University of Padua
Department of Statistical Sciences

Master's Degree in
Statistical Sciences

Differential detection and differential expression in single-cell RNA-seq data

Supervisor Prof. Davide Risso
Department of Statistical Sciences

Co-Supervisor Prof. Lieven Clement
Department of Applied Mathematics, Computer Science and Statistics,
University of Ghent

Co-Supervisor Drs. Jeroen Gilis
Department of Applied Mathematics, Computer Science and Statistics,
University of Ghent

Graduand Laura Perin
Serial number 2022770

Academic Year 2022/2023

Contents

Abstract

The advent of single-cell RNA sequencing, also known as scRNA-seq, has revolutionized the study of transcriptomes. Previously, the most used technology was bulk RNA-seq, where RNA is extracted from a sample made up of thousands of cells, however the RNA is grouped into one library before sequencing. This way, bulk RNA-seq data measures the average expression level of a gene within all the cells of a given sample. Often these bulk measurements work well but there are some cases in which average values are not enough, such as, for example, when studying complex and heterogeneous structures like cancers and for the identification of new cell types. In these cases gene expression needs to be studied separately for each cell type. This is possible with scRNA-seq data, which provides information about how many specific genes are present in each cell for each sample and it is able to show transcript heterogeneity at single cell level, underlining information that bulk data would otherwise not show. At the moment, differential analysis of scRNA-seq data is conducted with similar methods to those used to analyse bulk RNA-seq data. However, scRNA-seq produces a larger amount of data than bulk sequencing which leads to new challenges both computational and interpretation wise. The scRNA-seq count matrix is in fact a lot bigger than the RNA-seq matrix and it is highly sparse this leads to new problems that have not been completely solved.

In the last years many tools have been developed to analyse scRNA-seq data and some of them have the purpose of identifying genes that are differentially expressed in two or more groups; this process is in fact useful for understanding differences between cell groups.

scRNA-seq data are characterized by a highly spare matrix and literature still does not agree on how to address it; as underlined by Sarkar and Stephens (2021), this high proportion of zeros has lead to incorrect or imprecise terminology. Often zeros are in fact considered as missing values, this is misleading and incorrect since missing values do not give any information, whereas zero counts do; for instance they underline the fact that that gene is unlikely to be expressed.

Traditional scRNA-seq analysis does not account for the fact that genes that are not highly expressed can also provide useful information. According to Qiu (2020) zeros are in fact not a problem, but a useful signal; by binarizing counts one can obtain an expression profile which accurately reflects biological variation. Bouland, Mahfouz, and Reinders (2021) have demonstrated that the frequencies of zero counts are enough to capture biological variability and they are able to identify differentially expressed genes in different groups. It is therefore of interest to try and consider the high presence of zeros not as a problem but as an alternative way to discover information from scRNA-seq data.

Part of what this work aims to achieve is try and understand if the proportion of zeros alone is capable of capturing the biological variability and distinguish between differentially expressed groups.

Zimmerman, Espeland, and Langefeld (2021) suggest that there is another aspect that needs to be take into account when analysing scRNA-seq data: cells from the same individual share common genetic and environmental backgrounds, which implies they are not statistically independent and makes them pseudo-replicates. This means that results obtained with typical methods that do not account for this aspect are biased, have highly inflated type 1 error rates and reduced robustness and reproducibility. To account for these pseudo-replications one can aggregate all the counts of a gene belonging to the same patient, in this way pseudo-bulk counts are obtained. Part of this work will focus on comparing analysis conducted on single-cell level data with pseudo-bulk analysis to try and verify how sensitive aggregation methods are in detecting sub-population level differences.

Chapter 1

Introduction

1.1 Biological context

De Duve, a Nobel Prize-winning biochemist, hypothesised that life is one, and it is rooted in chemistry and information (De Duve (2002)). He believed that all living beings are made of cells, which are the basic unit of life, and that all living organisms have evolved from a common ancestor. De Duve emphasized that all living beings are constructed of the same basic biological building blocks. These building blocks include lipids, which make up the membranes of cells; carbohydrates, which are used to store energy; amino acids, which are the building blocks of proteins; and nucleic acids, which are the building blocks of DeoxyriboNucleic acid (DNA) and RiboNucleic acid (RNA). These nucleic acids carry information that is used to construct biomolecules, and to pass information on from generation to generation. De Duve's view of the unity of life was grounded in the understanding that all living beings share a common ancestry, and that the chemical and informational processes that underlie life are universal.

De Duve believes life is chemistry because a cell is made up of a complex network of interconnected chemical reactions. Most of these chemical reactions are initiated by proteins; this underlines how understanding changes in the abundance of proteins over time is fundamental in order to understand important biological processes.

Life can be seen as information because all the information needed for an organism to be self-organised is passed on from generation to generation thanks to DNA and RNA.

In this Introduction, DNA and RNA will be explained in further detail and state-of-the-art technologies for identifying and quantifying DNA and RNA will be presented. After this, the main bioinformatics data analysis methods and workflows used to extract knowledge from these technologies will be described and the data that will be used throughout this dissertation will be introduced.

1.2 Genome biology

The central paradigm of molecular biology outlines the fundamental processes by which genetic information is stored, processed and used to specify the traits and characteristics of an organism, from physical appearance to metabolic processes and behaviour (Figure 1.1). The paradigm describes how in a biological system information is transferred from DNA to RNA through a process called transcription, and the resulting RNA molecule is then translated into a protein through a process called translation. The field that investigates the process of transcription and translation, as described by the paradigm, is genome biology. In this paragraph, these different steps will be discussed in more detail.

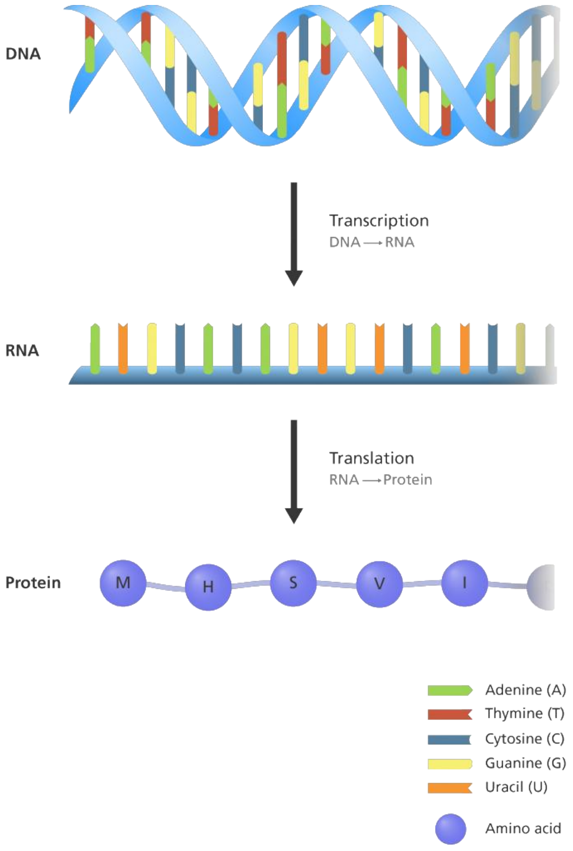

Figure 1.1: The central paradigm of biology describes the flow of genetic information from DNA to proteins, which are the functional building blocks of cells. RNA molecules are obtained from the DNA with a process called transcription. The RNA molecules are then translated into proteins, which are composed of long chains of amino acids. (Source: *yourgenome.org*)

1.2.1 Nucleic acids

DNA contains all the genetic information required to specify the traits and characteristics of an organism, from physical appearance to metabolic processes and behaviour. DNA is a long, double-stranded molecule that is composed of four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). These nucleotides are arranged in a specific sequence along the length of the DNA molecule, and the sequence determines the genetic information stored in the DNA. The two strands of DNA are held together by hydrogen bonds between the nucleotides, forming a double helix structure as can be seen in Figure 1.2. Both of these strands contain the same biological information, one is in fact the complementary of the other. The two strands are parallel but they are in the opposite direction, one is $5' \rightarrow 3'$ while the other is $3' \rightarrow 5'$; $5'$ represents the beginning of the strand while $3'$ the end. Every human cell contains about 2 metres of DNA which is compressed in each cell nucleus. The information necessary to produce a specific protein is contained in the gene which is a segment of DNA. Each gene contains a unique sequence of nucleotides, the fundamental elements of DNA, that provides the instructions for making a specific protein. The information contained in DNA is transferred into RNA through a process called transcription. Transcription involves the use of an enzyme called RNA polymerase, which reads the DNA sequence and synthesizes a complementary RNA molecule. RNA instead is a single-stranded molecule that is similar in structure to DNA. However, unlike DNA, RNA is not double-stranded and is typically shorter in length. The resulting RNA molecule, called pre-messenger RNA (pre-mRNA), contains the same genetic information as the DNA but in a different form. Messenger RNA then goes through a series of transformations, including the process of splicing, which removes non-coding parts of the sequence (introns) while retaining the coding parts (exons). The coding region contains genes that are responsible for producing functional products, such as proteins or RNA molecules, that are essential for various cellular processes whereas the non-coding region does not encode for proteins or functional RNA molecules.

Instead, this region contains regulatory elements that control gene expression, such as promoters, enhancers, and silencers. Additionally, non-coding regions can contain repetitive DNA sequences or transposable elements that do not have a known function but may contribute to genetic variation and evolution. In the last step mRNA is transported from the nucleus of the cell to the cytoplasm, where it is translated into a protein through a process called translation. Proteins are important because they are the primary functional units of cells and are responsible for a wide range of functions, including structural support, metabolic regulation, and communication. The genome is the full set of DNA molecules in a living organism and genomics is the study of the genome. The complete collection of all RNA in a cell, tissue

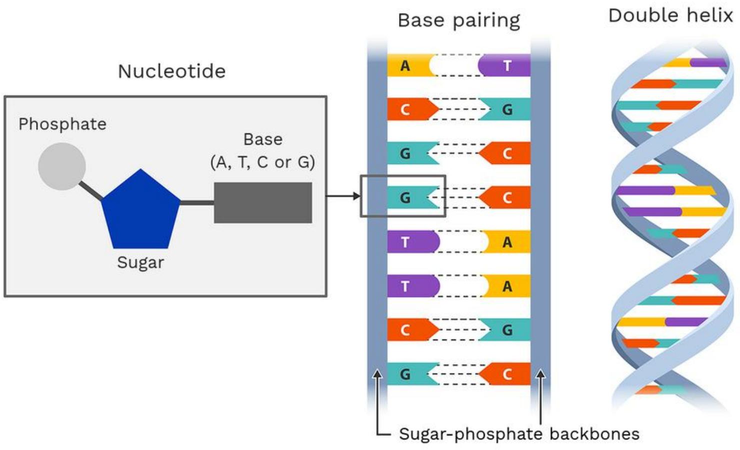

Figure 1.2: Basic structure of DNA, which consists of two complementary strands composed of nucleotides, each of which is made up of a base (adenine, thymine, guanine, or cytosine) attached to a sugar (deoxyribose) and a phosphate group, with the bases forming specific pairs (adenine with thymine, and guanine with cytosine) through hydrogen bonds. (Source: *theory.labster.com*)

or organism is also called the transcriptome and it's study is transcriptomics. The transcriptome reflects which and how active genes are being and is an important intermediate that determines the abundance of downstream proteins in a cell. In the next section we will introduce the technologies that can be used to unravel the genome and to quantify the transcriptome.

1.3 Quantification of gene expression

The process during which genetic information is converted into functional proteins is called gene expression and its quantification is useful in various biological contexts. Gene expression measurements can be obtained by sequencing the RNA molecules present in a sample; the expression levels of individual genes can then be estimated based on the number of RNA-seq reads that map to each gene.

There are various sequencing methods, but the most common datasets are obtained by second generation DNA sequencing machines, also known as Next Generation sequencing, or third generation sequencing machines. The second generation sequencing, which emerged in the mid-2000s, is a highly scalable technology that allows for sequencing the entire genome at once. There are various types of second generation sequencing technologies, the most common technology is "sequencing by synthesis" (SBS), commercialized by Illumina. One of the key features of this technology is the use of short reads. The genome is fragmented into small pieces, after which each piece is amplified and sequenced separately. This generates millions of small DNA fragments that are sequenced in parallel, which produces large amounts of data at a relatively low cost. Then, the RNA fragments are converted into cDNA (complementary DNA) using reverse transcription. This involves using an enzyme called reverse transcriptase to synthesize a complementary DNA strand from the RNA template. The cDNA is then amplified using PCR (polymerase chain reaction) to create many copies of each cDNA fragment. Once the cDNA fragments have been amplified, they are sequenced in parallel by first spatially separating them on a solid support such as a glass slide or a bead. DNA polymerase is then used to incorporate labeled nucleotides into the growing cDNA strand, one base at a time. The labeled nucleotides are detected optically, and the process of incorporation typically happens in cycles. During each cycle, a single base is added to the growing complementary strand, and the signal intensity of the labeled nucleotide is detected and converted into the appropriate nucleotide. After multiple cycles, millions of short DNA sequences are generated. These short DNA sequences are referred to as reads, and they can be used to assemble the entire genome or to analyze specific regions of interest. Since one can only sequence short reads, what has been sequenced has to then be aligned in order to identify to which gene each read belongs to. After alignment, one can count the number or reads for every gene, which provides information on its relative abundance. A general overview of the steps taken to perform RNA sequencing can be seen in Figure 1.3.

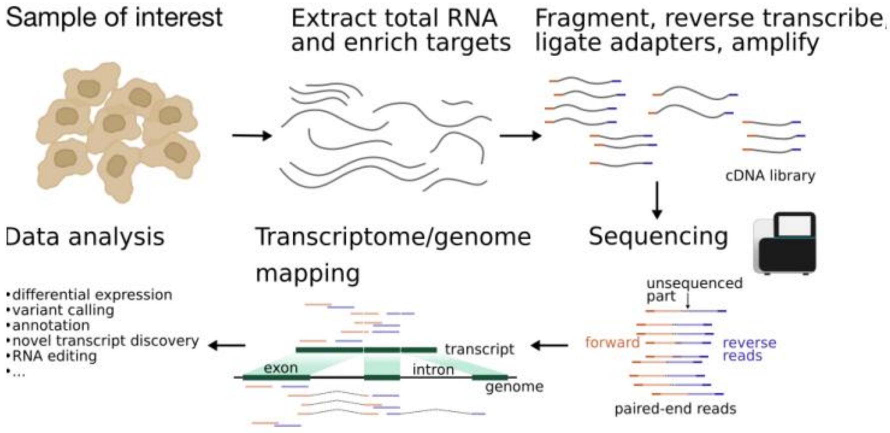

Figure 1.3: A sequencing protocol starts by extracting RNA from the sample of interest; the RNA is then fragmented, reversely transcribed and amplified. These fragments make up the cDNA library which is then sequenced. The raw RNA reads have to then be mapped on a reference genome and the number of reads that map on each gene are counted. These counts are then used to perform data analysis. (Source: Van den Berge et al. (2019))

Third generation sequencing, also known as long-read sequencing, is still under development and can produce considerably longer reads than second generation sequencing. This comes with a higher error rate compared to next-generation sequencers. Having longer reads can help face some computational challenges regarding genome assembly and transcript reconstruction.

In the next two paragraphs, we discuss the specific characteristics of bulk and single-cell sequencing in more detail.

1.3.1 Bulk RNA-seq

In bulk RNA extraction, RNA is extracted from a large number of cells and the resulting RNA mixture is sequenced. This way bulk RNA-seq methods give quantitative information about the expression of different genes in a given sample and represents the average expression of the gene in all the sequenced cells. This approach is useful for studying the RNA of populations of cells and for obtaining a general overview of the genetic material present in a sample.

While bulk RNA-seq is an established method for analyzing gene expression in a mixed population of cells, it may not provide sufficient resolution or accuracy for certain applications, such as studies of rare cell types or investigations into cellular heterogeneity, where expression differences between individual cells may be masked by the averaging effect of bulk analysis, making it difficult to identify important cell-to-cell variations or sub-populations that may play a key role in biological processes. In these scenarios, single-cell sequencing has emerged as a powerful tool, enabling the analysis of gene expression in individual cells, thereby providing a more detailed and accurate understanding of gene expression patterns, cell-to-cell variability, and cellular diversity.

1.3.2 Single-cell RNA sequencing

There are some cases, for example when studying heterogeneous tissues, where bulk measurements are not enough; single-cell RNA extraction is used instead and it enables the isolation and sequencing of RNA from individual cells. As shown in Figure 1.4, in scRNA-seq protocols a cell is extracted from a tissue and its RNA is retrotranscribed, amplified and then sequenced.

Plate-based protocols

Protocols have been widely available since 2014 and they are being continuously updated thus getting better and cheaper. In 2014 the SMART-seq2

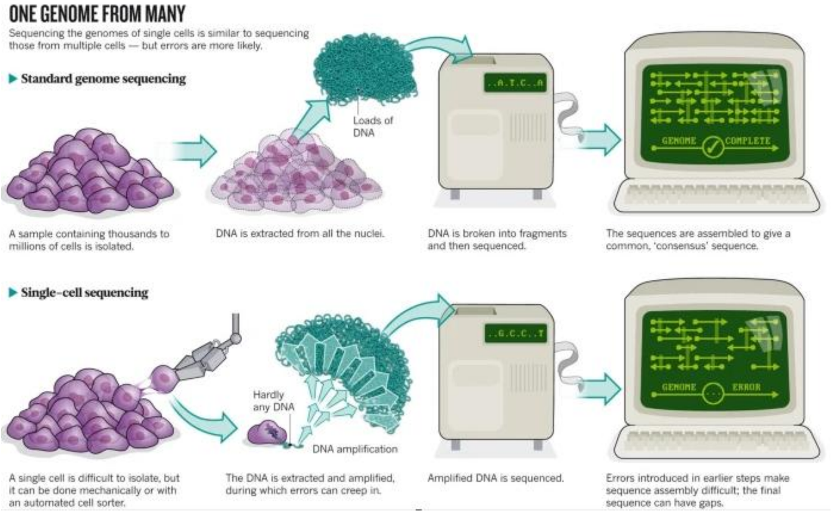
ONE GENOME FROM MANY
Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.
Standard genome sequencing
Loads of DNA
..A.T.C..A
GENOME COMPLETE
A sample containing thousands to millions of cells is isolated.
DNA is extracted from all the nuclei.
DNA is broken into fragments and then sequenced.
The sequences are assembled to give a common, 'consensus' sequence.
Single-cell sequencing
Hardly any DNA
DNA amplification
..G.C.C..T
GENOME ... ERROR
A single cell is difficult to isolate, but it can be done mechanically or with an automated cell sorter.
The DNA is extracted and amplified, during which errors can creep in.
Amplified DNA is sequenced.
Errors introduced in earlier steps make sequence assembly difficult; the final sequence can have gaps.

Figure 1.4: Bulk VS scRNA-seq protocol; the difference between the two protocols is mainly in the second column which shows how in bulk RNA-seq data DNA is extracted from all the nuclei whereas in scRNA-seq the DNA is extracted separately for each cell. (Source: *wycho.tistory.com*)

protocol was introduced (Picelli et al. (2014)); this protocol adopts manual separation of the cells into compartments and is still considered one of the most efficient protocols. In the same year Fluidgm C1 was introduced (Durruthy-Durruthy and Ray (2018)); this technology's main breakthrough is the fact that it is capable of automatically separating cells and amplifying their RNA.

In the micro-wells methods, cells are either isolated manually or sorted automatically and put into the micro-wells. The advantage of this method is that cells can be observed at the microscope before being sequenced, this allows to identify doubelts or cells that are dying. However it is a very hard and manual work so only few cells can be sequenced at the same time.

Droplet-based protocols

In 2015 there was an ulterior innovation and the Drop-seq technology was introduced (Macosko et al. (2015)). This method uses droplets to capture each cell and extract their RNA which allows for sequencing thousands of cells at the same time. Moreover, in 2017, 10X Genomics released Chromium (Weisenfeld et al. (2017)), a droplet based technology that allows for simultaneously sequencing thousands of cells. Droplet sequencing has indeed allowed a large profiling of transcriptomes, however assessing differential expression is hard due to inefficient sample processing and technical batch effects. To overcome these problems, Kang et al. (2018) proposed the demuxlet protocol, which takes advantage of the natural genetic variation to determine to which sample each cell belongs to and detect droplets containing two cells. It can sequence together cells of multiple patients, because there is genetic variation between patients and the sequence of every cell is given a barcode, as soon as a sequence with certain variability is found, it can be attributed to the correct patient. Demuxlet in fact implements a statistical model that uses maximum likelihood to determine the most likely donor for each cell. All these developments underline how it is getting easier and easier to obtain a very large amount of data often for a low price.

The Drop-seq method uses a droplet-based technology in which every cell is encapsulated in a gel drop together with a bead containing a unique barcode which allows to identify every droplet. The advantage of this method is that it allows to sequence thousands of cells from a sample, making this technology very cost-effective. However only few reads per cell are sequenced, which makes it hard to measure genes that are not highly expressed. These considerations lead to the conclusion that the method that one has to use depends on the biological interest; for example if one wants to analyse all the cell types in a tissue there need to be a lot of cells and so droplet methods are to be preferred, whereas if one wants to identify all the differentially expressed genes in a certain cell type there need to be a lot of reads so the SMART-seq2 method is the one that works better.

Challenges for data analysis

Bulk RNA-seq datasets are characterized by a high sequencing depth, which means that millions of reads are sequenced for each sample. This is possible since there are only few samples being sequenced. When dealing with scRNA-seq data instead, initially the RNA from a single cell is extracted, which means that there is a lot less RNA compared to the bulk procedure in which RNA is extracted from all the cells. This implies the need to amplify RNA, which leads to two consequences:

- Some genes are not amplified, which leads to a very high number of zeros;
- Some genes are amplified way to much resulting in amplification bias.

This occurs because some genes are easier to amplify than others. Gene amplification is performed with PCR (polymerase chain reaction) which is a relatively simple technique that amplifies a DNA template to produce specific DNA fragments in vitro. These problems result in an increase of the observed variance of the data.

With the droplet-based methods, it can happen that during the sequencing process two or more cells can occupy the same droplet, so when analysing the counts one thinks that they are observing the expression in one cell but in reality it is the sum of two or more cells. In the same way, it can also happen that one is sequencing empty droplets and so what is analysed is not the RNA of a cell but the ambient RNA. A solution to this problem was introduced by Kivioja et al. (2012) who proposed to insert Unique Molecular Identifiers (UMI) inside the library. UMI are barcodes made up of nucleotides that are added to the transcripts during reverse transcription. In this way, after the amplification, if two sequences with the same barcode are found, one can infer that they are a copy due to amplification and not two independent molecules of RNA. This means that one can count the number of UMI associated to every gene instead of counting the number of reads. The authors assert that this helps decrease bias and technical variability. However also this approach has its own problems since it often results in a very low number of counts and one could potentially remove biologically relevant cells which have low transcriptome complexity (Lönnberg et al. (2017)).

Another problem that needs to be addressed during the pre-processing of scRNA-seq data is that some cells can be of poor quality, for example they can be damaged or stressed. This problem is taken into consideration by leveraging the expression of mitochondrial genes; it is known that mitochondria are involved in stress response processes and cell death. This means that high expression of mitochondrial genes could be an indication of damaged or over-stressed cells.

1.4 Data analysis

Both for bulk and single-cell RNA data, the next step after sequencing is mapping the raw RNA sequencing reads onto a reference genome. During this process the reads are aligned to the genome and the number of reads that map to each gene are counted. The resulting count data represents the number of RNA molecules that were sequenced for each gene and can be used to quantify the expression levels of individual genes. It is to be noted that these counts represent relative abundance rather than absolute abundances; only a fraction of the RNA is in fact captured and then amplified with PCR, thus it is not the actual abundances but a proxy. After having mapped and obtained the counts, these are then normalized in order to account for technical variability in the sequencing process and differences in the amount of RNA that was initially isolated from the sample. Normalization methods such as TMM (trimmed mean of M-values) are commonly used to do so (Robinson and Oshlack (2010)). TMM normalization works by calculating a scaling factor for each sample that adjusts the library size to a common reference value. The effective library size is used to adjust the scaling factor to account for differences in library size between samples. The effective library size is calculated as the median of the ratios of the total read count for each sample to the geometric mean of the read counts across all samples. The effective library size is a crucial parameter in the TMM normalization process because it determines the degree of normalization required to adjust for library size differences between samples. After normalization, the count data can be used to perform differential expression (DE) analysis, which involves comparing the expression levels of individual genes across different samples or conditions. DE analysis can be used to identify genes that are significantly up- or down-regulated in response to specific conditions.

The main methods used to analyse single-cell RNA-seq data have their origin in methods developed to analyse bulk RNA-seq data; therefore, conventional methods for bulk RNA-seq data analysis will be first introduced.

1.4.1 Differential expression

The RNA-seq count matrix is a representation of the sequenced reads that have been aligned to a genome where the entries correspond to the number of reads that have been assigned to each gene for each sample. In the count matrix the rows typically represent the genes and the colums the samples or cells. Robinson, McCarthy, and Smyth (2010) highlight how the count matrix can be seen as a large multinomial distribution where each column in the matrix represents a separate trial of a multinomial experiment, with the number of trials being the total number of reads in the sample, and the number of categories being the number of genes in the genome. In this view, each entry in the count matrix is a count of a specific outcome of the multinomial experiment, and the sum of the entries in a row represents the total number of trials for that sample. However, because there are so many counts in the count matrix, often the total number of reads in a sample is large and this makes it appropriate to model the count data for each gene as a Poisson distribution, conditioning on the total count. In this way, the count data for each gene is modeled as a Poisson random variable, with the mean of the Poisson distribution being proportional to the mean expression level of the gene across all samples.

An important aspect is that the Poisson distribution assumes that the variance of the count data is equal to the mean, which is not always the case in RNA-seq data; the count data for many genes can in fact have a higher variance than the mean and this means that the data is overdispersed. Overdispersion is introduced due to biological variability and if not accounted for can result in incorrect inference and decreased statistical power. To account for this, it is common in the literature to use the negative binomial (NB) distribution to model the count data. EdgeR (Robinson, McCarthy, and Smyth (2010)) is a widely used software package for the analysis of differential gene expression in RNA-seq data.

Precisely, the edgeR method considers every row of the count matrix as a gene and every column as a sample, which means that given Y_{gi} the number of counts of gene g in sample i, edgeR fits a Generalized linear model (GLM) to each gene as follows:

$$\begin{gathered} Y_{gi} \sim NB(\mu_{gi}, \phi_g) \\ log(\mu_{gi}) = \eta_{gi} \\ \eta_{gi} = \beta_{0g} + \sum_{j=1}^{p} x_{ij}\beta_{jg} + O_i \end{gathered} \tag{1.1}$$

Where x_{ij} with $j = 1, ..., p$ and p the number of variables is the element of the design matrix made up of variables indicating for example to which group each observation belongs, and β_{jg} can represent the log2 fold change or the log2 fold change difference between groups for each variable. O_i represents an offset term, typically specified as the logarithm of the effective library size for each sample, and it is used in the model to adjust the mean expression level for each gene. μ_{gi} can also be written as $M_i p_{gj}$ where M_i represents the library size which is the total number of reads and p_{gj} is the relative abundance of gene g in group j to which sample i belongs. Since we are dealing with a negative binomial distribution the mean is $\mu_{gi} = M_i p_{gj}$ and the variance is $\mu_{gi}(1+\mu_{gi}\phi_g)$. In RNA-seq data, the dispersion parameter ϕ_g represents the coefficient of biological variation between samples, this allows the edgeR model to distinguish between technical and biological variation. edgeR estimates the dispersion parameters with an adjusted profile likelihood (Cox and Reid (1987)).

Genes are tested to see if they are significantly differentially expressed by testing for every gene g the following hypothesis:

$$\begin{aligned} H_0 &: C\beta_j = 0, \forall j \in 1, ..., p \\ H_1 &: \exists j : C\beta_j \neq 0 \end{aligned} \tag{1.2}$$

where C is a vector that allows testing linear combinations of β. Differential expression is assessed by using likelihood ratio tests (LRT) who are asymptotically χ^2 distributed.

An empirical Bayes procedure is then used to shrink the dispersions towards a consensus value by borrowing information across genes. Shrinkage is used to avoid the presence off big outliers and because in RNA-seq data it is common for $n - p$ where p is the number of parameters, to be small, a quasi-likelihood approach can thus be improved by sharing information across genes when estimating dispersion parameters. Lund et al. (2012) show how a scaled-inverse χ^2 prior distribution with d_0 degrees of freedom and a scaling factor ω_0 is used on each gene's dispersion obtaining $d_0\omega_0/\omega_g \sim \chi^2_{d_0}$. This produces an inverse-gamma posterior distribution $1/\omega_g|\hat{\omega_g} \sim Gamma(0.5(d_0 + n - p), 0.5(d_0\omega_0 + (n - p)\hat{\omega_g}))$. The hyper-parameters are estimated from the distribution of $\hat{\omega_g}$ by using the method of moments approach.

A quasi-likelihood approach (Tjur (1998)) is also commonly used when estimating the edgeR model; a model is specified for the mean and the variance for each observation as a function of its mean. Given Y_{gi} the observed count of gene g in sample i what is modeled is $E(Y_{gi}) = \mu_{gi}$ and $Var(Y_{gi}) = \omega_g^2 V_g(\mu_{gi})$ where $V_g()$ depends on the fitted distribution ($\mu_{gi} + \mu_{gi}^2\phi_g$ in the NB case and μ_{gi} if it's a Poisson). Both ω_g^2 and ϕ_g are dispersion parameters: the first one is a proportionality constant used in quasi-likelihood models, while the second is a parameter of the negative binomial distribution. In this case to test Equation 1.2 a weighted test is used.

1.4.2 Differential expression for single-cell RNA-seq

An important feature of single-cell RNA-seq is the possibility of looking at two aspects of the data distribution. Both differences in mean and differences in detection can in fact be assessed. By aggregating the data, differences in mean can in fact be analysed having the same or higher power of looking at differences in mean with bulk methods. This powerful aspect is currently not exploited but provides a promising avenue. A commonly used tool for visualizing differences in mean expression levels between different groups is the violin plot. A violin plot, like the one shown in Figure 1.5, is a type of density plot that shows the distribution of expression values for each group as a function of their density. The width of each violin corresponds to the density of expression values at different levels, with wider areas indicating higher densities of expression values. The height of the violin reflects the range of expression values, with the top and bottom indicating the maximum and minimum expression levels, respectively. Overall, the violin plot provides a useful way to compare the distribution of gene expression values between different groups and patients in single-cell RNA-seq data.

By examining the distribution of gene expression in different ways, scRNA-seq data can thus provide valuable insights into the biology of individual cells and the cellular heterogeneity of tissues and organs.

The majority of tools used to assess differential expression have been developed for bulk data. Single-cell data, however, also presents several challenges that need to be addressed in order to obtain accurate results. scRNA-seq typically have a lower sequencing depth compared to bulk RNA-seq, because the sequencing is performed on individual cells rather than on a pool of cells; they can have higher technical variability compared to bulk RNA-seq, due to the increased complexity of the sequencing process. Be-

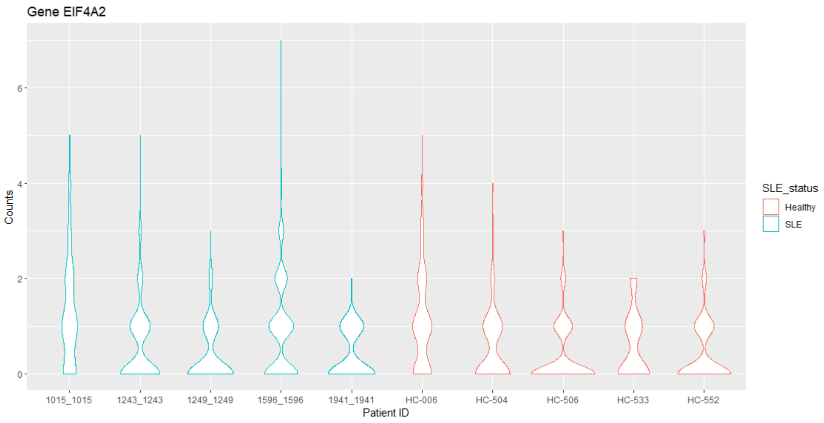

Figure 1.5: Violin plot of the expression of gene *EIF4A2* in the T4 naive cell type of the lupus study (see section 1.5) in 5 healthy patients and 5 patients affected by SLE

cause of both of these aspects, some genes may not be detected in individual cells. This aspect motivates us to focus on determining whether a gene is detected in a cell or not, rather than on the actual expression level. This type of analysis can provide useful insights into the presence or absence of genes in different cell types or conditions. While it is commonly believed that the highly sparse matrix, typical of scRNA-seq, is primarily caused by technical artifacts, there is growing evidence that these zeros also reflect biological variation. Zeros can in fact occur because of a technical dropout or because there is no or limited expression of a gene. Qiu (2020) has in fact demonstrated that by binarizing counts one obtains a binary expression profile of single-cell data, which accurately reflects biological variation and reveals the relative abundance of transcripts more robustly than counts; this happens because different sub-populations may have different dropout patterns and this can be a basis to detect cell types. Given these facts, part of what this work aims to do is try and understand if the proportion of zeros alone is capable of capturing the biological variability and distinguish between differentially expressed groups using a similar approach as Qiu (2020). However, the aim here is not to discover new cell types but to look at differential detection of genes across cell types or conditions.

As noted by Crowell et al. (2020), there is another aspect that typical modelling methods do not take into consideration, which is the fact that cells from the same individual share common genetic and environmental backgrounds. This implies that they are not statistically independent. In fact, all cells from the same cell type from the same patient have an expression more alike than cells from other patient and can therefore be seen as pseudo-replicates. This means that scRNA-seq data has a hierarchical structure that many methods do not consider and this leads to biased inference, highly inflated type 1 error rates and reduced robustness and reproducibility. Squair et al. (2021) have investigated the results obtained by several methods and have found that the most frequently used methods will identify differentially expressed genes even in the absence of biological differences. However, they conclude that false discoveries can be avoided by accounting for between-replicate variation. The dependence between cells of the same patient could be considered by using mixed effect models. This approach in fact takes into account the relationships between cells and can provide a more accurate representation of the underlying biology. Despite their potential advantages, mixed effect models can be computationally challenging to implement for scRNA-seq data, especially when working with large datasets. Zimmerman, Espeland, and Langefeld (2021) show that mixed models have the same performance as using pseudo-bulk data. Pseudo-bulk data refers to a method of aggregating transcriptomic data from individual cells into bulk-level expression data for a given cell type. This is accomplished by summing the gene expression values across cells within a defined group, such as cells from the same tissue or cell type. To conduct such analysis one can aggregate all the counts of a gene belonging to the same patient thus obtaining pseudo-bulk counts. In contrast, using pseudo-bulk data can make the analysis process quicker, as it reduces the complexity of the data.

1.5 Aims

The goal of this thesis is looking into what is the best approach to study differential detection is single-cell RNA-seq data. Pseudo-replication will be addressed as it is of interest to see if aggregating or not the data is of help in facing this problem; moreover, normalization strategies will be investigated, evaluating the need of adding an offset, and how data variability should be modeled will furthermore be taken into consideration.

Initially, the different strategies will be benchmarked on simulated data in order to assess type 1 error control, sensitivity and specificity. Then, the top performing methods will be used to analyse a case study; here it is of interest to see if performing a differential detection analysis can provide complementary information on top of a differential expression analysis.

Chapter 2

Methods

2.1 Notation

The original counts will be binarized in order to assess differential detection and the binarized counts of the same cell type and patient will be aggregated. In the coming paragraphs the following notations will be used: i indicates each of the N patients and g of the G genes. Thus the random variable Y_{gi} represents the number of times the gene g has been detected in patient i and y_{gi} its realization. n_i represents the number of cells for patient i and the total number of cells from all patients is M. The number of variables j is p and x_{ij} will be used to define the element of the design matrix.

2.2 Differential detection for single-cell RNA-seq data

As previously mentioned, scRNA-seq data have a very high amount of zero counts which can be due both to technical and biological reasons. A large amount of methods consider these zeros as a problem and aim to remove them from the analysis assuming that the gene in the droplet simply has not been detected. This procedure is not completely correct because it does not account for the fact that those zero counts can actually come from a sample that does not express that gene. Having taken this into consideration it is of interest to try and treat zeros not as a problem but as useful signal; to do so, the count matrix is binarized obtaining a matrix whose elements y_{gi} are equal to 1 if gene g was detected in sample i and equal to 0 if it was not detected. It is now of interest to see which genes are differentially detected in different conditions and to see how the results compare to the ones obtained by conducting a differential expression analysis.

2.2.1 Bernoulli regression

Given the binary nature of the data, it is logical to consider adapting a Bernoulli model to the binarized count matrix. The Bernoulli model is a statistical model that is commonly used to model binary outcomes and is well-suited to data that can be represented as a series of independent trials with two possible outcomes: success or failure. In the context of binary scRNA-seq data, the Bernoulli model can in fact be used to model the presence or absence of gene expression, with success representing the presence of expression and failure representing the absence of expression. This can be particularly useful for understanding the relationships between different genes and for identifying patterns of co-expression or co-regulation. However, as previously mentioned, the independence assumption requested by the Bernoulli model is violated due to the presence of pseudo-replicates. Mixed effect models, which account for dependence between observations, could be used but, as mentioned by Zimmerman, Espeland, and Langefeld (2021) they produce similar results as using pseudo-bulk data while being computationally more challenging.

It should also be taken into consideration that the independence assumption is always violated in transcriptomics data because genes do not act independently one from each other; however, none of the state-of-the-art methods account for this and also in this work this aspect is not taken into consideration.

2.3 Aggregation method

As previously stated, all cells that come from the same cell type and from the same patient have an expression that is more alike than cells from other patient and can therefore be seen as pseudo-replicates. When analyzing scRNA-seq data with pseudo-replicate observations, it's important to account for the correlation structure in the data to avoid overestimating the significance of observed differences. One way to do this is to use statistical methods that are designed to handle correlated observations, such as mixed-effects models or hierarchical modeling (Gelman and Hill (2006)). In hierarchical modeling, the data is assumed to have a hierarchical structure, where cells are nested within patients. This allows for the estimation of patient-level variation, which can be used to account for the correlations in the data. For example, in a hierarchical model for scRNA-seq data with pseudo-replicate observations, the expression levels of genes in individual cells are modeled as a function of patient-level effects and cell-level effects. The patient-level effects capture the variation between patients, while the cell-level effects capture the variation within patients. By incorporating patient-level effects into the model, it is possible to account for the correlation structure in the data and obtain more accurate estimates of differential gene expression. However, Zimmerman, Espeland, and Langefeld (2021) show how applying mixed models to single-cell RNA-seq data has the same performance as analysing pseudo-bulk data but, aggregating observations from the same cell type from each patient has a great computational gain. After having conditioned on patient and cell type, counts are independent and when Bernoulli data is aggregated, the data will be binomial. As a consequence, if only the mean detection is analysed there is no information loss and there is a gain in power by aggregating. So, in order to address this dependence, pseudo-bulk data is used. Pseudo-bulk data is obtained by aggregating all counts coming from the same patient and same cell type. The purpose of creating pseudo-bulk data is to account for the dependence between cells coming from the same patient, as this dependence can impact the validity of the results of traditional RNA-seq analysis methods. By creating a single sample that represents the entire sample, the dependence between observations is reduced, and the results of the analysis are more robust.

In this work, pseudo-bulk data was obtained with the *aggregateAcross-Cells()* function from the *scuttle* package which also aggregates metadata across cells.

2.3.1 Binomial regression

Given the nature of the observations contained in the newly created matrix, one method that can be used to assess differential detection between conditions is the Binomial model which is fitted for each gene in this way:

$$\begin{aligned} Y_{gi} &\sim Bin(\pi_{gi}, n_i) \\ logit(\pi_{gi}) &= log\frac{\pi_{gi}}{1-\pi_{gi}} = \eta_{gi} \\ \eta_{gi} &= \beta_{0g} + \sum_{j=1}^{p} x_{ij}\beta_{jg} \end{aligned} \tag{2.1}$$

Also in this case the regression coefficients β are estimated by using maximum likelihood. Inference is done with a Wald test which follows a t distribution with $n-p$ degrees of freedom under the null hypothesis of no association between the predictor and the outcome variable.

2.3.2 Quasi-binomial regression

Often RNA-seq data are overdispersed, which means that the variance of the response variable is higher than what is expected by the generalized linear model; this feature can be held into account by conducting a quasi-binomial analysis. The quasi-binomial models the first two moments (the mean and variance) according to the following expression:

$$\begin{aligned} E(Y_{gi}) &= \mu_{gi} \\ Var(Y_{gi}) &= \phi_g V(\mu_{gi}) \\ \mu_{gi} &= g^{-1}(\beta_{0g} + \sum_{j=1}^{p} x_{ij}\beta_{jg}) \end{aligned} \tag{2.2}$$

Here $V()$ represents the binomial variance, ϕ_g is the dispersion parameter, μ_{gi} is the expected value of Y_{gi} and $g()$ is the logistic link function. Inference is done by using a weighted test that incorporates weights based on the variance function of the working mode. The weighted test is typically implemented using a sandwich estimator, which involves estimating the variance of the estimated coefficients using a weighted sum of squared residuals.

The sandwich estimator can be used to calculate the standard errors of the estimated coefficients, which can then be used to perform hypothesis tests and calculate confidence intervals.

Compared to the canonical binomial GLM, this model adds an additional parameter ϕ_g to the binomial model which represents the dispersion parameter. When $\phi_g \to 0$ the estimation equations of the quasi-binomial likelihood coincide with the binomial one.

2.3.3 Offset

Often when modeling RNA-seq data an offset term is introduced as a normalization factor for every cell or sample. An offset is an additional term in a GLM model whose coefficient is not estimated but is considered equal to 1.

In this study, the impact of including an offset to both the binomial and quasi-binomial models will be assessed. In the first case we obtain the following model:

$$\begin{gathered} Y_{gi} \sim Bin(\pi_{gi}, n_i) \\ logit(\pi_{gi}) = \eta_{gi} = \beta_{0g} + \sum_{j=1}^{p} x_{ij}\beta_{jg} + logit(O_i) \\ O_i = \frac{1}{G}\sum_{g=1}^{G} \frac{y_{gi}}{n_i} \end{gathered} \tag{2.3}$$

This means that, on pseudo bulk data, the offset is the average detection of all the genes in one specific patient whereas in scRNA-seq data the offset is the average detection of all genes in one specific cell.

2.3.4 Shrinkage

Often in RNA-seq problems the dispersion parameter ϕ is highly variable between genes and the number of patients is often very small, and this results in imprecise estimates for ϕ. These problems can be solved by taking advantage of the fact that we are estimating G, with G the total number of genes, models at the same time and we can borrow information from other genes to stabilise the estimation using an empirical Bayes approach.

In an empirical Bayesian framework, statistical inference is performed using a hierarchical model where the hyper-parameters are estimated from the data themselves, rather than being priorly specified (George, Casella, et al. (1989)). Moreover, current methods do not use a posterior distribution, because this would be slow, they only use the posterior mode to then perform a frequentist analysis (Smyth (2004)).

In this study it is of interest to squeeze the dispersion parameter; to do so the *squeezeVar* function from the *limma* package is used. After having obtained the new dispersion estimate $\tilde{\phi_g}$ the moderate t-statistic is calculated $t_{gi} = \frac{\hat{\beta_{gi}}}{s_g \sqrt{v_{gi}} \tilde{\phi_g}}$ as proposed by Smyth (2004).

2.3.5 edgeR on binarized counts

A classical edgerR analysis was also conducted on the binarized count matrix. It was considered in two variants, one in which the dispersion was estimated resulting in a quasi-negative binomial model and one with $\phi \to 0$ which is a quasi-Poisson. Both models have a quasi likelihood because the *QLFit* function of the *edgeR* package estimates an additional over dispersion parameter in order to account for gene-specific biological and technical variability. Given y_{gi} the number of times gene g is detected in sample i, the quasi-negative binomial has the following features:

$$\begin{aligned} E(y_{gi}) &= \mu_{gi} \\ Var(y_{gi}) &= \omega_g(\mu_{gi} + \mu_{gi}^2\phi_g) \end{aligned} \tag{2.4}$$

where μ_{gi} represents the expected count for gene g in sample i given the sequencing depth and treatment conditions, ϕ_g is the NB dispersion parameter and ω_g is the quasi-likelihood dispersion parameter.

The quasi-Poisson instead has the following features:

$$\begin{aligned} E(y_{gi}) &= \mu_{gi} \\ Var(y_{gi}) &= \omega_g\mu_{gi} \end{aligned} \tag{2.5}$$

where again μ_{gi} represents the expected count for gene g in sample i given the sequencing depth and treatment conditions and ω_g is the quasi-likelihood dispersion parameter.

2.4 Stage-wise analysis

In the case study, a differential detection analysis as well as a differential expression analysis will be performed on the same data. As such, two hypotheses are tested for each gene; is there a difference in detection of the gene between samples, and is there a shift in mean expression between samples. To gain statistical power, these two hypotheses can be first tested jointly in the two stage testing paradigm proposed by Van den Berge et al. (2017).

The first stage, also called screening stage, considers an omnibus test that aggregates evidence across all the hypotheses that have been tested for every gene. This test thus indicates if a gene is either differentially expressed, differentially detected or both. To do so the p-values coming from the two tests have to be aggregated; the harmonic mean aggregation strategy, proposed by

Wilson (2019), was used to aggregate the p-values while accounting for the fact that the two tests are dependent. This stage increases the sensitivity of the effects that have a relatively low power by picking up DE and DD genes.

During the second stage, called confirmation stage, one sees if the genes that passed the screening test are either differentially expressed or differentially detected. This method allows for looking at different aspects of the distribution, difference in mean and difference in detection, while also controlling the false discovery rate at the gene level.

2.5 Simulation framework

To evaluate the performance of the different methods they were applied to simulated data obtained with the *swapper* package developed by Malfait (2022). *swapper* simulates differential expression based on feature swapping; in fact it randomly mixes a subset of features in one group of the data inducing DE signal. A very good feature of this simulation method is that it keeps the characteristics of the original data and does not rely on any modeling assumptions. However, a negative side of this approach occurs in extremely sparse datasets. In fact, if the counts of the two genes for which the counts are being swapped between samples of groups a and b are primarily zeros, this will result in only zeros but that gene will still be flagged as differentially expressed.

The simulation data set originates from the Systemic lupus erythematosus data analysed by Perez et al. (2022) (see Chapter 4 for further details). The data set was first filtered in order to obtain a homogeneous sample retaining only healthy European women who are less than 50 years old. Moreover, only subjects coming from three sequencing batches were kept and, in order to have a more homogeneous sample, only cells coming from T4 lymphocytes were considered. This leads to a data set with 4767 genes and 44 patients. Differential expression will be performed by randomly splitting these patients in two groups and since it is a homogeneous group, no differential detection or differential expression is expected. Two main cases were considered: one in which there are no differentially expressed genes and one in which 5% of genes are DE. Patients were randomly assigned to two mock groups; for the first case it is expected that, thanks to randomization, there is no DE, while to obtain the second dataset, 5% of gene counts were randomly swapped between the two mock groups hence introducing differential expression. Moreover, since 22 vs 22 studies are very rare, the data set was downsampled to obtain two new data sets in which the are 5 vs 5 and 10 vs 10 comparisons. In this way, not only is the data more realistic, but one can also see how each methods' performance varies according to sample size.

2.6 Evaluation criteria

The performance of each method will be evaluated by calculating both the True Positive Proportion (TPP) and the False Discovery Proportion (FDP).

The TPP, also known as sensitivity, is a measure of the proportion of positive cases that are correctly identified as such. In other words, it is the number of true positive results divided by the number of all positive cases.

The FDP, on the other hand, is a measure of the proportion of false positive results among all positive results. It is defined as the number of false positive results divided by the number of all positive results, both true and false.

Both TPP and FDP are commonly used performance measures in the field of statistics and data analysis, particularly in the context of binary classification problems. They provide complementary information about the accuracy and reliability of a given method and are used to compare different methods and select the best one for a particular problem.

In this work both TPP and FDP will be obtained thanks to the *iCOBRA* package developed by Lisa and Bot (2017) and will be shown together on a FDP-TPP curve. Each curve shows the performance of each method by evaluating the sensitivity with respect to the false discovery proportion. The three circles on each curve represent the points when the FDP level is set at nominal levels of 1%, 5% and 10%, respectively.

Chapter 3

Simulation study

In order to benchmark the performance of the different modeling approaches described in the Methods section, each method will be evaluated on a simulated data-set with known ground truth. Initially, the ability of each method to control the type 1 error proportions at the desired level on a simulated mock data-set without differential expression and differential detection signal will be analysed. Next, the sensitivity and specificity of each method will be analysed on simulated data in which a differential expression and detection signal will be introduced artificially.

The methods used to assess differential detection are the binomial regression (as defined in Equation 2.1), the quasi-binomial regression (as defined in Equation 2.2), a binomial regression with an offset (as defined in Equation 2.3), a quasi-binomial regression with an offset, a squeezed quasi-binomial (as defined in Section 2.3.4), a squeezed quasi-binomial with an offset, a quasi-negative binomial (as defined in Equation 2.4) and a quasi-Poisson (as defined in Equation 2.5). These methods will be tested on aggregated data and, as a reference, also on non aggregated data.

3.1 No differentially expressed genes

The analysed data comes from the Perez et al. (2022) study in which a filtering procedure was first conducted in order to obtain a homogeneous sample in which only healthy European women were retained. Patients were then randomly assigned to two mock groups; thanks to randomization, no differential expression is expected on average. Assuming that the statistical assumptions of the statistical model hold, the p-values obtained from the test should follow a uniform distribution under the null hypothesis. This means that the probability of obtaining a p-value below the 0.05 threshold is equal to the threshold itself which is 5% in this case. This means that 5% of the genes are expected to have a non-adjusted p-value below 0.05.

3.1.1 22 vs 22 patients comparison

Differential detection was initially tested on a data-set containing 22 patients per group and the methods were applied both on aggregated and non aggregated binarized counts.

As a reference differential expression was first evaluated by using the conventional edgeR analysis on the aggregated count matrix which lead to a p-value distribution shown in Figure 3.1. This analysis detected 3.36% of differentially expressed genes; this means that the method had a non-adjusted p-value of 0.05 or smaller for 3.36% of the genes.

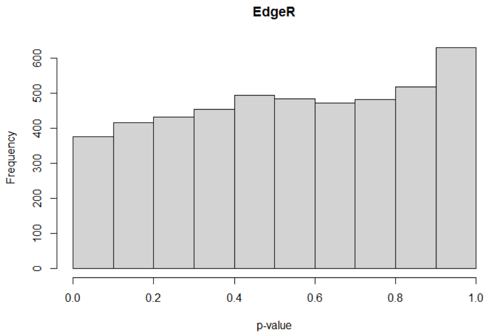

Figure 3.1: p-values of the differential expression analysis conducted with edgeR on the 22 vs 22 patients comparison in which no differential expression is expected.

The analysis on non aggregated binarized counts was then conducted and the resulting p-values of the single-cell level analysis are shown in Figure 3.2. This type of analysis resulted in a higher number of genes found as differentially detected in all methods compared to the edgeR analysis for differential expression. The binomial regression finds 7.43% of differentially detected genes, the quasi-binomial 7.49%, the binomial with an offset 7.34%, the quasi-binomial with an offset 7.85%, the squeezed quasi-binomial 7.47% and the the squeezed quasi-binomial with an offset 7.47%. This shows that the methods have similar type 1 error control but their p-value distribution are overly liberal. This high amount of genes that are found as differentially detected is probably due to that fact that pseudo-replication in not take into account.

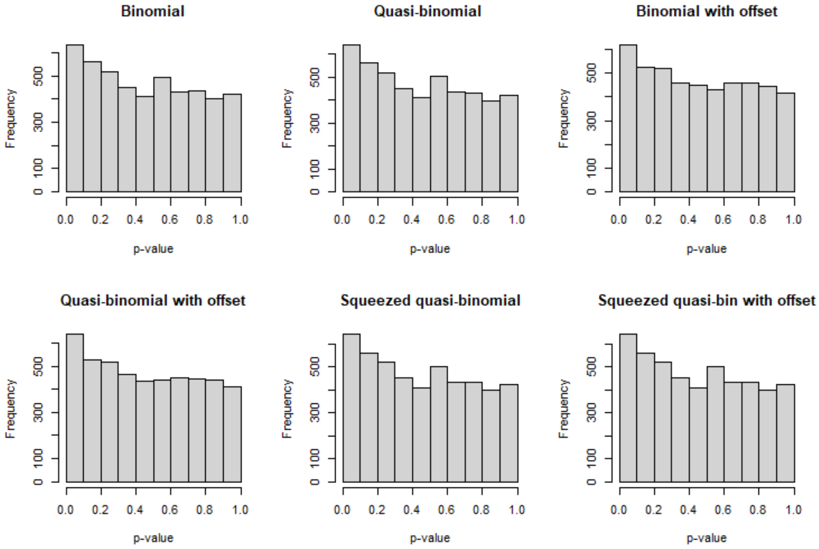

Figure 3.2: p-values of the differential detection analysis conducted on the non aggregated counts of the 22 vs 22 patients comparison in which no differential expression is expected.

The methods were subsequently applied to the binarized aggregated counts; the resulting p-values are shown in Figure 3.3. Compared to the p-values obtained with the analysis on the non aggregated data, apart from the binomial distribution, the p-values in Figure 3.3 have a much more uniform distribution. This can also be noted analysing the percentage of genes with a p-value smaller than 0.05. With the analysis conducted on the aggregated data, in fact, the percentages are 7.43 for the binomial model, 4.15 for the quasi-binomial, 7.15 for the binomial with an offset, 4.45 for the quasi-binomial with an offset, 4.34 for the squeezed quasi-binomial and 3.04 for the squeezed quasi-binomial with an offset. From these results, it is clear that accounting for overdispersion and aggregating the counts per patient is crucial and helps reduce the false positive proportion thus obtaining a uniform p-value distribution.

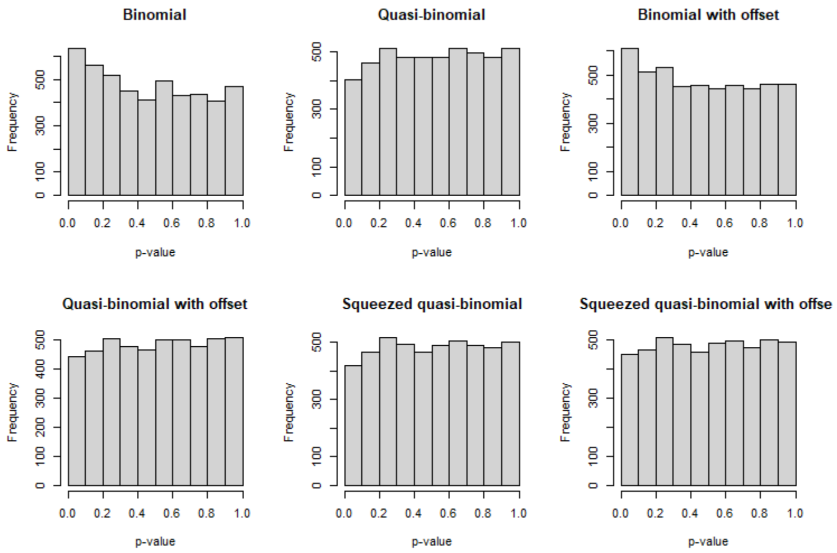

Figure 3.3: p-values of the differential detection analysis conducted on the aggregated counts of the 22 vs 22 patients comparison in which no differential expression is expected.

As described in Section 2.3.5, edgeR was further tailored in order to perform differential detection analyses, by fitting a quasi-negative binomial and a quasi-Poisson on the binarized aggregated counts. The resulting p-values are shown in Figure 3.4; the two approaches respectively produce 3.04% and 3.36% of differentially detected genes.

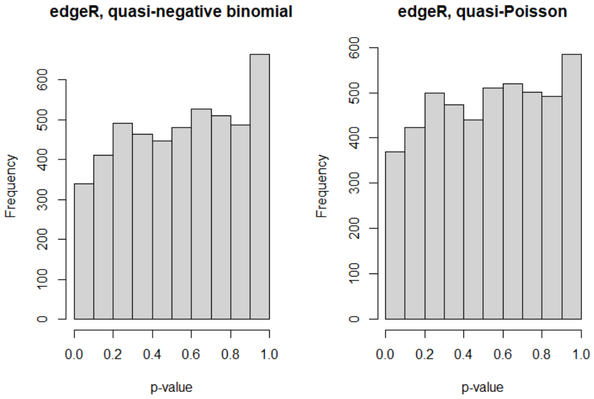

Figure 3.4: p-values of the differential detection analysis conducted with Quasi-NB and quasi-Poisson on the aggregated counts of the 22 vs 22 patients comparison in which no differential expression is expected.

By looking at the p-value distributions obtained by adapting all methods, the quasi-binomial with an offset and the shrunken quasi-binomial with an offset applied on pseudo-bulk level are the methods who produce the most uniformly distributed p-values.

3.1.2 10 vs 10 patients comparison

To investigate how sample size affects the methods' results, 10 patients were sampled and analysed per group. As for the previous case, an edgeR analysis was conducted on the aggregated counts and the results, as can be seen in Figure 4.4 in the Appendix, were very similar to the ones obtained on the 22vs22 data-set.

The analysis was then conducted on single-cell level and resulted in the p-value distributions that can be seen in Figure 3.5. It is interesting to note how all the methods have a spike on p-values equal to zero even though, since it is assumed there is no signal, their distribution should be uniform. The binomial model has in fact signaled 26.92% of the genes as differentially detected, the quasi-binomial 26.92%, the binomial with an offset 13.56%, the quasi-binomial with an offset 15.18%, the squeezed quasi binomial 26.92% and the squeezed quasi-binomial with an offset 14.33%. These percentages are a lot higher than the ones found with the analysis on the 22 vs 22 sample showing how higher sample size probably produces more accurate estimates.

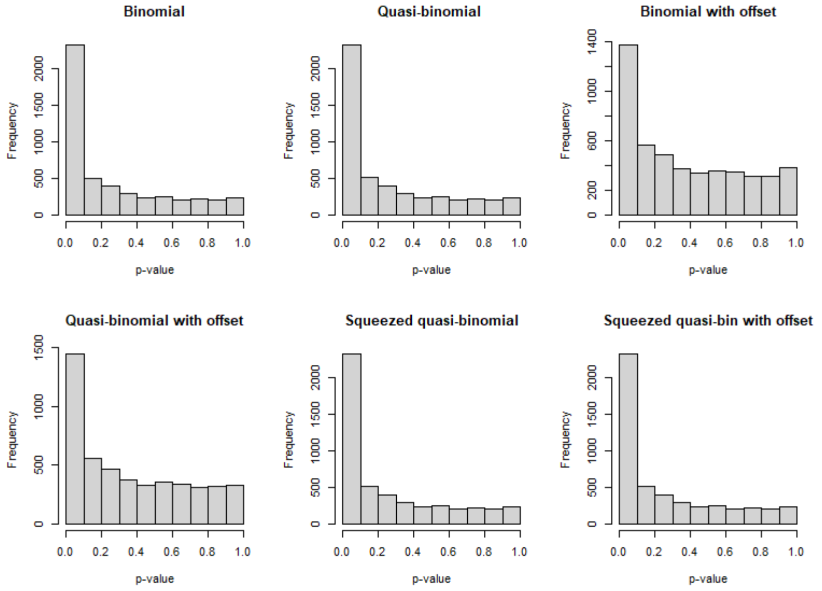

Figure 3.5: p-values of the differential detection analysis conducted on the non aggregated counts of the 10 vs 10 patients comparison in which no differential expression is expected.

The analysis was then conducted on the aggregated binarized counts; the resulting p-values can be seen in Figure 3.6. The percentages of genes flagged by the methods were 26.92% for the binomial model, 3.03% for the quasi-binomial, 13.33% for the binomial with an offset, 4.90% for the quasi-binomial with an offset, 2.20% for the squeezed quasi-binomial and 3.82% for the squeezed quasi-binomial with an offset. What emerges from these analysis, just as for the 22 vs 22 case, is that models that account for over-dispersion have a lower number of differentially detected genes than those who do not account for it. However, compared to the 22 vs 22 analysis the quasi-binomial and squeezed quasi-binomial models seem to have a more conservative p-value distribution while in this case, adding an offset seems to improve the results; the quasi-binomial model with an offset and the squeezed quasi-binomial with an offset in fact have a uniform p-value distribution.

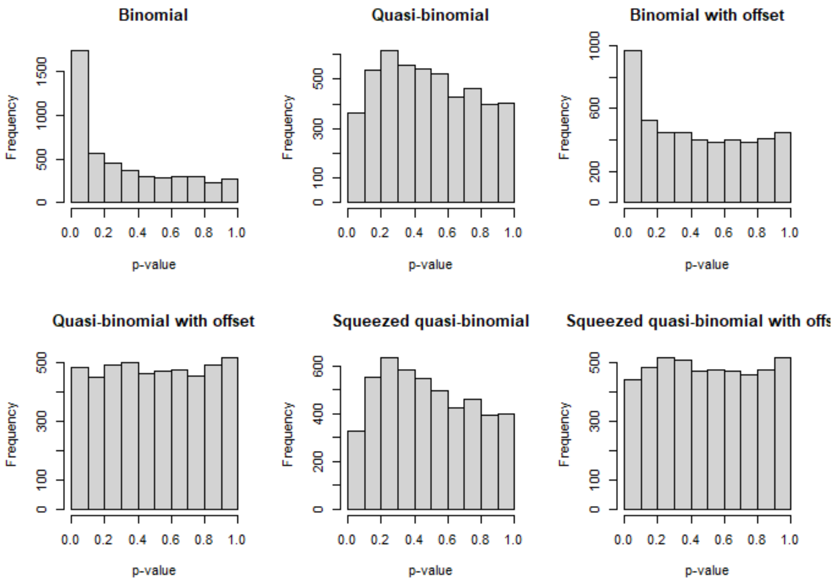

Figure 3.6: p-values of the differential detection analysis conducted on the aggregated counts of the 10 vs 10 patients comparison in which no differential expression is expected.

Quasi-negative binomial and quasi-Poisson models were fitted on the binarized aggregated count matrix and produced coherent results with the 22 vs

22 case, their p-value distribution can be seen in Figure 4.5 in the Appendix.

3.1.3 5 vs 5 patients comparison

In order to further investigate the role that sample size has on the results, 5 patients per group were sampled and analysed. The edgeR analysis again performs similarly to the 22 vs 22 case as can be seen in Figure 4.6 in the Appendix.

Analysis on single-cell level was again conducted and the p-value distributions can be seen in Figure 4.7 in the Appendix. As for the previous data-sets, single-cell level analysis produces a high amount of differentially detected genes underlining the need of working on pseudo-bulk data. The methods were then applied to the binarized aggregated matrix and the resulting p-values are shown in Figure 3.7. Like the previous analysis, models that account for over-dispersion have a lower percentage of DD genes than those who do not account for it.

A quasi-negative binomial and quasi-Poisson model was adapted on the binarized aggregated counts and the resulting p-value distributions are similar to those obtained in the 22 vs 22 analysis; their p-value distributions can be seen in Figure 4.8 in the Appendix.

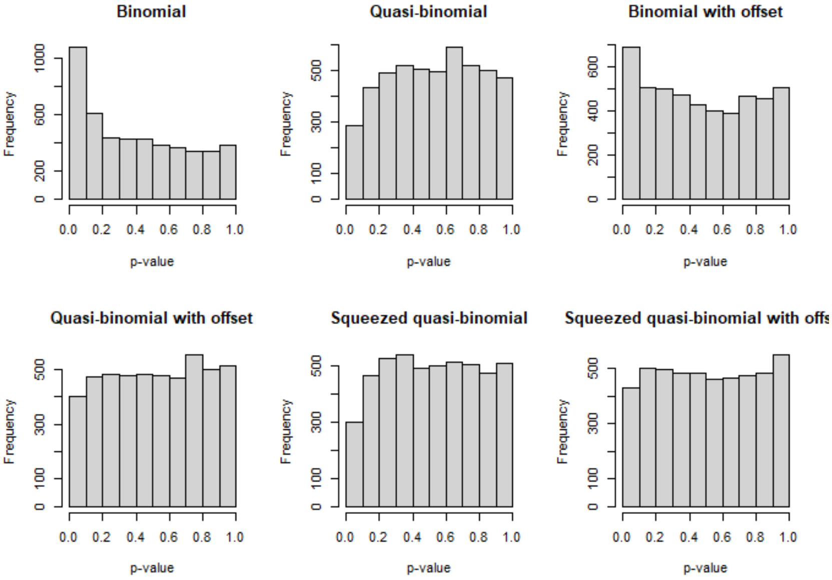

Figure 3.7: p-values of the differential detection analysis conducted on the aggregated counts of the 5 vs 5 patients comparison in which no differential expression is expected.

The percentage of differentially detected genes in all methods and of differentially expressed genes found by the edgeR model, can be easily seen in Table 3.1. In all three sample sizes pseudo-bulk level analysis has better results than the analysis conducted on single-cell level, this is probably due to the fact that pseudo-replications are take into account. Moreover, accounting for over dispersion also helps improve results. In all three data-sets the best performing method is the quasi-binomial with an offset because it has the best type 1 error control and the p-value distribution is uniform.

Method	**22 vs 22 non agg**	**22 vs 22 agg**	**10 vs 10 non agg**	**10 vs 10 agg**	**5 vs 5 non agg**	**5 vs 5 agg**
edgeR	/	3.36	/	3.32	/	4.09
Binomial	7.43	7.43	26.92	26.92	14.70	14.70
Quasi-binomial	7.49	4.15	26.92	3.03	14.64	2.26
Binomial with offset	7.34	7.15	13.56	13.33	8.18	8.04
Quasi-bin with offset	7.85	4.45	15.18	4.90	8.49	3.99
Squeezed quasi-bin	7.47	4.34	26.92	2.20	14.64	2.10
Sq. quasi-bin with offset	7.47	3.04	14.33	3.82	8.49	4.26
Quasi-NB	/	3.04	/	1.72	/	3.16
Quasi-Poisson	/	3.36	/	3.63	/	2.62

Table 3.1: Percentage of differentially detected and differentially expressed (edgeR) genes in all methods according to the various sample sizes. It is expected that 5% of genes have a p-value below 0.05.

3.2 Differentially expressed genes

The analysis will now be repeated on the data coming from the Perez et al. (2022) study in which a filtering procedure was first conducted in order to obtain a homogeneous sample in which only healthy European women were retained. Patients were then randomly assigned to two mock groups and thanks to randomization, no differential expression is expected on average. Signal was then induced in 5% of the genes by randomly swapping them between the two groups as described in the Section 2.5. It is expected that the p-value distribution has a spike on zero but, apart from that, it has a uniform distribution. This means that, if we start from a mock comparison with 4767 genes and 5% of them are then swapped (238 genes), 4529 non-swapped genes are left. It is expected to have 5% of the p-values below 0.05 due to random chance (226 genes) which means that 464 genes are expected to be found with a p-value below 0.05. This means that a non-adjusted p-value below 0.05 is expected in 9.73% of genes.

The performance of each method will be evaluated by calculating TPP and FDP as explained in Section 2.6.

3.2.1 22 vs 22 patients comparison

A differential expression analysis initially was conducted on the aggregated count matrix of the 22 vs 22 data-set with the edgeR method; the resulting p-value distribution can be seen in Figure 3.8 and the percentage of genes that the method found as differentially expressed is 7.30%.

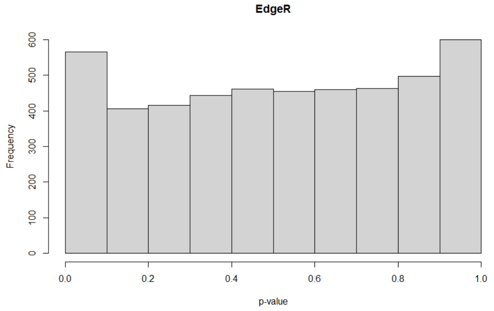

Figure 3.8: p-values of the differential expression analysis conducted with edgeR on the 22 vs 22 patients comparison in which differential expression is expected in 5% of genes.

The analysis was then conducted on single-cell level on the binarized counts; the p-value distributions can be seen in Figure 3.9. The binomial regression flagged 11.41% of genes as differentially detected, the quasi-binomial 11.47%, the binomial with an offset 11.22%, the quasi-binomial with an offset 11.73% and both the squeezed quasi-binomial and the squeezed quasi-binomial with an offset 11.45%.

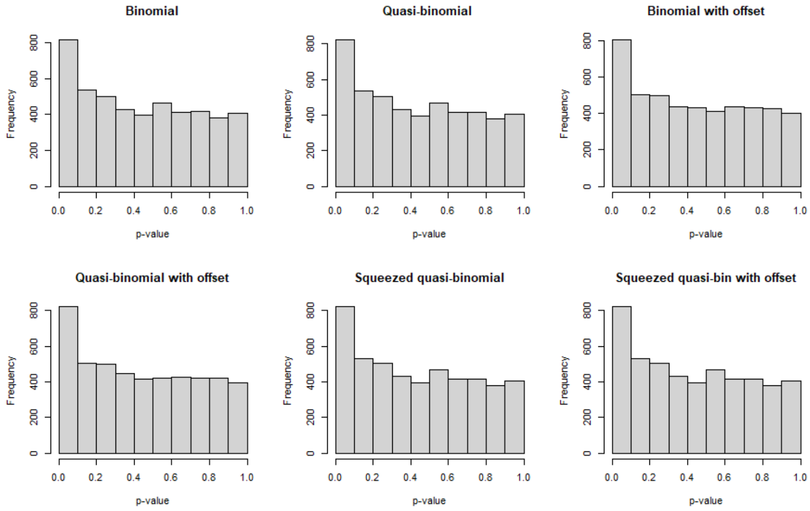

Figure 3.9: p-values of the differential detection analysis conducted on the non aggregated counts of the 22 vs 22 patients comparison in which differential expression is expected in 5% of genes.

The performance of each method was then evaluated by comparing each model's False Discovery Proportion - True Positive Proportion curves. As mentioned in Section 2.6, a method performs well if it has low levels of and high levels of TPP, this means that the curves in the corresponding plot should be in the top-left part. The edgeR analysis is shown as a benchmark to see how single-cell level analysis compare to it. It is clear that there is more statistical power to detect differential expression than differential detection. The methods applied to the binarized non aggregated count matrix all have similar perfromance with the squeezed quasi-binomial performing marginally better.

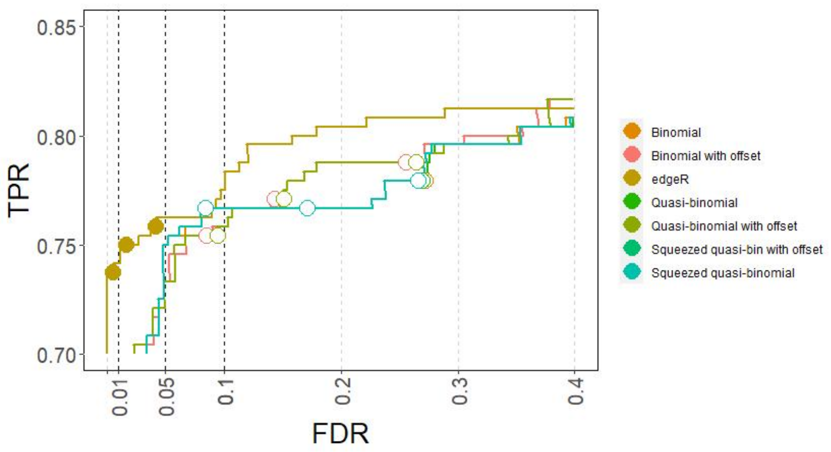

Figure 3.10: Performance evaluation of the methods with false discovery proportion and true positive proportion curves. Differential detection results applied to non aggregated data and differential expression assessed with edgeR are compared in the 22 vs 22 patients comparison.

The analysis was then conducted on the aggregated binarized counts; the p-value distributions can be seen in Figure 3.11. The binomial model presents 11.41% of genes as differentially detected, the quasi-binomial 8.03%, the binomial with an offset 11.08%, the quasi-binomial with an offset 8.31%, the squeezed quasi-binomial 7.64% and the squeezed quasi-binomial with an offset 7.89%. As already suggested by the analysis conducted on the data sets with no differentially expressed genes, it is clear that analysis conducted on aggregated counts have lower percentages of genes that are found as differentially detected than the analysis on single-cell level, thus likely having better type 1 error control. Moreover, methods that account for over-dispersion help reduce this percentage.

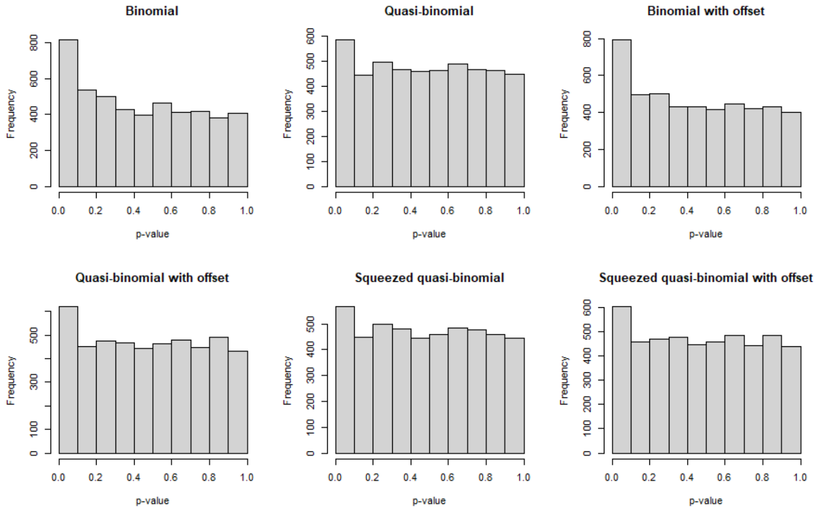

Figure 3.11: p-values of the differential detection analysis conducted on the aggregated counts of the 22 vs 22 patients comparison in which differential expression is expected in 5% of genes.

Both the quasi-negative binomial and the quasi-Poisson were fitted to the aggregated binarized data and the p-values shown in Figure 3.12 were obtained. The quasi-negative binomial produced 6.96% of differentially detected genes whereas the quasi-Poisson 7.32%.

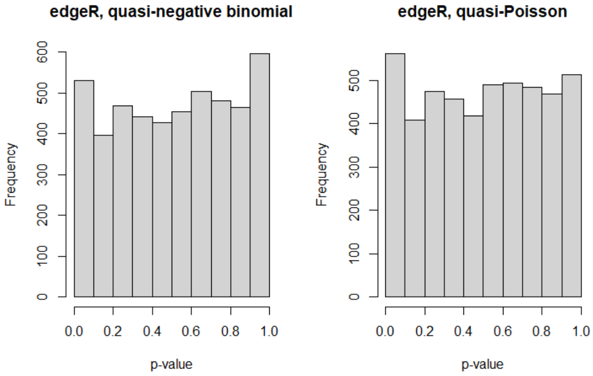

Figure 3.12: p-values of the differential detection analysis conducted with Quasi-NB and quasi-Poisson on the aggregated counts of the 22 vs 22 patients comparison in which no differential expression is expected.

The performance of each method was evaluated by comparing the False-Discovery Proportion - True Positive Proportion curves shown in Figure 3.13. As for the non aggregated data, also in this case all methods seem to have similar performances except for the binomial regression strategy that has a higher False Discovery Proportion. Methods that have the highest True Positive Proportions and lowest False Discovery Proportions are the quasi-negative binomial and quasi-Poisson.

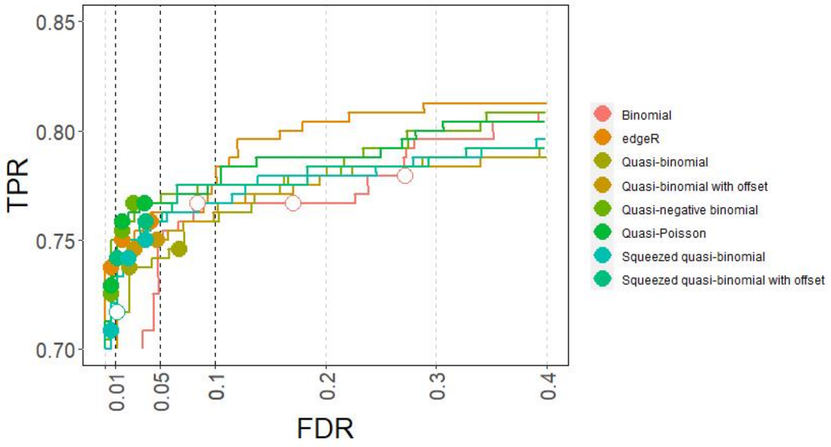

Figure 3.13: Performance evaluation of the methods with false discovery proportion and true positive proportion curves. Differential detection results applied to aggregated data and differential expression assessed with edgeR are compared in the 22 vs 22 patients comparison.

To see how the methods compare on aggregated and non aggregated data a FDP-TPP curve was obtained with the top performers of both cases. Since quasi-Poisson and quasi-negative binomial had the same performance, the quasi-binomial model with an offset was considered for the aggregated data. The quasi-Poisson model has a slightly better TPP and . Both edgeR and the quasi-binomial with an offset adapted on pseudo-bulk level have in fact very similar results suggesting that if aggregation and over-dispersion are take into account the performance is good. The squeezed quasi-binomial adapted on non aggregated data has a higher FDP compared to the other methods. These results seem to indicate that aggregation helps to control false discovery proportion without losing the ability of detecting true positive genes.

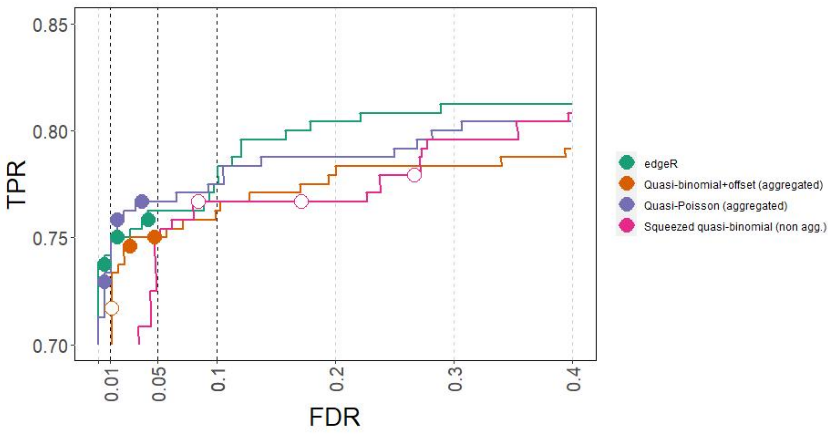

Figure 3.14: Performance evaluation of the methods with false discovery proportion and true positive proportion curves. Methods with the best performance at single-cell level and at pseudo-bulk level were compared in the 22 vs 22 patients comparison. The quasi-binomial with an offset and quasi-Poisson were modeled on the binarized aggregated counts, the squeezed quasi-binomial was modeled on the binarized count matrix and edegR was used to assess differential expression on the pseudo-bulk data.

3.2.2 10 vs 10 patients comparison

10 patients per group were sampled in order to obtain a smaller data set and see how the methods perform with a smaller sample size. As a benchmark, differential expression was analysed with edgeR on the aggregated count matrix producing similar results as for the 22 vs 22 case.

The methods were then applied to the binarized count matrix. Like in the analysis on the 10 vs 10 data-set in which no differential expression was assumed, the analysis on single-cell level produces a high peak on zero p-values as shown in Figure 4.10 in the Appendix, thus suggesting these methods provide overly liberal results. The performance of the methods (Figure 4.11 in Appendix) was similar to the one obtained in the 22 vs 22 study underlining how the edgeR method performs better then those applied on single-cell level; amongst the models fitted on the aggregated data the quasi-binomial regression seems to have better true positive proportions and lower false discovery proportions than the remaining methods.

The methods were then applied to the binarized aggregated matrix and the resulting p-values can be seen in Figure 3.15. Also in this case, it is clear that accounting for over-dispersion and working on pseudo-bulk data helps reduce the proportion of differentially detected genes.

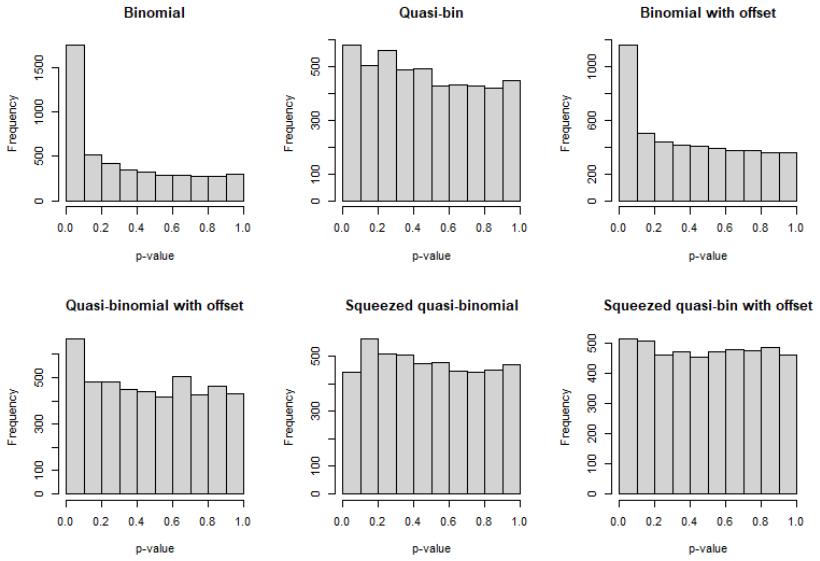

Figure 3.15: p-values of the differential detection analysis conducted on the non aggregated counts of the 10 vs 10 patients comparison in which differential expression is expected in 5% of genes.

The quasi-negative binomial and the quasi-Poisson were also fitted to the binarized aggregated matrix and produced similar results to those obtained in the 22 vs 22 case. Their p-value distributions can be seen in Figure 4.12 in the Appendix.

Each method's performance was evaluated with the FDP-TPP curve shown in Figure 3.16; also in this case the performance was coherent with what emerged in the 22 vs 22 analysis where the edgeR is the method with the best TPP even though it just slightly over-performs both the quasi-binomial model with an offset and the squeezed quasi-binomial with an offset. Also in this case, aggregation based methods have both better FDP and TPP than analysis conducted on single-cell level.

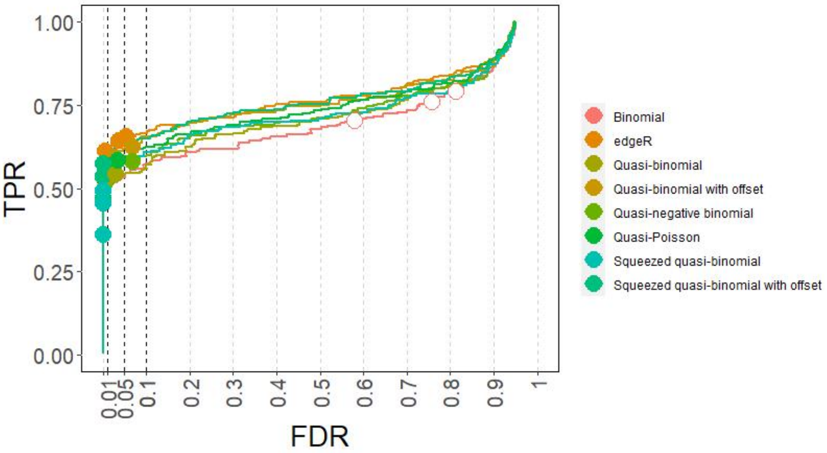

Figure 3.16: Performance evaluation of the methods with false discovery proportion and true positive proportion curves. Differential detection results applied to aggregated data and differential expression assessed with edgeR are compared in the 10 vs 10 patients comparison.

Figure 3.17 shows the FDP-TPP curves of the differential detection analysis conducted on single-cell level with a squeezed quasi-binomial and on pseudo-bulk level with both a quasi-Poisson and a quasi-binomial with an offset. The differential expression analysis assesses with edgeR was also conducted as a reference. It is clear how pseudo-bulk analysis has a much better performance that differential detection analysis conducted on single-cell level since it has much higher FDP; this happens because pseudo-replicates are not taken into consideration. As for the 22 vs 22 comparison, the three methods applied on pseudo-bulk level have very similar performances; in this case for the differential detection analysis the quasi-binomial with an offset is to be preferred to the quasi-Poisson.

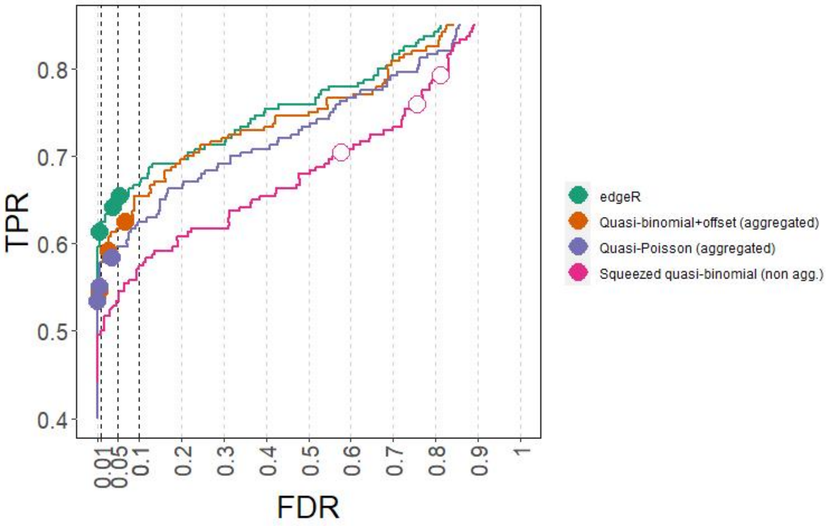

Figure 3.17: Performance evaluation of the methods with false discovery proportion and true positive proportion curves. Differential detection was obtained in the 10 vs 10 patients comparison with a quasi-Poisson and a quasi-binomial with an offset on pseudo-bulk level and with a squeezed quasi-binomial on single-cell level and differential expression was assessed with edgeR.

3.2.3 5 vs 5 patients comparison

To further investigate the role played by sample size, 5 patients were sampled per group and then analysed. Differential expression was first analysed with the edgeR model adapted on the aggregated count matrix and similar results to the 22 vs 22 analysis were obtained. The binarized aggregated matrix was then analysed and, as for the bigger sample sizes, all methods found a high percentage of differentially detected genes between the two groups. As can be seen in Table 3.2 adding an offset helps reduce the percentage of DD genes. FDP-TPP curves were used to compare each model's performance and, as for the previous analysis conducted on single-cell level, edgeR over performs the other models; amongst the methods applied to the binarized count matrix binomial with an offset and the quasi-binomial with an offset have the best results both as regards the true positive proportion and the false discovery proportion (Figure 4.15 in Appendix).

The analysis was then conducted on pseudo-bulk level; the p-value distributions can be seen in Figure 3.18. These results underline the importance of aggregation and how, with a low sample size, it it fundamental to account for both over-dispersion and normalization.

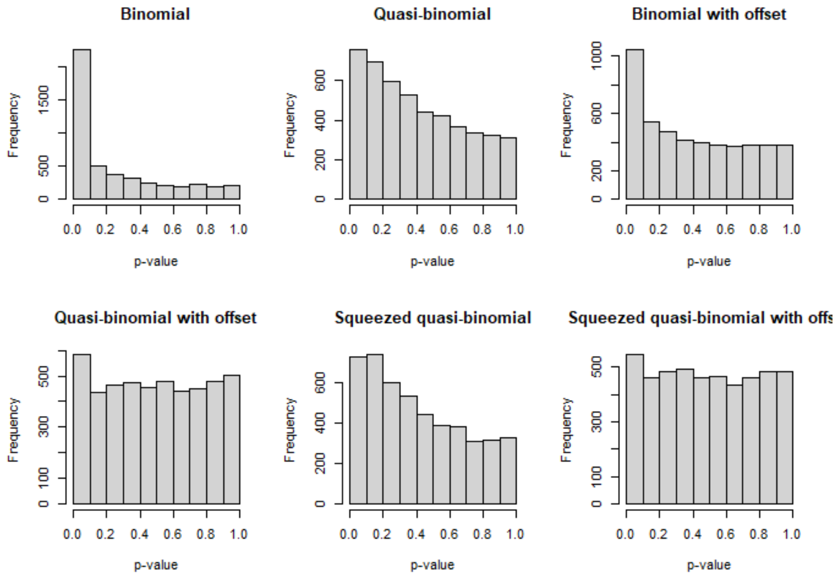

Figure 3.18: p-values of the differential detection analysis conducted on the aggregated counts of the 5 vs 5 patients comparison in which differential expression is expected in 5% of genes.

Both the quasi-negative binomial and the quasi-Poisson were fitted to the aggregated binarized data and have similar p-value distributions as in the 22 vs 22 case (Figure 4.16 in Appendix).

The performance of each method was evaluated by comparing the false-discovery proportion - true positive proportion curves shown in Figure 3.19. The edgeR analysis slightly performs better than the other methods. Apart from the binomial model and the quasi-binomial model, all methods seem to have similar performances.

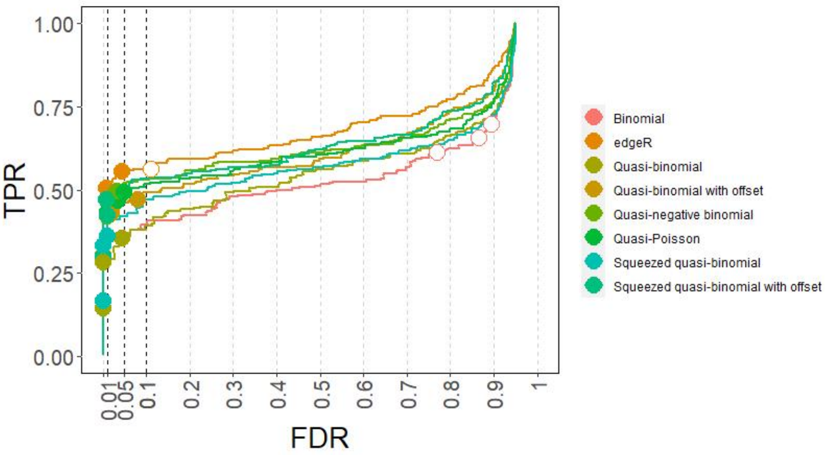

Figure 3.19: Performance evaluation of the methods with false discovery proportion and true positive proportion curves. Differential detection results applied to aggregated data and differential expression assessed with edgeR are compared in the 5 vs 5 patients comparison.

Methods applied on single-cell level and on pseudo-bulk level are compared in Figure 3.20. Differential detection was asses on non aggregated data with a squeezed quasi-binomial whereas on aggregated data a quasi-Poisson and a quasi-binomial with an offset were used. Differential expression was also tested with the edgeR model. It can clearly be seen how not accounting for the presence of pseudo-replicates in the data leads to a high false discovery proportion which indicates that the analysis should be conducted on pseudo-bulk level. Also in this case the methods adapted on aggregated data have a very similar performance.

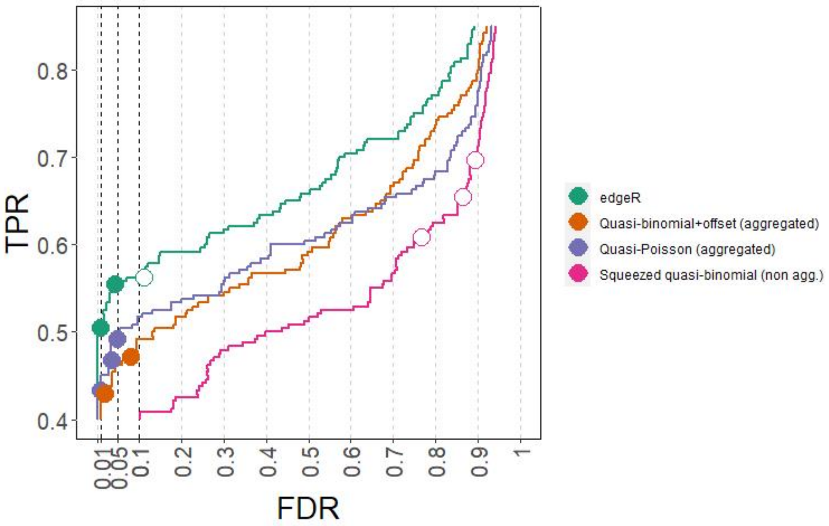

Figure 3.20: Performance evaluation of the methods with false discovery proportion and true positive proportion curves. Differential detection was obtained in the 5 vs 5 patients comparison with a quasi-Poisson and a quasi-binomial with an offset on pseudo-bulk level and with a squeezed quasi-binomial on single-cell level and differential expression was assessed with edgeR.

The percentage of differentially detected genes in all methods and of differentially expressed genes found by the edgeR model, can be easily seen in Table 3.2. Since the information of 5% of genes was swapped, 5% of genes are expected to be found on average as differentially expressed. As for the analysis conducted on the data-set in which no differential expression was induced, pseudo-bulk level analysis has a better type 1 error control than the analysis performed on single-cell level, this is probably due to the fact that pseudo-replication is accounted for. In the 10 vs 10 patient comparison the quasi-binomial with an offset was the top performing method whereas in the

22 vs 22 and 5 vs 5 comparisons it was the quasi-Poisson. This suggests that they probably have a similar performance in general.

Method	**22 vs 22 non agg**	**22 vs 22 agg**	**10 vs 10 non agg**	**10 vs 10 agg**	**5 vs 5 non agg**	**5 vs 5 agg**
edgeR	/	7.30	/	8.47	/	7.43
Binomial	11.41	11.41	28.72	28.72	38.14	38.14
Quasi-binomial	11.47	8.03	28.72	7.11	37.99	8.27
Binomial with offset	11.22	11.08	17.64	17.47	15.84	15.61
Quasi-bin with offset	11.73	8.31	18.25	8.50	16.55	7.70
Squeezed quasi-bin	11.45	7.64	28.72	4.85	37.99	7.36
Sq. quasi-bin with offset	11.45	7.89	18.27	6.10	16.34	6.65
Quasi-NB	/	6.96	/	8.77	/	8.03
Quasi-Poisson	/	7.32	/	7.30	/	9.08

Table 3.2: Percentage of differentially detected and differentially expressed (edgeR) genes in all methods according to the various sample sizes. It is expected that 9.73% of genes have a p-value below 0.05.

3.3 Stage-wise analysis

As mentioned throughout this work, single-cell RNA-seq data allows to look at two aspects of the distribution: differences in mean and in detection. Thanks to this, both differential detection and differential expression were tested on the genes. It is now of interest to try and aggregate the information obtained by both methods in order to try and increase power. To do so a stage-wise analysis was implemented as proposed by Van den Berge et al. (2017). This analysis was conducted twice for every data-set, one time in which differential detection was assessed by the best performing model on the non aggregated data and once with the model that resulted the best on the aggregated data. Differential expression, in both cases, was obtained with the edgeR analysis.

For the 22 vs 22 analysis differential detection on single-cell level was assessed with the squeezed quasi-binomial. 214 genes (4.50%) resulted as having difference in signal between the two groups, of these 211 were differentially detected and 183 differentially expressed. The overlap between the results can be seen in Figure 4.19; the majority of genes were detected by the three methods. It can also be seen that the squeezed quasi-binomial picks-up a higher number of genes than the edgeR model, this result was to be expected since, as can be seen in the first plot of Figure 3.9, it is overly liberal, thus producing many false positives. Furthermore, 53 genes that are known to be differentially expressed, were not identified by any model. Analysing these specific genes it shows that they mainly have zero counts and so, even though the counts were swapped between the two conditions, they will essentially remain unaltered. Looking at the first FDP-TPP curve in Figure 4.20, it emerges that the edgeR analysis still has the lowest false discovery proportion and the stage-wise analysis does not seem to improve either FDP or TPP. This, and the fact that methods do not have a big difference, can be due to the fact that the simulation strategy induces big differences between the groups so the effect is easy to detect. The stage-wise analysis does not help because, with swapping, usually both differential detection and expression is induced.

Differential detection was then tested with the quasi-binomial on the aggregated binarized data. In this case the stage-wise analysis determined that there is signal in 186 genes, 181 of which are differentially detected and 182 differentially expressed. The fact that modeling pseudo-bulk data helps reduce the number of false positive genes can be seen in Figure 4.20, where the number of these genes is a lot lower than the ones found in the previous analysis. Figure 4.20 shows that the screening analysis, in this case, manages to detect a higher number of genes while controlling FDP.

The analysis were also performed on the 10 vs 10 and 5 vs 5 data-sets (see Appendix) which gave similar results therefore suggesting that, due to the strong signal produced by swapping genes, stage-wise analysis does not result in an increased performance compared to testing for DE and DD with separate tests.

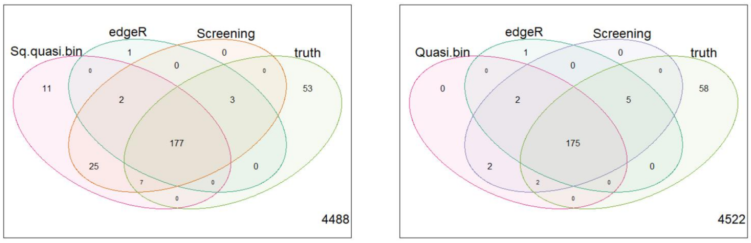

Figure 3.21: The graphs show Venn diagrams in which the overlap of the methods analysed are shown. In the left panel the overlap between differential detection assessed with the squeezed quasi-binomial on non aggregated data, differential expression tested with edgeR, the screening results and the real DE genes is shown. In the right panel differential detection is instead assessed with the quasi-binomial adapted on pseudo-bulk level.

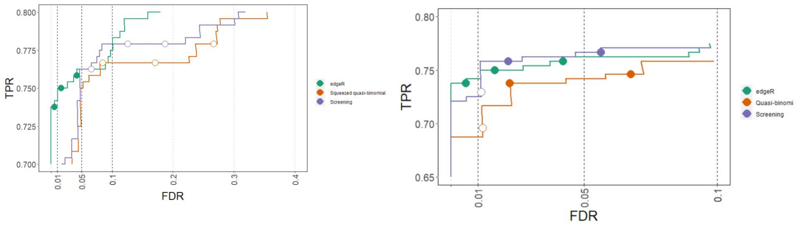

Figure 3.22: Performance evaluation of the methods with false discovery proportion and true positive proportion curves. The left panel compares the screening result with differential expression assessed with edgeR and differential detection tested with a squeezed quasi-binomial adapted on single-cell level. The right panel instead used a quasi-binomial adapted on pseudo-bulk level to assess differential detection.

Chapter 4

Application on real data

4.1 Reference data

Systemic lupus erythematosus (SLE) is the most common type of lupus; it is an autoimmune disease which causes the immune system to attack its own tissues, causing tissue inflammation and organ damage. SLE can affect joints, skin, brain, lungs, kidneys and blood vessels. To date, there still is no cure for this disease but there are medical interventions that can help control it (Carter, Barr, and Clarke (2016)). Moreover SLE is hard to diagnose because it has a wide range of symptoms. As such, it is of interest to try and identify a genetic component linked to lupus susceptibility.

Different approaches have been used to try and identify the genetic component linked to SLE. Flow cytometry analysis was applied to quantify the composition on the basis of known cell surface markers; this approach reported B and T cell lymphopenia (a disorder in which the blood does not have enough white blood cells Ducloux et al. (2010)). Moreover, a bulk trascriptomic analysis was preformed on peripheral blood mononuclear cells (PBMCs), which reported elevated expression of interferon-stimulated genes.

However, neither methods are considered optimal since flow cytometry is biased, as it uses a limited set of markers, whereas bulk trascriptomic analysis does is not able to detect cell type specific differences in expression. This implies that single-cell RNA sequencing could provide an unbiased approach for detecting cell type specific transcriptional states of circulating immune cells.

The data that will be used throughout this dissertation was collected from systemic lupus erythematosus cases in the California Lupus Epidemiological Study (CLUES) cohort, matching healthy controls from the UCSF Rheumatology Clinic, and additional controls from the Immune Variation Project (ImmVar). As described by Perez et al. (2022), mux-seq was used to profile 1.2 million human peripheral blood mononuclear cells (PBMCs) coming from 264 samples, of which 162 corresponded to SLE cases and the remaining were healthy controls. The majority of patients were women either of European or Asian descent. This is because it is known that SLE mainly affects women and those with Asian, African, and Hispanic ancestries (Carter, Barr, and Clarke (2016)).

PBMCs were pooled and profiled using 10x Genomics' Chromium Single Cell 3' V2 chemistry and processed using the 10x Cell Ranger pipeline. The cells were assigned to their donor with the Freemuxlet technology (Kang et al. (2018)) and quality control and doublet removal was performed with Scrublet (McGinnis, Murrow, and Gartner (2019)). Moreover platelets, megakaryocytes and red blood cells were removed using gene markers yielding a total of 1,263,676 cells remaining in the final dataset. Cell types were annotated using canonical marker genes and for each cell type, a percentage was calculated as the number of cells divided by the total number of cells assigned to the sample.

The analysis conducted by Perez et al. (2022) suggests that the decrease of $CD4^{+}T$ cells in Asian women explains the lymphopenia observed in patients with SLE and was not associated with immunosuppressant treatment. Moreover *ncMs* (non-classical myeloid cells) produced the biggest type 1 signature. Although *cDCs* (classical dendritical cells) and *pDCs* (plasmacytoid dendritic cells) expressed interferon signaling, their scarcity in circulation limited their contribution to the overall signal. *IFNB1* and *IFNA* were neither detected in *pDCs* nor in other myeloid cell types which underlines how type 1 interferons are likely not among circulating immune cells. An intereseting result is the expansion of *GZMH+* but not *GZMK+* cytotoxic *CD8+* T cells in SLE. The significant expansion of *GZMH+ CD8+* T cells suggests a pathogenic role for these cells in SLE. These results allow the authors to hypothesize a model for the initiation and worsening of SLE: an adaptive immune response is initiated by foreign and auto-antigens followed by chronic exposure to antigens in damaged tissue. This results in epitope spreading, where new auto-antigens are introduced into the immune system and become future targets of the autoimmune response.

The analysed data comes from the study conducted by Perez et al. (2022) on systemic lupus erythematosus. The original data-set contains information about 32738 genes and 1263676 cells coming from 261 patients of which 149 are European women, 107 are Asian and the remaining are African American.

Before being analysed, we filtered the data-set in order to account for possible confounders. Only European women were retained because they were the largest group. Moreover, samples of only five particular batches were retained, so that all data comes from either the CLUES or LupCon study and each batch has cells of either study allowing for incorporating the batch variable in the model. Gene level filtering was subsequently applied:

only genes that have some expression in at least 200 cells per cell type were retained. A differential detection analysis was conducted separately for each cell type: B memory cells, T4 naive and non-classical myeloid cells. These three cell types were selected because the study contains a high number of cells coming from the T4 naive cell type, a medium number of B memory cells and a relatively low number of non-classical myeloid cells.

Due to the high dimensions of the data set and because all methods performed better on aggregated data in the simulation study, only pseudobulk analysis were conducted on the lupus data set.

4.2 B memory cell type

The first cell type to be analysed are memory B cells (*Bmem*). B cells are a type of white blood cells that are crucial for the immune system; they are responsible for producing antibodies which help to identify and neutralize foreign substances but also recognize and remember specific antigens; this allows a more rapid and effective immune response to subsequent infections. B memory cells specifically have the function of memorizing the characteristics of the antigen that activated their parent B cell during initial infection allowing the immune system to recognize it in case another infection occurs.

The data-set contains 13596 cells coming from 47 patients of which 22 are healthy and 25 are affected by lupus. After gene-level filtering 2556 genes are retained.

As a reference, a differential expression analysis was conducted on the aggregated counts with edgeR and 19.01% of genes resulted differentially expressed between healthy women and those with lupus.

A differential detection analysis was then conducted on the binarized aggregated counts; the binomial regression found 34.59% of genes to be differentially detected, the quasi-binomial 21.67%, the binomial with an offset 30.59%, the quasi-binomial with an offset 20.62%, the squeezed quasi-binomial 21.99%, the squeezed quasi-binomial with an offset 20.74%, the quasi-negative binomial 16.16% and the quasi-Poisson 20.66%. The p-value distributions of each method can be seen in Figure 4.1.

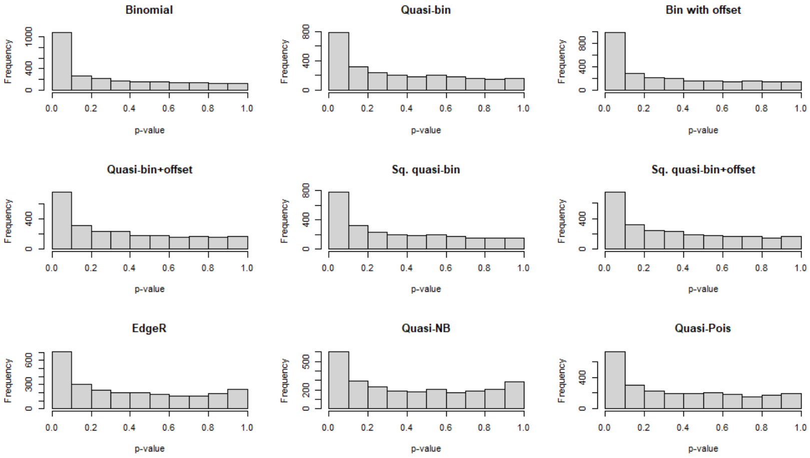

Figure 4.1: p-value distribution of the differential detection analysis conducted on pseudo-bulk level on the B memory cell type. Differential expression results assess with edgeR are added as a reference.

It can be seen how both the binomial regression and the quasi-binomial with an offset have found a higher number of genes as differentially detected compared to the other methods. This is coherent with the results of the simulation study and thus underlining how these two models have a worse control of type 1 error rate. The remaining methods instead have similar percentages and, as suggested by the simulation study results, similar type 1 error rate control.

For this cell type, both the quasi-Poisson and the quasi-binomial with an offset have very similar percentages of genes that are found as differentially detected. These two methods were the ones that resulted as top performing in the simulation study both in term of of TPP and FDP. Since the two methods have a very similar performance but the quasi-Poisson is faster, it will be here used to compare differentially detected genes with those found as differentially expressed by the edgeR model. As previously mentioned in fact, single-cell RNA-seq data provides the opportunity to examine two different aspects of gene expression distributions: differences in mean expression levels, which are here assessed using edgeR, and differences in detection rates, which are here analyzed by using the quasi-Poisson model. Combining these two aspects can provide a more comprehensive understanding of gene expression patterns and their biological significance. Therefore, these two analyses will be integrated in order to gain further biological knowledge and to do so a stage-wise analysis is conducted. 146 genes resulted as having difference in signal between the two groups, of these 113 are differentially detected and 103 differentially expressed. It is therefore clear that by looking at these two aspects a greater number of genes is taken into consideration therefore gaining additional biological insights.

4.3 T4 naive cell type

T4 cells, also known as CD4+ T cells, are a type of T cell (a type of white blood cell) that are present in the adaptive immune system. T4 cells play a role in the regulation of other immune cells and in the development of immunity and tolerance of self-antigens (Caza and Landas (2015)). Naive T4 cells are a sub-population of CD4+T cells that have not yet encountered antigens; it is believed that they play a key role in the initiation of the adaptive immune response since they have the potential to differentiate into different sub-types of T cells (Young and Geha (1986)).

The data-set contains 18959 T4 naive cells, originating from 48 patients of which 22 are healthy and 26 are affected by lupus. After gene-level filtering, 3870 genes are retained.

As a reference, differential expression analysis was conducted on the aggregated counts with edgeR and 15.87% of genes resulted differentially expressed between healthy women and those with lupus.

A differential detection analysis was then conducted on the binarized aggregated counts; the binomial regression found 38.48% of genes to be differentially detected, the quasi-binomial 26.05%, the binomial with an offset 25.25%, the quasi-binomial with an offset 16.64%, the squeezed quasi-binomial 25.99%, the squeezed quasi-binomial with an offset 16.82%, the quasi-negative binomial 26.05% and the quasi-Poisson 25.25%. The p-value distributions of each method can be seen in Figure 4.2.

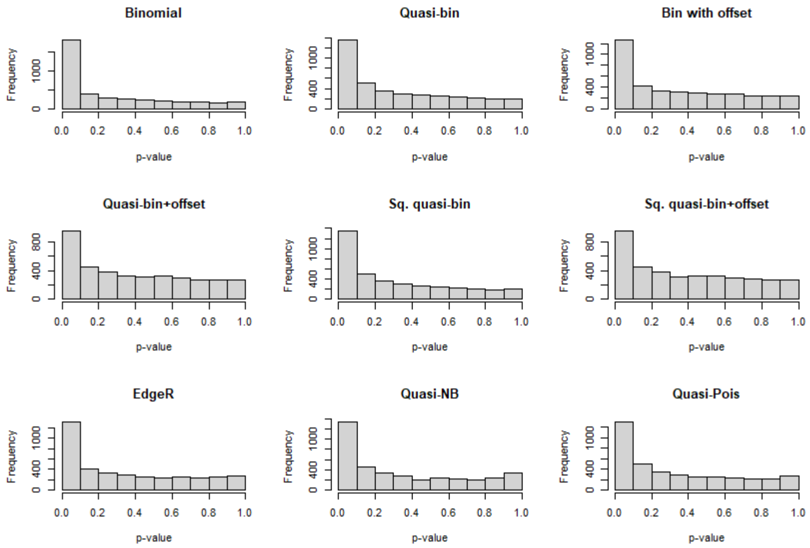

Figure 4.2: p-value distribution of the differential detection analysis conducted on pseudo-bulk level on the T4 naive cell type. Differential expression results assess with edgeR are added as a reference.

It can be seen how the binomial regression has a much higher percentage of differentially detected genes than the other methods. As shown in the simulation study, it is very likely that these genes are false positives since the binomial model alone can not control type 1 error rate. As for the previous cell type, since the quasi-Poisson is faster than the quasi-binomial with an offset it will be used during the upcoming analysis. A stage-wise procedure will in fact be implemented in order to combine the results obtained when looking at differences in mean and when looking at differences in detection. 319 genes resulted as having difference in signal between the two groups, of these 233 are differentially detected and 176 differentially expressed.

4.4 Non-classical myeloid cells

Non-classical myeloid cells (ncMs) are a group of immune cells that are distinct from classical myeloid cells, such as neutrophils, monocytes, and macrophages. They are heterogeneous and include a variety of cell types, such as dendritic cells, eosinophils, basophils, and mast cells, as well as subsets of monocytes and macrophages. ncMs are involved in various physiological and pathological processes, such as tissue repair, immune regulation, and inflammation. They have unique functional properties that allow them to perform specialized roles in these processes. For example, dendritic cells are specialized in presenting antigens to T cells, eosinophils and basophils are important in allergic responses, and mast cells are involved in host defense against parasites and in allergic reactions.

The data-set contains 7020 ncM cells coming from 48 patients of which 22 are healthy and 26 are affected by lupus. After gene-level filtering, 2270 genes are retained.

As a reference, differential expression analysis was conducted on the aggregated counts with edgeR and 22.37% of genes were flagged as being differentially expressed between healthy women and those with lupus.

Differential detection analysis was then conducted on the binarized aggregated counts; the binomial regression found 33.29% of genes to be differentially detected, the quasi-binomial 22.11%, the binomial with an offset 30.18%, the quasi-binomial with an offset 21.87%, the squeezed quasi-binomial 22.00%, the squeezed quasi-binomial with an offset 21.90%, the quasi-negative binomial 19.41% and the quasi-Poisson 20.00%. The p-value distributions of each method can be seen in Figure 4.3.

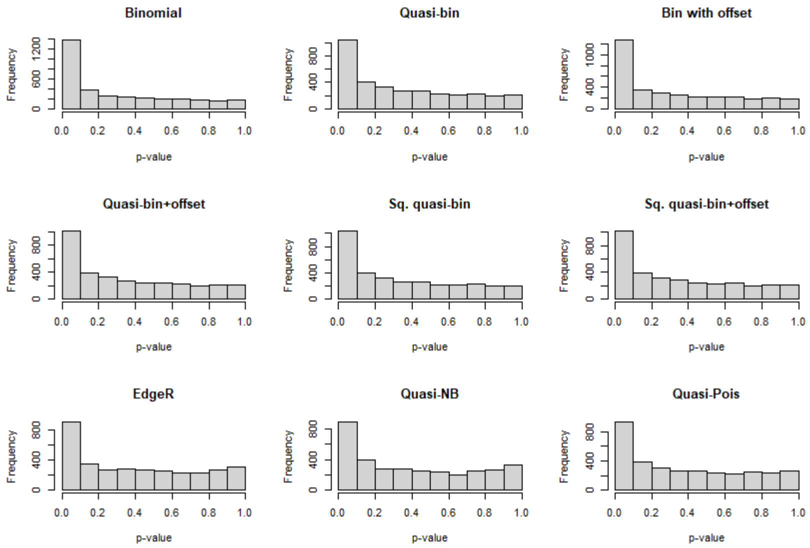

Figure 4.3: p-value distribution of the differential detection analysis conducted on pseudo-bulk level on the non-classical myeloid cell type. Differential expression results assess with edgeR are added as a reference.

Since both differences in mean and in detection were assessed it is of interest to combine these results in order to obtain further biological knowledge. Differential expression was assessed with edgeR whereas for differential detection numerous methods were applied. Again, the quasi-Poisson model will be used during the stage-wise analysis. 355 genes were found as presenting differences between healthy women and those affected by lupus. Of these genes, 275 are differentially detected and 316 are differentially expressed. It can clearly be seen that if both aspects of the distribution are taken into consideration a much higher number of genes who present differences between the two health statuses are found, potentially resulting in additional biological insights.

The resulting percentage of genes found as differentially detected by all methods and the percentage of differentially expressed genes in all cell types are summarized in Table 4.1.

Cell type	Bin	Quasi-bin	Offset	Quasi+off	Sq. quasi	Sq.quasi+off	Quasi-NB	Quasi-Pois	edgeR
B memory	34.59	21.67	30.59	20.62	21.99	20.74	16.16	20.66	19.01
T4 naive	38.48	26.05	25.25	16.64	25.99	16.82	26.05	25.25	15.87
ncM	33.29	22.11	30.18	21.87	22.00	21.90	19.41	20.00	22.37

Table 4.1: Percentage of genes found as differentially detected and differentially expressed (edgeR) for every method in all cell types.

Table 4.1 shows how the binomial regression and the binomial regression with an offset have a much higher percentage of genes, this is however coherent with what had emerged in the simulation study. Methods that do not account for over-dispersion have in fact a higher rate of differentially detected genes thus having worse control of type 1 error rate.

4.5 Biological interpretation

A gene set enrichment analysis (GSEA) will be conducted to determine whether a set of genes is significantly enriched in a particular biological function, pathway, or phenotype. GSEA can be used to identify biologically relevant pathways or processes that are differentially expressed in two or more groups of samples, such as disease vs. control. In this master thesis, the online platform MSigDB (Molecular Signaltures Database, Subramanian et al. (2005)) was used to perform GSEA.

The Gene Set Enrichment Analysis procedure involves comparing a user-supplied set of genes with the gene sets in MSigDB, which contains pre-defined sets of genes that are associated with various biological processes, pathways, and diseases. The goal is to identify whether the user-supplied set of genes is significantly enriched in any of the pre-defined gene sets. To perform GSEA, the expression data for the user-supplied set of genes and the gene sets in MSigDB are first ranked based on their correlation with a specific phenotype, such as disease status. An enrichment score is then computed for each gene set in MSigDB, reflecting the degree to which the genes in the set are overrepresented at the top or bottom of the ranked list. To assess the statistical significance of the enrichment score, a Fisher's exact test is performed to determine the probability of observing the overlap between the user-supplied gene set and the gene set in MSigDB by chance. The p-value obtained from the Fisher's exact test measures the probability of observing an overlap as extreme or more extreme than the observed overlap, assuming that there is no true association between the user-supplied gene set and the gene set in MSigDB.

In this study, Gene Set Enrichment Analysis will be performed separately for each cell type. First, GSEA will be performed using the set of genes that resulted significantly different with the stage-wise procedure. Additionally, GSEA will be performed on the genes that resulted differentially expressed with the edgeR analysis. The results of the two GSEA analyses will be compared to determine if performing stage-wise testing provides additional biological insight.

4.5.1 B memory cells

In the GSEA analysis performed using the set of differentially expressed genes, the three most significant gene sets are cytosolic ribosome, cytoplasmic translation and the ribosome. Instead, when conducting GSEA on the genes found with the stage-wise testing analysis, the top three gene sets identified were those related to immune response, RNA binding and innate immune response. These differences underline how stage-wise testing provides further biological knowledge than only performing a differential expression analysis. The immune response gene set can be related to lupus because lupus is an autoimmune disease that occurs when the immune system attacks healthy tissues in the body. In lupus, there is often an overactive immune response and dysregulation of immune cells and molecules. The immune response gene set identified in GSEA analysis may contain genes involved in various aspects of the immune response, such as immune cell activation, inflammation, and cytokine signaling, that are known to play a role in the pathogenesis of lupus (Rönnblom and Pascual (2008)). Therefore, the identification of the immune response gene set in GSEA analysis suggests a potential association with the pathogenesis of lupus.

4.5.2 T4 naive cells

In the T4 naive cell type instead, there were no differences between the gene sets obtained when performing GSEA on the genes found with the stage-wise testing analysis and those differentially expressed. This highlights how stage-wise testing does not give substantially different results than those obtained with edgeR.

4.5.3 Non-classical myeloid cells

GSEA was performed separately for two groups of genes, those obtained with stage-wise testing and those resulting differentially expressed with edgeR. Upon comparing the gene sets obtained from the two groups using GSEA, they were found to be similar. However, the results based on the genes from stage-wise analysis contained the peptide biosynthetic process gene set, which was not present in the results based on the genes that were only differentially expressed. This result is of interest because there is evidence suggesting that peptide biosynthesis may be involved in the pathogenesis of lupus. Peptides derived from self-antigens can be presented to T cells by antigen-presenting cells, leading to the activation of autoreactive B cells and the production of autoantibodies. In particular, there is some evidence that abnormal peptide biosynthesis and presentation may contribute to the development of lupus. For example, studies have shown that autoantibodies in lupus patients can recognize and bind to peptides derived from self-antigens, suggesting a role for abnormal peptide presentation in the pathogenesis of the disease. Additionally, genetic variations in genes involved in peptide biosynthesis and processing, such as the HLA genes, have been associated with an increased risk of developing lupus (Klein and Sato (2000)). While the relationship between peptide biosynthesis and lupus is not yet fully understood, these findings suggest that abnormal peptide presentation may play a role in the development of autoimmunity and the pathogenesis of lupus and stage-wise testing provides further biological knowledge than simply performing and edgeR analysis.

Discussion

The aim of this master dissertation was to leverage single-cell RNA-seq data to examine two distinct aspects of the distribution, i.e., differences in mean and in detection. Canonical, differential expression analyses were performed using the popular edgeR package. To additionally test for differential detection in scRNA-seq data multiple techniques were implemented and compared. The outcomes of the differential expression and differential detection analysis were then integrated by using a stage-wise testing procedure, to gain more statistical power and yield a more comprehensive understanding of the biological mechanisms involved. Furthermore, an additional aim of the study was to address the problem of the presence of pseudo-replicates present in scRNA-seq data by aggregating single-cells for each cell type and patient combination, thus creating pseudo-bulk data.

Various methods were presented for assessing differential detection at both the single-cell and pseudo-bulk levels. To assess differential detection, we first binarise the count data, and given the binary nature of the data, it is logical to consider applying a logistic regression. However, as discussed below, the logistic model alone did not account for certain aspects of the data, which led to the proposal of modifications. These included the use of a quasi-likelihood approach that accounts for over-dispersion, the inclusion of an offset term that acts as a normalization factor, and adopting a quasi-binomial model with a shrunk dispersion parameter; this last modification allows for borrowing information between genes and shrinks the dispersion parameter towards a common value using an empirical Bayesian approach. In addition, two modifications of the edgeR method were implemented allowing to fit a quasi-Poisson and quasi-negative binomial to the data.

To evaluate the performance of each method, a simulation study was carried out. The simulated data was based on a real case study conducted by Perez et al. (2022) on lupus. A differential detection analysis was performed on two datasets: one in which it was assumed that there was no differential expression, and another in which 5% of genes were assumed to be differentially expressed. By comparing the results obtained from the simulated data to the ground truth, the performance of each method was assessed.

In order to investigate the need of accounting for the presence of pseudo-replicates, the analysis was conducted both on single-cell level and on pseudo-bulk level. The results of the simulation study show that not accounting for the presence of pseudo-replicates leads to an inability to control type 1 error thus leading to a high presence of false positive genes. Additionally, accounting for over-dispersion using quasi-likelihood further improved controlling the type 1 error. This is probably due to the fact that the data is over-dispersed and not accounting for this aspect lead to an incorrect statistical inference and the model will not be able to capture all the variability in the data, thus leading to inaccurate predictions. Moreover, the insertion of an offset parameter, which serves as a normalization factor, helps both in terms of FDP and TPP. These results therefore underline the need to account for differences in sequencing depth. The top performing methods in terms of TPP and FDP are the quasi-binomial with an offset and the quasi-Poisson both adapted on pseudo-bulk level.

In the simulation study, the usage of a stage-wise testing paradigm did not improve the performance; we hypothesise that this is due to the fact that the simulation strategy introduced a very strong signal, thus inducing both differences in mean and in detection. In the case study however, using stage-wise testing did seem to provide further biological information given that the differential detection analysis provided complementary information to the differential expression analysis.

In order to further investigate the performance of each method it would be useful to conduct the analysis also on an imbalanced data-set where the imbalance can be both between the number of patients in each mock group and in the number of cells per patient. This analysis is more challenging so it will highlight further problems and probably major differences between the performance of the different strategies to test for differential detection.

Another aspect that should be taken into consideration in order to improve the results is the development of a different simulation strategy. In this thesis, data were simulated by simply swapping the original counts between two genes in one of the treatment arms. However, given the sparsity of the data, this may induce limited changes for some genes. Indeed, if the counts of the two genes for which the counts are being swapped between samples of groups a and b are primarily zeros, this will result in only zeros but that gene will still be flagged as differentially expressed.

In conclusion, this study highlights the importance of accounting for pseudo-replication and over-dispersion in scRNA-seq data analysis. The study recommends the use of either a quasi-Poisson model or a quasi-binomial model with an offset to assess differential detection, as these were found to have the best balance of true positive rate and false discovery rate. Furthermore, the study demonstrates the potential benefits of a stage-wise testing procedure in scRNA-seq data analysis, which allows for examining differences in mean and detection while controlling the false discovery rate at the gene level. This approach provides a more comprehensive analysis of the differences in gene expression between healthy individuals and those with lupus, and can facilitate a deeper understanding of the biological mechanisms underlying the disease. Overall, this master dissertation provides valuable insights and recommendations for scRNA-seq data analysis, which can help to improve the accuracy and robustness of differential gene expression analysis in complex diseases such as lupus.

Bibliography

Bouland, Gerard A, Ahmed Mahfouz, and Marcel J T Reinders. 2021. “Differential analysis of binarized single-cell RNA sequencing data captures biological variation.” *NAR Genomics and Bioinformatics* 3 (4). ISSN: 2631-9268.

Carter, Erin E, Susan G Barr, and Ann E Clarke. 2016. “The global burden of SLE: prevalence, health disparities and socioeconomic impact.” *Nature Reviews Rheumatology* 12 (10): 605–620. ISSN: 1759-4804.

Caza, Tiffany, and Steve Landas. 2015. “Functional and phenotypic plasticity of CD4+ T cell subsets.” *BioMed Research International* 2015.

Cox, David Roxbee, and Nancy Reid. 1987. “Parameter orthogonality and approximate conditional inference.” *Journal of the Royal Statistical Society: Series B (Methodological)* 49 (1): 1–18.

Crowell, Helena L, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra, and Mark D Robinson. 2020. “Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data.” *Nature Communications* 11 (1): 6077.

De Duve, Christian. 2002. *Life evolving: Molecules, mind, and meaning.* 114. Oxford University Press on Demand.

Ducloux, Didier, Cécile Courivaud, Jamal Bamoulid, Bérengère Vivet, Aline Chabroux, Marina Deschamps, Jean-Michel Rebibou, Christophe Ferrand, Jean-Marc Chalopin, Pierre Tiberghien, et al. 2010. "Prolonged CD4 T cell lymphopenia increases morbidity and mortality after renal transplantation." *Journal of the American Society of Nephrology* 21 (5): 868–875.

Durruthy-Durruthy, Robert, and Manisha Ray. 2018. "Using Fluidigm C1 to generate single-cell full-length cDNA libraries for mRNA sequencing." *Disease Gene Identification: Methods and Protocols,* 199–221.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

George, Edward I, George Casella, et al. 1989. "Empirical Bayes confidence estimation."

Kang, Hyun Min, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. 2018. "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation." *Nature Biotechnology* 36 (1): 89–94.

Kivioja, Teemu, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. 2012. "Counting absolute numbers of molecules using unique molecular identifiers." *Nature Methods* 9 (1): 72–74. ISSN: 1548-7105.

Klein, JAN, and Akie Sato. 2000. “The HLA system.” *New England Journal of Medicine* 343 (10): 702–709.

Lönnberg, Tapio, Valentine Svensson, Kylie R James, Daniel Fernandez-Ruiz, Ismail Sebina, Ruddy Montandon, Megan SF Soon, Lily G Fogg, Arya Sheela Nair, Urijah N Liligeto, et al. 2017. “Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria.” *Science Immunology* 2 (9).

Lund, Steven P, Dan Nettleton, Davis J McCarthy, and Gordon K Smyth. 2012. “Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates.” *Statistical Applications in Genetics and Molecular Biology* 11 (5).

Macosko, Evan Z, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. 2015. “Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets.” *Cell* 161 (5): 1202–1214.

Malfait, Milan. 2022. “swapper: Simulate DE Signal By Feature Swapping.” Https://github.com/milanmlft/swapper.

McGinnis, Christopher S, Lyndsay M Murrow, and Zev J Gartner. 2019. “DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors.” *Cell Systems* 8 (4): 329–337.

Perez, Richard K, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C Hartoularos, Sasha Targ, Yang Sun, et al. 2022. “Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus.” *Science* 376 (6589).

Picelli, Simone, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. 2014. “Full-length RNA-seq from single cells using Smart-seq2.” *Nature Protocols* 9 (1): 171–181.

Qiu, Peng. 2020. “Embracing the dropouts in single-cell RNA-seq analysis.” *Nature Communications* 11 (1): 1–9. ISSN: 2041-1723.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” *Bioinformatics* 26 (1): 139–140. ISSN: 1367-4803.

Robinson, Mark D, and Alicia Oshlack. 2010. “A scaling normalization method for differential expression analysis of RNA-seq data.” *Genome Biology* 11 (3): 1–9.

Rönnblom, Lars, and V Pascual. 2008. “The innate immune system in SLE: type I interferons and dendritic cells.” *Lupus* 17 (5): 394–399.

Sarkar, Abhishek, and Matthew Stephens. 2021. “Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis.” *Nature Genetics* 53 (6): 770–777. ISSN: 1546-1718.

Smyth, Gordon K. 2004. “Linear models and empirical bayes methods for assessing differential expression in microarray experiments.” *Statistical Applications in Genetics and Molecular Biology* 3 (1).

Squair, Jordan W, Matthieu Gautier, Claudia Kathe, Mark A Anderson, Nicholas D James, Thomas H Hutson, Rémi Hudelle, et al. 2021. "Confronting false discoveries in single-cell differential expression." *Nature Communications* 12 (1): 5692. ISSN: 2041-1723.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. 2005. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences* 102 (43): 15545–15550.

Tjur, Tue. 1998. "Nonlinear regression, quasi likelihood, and overdispersion in generalized linear models." *The American Statistician* 52 (3): 222–227.

Van den Berge, Koen, Katharina M Hembach, Charlotte Soneson, Simone Tiberi, Lieven Clement, Michael I Love, Rob Patro, and Mark D Robinson. 2019. "RNA sequencing data: hitchhiker's guide to expression analysis." *Annual Review of Biomedical Data Science* 2:139–173.

Van den Berge, Koen, Charlotte Soneson, Mark D Robinson, and Lieven Clement. 2017. "stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage." *Genome Biology* 18 (1): 1–14.

Weisenfeld, Neil I, Vijay Kumar, Preyas Shah, Deanna M Church, and David B Jaffe. 2017. "Direct determination of diploid genome sequences." *Genome Research* 27 (5): 757–767.

Wilson, Daniel J. 2019. “The harmonic mean p-value for combining dependent tests.” *Proceedings of the National Academy of Sciences* 116 (4): 1195–1200.

Young, Michael, and Raif S Geha. 1986. “Human regulatory T-cell subsets.” *Annual Review of Medicine* 37 (1): 165–172.

Zimmerman, Kip D, Mark A Espeland, and Carl D Langefeld. 2021. “A practical solution to pseudoreplication bias in single-cell studies.” *Nature Communications* 12 (1): 738. ISSN: 2041-1723.

Appendix

4.6 Simulation study

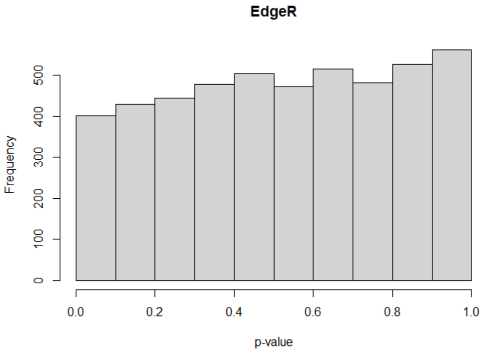

Figure 4.4: p-values of the differential expression analysis conducted with edgeR on the 10 vs 10 patients comparison in which no differential expression in expected.

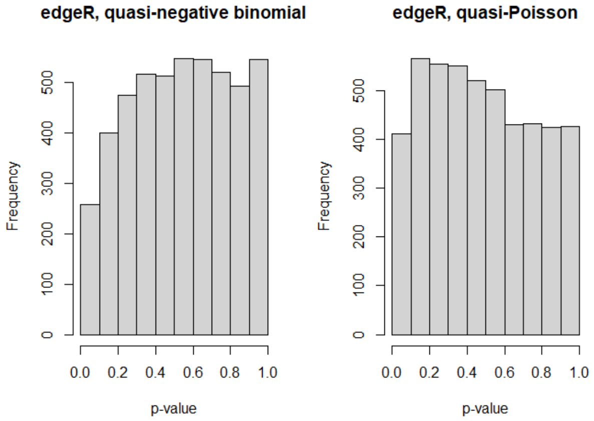

Figure 4.5: p-values of the differential detection analysis conducted with Quasi-NB and quasi-Poisson on the aggregated counts of the 10 vs 10 patients comparison in which no differential expression is expected.

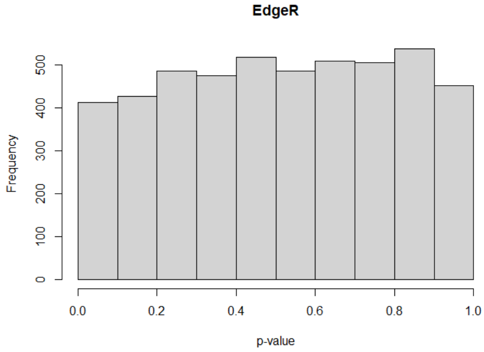

Figure 4.6: p-values of the differential expression analysis conducted with edgeR on the 5 vs 5 patients comparison in which no differential expression is expected.

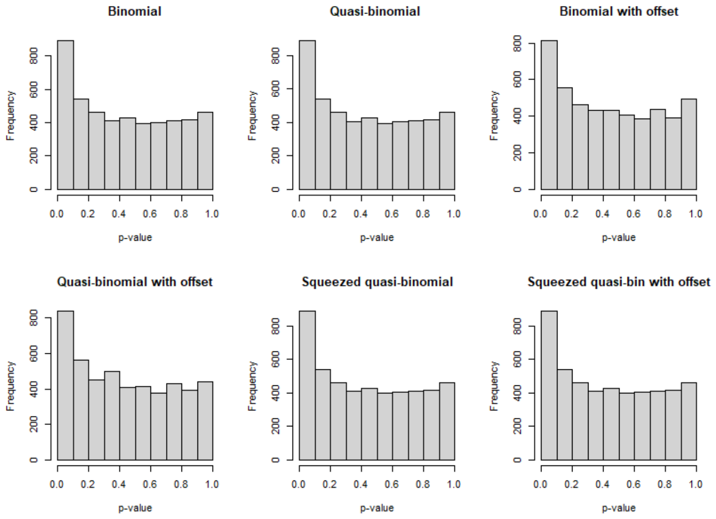

Figure 4.7: p-values of the differential detection analysis conducted on the non aggregated counts of the 5 vs 5 patients comparison in which no differential expression is expected.

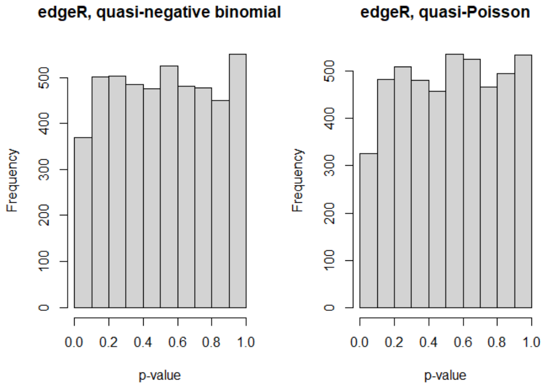

Figure 4.8: p-values of the differential detection analysis conducted with Quasi-NB and quasi-Poisson on the aggregated counts of the 5 vs 5 patients comparison in which no differential expression in expected.

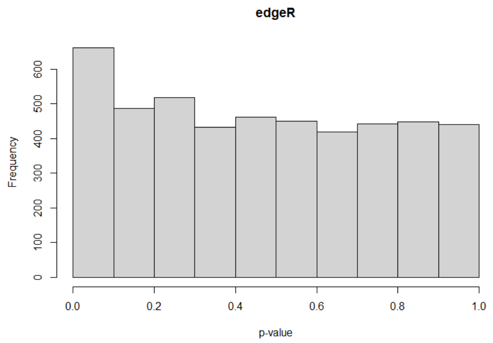

Figure 4.9: p-values of the differential expression analysis conducted with edgeR on the 10 vs 10 patients comparison in which differential expression in expected in 5% of genes.

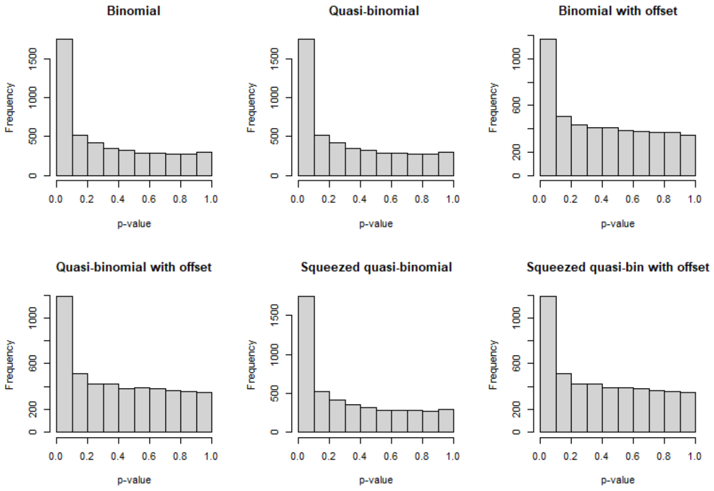

Figure 4.10: p-values of the differential detection analysis conducted on the non aggregated counts of the 10 vs 10 patients comparison in which differential expression in expected in 5% of genes.

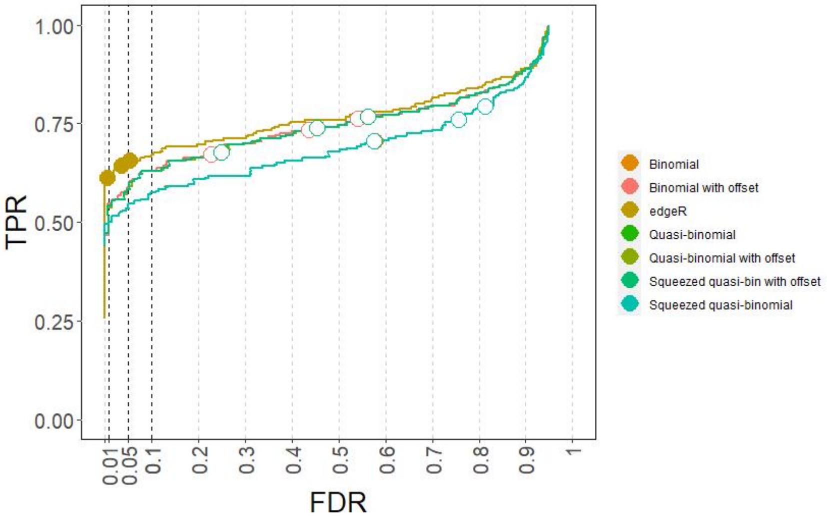

Figure 4.11: Performance evaluation of the methods with false discovery proportion and true positive proportion curves. Differential detection results applied to non aggregated data and differential expression assessed with edgeR are compared in the 10 vs 10 patients comparison.

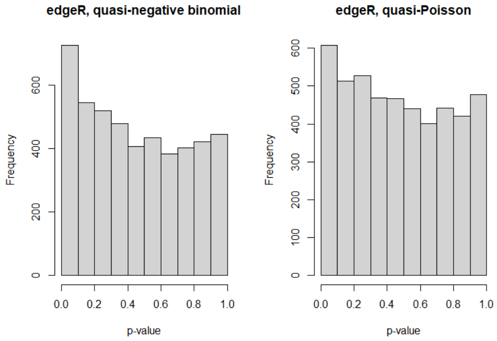

Figure 4.12: p-values of the differential detection analysis conducted with Quasi-NB and quasi-Poisson on the aggregated counts of the 10 vs 10 patients comparison in which no differential expression in expected.

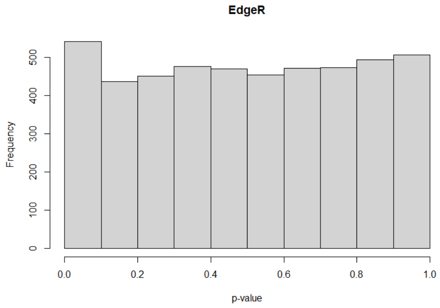

Figure 4.13: p-values of the differential expression analysis conducted with edgeR on the 5 vs 5 patients comparison in which differential expression in expected in 5% of genes.

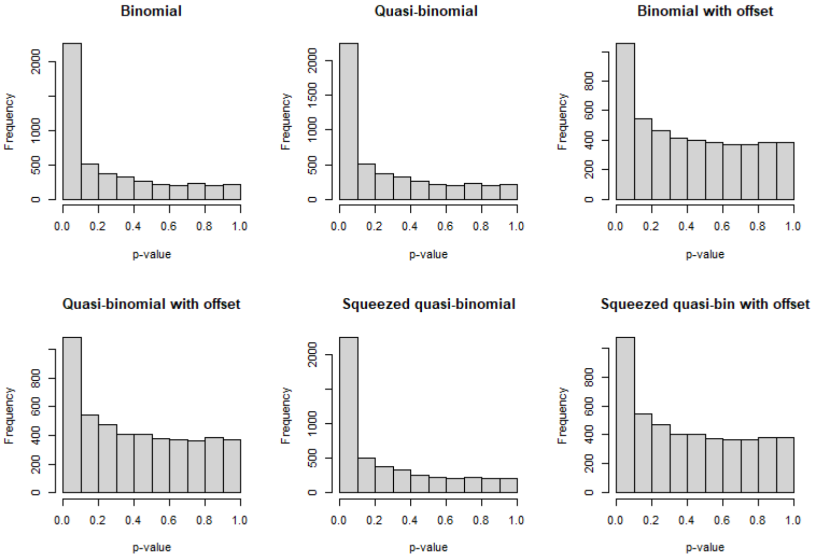

Figure 4.14: p-values of the differential detection analysis conducted on the non aggregated counts of the 5 vs 5 patients comparison in which differential expression in expected in 5% of genes.

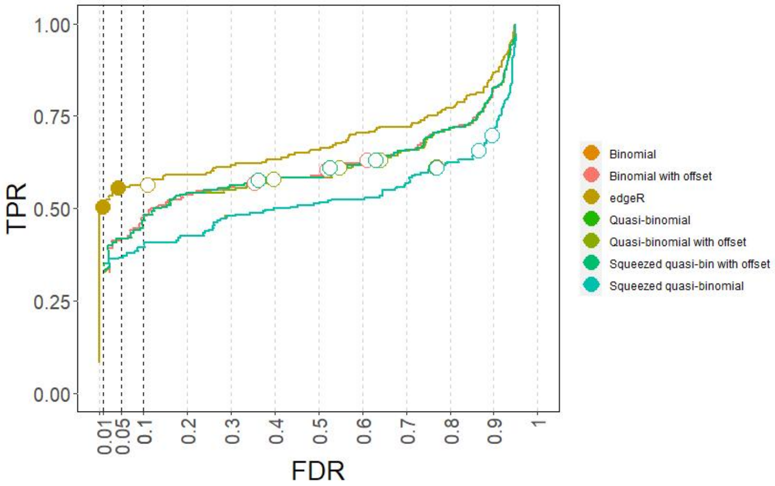

Figure 4.15: Performance evaluation of the methods with false discovery proportion and true positive proportion curves. Differential detection results applied to non aggregated data and differential expression assessed with edgeR are compared in the 5 vs 5 patients comparison.

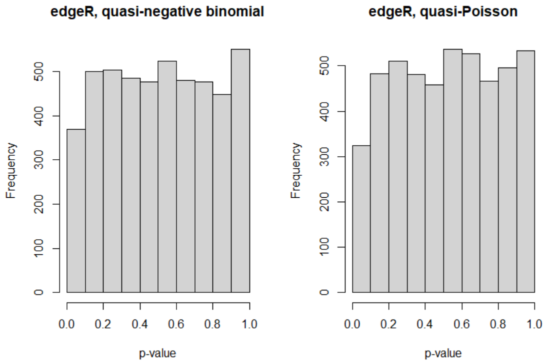

Figure 4.16: p-values of the differential detection analysis conducted with Quasi-NB and quasi-Poisson on the aggregated counts of the 5 vs 5 patients comparison in which no differential expression in expected.

4.6.1 Stage-wise analysis

Stage-wise analysis was then performed on the 10vs10 data-set; initially differential detection was assessed with the quasi-binomial adapted on single-cell level. The screening stage flagged 574 genes, 566 of which were differentially detected and 178 differentially expressed. It is clear how, with low sample size, models fitted at single-cell level have a huge amount of false positive genes as can be seen in Figure 4.19. Also in this case there are some differentially expressed genes that are not picked up by any method this is, as previously explained, because only zero counts were swapped. FDP-TPP curves were used to assess the method's power and it can be seen that the edgeR analysis clearly has higher TPP and better false discovery proportions. Differential detection was then investigated by adapting a squeezed quasi-binomial with an offset on the binarized aggregated counts. The stage-wise analysis picked-up 163 genes, 144 of which are differentially detected and 160 differentially expressed. Aggregation helps contain the number of false positive genes, as can be seen in Figure 4.19. Figure 4.20 shows how stage-wise analysis does not add any power since it again performs like the edgeR analysis.

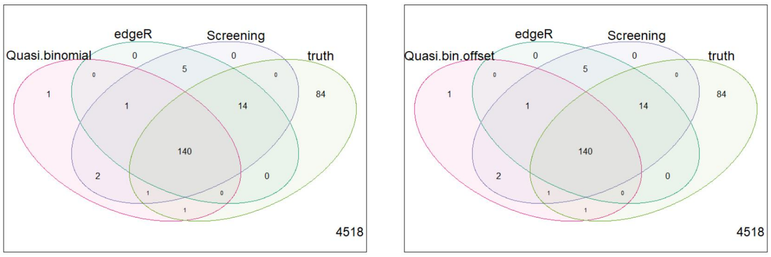

Figure 4.17: The graphs show Venn diagrams in which the overlap of the methods analysed on the 10 vs 10 comparison are shown. In the left panel the overlap between differential detection assessed with the quasi-binomial on non aggregated data, differential expression tested with edgeR, the screening results and the real DE genes is shown. In the right panel differential detection is instead assessed with the quasi-binomial with an offset adapted on pseudo-bulk level.

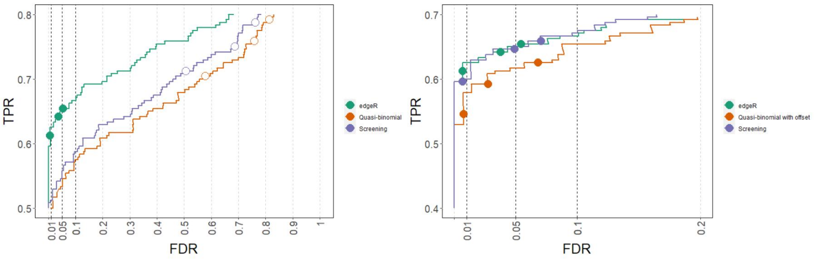

Figure 4.18: Performance evaluation of the methods with false discovery proportion and true positive proportion curves in the 10 vs 10 comparison . The left panel compares the screening result with differential expression assessed with edgeR and differential detection tested with a quasi-binomial adapted on single-cell level. The right panel instead used a quasi-binomial with an offset adapted on pseudo-bulk level to assess differential detection.

In the 5vs5 data-set the stage wise-analysis was first conducted by using the quasi-binomial with an offset to evaluate differential detection on single-cell level. The screening stage picked-up 292 genes, 286 of which resulted as differentially detected and 149 differentially expressed. Just like in the analysis conducted on larger sample sizes, the single-cell level analysis has a large amount of false positive genes. The FDP-TPP curves show how the edgeR analysis over performs both the screening methods and the quasi-binomial with an offset. To evaluate differential detection on the binarized aggregated matrix, a quasi-Poisson model was used. In this case the screening stage signaled 138 genes, 115 of which are differentially detected and 137 differentially expressed. In this case aggregation did not help reduce the number of false positive genes detected by the quasi-Poisson. Also in this case the screening method does not help to gain power and performs as the edgeR model.

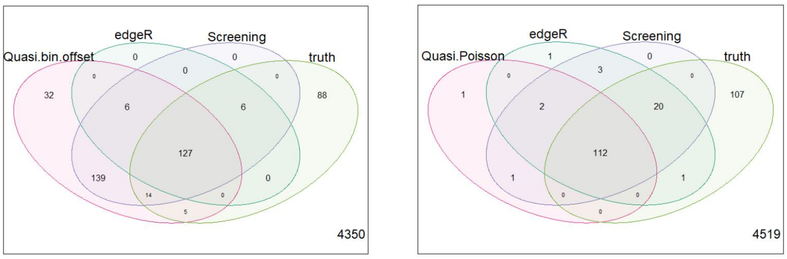

Figure 4.19: The graphs show Venn diagrams in which the overlap of the methods analysed on the 5 vs 5 comparison are shown. In the left panel the overlap between differential detection assessed with the quasi-binomial with an offset on non aggregated data, differential expression tested with edgeR, the screening results and the real DE genes is shown. In the right panel differential detection is instead assessed with the quasi-Poisson adapted on pseudo-bulk level.

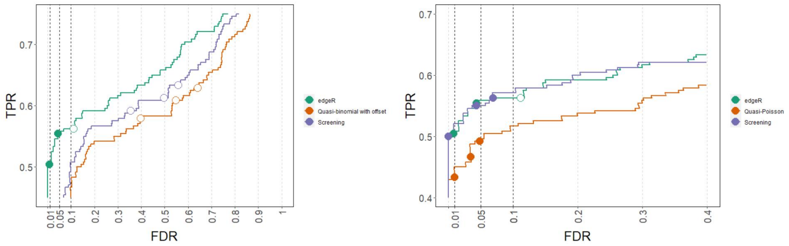

Figure 4.20: Performance evaluation of the methods with false discovery proportion and true positive proportion curves in the 5 vs 5 comparison . The left panel compares the screening result with differential expression assessed with edgeR and differential detection tested with a quasi-binomial with an offset adapted on single-cell level. The right panel instead used a quasi-Poisson adapted on pseudo-bulk level to assess differential detection.