FACULTY OF SCIENCE
Charles University

BACHELOR THESIS

Matěj Zátko

Detection of subpopulation-specific neuronal membrane molecules using single-cell expression data

Department of Cell Biology

Supervisor of the bachelor thesis:	Mgr. Martin Modrák, Ph.D.
Study programme:	Bioinformatika
Study branch:	B-BINF

Prague 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V .... dne ....

Podpis autora

First and foremost, I would like to thank my family for their immense support throughout my entire bachelor studies. I also thank my supervisor Martin Modrák and my advisor Helena Janíčková for answering all my questions, giving out their opinions and guiding me towards the goal of creating a good thesis.

Title: Detection of subpopulation-specific neuronal membrane molecules using single-cell expression data

Author: Matěj Zátko

Department: Department of Cell Biology

Supervisor: Mgr. Martin Modrák, Ph.D., Institute of Microbiology of the Czech Academy of Sciences

Advisor: MUDr. Mgr. Helena Janíčková, Ph.D., Institute of Physiology of the Czech Academy of Sciences

Abstract: Single-cell RNA sequencing is a powerful technology that allows the investigation of gene expression at an unprecedented level. Insights into gene expression in individual cells can help biologists uncover cellular heterogeneity and identify previously unknown cell types. Here, we use single-cell RNA sequencing datasets that reveal subtypes of mouse neurons to find population-specific membrane proteins. These proteins could potentially serve as entry points for targeted drug distribution, allowing for drugs to act only on selected neuronal populations. We start by identifying five suitable single-cell mouse neuron datasets. Next, we present an overview and a comparison of currently available methods for differential gene expression analysis, an approach that involves quantifying variations in gene expression between groups and/or conditions, based on previous benchmarks. Lastly, we apply the Wilcoxon rank-sum test to selected datasets in order to identify population-specific membrane proteins.

Abstrakt: Single-cell RNA sekvenování nám umožňuje zkoumat genovou expresi na nebývalé úrovni. Informace o transkripci genů v jednotlivých buňkách může poukázat na dříve nerozpoznatelné rozdíly a pomoci nám odhalit nové buňečné typy. Zde jsme použili existující single-cell RNA datasety k nalezení populačně-specifických membránových proteinů u populací myších neuronů. Tyto membránové proteiny by mohly být později využity k zacílení léčby na konkrétní neuronové populace. Nejprve jsme identifikovali 5 vhodných single-cell datasetů. Následně jsme, na základě předchozích testů, porovnali stávající metody pro analýzu diferenciální exprese, techniku, která zkoumá rozdíly v expresi genů mezi buňkami. Na závěr jsme k identifikaci populačně-specifických membránových proteinů v datasetech použili Wilcoxonův test.

Keywords: bioinformatics, RNA-seq, neuron, membrane proteins

Klíčová slova: bioinformatika, RNA-seq, neuron, membránové proteiny

Contents

Introduction

Single-cell RNA sequencing (scRNA-seq) is a powerful technology that enables the investigation of gene expression in individual cells. Traditional RNA sequencing measures gene expression in a bulk sample of cells and thus provides an average of gene expression across the entire cell population. In contrast, scRNA-seq measures gene expression in each individual cell, providing a more detailed view of heterogeneity within a population. This technology can be used to identify previously unknown cell types, study developmental trajectories of cells, identify novel drug targets or investigate disease mechanisms at the single-cell level.

This thesis aims to use publicly available scRNA-seq datasets to identify membrane molecules specific to individual neuronal populations in mice, as a part of an ongoing research project. Today's drugs target surface molecules/receptors but do not target a specific neuronal population. If we were to find a population-specific membrane molecule with suitable properties, we could select a suitable ligand to bind to it. Through the process of internalization, the membrane molecule alongside the ligand and the drug bound to it would then be transferred inside the cell. Thus, enabling us to target drugs to a particular population of neurons.

The goal of the work is to, first, identify suitable publicly available datasets of single-cell gene expression data in mouse neurons and their populations. Second, present a brief overview of published bioinformatics methods for determining genes that are specifically expressed by different cell populations. And third, apply one of the methods to selected datasets and use it to identify membrane proteins expressed by specific neuronal populations.

The thesis is composed of three chapters. Chapter 1 discusses how the datasets were selected and introduces each one. Chapter 2 aggregates previous benchmarks to compare differential gene expression analysis methods, which can be used to identify population-specific genes. Finally, in Chapter 3, we perform the analysis and identify population-specific membrane proteins in the selected datasets using the Wilcoxon rank-sum test.

1. Single-cell mouse neuron datasets

To identify appropriate single-cell RNA-seq datasets that describe murine neuronal populations, PubMed was used with the search queries "revealed mouse neuron types, single cell RNA sequencing" and "mouse neuron populations, single cell RNA sequencing" along with the "Associated data" filter enabled. In total, over 50 articles were considered but most of them were either irrelevant (not about scRNA-seq and cell types), about sensory neurons (which are not of interest to us) or only contained raw sequencing data without cluster assignment. In the end, three relevant datasets were identified and used for the analysis (Zeisel et al., 2015; Tasic et al., 2016; Chen et al., 2017), along with two additional datasets as per request of the advisor (Campbell et al., 2017; Bhattacherjee et al., 2019). Most datasets were available from the Gene Expression Omnibus database (Barrett et al., 2012). However, the files lacked any annotation, which made finding the required data difficult, as the only cue about what the files contained were their cryptic names. This shortcoming could be overcome by annotating each supplementary file and describing its contents. This would make the database more clear and user-friendly.

1.1 Overview of selected datasets

Zeisel et al. constructed a cellular taxonomy of the mouse cortex and hippocampus. Single-cell RNA sequencing was used to analyze 3005 single cells from the primary somatosensory cortex (S1) and the hippocampal CA1 region. Because traditional hierarchical clustering led to fragmented clusters, Zeisel et al. developed their own clustering method BackSPIN that identified 9 major cell types: S1 and CA1 pyramidal neurons, interneurons, oligodendrocytes, astrocytes, microglia, vascular endothelial cells, mural cells (i.e. pericytes and vascular smooth muscle cells) and ependymal cells. The clustering was then repeated on the 9 major populations to reveal 47 distinct cell subpopulations in total. The three major neuronal populations, S1 pyramidal neurons, CA1 pyramidal neurons and interneurons, had 8, 5 and 16 subtypes respectively.

Tasic et al. studied cell types in the primary visual cortex (V1) of adult mice. After single-cell RNA sequencing of more than 1600 cells, two iterative clustering methods were used to reveal 49 different cell subpopulations, 23 of which were GABAergic, 19 glutamatergic and 7 non-neuronal. A machine-learning method was used for validation and to find "intermediate" cells (cells that belong to multiple clusters). Out of the 23 GABAergic cell types, 18 belonged to previously described classes (Vip, Pvalb and Sst) and 5 were novel. Among the glutamatergic neurons, 6 major layer-specific types were identified in correspondence to previous studies (L2/3, L4, L5a, L5b, L6a and L6b), but further subdivision revealed a total of 19 glutamatergic subpopulations. The 7 non-neuronal cell populations were astrocytes, microglia, oligodendrocyte precursor cells, two types of oligodendrocytes, endothelial cells and smooth muscle cells. Correspondence with cell types identified by Zeisel et al. was also provided.

Chen et al. analyzed the transcriptomes of over 14,000 single cells from the hypothalamus. From the 14,000 cells, only those that expressed more than 2000 genes were used for clustering. Semi-supervised clustering of 3,319 cells that met the above condition revealed 45 distinct cell populations. Based on known markers, 34 of these populations were classified as neuronal and the remaining 11 as non-neuronal. The neuronal types were further divided into 15 glutamatergic and 18 GABAergic subtypes and one additional histaminergic subtype. The neuronal cell types were found to be largely hypothalamus-specific, while the non-neuronal populations were similar to other brain regions.

Campbell et al. profiled gene expression of 20,921 cells from and around the hypothalamic arcuate-median eminence complex (Arc-ME). In total, 34 neuronal (24 of which were from Arc-ME) and 36 non-neuronal (26 from Arc-ME) populations were revealed. After the first round of clustering, each cluster was assigned one of the following identities based on the expression of marker genes: neurons, ependymocytes, tanycytes, oligodendrocyte lineage cells, oligodendrocyte precursor cells, macrophages, endothelial cells, mural cells and astrocytes. Further sub-clustering of the neuronal cells identified 34 clusters, with most of them associated with unique candidate markers. However, some did not have a distinct marker and were instead defined by a combination of markers or lacked any markers altogether.

Bhattacherjee et al. performed scRNA-seq of 29,864 single cells from the prefrontal cortex (PFC). Clustering revealed 8 major cell populations, 2 of which were neuronal: excitatory and inhibitory neurons. The excitatory neurons, largest population in the PFC, were found to be composed of 13 distinct subtypes. The inhibitory neurons, which form a much smaller population compared to the excitatory neurons, were classified into 12 clusters. To enable a better comparison of the PFC neurons with other cortical regions, the excitatory neurons were clustered once again, but at a higher resolution, resulting in 26 subtypes. Most subtypes had corresponding clusters in other cortical regions, but some were unique to the PFC. Despite the similarities across regions, there were still significant differences in gene expression between corresponding groups.

2. Identification of population-specific genes

To identify genes that are specific to a certain population of cells, so-called marker genes, differential gene expression analysis (DGE) can be employed. Differential expression analysis is a widely used method in bioinformatics that compares levels of gene expression between biological groups or conditions. The goal of this analysis is to find genes that are up- or down-regulated in response to a certain stimulus or between groups. This is important to gain insight into the biological differences between cells under different conditions (e.g. healthy and diseased) or between groups, such as different cell types.

It is, however, important to note that not all differentially expressed genes (DEGs) are marker genes. While some DEGs may be specific to certain cell types, many have functions that are not related to cell type or may be involved in common biological processes shared by multiple cell types. A gene can also be upregulated in one cell population, but still expressed in others, meaning that it is differentially expressed but may not be considered a marker in a conventional sense. It is important to validate the expression of differentially expressed genes using additional methods to ensure their relevance for research-specific purposes.

Numerous DGE tools and techniques are available. To provide an overview and a comparison of the most important ones, Google Scholar was used with the search query "comparison of differential expression analysis tools for single cell" and the top 4 relevant review articles were selected (Dal Molin et al., 2017; Soneson and Robinson, 2018; Wang et al., 2019; Das et al., 2021). Furthermore, because only one of the aforementioned articles covered the popular single-cell analysis platform Seurat (Hao et al., 2021) and none covered Scanpy (Wolf et al., 2018) or scran (Lun et al., 2016), one additional article was taken into consideration (Pullin and McCarthy, 2022).

Some tools were developed specifically for analyzing single-cell data, such as MAST, SCDE, scDD, DEsingle or D^{3}E, while others like edgeR, DESeq2 and limma were initially designed for bulk RNA-seq data, but are now also being utilized for single-cell data. The methods can be also classified as either parametric, meaning they assume a certain distribution of the data (such as Poisson or negative binomial) or non-parametric, which are distribution-free. Additionally, the methods differ in the test statistics they use. An overview of discussed DGE methods is provided in Table 2.1.

In contrast to methods based on differential expression, some methods employ a different approach. RankCorr (Vargo and Gilbert, 2020) uses a feature selection algorithm to identify molecular markers, NSForest (Aevermann et al., 2021) uses a machine learning-based approach and Cepo (Kim et al., 2021) is based on the idea that stable gene expression is a key indicator of cell identity.

Method	Origin	Modeling paradigm	Citation
BPSC	single-cell	parametric	Vu et al. (2016)
D^{3}E	single-cell	non-parametric	Delmans and Hemberg (2016)
DECENT	single-cell	parametric	Ye et al. (2019)
DESeq	bulk	parametric	Anders and Huber (2010)
DESeq2	bulk	parametric	Love et al. (2014)
DEsingle	single-cell	parametric	Miao et al. (2018)
EBSeq	bulk	parametric	Leng et al. (2013)
edgeR	bulk	parametric	Robinson et al. (2010)
EMDomics	single-cell	non-parametric	Nabavi et al. (2016)
limma	bulk	parametric	Ritchie et al. (2015)
MAST	single-cell	parametric	Finak et al. (2015)
Monocle	single-cell	parametric	Trapnell et al. (2014)
Monocle 2	single-cell	parametric	Qiu et al. (2017)
NODES	single-cell	non-parametric	Sengupta et al. (2016)
Presto	single-cell	non-parametric	Korsunsky et al. (2019)
ROTS	bulk	non-parametric	Seyednasrollah et al. (2016)
SAMseq	bulk	non-parametric	Li and Tibshirani (2013)
scDD	single-cell	non-parametric	Korthauer et al. (2016)
SCDE	single-cell	parametric	Kharchenko et al. (2014)
SigEMD	single-cell	non-parametric	Wang and Nabavi (2018)
SINCERA	single-cell	non-parametric	Guo et al. (2015)
Welch's t-test	general	parametric	Welch (1947)
Wilcoxon test	general	non-parametric	Wilcoxon (1945)

Table 2.1: Overview of DGE methods

2.1 Comparison of differential gene expression methods

2.1.1 Bulk vs. Single-cell analysis

Single-cell RNA analysis poses several new challenges in comparison to bulk RNA. One of the primary ones is increased technical noise due to the lower amount of RNA in individual cells. Another challenge is so-called "dropout events", where genes may not be detected at all because of low expression levels. Other challenges include high heterogeneity, low library sizes, large zero-read counts and multimodality (the presence of multiple subpopulations within a larger cell population). Because of these challenges, it was unclear whether traditional DGE methods developed for bulk RNA-seq would also perform well for scRNA-seq data.

Surprisingly, all reviews found that bulk methods perform just as well on single-cell data despite their specific characteristics, while also being generally faster, because of their simpler design. However, Soneson and Robinson notes that some bulk RNA methods showed a stronger dependence on data prefiltering and Pullin and McCarthy found them to be more memory intensive in comparison to single-cell tools.

2.1.2 Datasets

Usually, a combination of real and simulated datasets was used to assess the performance of the various DGE methods. Simulated datasets were used because, unlike real data, they can be used to model different data distributions corresponding to different scenarios as well as provide complete knowledge about which genes are differentially expressed and which are not.

Dal Molin et al. simulated 10,000 genes across two conditions with a sample size of 100 cells each. Out of the 10,000 genes, 2000 of them were simulated as differentially expressed according to four distinct data distributions that aim to model multimodality. This resulted in four groups of 500 DEGs each and 8000 non-DEGs. The procedure was repeated 10 times in order to generate 10 independent datasets. All datasets were created using scripts provided in the scDD package.

Wang et al. also used the scDD package to generate 10 simulated datasets. They, however, used 75 single cells for each of the two conditions and 20,000 genes. Again, 2000 genes were simulated with different distributions and equally divided into four groups. The rest were modeled as non-differentially expressed genes. Apart from multimodality, dropout events were also simulated by introducing a large number of zeros. When assessing the effect of sample size, they used a range of 10 to 400 cells, instead of the fixed 75.

On the other hand, Pullin and McCarthy used the splat simulation model from the Splatter package (Zappia et al., 2017) to create over 170 simulated datasets. Parameters for the model were estimated from real data. Ten simulation scenarios were considered, each with 2000 genes, 2000 cells and 5 clusters. The different simulation scenarios had their parameters estimated from different real datasets.

Soneson and Robinson utilized a modified version of the powsim R package (Vieth et al., 2017) to create 3 two-group simulated datasets based on 3 real datasets. These were then subsampled to generate multiple dataset instances at different sample sizes. Each instance was modeled to have 10% of the genes either up- or downregulated. Mean and dispersion were estimated from the respective real datasets using edgeR.

2.1.3 Performance

True positive (TP) is defined as a truly differentially expressed gene that was marked by the tool as such. However, if the tool incorrectly identified a gene that isn't differentially expressed as being differentially expressed, this is referred to as a false positive (FP). On the other hand, a gene that the tool correctly identifies as not being differentially expressed is referred to as a true negative (TN), while a gene that is differentially expressed but is not identified as such by the tool is a false negative (FN).

The reviews evaluated the performance of the investigated tools using a multitude of metrics, such as true positive rate ($\frac{TP}{TP+FN}$), false positive rate ($\frac{FP}{FP+TN}$), false discovery rate ($\frac{FP}{FP+TP}$), precision ($\frac{TP}{TP+FP}$), accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$), F1 score ($\frac{2TP}{2TP+FP+FN}$) and more.

Simulated data

The reviews agree that there is a trade-off between precision and the number of true positives identified. This is a common phenomenon in any classification task. Dal Molin et al. found that tools which were able to identify a large number of truly differentially expressed genes, like D^{3}E and Monocle, were also the least precise, introducing a great number of false positives. This is consistent with Soneson and Robinson, who rank Monocle amongst the worst in false discovery rate. In contrast, the most precise tools (MAST, SCDE and DESeq) identified lower numbers of TPs.

Wang et al. found that Monocle 2 was able to identify the greatest number of true positives, but was also the least precise. Generally, the non-parametric methods were able to identify more TPs at the cost of precision, while the parametric methods (like MAST, SCDE, edgeR and SINCERA) did the opposite: high precision and low number of TPs. In terms of accuracy and F1 score, DEsingle and SigEMD performed the best, reporting a high number of TPs and not many FPs.

Pullin and McCarthy ranked scran's Binomial-any method, Wilcoxon rank-sum based methods (in Seurat and Scanpy), edgeR, as well as Student's t-test as the best tools regarding true positive rate (TPR). The worst methods in terms of TPR were NSForest, Cepo and RankCorr. On the contrary, RankCorr and NSForest as well as scran's Binomial-any method were the most precise. When ranked by the F1 score, scran's Binomial-any method, the Wilcoxon rank-sum-based methods and edgeR came out on top. NSForest, Cepo and scran's other Binomial methods showed the worst performance. In general, the best-performing methods were the ones based on the Wilcoxon rank-sum test, Student's t-test or logistic regression.

Soneson and Robinson evaluated the false discovery rate (FDR) and true positive rate of each method. Methods that performed well in terms of both FDR and TPR after filtering include edgeR/QLF (edgeR with Quasi-Likelihood F-test), SAMseq, DEsingle and voom/limma. Methods that were best able to rank truly differentially expressed genes ahead of non-differential ones, were edgeR, followed by MAST, limma, SCDE, DEsingle, DESeq2, SeuratBimod (Seurat's likelihood-ratio test) without filtering and the non-parametric methods like the t-test, Wilcoxon test and D^{3}E. Prefiltering improved the performance of most methods.

Real data

On real data, Wang et al. discovered that Monocle 2, EMDomics, SINCERA, D^{3}E and DEsingle had the highest TPRs, whereas SCDE, scDD and MAST the lowest. Once again this came at the cost of precision. Gene set enrichment analysis was also performed. Surprisingly, scDD and SCDE were unable to recover stem cell biology terms relevant to the dataset used. Other methods performed well. This indicates that certain methods are better at finding biologically significant genes.

Across all datasets, Das et al. found EBSeq and DECENT followed by edgeR/QLF to be the best tools in terms of F1 score, accuracy and FDR, whereas scDD, NODES, EMDomics, ROTS and DEsingle consistently performed the worst. Regarding sensitivity and specificity, DECENT and EBSeq once again performed the best, while other methods had high sensitivity and low specificity or vice versa.

It is important to note that when the performance of tools is assessed using real data, we do not have complete knowledge about the expression of genes. The reviews rely on lists of differentially expressed genes that were obtained through experimental validation. However, it is possible that some DEGs might have been overlooked, or some genes might have been incorrectly marked as differentially expressed.

Negative control

Most reviews also used some dataset(s), which contained no DEGs, as a negative control to test tool performance, expecting no DEGs reported. According to Dal Molin et al., all tools perform well, detecting 0 DEGs, except for D^{3}E, which consistently detected over 250 DEGs across the datasets.

Soneson and Robinson found that without prior filtering, the best-performing tools were ROTS and SeuratTobit (no longer included in the Seurat package), whereas edgeR/QLF and SeuratBimod introduced the largest number of FPs. When genes with low expression were filtered out, the performance of most methods improved. They also observed that the FPs of different methods had distinct characteristics. False positives of NODES, ROTS, SAMseq and SeuratBimod were highly expressed genes that had few zeros, while false positives of edgeR/QLF, SeuratTobit and MAST had many zeros.

Wang et al. states that MAST, SCDE, edgeR, and SINCERA did not find any DEGs, as expected. In contrast, DEsingle, scDD, DESeq2, SigEMD, D^{3}E, EMDomics, and Monocle 2 all found some DEGs, with EMDomics and Monocle 2 (which performed best in terms of TPR) introducing the most false positives.

2.1.4 Consensus

All review articles found low consensus among the analyzed tools. Dal Molin et al. revealed a large discrepancy between the different methods. The number of differentially expressed genes (DEGs) reported by the tools varied between 271 and 8,401 when using real data. The highest number of genes was reported by D^3E, but it also reported the most false positives among all tools.

The eleven tools tested by Wang et al. had only 92 DEGs in common when considering the top 1000 reported genes of each method. Moreover, only 41 of them were included in the gold standard list. Pairwise agreement between tools was not observed either.

Soneson and Robinson and Das et al. both state that the level of similarity between different methods varied significantly across different datasets. The performance of the tools was also highly inconsistent across the datasets.

2.1.5 Effect of sample size

As reported by Soneson and Robinson and Wang et al., all methods show increased TPRs with increased sample size, as expected. Wang et al. observed a large increase in precision when they increased the number of cells from 10 to 75. Further increase from 75 to 400 still yielded an improvement but less dramatic. This highlights the importance of sample size in DGE analysis. With the increased sample size, Monocle 2, EMDomics and DESeq2 were able to achieve TPRs near 100%, but showed bad FPRs. However, DEsingle and SigEMD were able to score well both in terms of TPR and FPR. Interestingly, Das et al. found that some tools, including edgeR and DESeq, performed better with datasets that contained a relatively small number of cells.

2.1.6 Effect of multimodality and dropout events

The authors that modeled multimodality into their simulated data, like Dal Molin et al. and Wang et al., found that in scenarios with a high degree of multimodality, the tools generally perform worse, with lower TPR and precision, than in the cases of low multimodality or no multimodality at all. Wang et al. who also modeled dropout events into their simulated datasets found that all tools perform worse when large amount of zero counts is introduced.

2.1.7 Speed

Another important aspect of DGE tools, and any computational software for that matter, is processing speed. This becomes especially important as the number of cells in a dataset increases.

Dal Molin et al. tested the tools in both serial and parallel settings. Predictably, tools that support parallel execution saw a substantial speed increase.

In general, all tools performed well, except for D^{3}E, which took 4 days to finish. Also, methods made for bulk RNA-seq performed generally faster than the more complicated scRNA-seq methods.

Soneson and Robinson ran all methods in serial mode, only utilizing a single core, for a fair comparison. BPSC, DEsingle, D^{3}E and SCDE, were the slowest and tools for bulk RNA-seq were generally faster.

Wang et al. found that simpler methods, like SINCERA, edgeR, MAST, Monocle 2 and DESeq2 ran fast, while scDD was the slowest. Additionally, the non-parametric methods, such as SigEMD, EMDomics and D^{3}E, ran slower than model-based methods.

According to Das et al., DECENT was the best-performing method, but it was also the most computationally intensive. EBSeq, ROTS, EMDomics, and NODES were also among the slower methods. Simple methods, like the t-test and Wilcoxon test, performed relatively well and were also the fastest.

Pullin and McCarthy found edgeR, Seurat's Negative Binomial, MAST and NSForest to be the most time-consuming, while most of Scanpy's methods, Presto, Cepo and RankCorr were the fastest. Interestingly, it was also revealed that Seurat's methods are much slower than scran or Scanpy's even when employing the same statistical tests. Memory usage was also measured and it was found that bulk RNA-seq methods are generally more memory intensive than single-cell-specific ones.

2.1.8 Ease of use

Most methods are implemented in R and are usually available via Bioconductor, CRAN or a GitHub repository. Additionally, Bioconductor packages include documentation and guides to help the users familiarize themselves with the tool. Some tools (like Seurat, Scanpy or SCDE) even have their own websites with documentation, tutorials, forums and more.

Some reviews ran into technical problems, which made it impossible to test certain methods and which should be addressed. Dal Molin et al. had to exclude BASiCS (Vallejos et al., 2016) because of its unconventional input and scDD because it required an unstable version of R. Pullin and McCarthy were unable to include several tools. SMaSH (Nelson et al., 2022), because of issues with parallelization. Scanpy's Logistic regression, because it did not return what was documented. DESeq2 included in Seurat, because of persistent errors. COMET (Delaney et al., 2019), because it did not allow processing in memory and lastly Venice (Vuong et al., 2020), because of an implementation error.

2.2 Conclusions

The review articles agree that there is no single tool better than the others in all situations. The tools usually trade their ability to detect truly differentially expressed genes for the introduction of false positive results. There is also a lack of agreement among the methods as each of them marks different genes as differentially expressed in the same dataset. It was surprising to find that tools made specifically for scRNA analysis did not outperform the traditional bulk methods. Almost all tools performed better with increased sample size and less multimodal data. In terms of efficiency, bulk RNA methods and parametric methods usually ran faster compared to the non-parametric single-cell tools.

Most reviews, however, provide overall rankings of the tested methods. Soneson and Robinson ranks edgeR/QLF, MAST, limma, t-test and Wilcoxon test among the top methods. Wang et al. states that the non-parametric methods generally perform better than the parametric ones. Das et al. found DECENT followed by EBSeq, to perform the best and give more robust results than other tools and Pullin and McCarthy highlight the performance of methods based on logistic regression, Student's t-test and the Wilcoxon rank-sum test. Thus, there isn't a consensus among the reviews about which tool is best overall, but since the Wilcoxon rank-sum test showed good overall performance in both Soneson and Robinson and Pullin and McCarthy (two out of the three reviews where it was tested) and is non-parametric (which is highlighted by Wang et al.), it was chosen for the analysis.

3. Practical implementation

In order to perform the data analysis and identify population-specific membrane molecules in the selected datasets, R version 4.2.0 was used (R Core Team, 2022), along with R Markdown (Allaire et al., 2022). R Markdown allows the user to combine text, code, and output from R and to generate (knit) documents in the HTML format, among others. Packages used are listed in Table 3.1. A complete list of all installed packages including dependencies can be found in Attachment 3.4. The R script itself can be found in Attachment 3.4.

Package	Version	Citation
Seurat	4.1.1	Hao et al. (2021)
tidyverse	1.3.2	Wickham et al. (2019)
org.Mm.eg.db	3.15.0	Carlson (2022)
AnnotationDbi	1.58.0	Pagès et al. (2022)
gtools	3.9.3	Bolker et al. (2022)
VennDiagram	1.7.3	Chen (2022)
details	0.3.0	Sidi (2022)

Table 3.1: Used R packages

3.1 Pipeline

3.1.1 Loading and integration

The datasets were usually made up of two files. First, a counts matrix, where one axis represents genes and the other the individual cells. The elements of the matrix contain the number of transcripts detected in each cell for each gene. Second, a metadata file containing additional information about each cell, like mouse line, sample, sex, and most importantly the cell type. Typically two cell-type resolutions were provided. A coarser level and a more detailed level, where each cluster was further subdivided. The only exception to this was the dataset by Chen et al., which only contained detailed clustering and the coarse level was reconstructed manually by combining corresponding clusters.

The Seurat package was used to perform the data analysis. Seurat is a widely used computational tool specifically developed to analyze scRNA-seq data. It provides a user-friendly interface for exploring and visualizing scRNA-seq data, identifying subpopulations of cells, detecting differentially expressed genes and integrating data from multiple samples.

After downloading, the counts matrix and the metadata were transformed to the correct format and loaded into Seurat. Next, normalization was performed. Normalization was done by dividing the gene counts of each cell by the total counts for that cell and then multiplying by a factor of 10,000. The result was then natural-log transformed after adding 1 (to avoid taking the log of 0).

3.1.2 Filtering

For each gene, a list of Gene Ontology (Ashburner et al., 2000; Gene Ontology Consortium, 2021) terms was retrieved. Gene Ontology (GO) is a bioinformatics project that provides a hierarchical vocabulary for describing genes and their functions across different organisms. The three main domains covered by GO are Molecular Function, Cellular Component and Biological Process. Each gene is assigned different GO terms based on experimental evidence or predictions.

In our case, only genes with the term "external side of plasma membrane" were kept, in order to find genes that encode proteins embedded or anchored in the plasmatic membrane. This reduced the number of genes by about 97%, meaning only 3% of the original genes are known to encode proteins on the external side of the plasma membrane.

Alongside gene filtering, the cells were also filtered by population. Non-neuronal populations, like astrocytes, endothelial cells, microglia and oligodendrocytes, were filtered out with only neuronal populations remaining.

3.1.3 Differential gene expression analysis

We used the Wilcoxon rank-sum test to perform differential gene expression analysis, since it demonstrated good performance in the comparison (Section 2.2). The threshold of $\log_2$ fold change was set to 0.25 (meaning only markers with around 1.2 fold change in expression remained). The results were then filtered by removing genes that were classified as markers of more than one population due to Seurat's "one vs all" approach to marker identification. Next, a threshold of 5% was set on the adjusted p-value. Finally, an additional column containing $\log_2$ fold change between each marker's cluster and the cluster with the highest expression from all other clusters was added. After filtering, the marker genes were visualized using a dot plot. The same procedure was used for both the coarse and detailed level cell types. An overview of the entire pipeline is depicted in Figure 3.1.

3.2 Results

Population-specific markers were detected in each dataset (at both the coarse and detailed level) and are visualized in Figures 3.2–3.11 in the form of a dot plot. Dot plots give great visual insight into the characteristics of each detected marker and can help you identify suitable ones. Note that each "column" of the dot plot (representing genes) is scaled for better visualization. This means that the color gradient reflects the relative expression levels of the genes, rather than absolute values. This can lead to misleading results in plots with few clusters but we opted to use it to improve overall readability. The size of the dot represents the percentage of cells in a cluster, which express a given gene.

In total 259 potential markers were found across all datasets. Results of the analysis from each dataset were aggregated into one table, which can be found in Attachment 3.4. The explanation of each column of the table is available in Table 3.2.

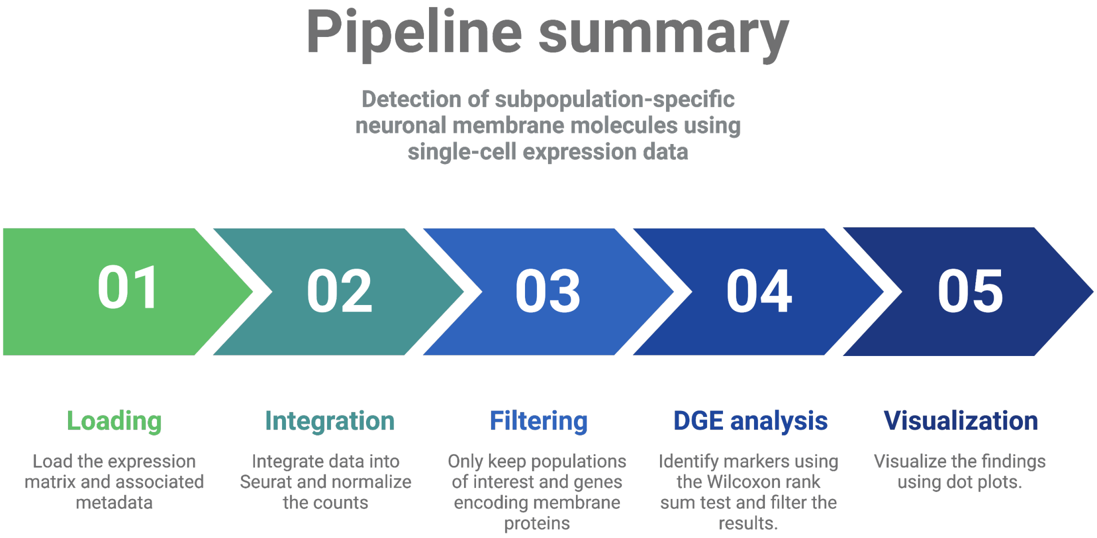
Pipeline summary
Detection of subpopulation-specific neuronal membrane molecules using single-cell expression data
01
02
03
04
05
Loading
Load the expression matrix and associated metadata
Integration
Integrate data into Seurat and normalize the counts
Filtering
Only keep populations of interest and genes encoding membrane proteins
DGE analysis
Identify markers using the Wilcoxon rank sum test and filter the results.
Visualization
Visualize the findings using dot plots.

Figure 3.1: Summary of the pipeline (Created with BioRender.com)

Column	Explanation
p_val	the p-value of the statistical test used to determine differential expression
avg_log2FC	the average $\log_2$ fold change in expression between the cells from the marker's cluster and cells from all other clusters
avg_log2FC_to_second	the average $\log_2$ fold change in expression between the cells from the marker's cluster and cells from the cluster with the highest expression other than marker's cluster
pct.1	the percentage of cells in marker's cluster that express the gene
pct.2	the percentage of cells in all other clusters that express the gene
p_val_adj	adjusted p-value, based on Bonferroni correction using all features in the dataset
cluster	the marker's cluster
gene	the name or identifier of the gene
level	the detail level at which the marker was found
dataset	the dataset in which the marker was found

Table 3.2: Explanation of the results table

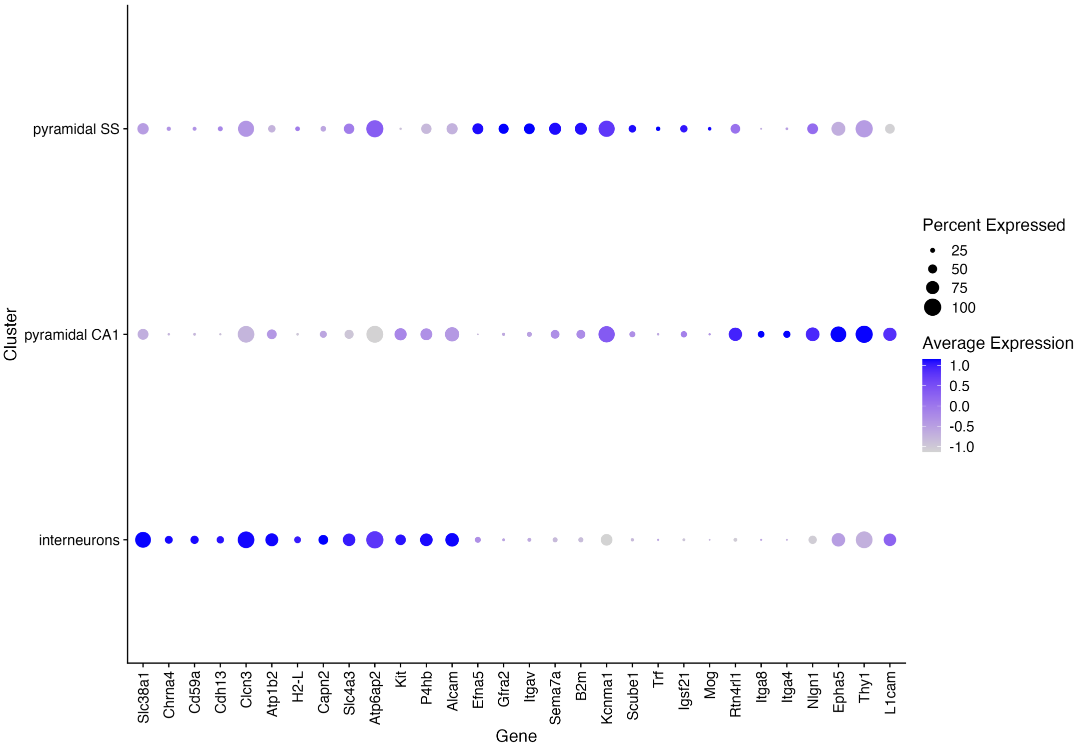

Figure 3.2: Coarse level markers in Zeisel et al.. The size of the dot encodes the percentage of cells within a class, while the color encodes the average expression level across all cells within a class.

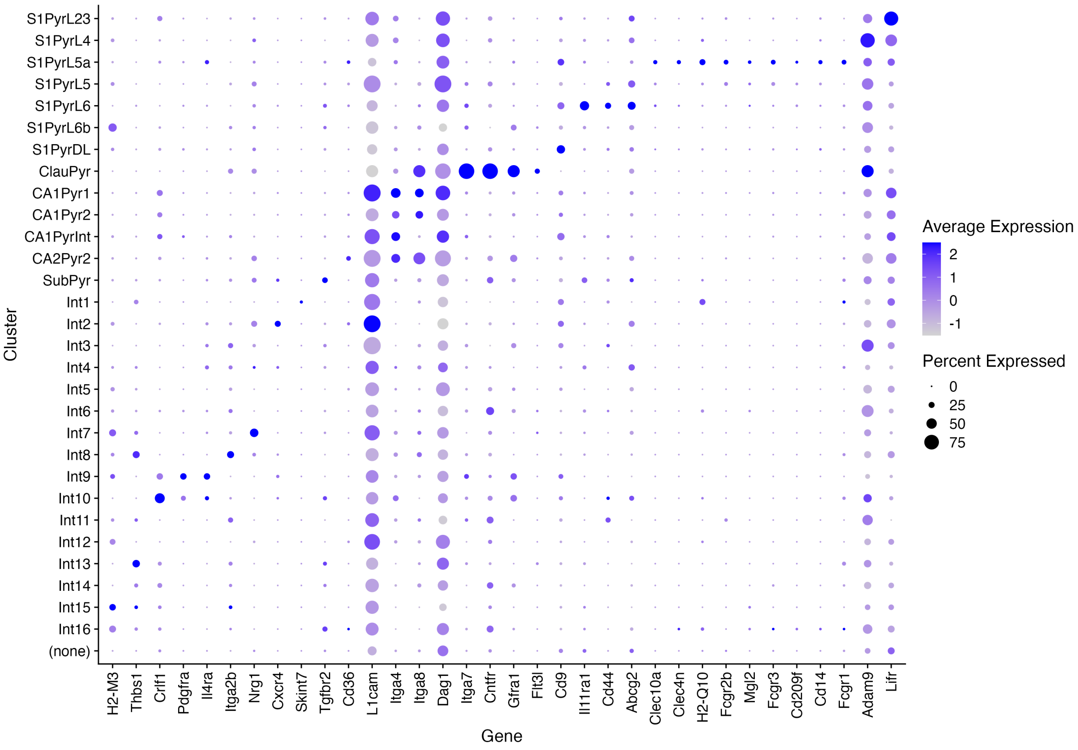

Figure 3.3: Detailed level markers in Zeisel et al.. The visualisation is the same as in Figure 3.2.

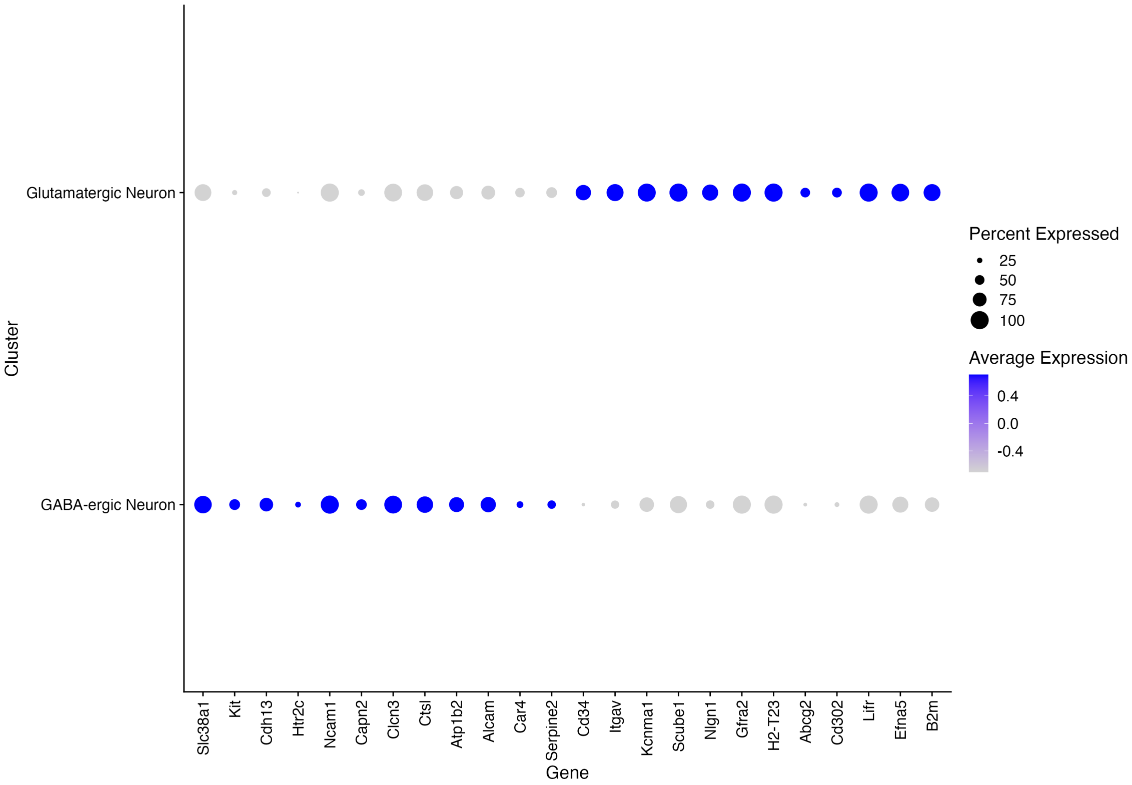

Figure 3.4: Coarse level markers in Tasic et al.. The visualisation is the same as in Figure 3.2.

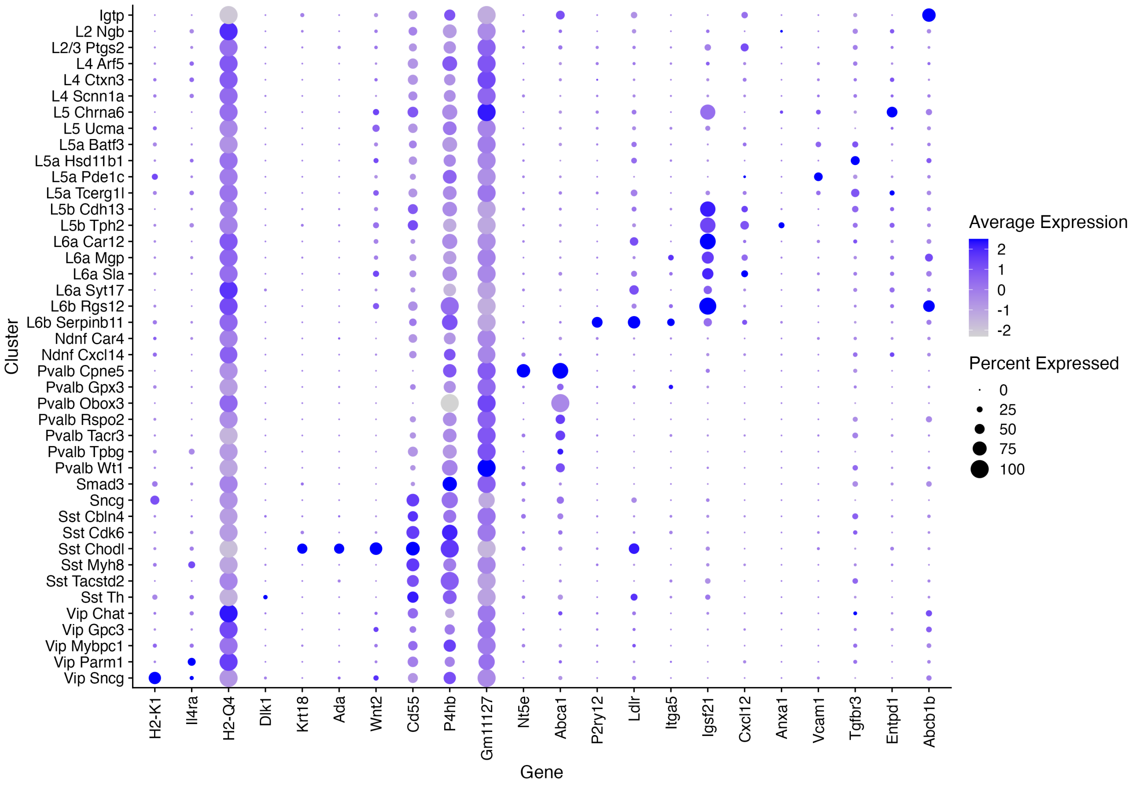

Figure 3.5: Detailed level markers in Tasic et al.. The visualisation is the same as in Figure 3.2.

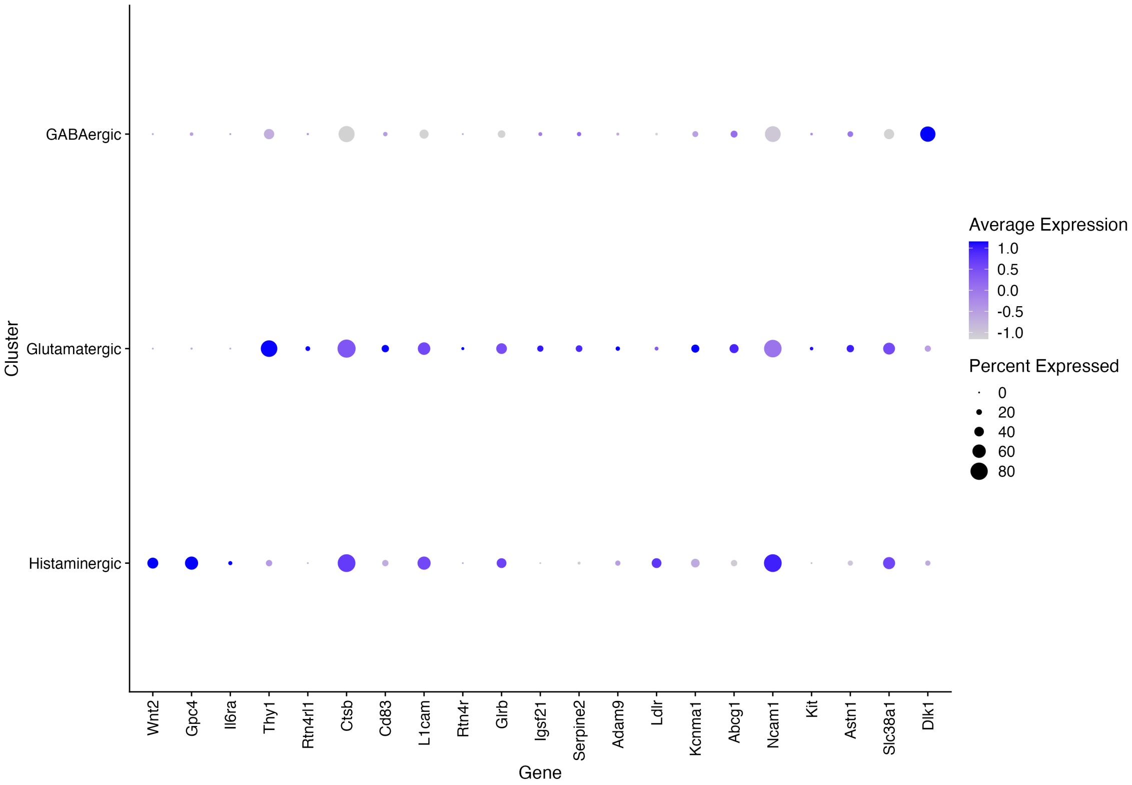

Figure 3.6: Coarse level markers in Chen et al.. The visualisation is the same as in Figure 3.2.

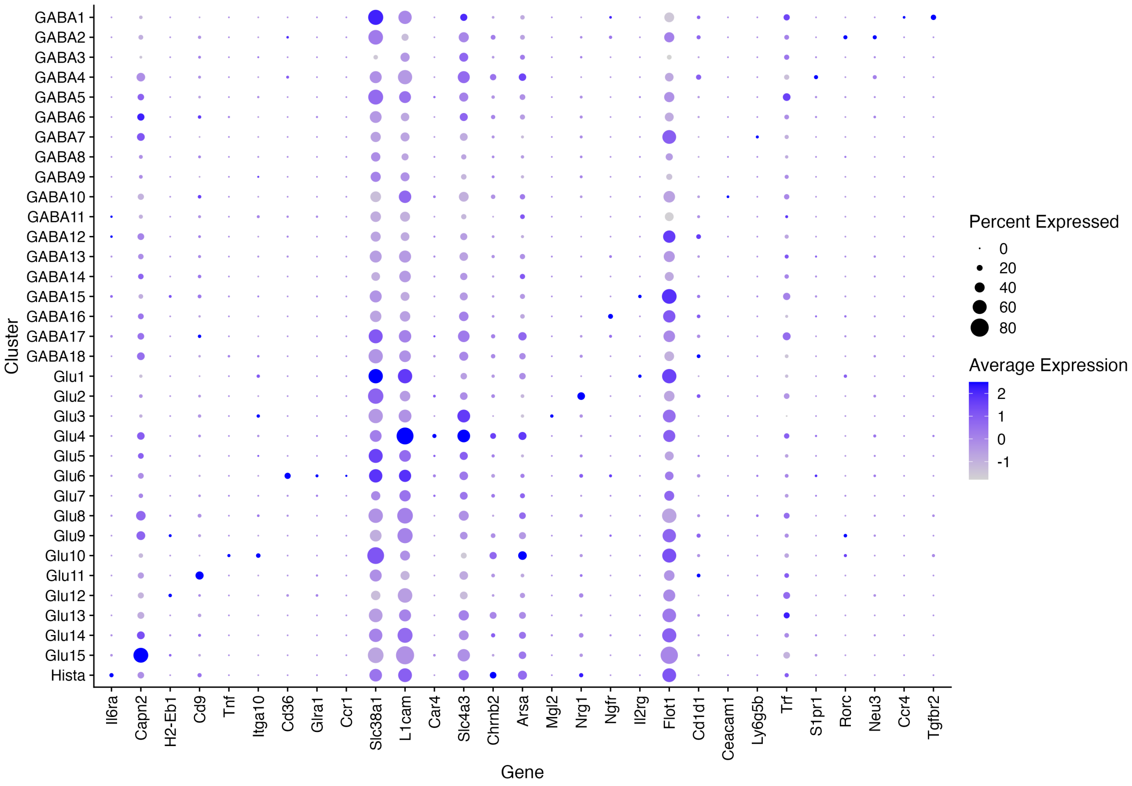

Figure 3.7: Detailed level markers in Chen et al.. The visualisation is the same as in Figure 3.2.

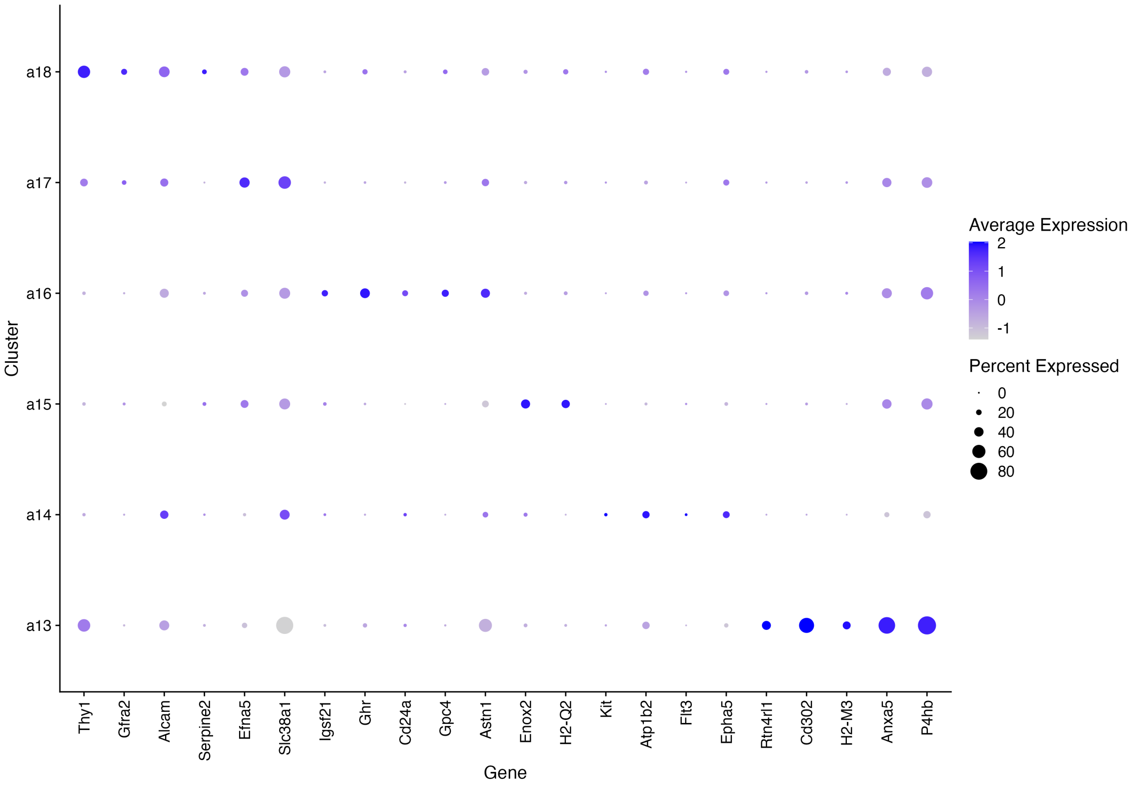

Figure 3.8: Coarse level markers in Campbell et al.. The visualisation is the same as in Figure 3.2.

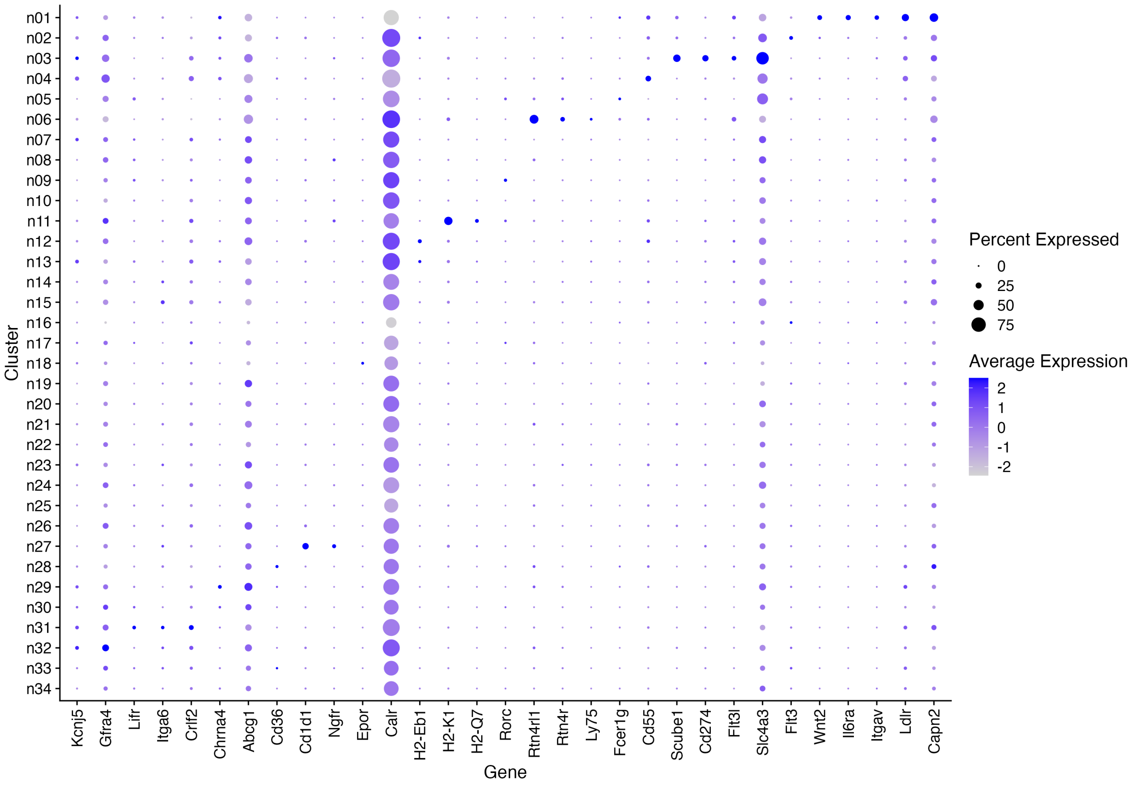

Figure 3.9: Detailed level markers in Campbell et al.. The visualisation is the same as in Figure 3.2.

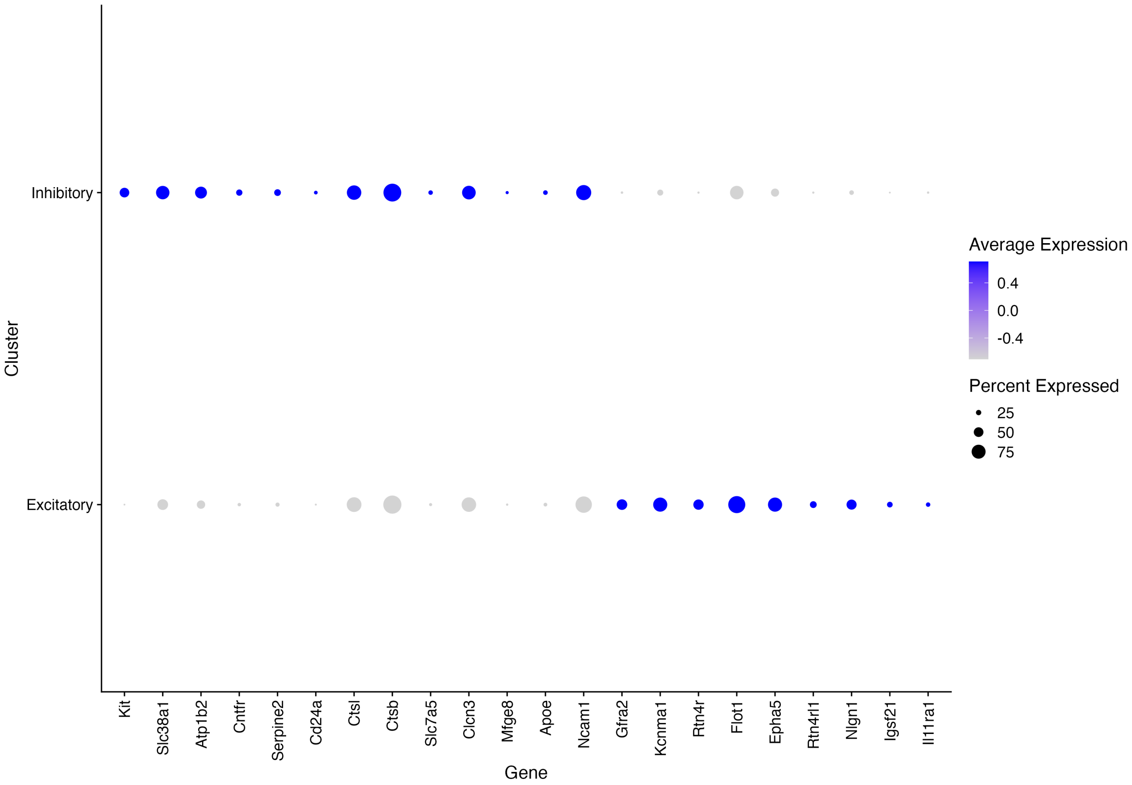

Figure 3.10: Coarse level markers in Bhattacherjee et al.. The visualisation is the same as in Figure 3.2.

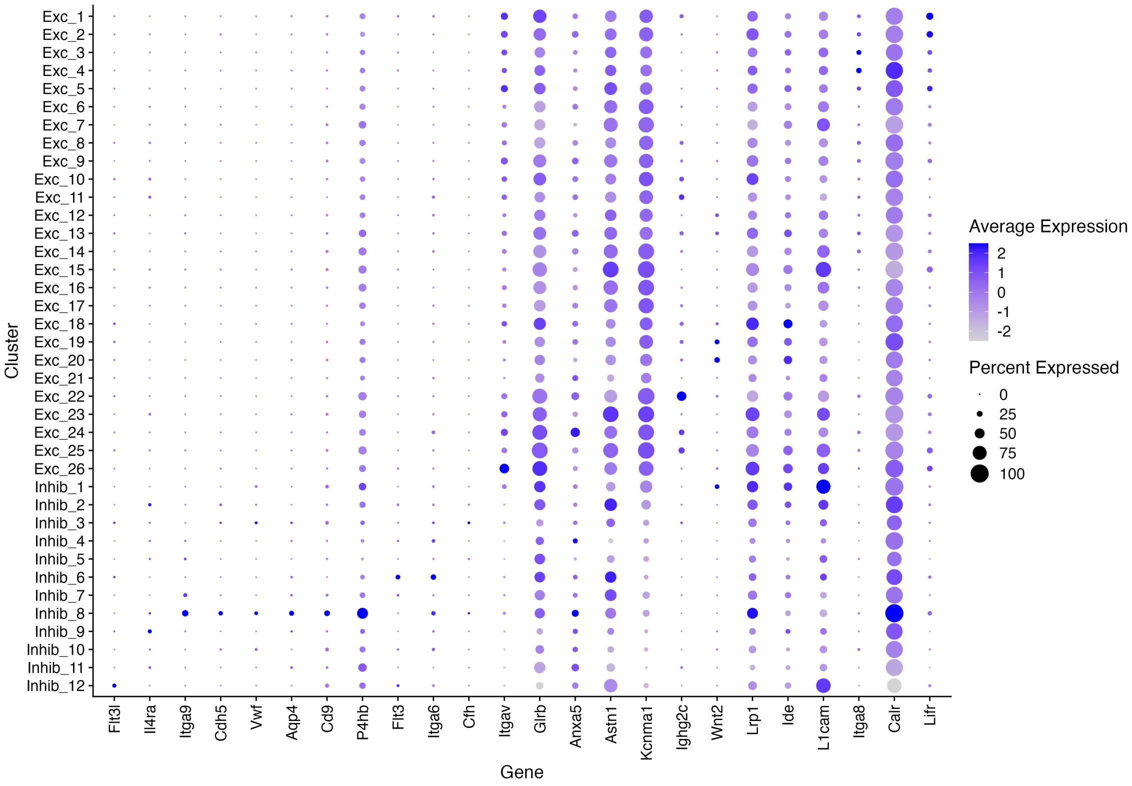

Figure 3.11: Detailed level markers in Bhattacherjee et al.. The visualisation is the same as in Figure 3.2.

3.3 Consensus

Since Tasic et al. provided a mapping between their clusters and clusters detected by Zeisel et al., a consensus in terms of detected population-specific membrane molecules could be inferred. Marker genes of corresponding clusters were compared, however, no intersection was found. Another type of comparison was performed. The expression of marker genes from Tasic et al. was analyzed in Zeisel et al. and each gene was labeled as having correct "cross-max" if it had maximum expression in the correct corresponding cluster, even if it was not among the detected markers. This comparison also revealed poor consensus with only 1 gene out of 22 corresponding between the datasets.

A similar comparison was performed between Tasic et al., Chen et al. and Bhattacherjee et al.. GABAergic population of Tasic et al., GABAergic population of Chen et al. and inhibitory population of Bhattacherjee et al. were compared (since GABAergic neurons are inhibitory). The same was done for glutamatergic and excitatory neurons. This time intersection between the three datasets was found and is visualized in Figures 3.12 and 3.13. "Cross-max" consensus was also determined. GABAergic neurons had a consensus of 55.8% and glutamatergic 61.8%. This is a poor consensus given that there are only two groups and a score of 50% can be achieved simply by chance.

The low consensus could be an effect of noise in the data or experimental artifacts. It is also possible that the compared populations are heterogenous and the poor consensus is inherent in the data.

3.4 Conclusions

We used Seurat, a popular single-cell analysis R package, to perform differential gene expression analysis using the Wilcoxon rank-sum test on five mouse neuron datasets. Altogether, we identified 259 population-specific membrane genes, which are provided in Attachment 3.4. These potential markers may be further investigated and experimentally validated. Consensus across datasets was also studied where possible, however, poor-to-no consensus was found and thus the results should be approached with great caution. We also do not expect a TPR of 100%, which implies that there could be some markers that we might have missed.

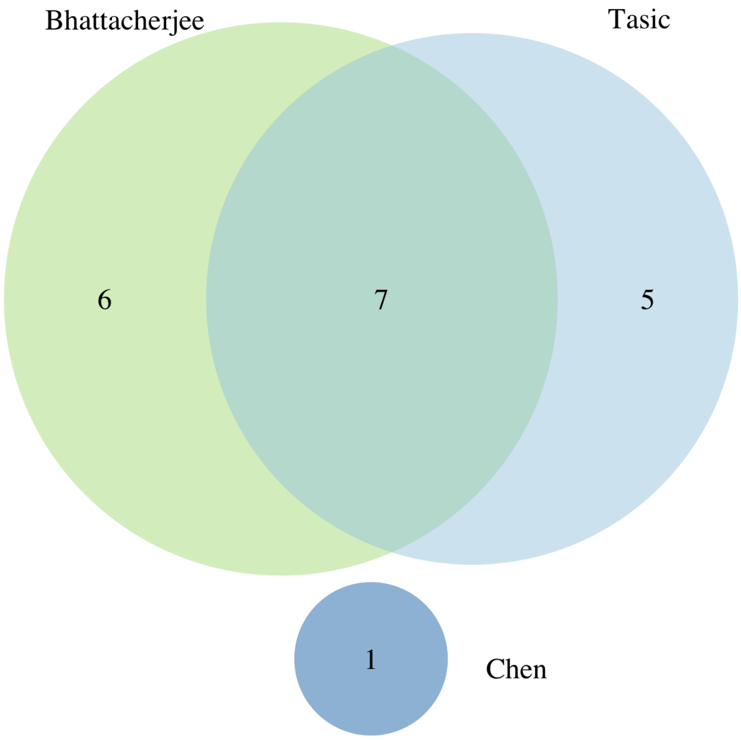

Figure 3.12: Consensus of GABAergic populations

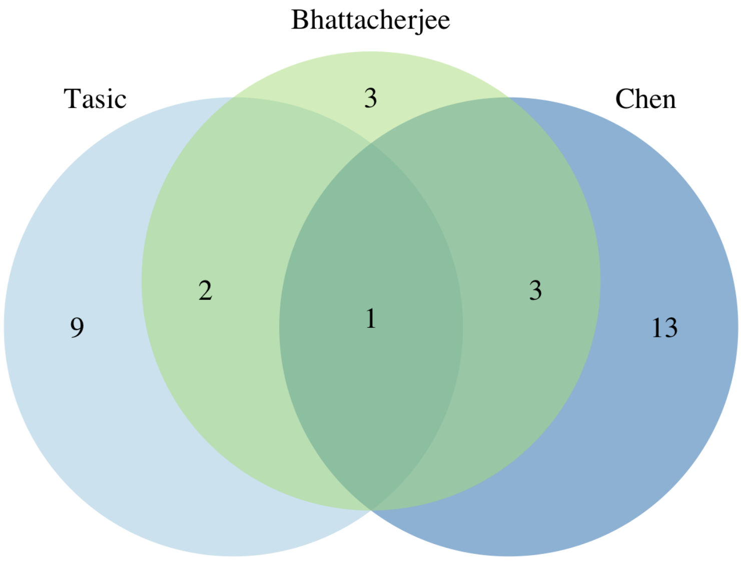

Figure 3.13: Consensus of glutamatergic populations

Conclusion

In this thesis, we first identify publicly available single-cell RNA sequencing datasets about murine neuronal cell types. The five chosen datasets cover cell types in different parts of the brain, like the cortex, hippocampus or hypothalamus.

Next, we present an overview of methods for performing differential gene expression analysis, where differences in the expression of genes between cell groups are studied, and provide a comparison of these methods from different perspectives by combining several previous benchmarks. It was found that there isn't an agreement among the benchmarks about which tool is best overall, as they each highlight different methods. Most tools showed a tradeoff between identifying true positives and introducing false negatives. The tools also showed low consensus between each other, as each tool identifies different DEGs. Interestingly, tools made for scRNA data did not perform better than the traditional bulk RNA tools. Small sample size and multimodality in data hindered the performance of most methods. The bulk-parametric category of DGE tools was found to be the least CPU-intensive. Since the Wilcoxon rank-sum test showed good performance, it was chosen for the analysis.

Lastly, we perform DGE analysis using the Wilcoxon rank-sum test on the five selected datasets at two resolutions each. We discovered 259 population-specific genes embedded in the plasma membrane throughout the datasets. These candidate marker genes could be experimentally validated and later potentially used to design drugs that would target specific neuronal populations.

Bibliography

Brian Aevermann, Yun Zhang, Mark Novotny, Mohamed Keshk, Trygve Bakken, Jeremy Miller, Rebecca Hodge, Boudewijn Lelieveldt, Ed Lein, and Richard H Scheuermann. A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell rna sequencing. *Genome research*, 31(10):1767–1780, 2021.

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2022. URL `https://github.com/rstudio/rmarkdown`. R package version 2.16.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.

Aritra Bhattacherjee, Mohamed Nadhir Djekidel, Renchao Chen, Wenqiang Chen, Luis M Tuesta, and Yi Zhang. Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction. *Nature communications*, 10(1):4169, 2019.

Ben Bolker, Gregory R. Warnes, and Thomas Lumley. *gtools: Various R Programming Tools*, 2022. URL `https://CRAN.R-project.org/package=gtools`. R package version 3.9.3.

John N Campbell, Evan Z Macosko, Henning Fenselau, Tune H Pers, Anna Lyubetskaya, Danielle Tenen, Melissa Goldman, Anne MJ Verstegen, Jon M Resch, Steven A McCarroll, et al. A molecular census of arcuate hypothalamus and median eminence cell types. *Nature neuroscience*, 20(3):484–496, 2017.

Marc Carlson. *org.Mm.eg.db: Genome wide annotation for Mouse*, 2022. R package version 3.15.0.

Hanbo Chen. *VennDiagram: Generate High-Resolution Venn and Euler Plots*, 2022. URL `https://CRAN.R-project.org/package=VennDiagram`. R package version 1.7.3.

Renchao Chen, Xiaoji Wu, Lan Jiang, and Yi Zhang. Single-cell rna-seq reveals hypothalamic cell diversity. *Cell reports*, 18(13):3227–3241, 2017.

Alessandra Dal Molin, Giacomo Baruzzo, and Barbara Di Camillo. Single-cell rna-sequencing: assessment of differential expression analysis methods. *Frontiers in genetics*, 8:62, 2017.

Samarendra Das, Anil Rai, Michael L Merchant, Matthew C Cave, and Shesh N Rai. A comprehensive survey of statistical approaches for differential expression analysis in single-cell rna sequencing studies. *Genes*, 12(12):1947, 2021.

Conor Delaney, Alexandra Schnell, Louis V Cammarata, Aaron Yao-Smith, Aviv Regev, Vijay K Kuchroo, and Meromit Singer. Combinatorial prediction of marker panels from single-cell transcriptomic data. *Molecular systems biology*, 15(10):e9005, 2019.

Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (d 3 e)-a tool for gene expression analysis of single-cell rna-seq data. *BMC bioinformatics*, 17:1–13, 2016.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13, 2015.

Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. *Nucleic acids research*, 49(D1):D325–D334, 2021.

Minzhe Guo, Hui Wang, S Steven Potter, Jeffrey A Whitsett, and Yan Xu. Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLoS computational biology*, 11(11):e1004575, 2015.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13): 3573–3587, 2021.

Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

Hani Jieun Kim, Kevin Wang, Carissa Chen, Yingxin Lin, Patrick PL Tam, David M Lin, Jean YH Yang, and Pengyi Yang. Uncovering cell identity through differential stability with cepo. *Nature Computational Science*, 1(12): 784–790, 2021.

Ilya Korsunsky, Aparna Nathan, Nghia Millard, and Soumya Raychaudhuri. Presto scales wilcoxon and auroc analyses to millions of observations. *BioRxiv*, page 653253, 2019.

Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome biology*, 17:1–15, 2016.

Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart MG Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013.

Jun Li and Robert Tibshirani. Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data. *Statistical methods in medical research*, 22(5):519–536, 2013.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15 (12):1–21, 2014.

Aaron TL Lun, Davis J McCarthy, and John C Marioni. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. 2016.

Zhun Miao, Ke Deng, Xiaowo Wang, and Xuegong Zhang. Desingle for detecting three types of differential expression in single-cell rna-seq data. *Bioinformatics*, 34(18):3223–3224, 2018.

Sheida Nabavi, Daniel Schmolze, Mayinuer Maitituoheti, Sadhika Malladi, and Andrew H Beck. Emdomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics*, 32(4):533–541, 2016.

Michael E Nelson, Simone G Riva, and Ana Cvejic. Smash: a scalable, general marker gene identification framework for single-cell rna-sequencing. *BMC bioinformatics*, 23(1):328, 2022.

Hervé Pagès, Marc Carlson, Seth Falcon, and Nianhua Li. *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*, 2022. URL `https://bioconductor.org/packages/AnnotationDbi`. R package version 1.58.0.

Jeffrey M Pullin and Davis J McCarthy. A comparison of marker gene selection methods for single-cell rna sequencing data. *bioRxiv*, pages 2022–05, 2022.

Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. Single-cell mrna quantification and differential analysis with census. *Nature methods*, 14(3):309–315, 2017.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

Debarka Sengupta, Nirmala Arul Rayan, Michelle Lim, Bing Lim, and Shyam Prabhakar. Fast, scalable and accurate differential expression analysis for single cells. *BioRxiv*, page 049734, 2016.

Fatemeh Seyednasrollah, Krista Rantanen, Panu Jaakkola, and Laura L Elo. Rots: reproducible rna-seq biomarker detector—prognostic markers for clear cell renal cell cancer. *Nucleic acids research*, 44(1):e1–e1, 2016.

Jonathan Sidi. *details: Create Details HTML Tag for Markdown and Package Documentation*, 2022. URL `https://CRAN.R-project.org/package=details`. R package version 0.3.0.

Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, 15(4):255–261, 2018.

Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience*, 19(2):335–346, 2016.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

Catalina A Vallejos, Sylvia Richardson, and John C Marioni. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome biology*, 17(1):1–14, 2016.

Alexander HS Vargo and Anna C Gilbert. A rank-based marker selection method for high throughput scrna-seq data. *BMC bioinformatics*, 21(1):1–51, 2020.

Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: power analysis for bulk and single cell rna-seq experiments. *Bioinformatics*, 33(21):3486–3488, 2017.

Trung Nghia Vu, Quin F Wills, Krishna R Kalari, Nifang Niu, Liewei Wang, Mattias Rantalainen, and Yudi Pawitan. Beta-poisson model for single-cell rna-seq data analyses. *Bioinformatics*, 32(14):2128–2135, 2016.

Hy Vuong, Thao Truong, Tan Phan, and Son Pham. Venice: A new algorithm for finding marker genes in single-cell transcriptomic data. *bioRxiv*, pages 2020–11, 2020.

Tianyu Wang and Sheida Nabavi. Sigemd: A powerful method for differential gene expression analysis in single-cell rna sequencing data. *Methods*, 145:25–32, 2018.

Tianyu Wang, Boyang Li, Craig E Nelson, and Sheida Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC bioinformatics*, 20(1):1–16, 2019.

Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Frank Wilcoxon. *Individual comparisons by ranking methods*. International Biometric Society, 1945.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Chengzhong Ye, Terence P Speed, and Agus Salim. Decent: differential expression with capture efficiency adjustment for single-cell rna-seq data. *Bioinformatics*, 35(24):5155–5162, 2019.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

List of Figures

List of Tables

List of Abbreviations

scRNA-seq = single-cell RNA sequencing

S1 = primary somatosensory cortex

V1 = primary visual cortex

Arc-ME = arcuate-median eminence complex

PFC = prefrontal cortex

DGE = differential gene expression

DEG = differentially expressed gene

TP = true positive

TN = true negative

FP = false positive

FN = false negative

TPR = true positive rate

FPR = false positive rate

FDR = false discovery rate

GO = Gene Ontology

Attachments

script.Rmd

R Markdown script that performs the analysis.

script.html

Knitted HTML document generated by the script.

markers_combined.csv

CSV file containing the identified markers. The meaning behind each column is explained in Table 3.2.