathan Ackermann Amanda M Wang Zhiping et al. «Integration of ATAC-seq and RNA-seq Identifies Human Alpha Cell and Beta Cell Signature Genes.» In: *Molecular Metabolism (Germany)* 5.3 (2016), pp. 233–44 (cit. on p. 53).

[61] *Uniprot*. URL: https://www.uniprot.org/ (cit. on p. 60).

[62] Haim Lucile Ben and Rowitch David H. «Functional Diversity of Astrocytes in Neural Circuit Regulation.» In: *Nature Reviews. Neuroscience* 18.1 (2016), pp. 31–41 (cit. on p. 60).

[63] Kaushal A Cahoy J. D Emery B et al. «A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function.» In: *The Journal of Neuroscience* 28.1 (2008), pp. 264–78 (cit. on p. 60).

[64] Jain Ashish Starks Rebekah R Biswas Anilisa et al. «Combined Analysis of Dissimilar Promoter Accessibility and Gene Expression Profiles Identifies Tissue-specific Genes and Actively Repressed Networks.» In: *Epigenetics Chromatin* 12.1 (2019), p. 16 (cit. on p. 60).

[65] Bandler Rachel C Mayer Christian Hafemeister Christoph et al. «Developmental Diversification of Cortical Inhibitory Interneurons.» In: *Nature (London)* 555.7697 (2018), pp. 457–62 (cit. on p. 64).