THE UNIVERSITY OF CHICAGO

STATISTICAL METHODS FOR SINGLE-CELL RNA DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

CHIH-HSUAN WU

CHICAGO, ILLINOIS
AUGUST 2024

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to my advisor, Mengjie Chen, for her unwavering support and guidance throughout my journey in the field of statistical methodology for single-cell RNA data. Her expertise and dedication have been instrumental in building my knowledge and skills. During the challenging times of the COVID-19 pandemic, when my research progress was hindered, she encouraged me and helped me overcome my frustration. I have learned a great deal from her, not only from her vast knowledge and experience but also from her positive attitude and resilience. Without her, I would not have completed this work. Thank you, Mengjie, for being an inspiring mentor and a steadfast source of support. I also thank Mary Sara McPeek and Jingshu Wang for participating in my candidate proposal presentation and dissertation defense. Their insightful feedback and constructive criticism greatly contributed to the improvement and refinement of my research.

I thank all the friends I met at UChicago. The company of my Taiwanese friends on countless Saturday nights made me feel at home and provided warmth and comfort while studying abroad. I am also grateful to my PhD cohorts for their support and encouragement during tough times. In particular, I want to thank You-Lin Chen, with whom I spent the most time. His support during the low points in my life was invaluable, and he broadened my horizons both in my career and through our many adventures. Thank you all for making this journey memorable and enriching.

I would like to express my deepest gratitude to my friends and family in Taiwan. I thank my parents, Su-Min Lin and Chun-Chuan Wu, for encouraging me to explore the world and for their unwavering love and care. Special thanks to my mom, who has held our family together and been my life mentor. I also thank my sister, my brother, and my parents' friends for taking care of my parents and bringing them joy. Additionally, I am grateful to Zou-An Lai and Kai-Qi Xiong for sharing in my happiness and sadness throughout my PhD journey. Your support has been invaluable to me.

ABSTRACT

Recent advancements in single-cell RNA sequencing (scRNA-seq) have revolutionized transcriptomic research by enabling the study of gene expression at unprecedented resolution, revealing intricate cellular heterogeneity and dynamics. Despite these advancements, analyzing scRNA-seq data poses significant challenges, including normalization biases, excessive zeros, and donor effects, which can confound differential expression (DE) analysis. This thesis addresses these challenges through innovative methodologies.

Chapter 2 critically evaluates existing DE analysis methods, highlighting limitations such as normalization biases and the impact of excessive zeros on statistical models. To overcome these issues, we introduce a novel paradigm using a generalized linear mixed model (GLMM) that leverages raw unique molecular identifier (UMI) counts for robust DE analysis.

In Chapter 3, we adopt the HIPPO framework, a clustering algorithm that prioritizes zero proportion as a primary indicator. We examine the potential of higher order counts (proportions) to extract additional information beyond zero proportion, aiming to enhance the HIPPO algorithm. To achieve this, we introduce the k-inflation test for identifying k-inflated genes and develop a Poisson proportion t-test for further analysis.

Chapter 4 shifts focus to cell-type specific differential methylation analysis, recognizing RNA methylation, particularly N6-methyladenosine (m6A), as pivotal in RNA regulation. Building on the RADAR framework, a novel methodology is proposed using a mixture Poisson GLMM to analyze methylation data integrated with scRNA-seq. This approach utilizes cellular composition estimates to uncover differential methylation patterns across cell types without direct measurement of cell-type specific methylation.

This dissertation proposes novel frameworks for DE and methylation analysis in scRNA-seq data, enhancing our understanding of cellular diversity and gene regulation. These methodologies contribute to advance biological research and pave the way for new discoveries in precision medicine and therapeutic development.

CHAPTER 1

INTRODUCTION

1.1 Technological advancements in single cell RNA sequencing

Over the past decade, the field of transcriptomics has been revolutionized by the emergence of single-cell RNA sequencing (scRNA-seq) technologies. These advancements have enabled researchers to analyze and break down cellular heterogeneity with unprecedented resolution, revealing the complex landscape of gene expression within individual cells. Unlike traditional bulk RNA sequencing, which averages gene expression profiles across thousands of cells, scRNA-seq allows for the examination of transcriptomic variations at the single-cell level. This capability is crucial for understanding diverse biological processes, such as development, differentiation, and disease progression, where cell-to-cell variability plays a critical role.

The transition from bulk RNA sequencing to single-cell RNA sequencing marks a significant leap in the resolution and depth of transcriptomic analysis. Bulk RNA sequencing methods, while powerful, provide only an average expression profile across a population of cells, which may mask the underlying heterogeneity and obscure the contributions of individual cell types. This limitation is particularly problematic in tissues composed of diverse cell populations, where significant differences between cell types can drive biological functions and disease mechanisms. In contrast, scRNA-seq allows for obtaining detailed information of these complex tissues, enabling the identification of distinct cell types and states. This detailed analysis not only improves our understanding of cellular diversity but also helps find new biomarkers and therapeutic targets.

1.2 Differential expression in different cell types

One of the key applications of scRNA-seq is the identification of differentially expressed genes (DEGs) across different cell types or conditions. Differential expression analysis helps

to uncover the molecular signatures that define specific cell populations and their functional states.

However, traditional differential expression methods are not well suited for scRNA-seq data. Some methods are adapted from approaches originally designed for bulk RNA sequencing data, which do not utilize the unique characteristics of scRNA-seq. Most methods rely on normalization procedures that may introduce biases. For scRNA-seq data, gene expression levels are likely to have a higher proportion of zeros compared to traditional bulk RNA-seq data, so scRNA-seq data sets have been analyzed mostly through zero-inflated models or normalized procedures that may introduce biases. In state-of-the-art protocols, unique molecular identifiers (UMIs) provide a more accurate measure of transcript abundance by mitigating the effects of amplification biases. Studies have shown that cell-type heterogeneity is the major driver of zeros observed in 10X UMI data. Pre-processing the data with normalization procedures may discard an important feature for differential expression analysis between different cell types.

1.3 Outline

This thesis aims to address several critical aspects of differential expression and methylation analysis in single-cell data.

Chapter 2 discusses the problems and limitations of existing methods for differential expression analysis. We dissect four major challenges in single-cell DE analysis: normalization, excessive zeros, donor effects, and cumulative biases. These challenges highlight the limitations and conceptual pitfalls in current workflows. In response, we propose a novel approach to address several of these issues—a generalized linear mixed model (GLMM) for differential expression analysis in single-cell RNA data. This model utilizes raw UMI counts without normalization, thus preserving the integrity of the original data and avoiding potential biases introduced by normalization procedures. The GLMM framework includes random effects to

account for the inherent variability among donors, offering a robust and flexible approach to identify DEGs.

In Chapter 3, we adopt the HIPPO framework, a clustering algorithm that prioritizes zero proportion as a primary indicator. We explore the potential of higher order counts (proportions) to extract additional information beyond zero proportion, aiming to enhance the HIPPO algorithm. To achieve this, we introduce the k-inflation test for identifying k-inflated genes and develop a Poisson proportion t-test for further analysis.

Chapter 4 explores cell-type specific differential methylation analysis. RNA methylation, particularly N6-methyladenosine (m6A), is a critical post-transcriptional modification influencing RNA metabolism and function. Integrating methylation data with scRNA-seq enhances our understanding of gene regulation at the single-cell level. Building on the RADAR framework, we propose a novel methodology for cell-type specific differential methylation analysis using a mixture Poisson generalized linear mixed model. This approach utilizes estimates of cellular compositions to identify differential methylation patterns without the requirement for direct experimental measurement of cell-type specific methylation.

CHAPTER 2

THE CURSES OF PERFORMING DIFFERENTIAL EXPRESSION ANALYSIS USING SINGLE-CELL DATA

2.1 Introduction

Differential expression (DE) analysis in single-cell transcriptomics provides essential insights into cell-type–specific responses to internal and external stimuli (Saliba et al. [2014], Greenwald et al. [2019], Grubman et al. [2019], Lawlor et al. [2017]). While many methods are available to identify differentially expressed genes from single-cell transcriptomics, recent studies raise important concerns about the performance of state-of-the-art methods, including both methods tailored to single cell data and techniques that work well in bulk (Squair et al. [2021] , Das et al. [2021], Das et al. [2022]). As population-level single-cell studies rapidly become more feasible, powerful and accurate analytical methods will be essential for obtaining meaningful results. In this context, we discuss the four "curses" that currently plague the differential expression analysis of single-cell data: normalization, zeros, donor effects, and cumulative biases, highlighting the various limitations and conceptual flaws in the current workflows. We demonstrate these limitations using real data from 10X single-cell RNA-seq (sRNA-seq) data from post-menopausal fallopian tubes (Lengyel et al. [2022]). Finally, we present a new paradigm that offers a potential solution to some of these issues and illustrate its performance using two case studies.

2.1.1 The curse of normalization

The term "normalization" has been used to denote multiple distinct approaches in genomics (Li et al. [2015], Zyprych-Walczak et al. [2015]). For example, it can refer to the process of correcting PCR amplification biases introduced during sequencing library preparation (library size normalization) (Dillies et al. [2013], Robinson and Oshlack [2010], Lytal et al.

[2020]), the process of harmonizing data across different experimental batches (batch normalization) (Leek et al. [2010], Korsunsky et al. [2019], Chen and Zhou [2017], Chen et al. [2021], Hu et al. [2022], or to the process of transforming the data to adhere to a normal distribution (data distribution normalization) (Schmid et al. [2010]). All three have been introduced to handle both bulk and single cell RNA-seq data, aiming to minimize unwanted technical variations. Choosing appropriate normalization techniques for DE analysis of scRNA-seq data is clearly important to maintain the integrity of the data, but the field has yet to establish a definitive gold standard outlining the circumstances for which different normalizations should be performed.

Library size normalization is critical in bulk RNA-seq analysis, as it is impossible to track the absolute abundance of RNA molecules in typical bulk RNA-seq protocols due to an unknown fold of amplification introduced by PCR during library construction. Normalization, in this instance, focuses on estimating and subsequently correcting for a sample-specific size factor. This process allows bulk RNA-seq to estimate relative RNA abundances. Post-normalization, samples are calibrated against a common reference, resulting in most genes displaying similar expression levels across samples. When performing differential expression analysis with bulk RNA-seq data, genes are classified as either up-regulated or down-regulated, based on the assumption that the majority remain unchanged across groups. While this size-factor based normalization technique is suitable for bulk RNA-seq, it does not translate effectively to scRNA-seq. Protocols in scRNA-seq, such as the 10X, employ unique molecular identifiers (UMIs) which discern between genuine RNA molecules and those generated via PCR. This enables the absolute quantification of RNA levels. Unfortunately, size-factor–based normalization methods, like counts per million reads mapped (CPM) convert data into relative abundances erasing useful data provided by the UMIs. Furthermore, CPM-normalized data does not account for competition among genes for cellular resources because the uniform number of molecules found in CPM-normalized data does not accurately

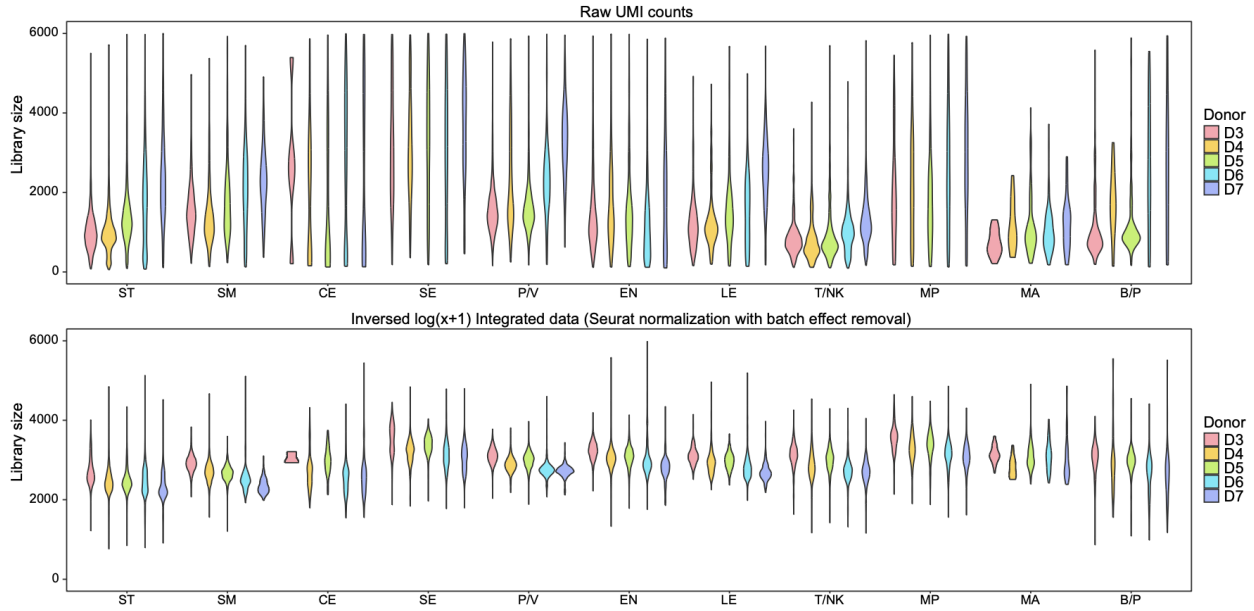
represent true expression levels, which ultimately leads to suboptimal DE analysis results.

In batch effect normalization, dimension reduction methods pinpoint genes with consistent expression patterns across various batches; these genes act as anchors, guiding the alignment and integration of data (Tran et al. [2020]). However, in scRNA-seq analysis, only highly expressed or highly variable genes are retained for estimating batch effects and subsequent integration. As a result, gene numbers in integrated scRNA-seq datasets are noticeably reduced compared to the raw UMI data.

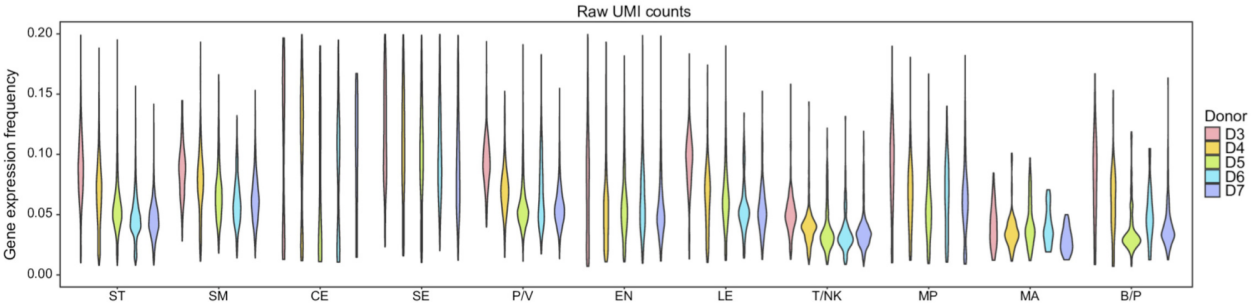
For data distribution normalization, the field offers both straightforward (e.g., log-transformation) and advanced strategies (e.g., variance stabilizing transformation, or VST). A notable implementation for scRNA-seq of VST is sctransform (Hafemeister and Satija [2019]), which employs a regularized negative binomial regression model, preserving the Pearson residuals for future analytical steps, including DE analysis (Lause et al. [2021]). However, if the underlying data distribution deviates significantly from the assumed model, the application of VST may introduce bias into the analysis.

To demonstrate the effects of various normalization methods on single-cell data, we compared the raw UMI counts of 10x scRNA-seq data obtained from post-menopausal fallopian tubes (see Methods) with data normalized using one of three methods: 1) CPM; 2) integrated log-normalized counts after removing batch effects using the Seurat CCA model (Argelaguet et al. [2020]); and 3) VST data using sctransform (Hafemeister and Satija [2019]). As a result, we see the total UMI counts revealed substantial variations in library sizes across different cell types; notably macrophages (MP) and secretory epithelial (SE) cells exhibited significantly higher RNA content than other cell types (Fig. 2.1a). Furthermore, SE cells exhibited larger mean library sizes than mast (MA) cells across all donors. These findings align with the understanding that the main active cell types in post-menopausal fallopian tubes are MP and SE cells, with other cell types remaining dormant post-menopause. However, in the integrated data, the disparities in library size distribution were mitigated, even

within cell types (Fig. 2.1a). While integration reduced differences across donors, it came at the cost of diminishing variation across cell types. It is worth mentioning that CPM normalization equalizes library sizes across all cell types; such normalizations may potentially obscure differences between cell types that are vital for understanding their unique biological functions.



(a)



(b)

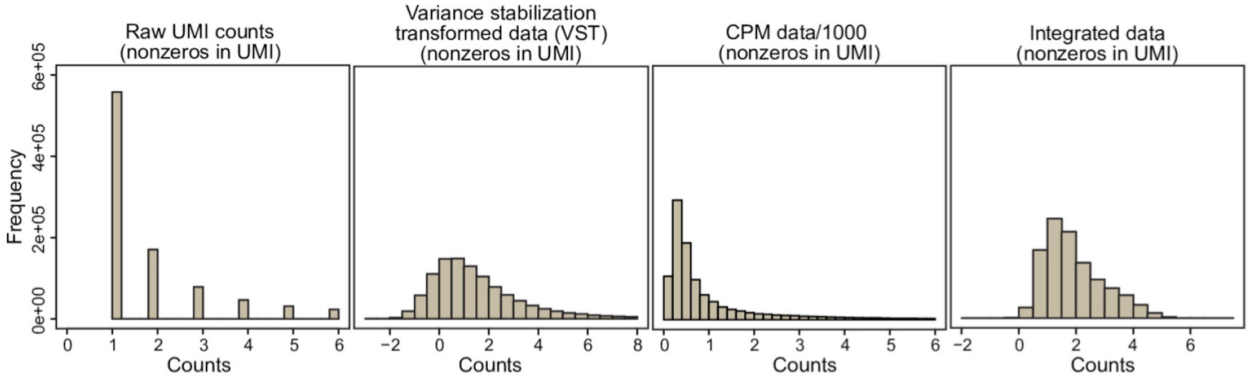
Figure 2.1: Effects of normalization on library size and zero frequency. a) Violin plots display library sizes based on raw UMI counts (top) and after data integration (bottom), categorized by cell types and donors. b) Violin plot illustrating the frequency of gene expression (non-zero counts) in raw UMI data.

2.1.2 The cures of zeros

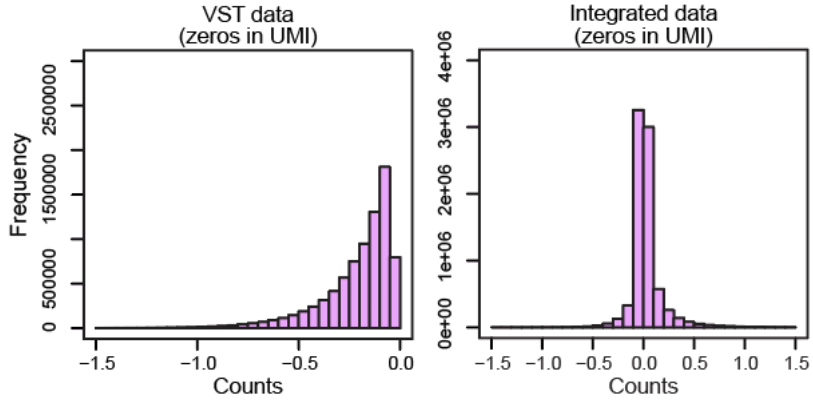
Bulk RNA-seq provides the average transcriptional output of each gene expressed within a population of heterogenous cell types (Wang et al. [2019], Yang et al. [2021]). Even a moderate sequencing depth can yield information about many thousands of different genes. In comparison, scRNA-seq data is much sparser in comparison, with fewer genes expressed per sample and a high proportion of genes with zero UMI counts. Zeros in UMI counts for a gene can arise from three scenarios: a genuine zero, indicating that the gene is not expressed, or a sampled zero, indicating that the gene is expressed at a low level, or a technical zero, indicating that the gene is expressed at a high level but not captured by the assay. Despite an increasing body of evidence suggesting that cell-type heterogeneity is the major driver of zeros observed in 10X UMI data (Kim et al. [2020], Qiu [2020], Svensson [2020]), the prevailing notion within the single-cell community is that zeros are largely uninformative technical artifacts caused by "drop-out" genes (i.e., technical zeros).

Accordingly, many single-cell DE studies include pre-processing steps aimed at removing so-called zero inflation. Several popular pre-processing methods include: 1) performing feature selection by aggressively removing genes based on their zero detection rates, such as requiring non-zero values in at least 10% of total cells and restricting DE analysis to a smaller gene set; 2) imputing zeros and performing DE on imputed values (Gong et al. [2018], Li and Li [2018], Tracy et al. [2019], Chen and Zhou [2018]); or 3) modeling zeros explicitly as an extra component and essentially performing DE on non-zero values only (Pierson and Yau [2015], Finak et al. [2015]).

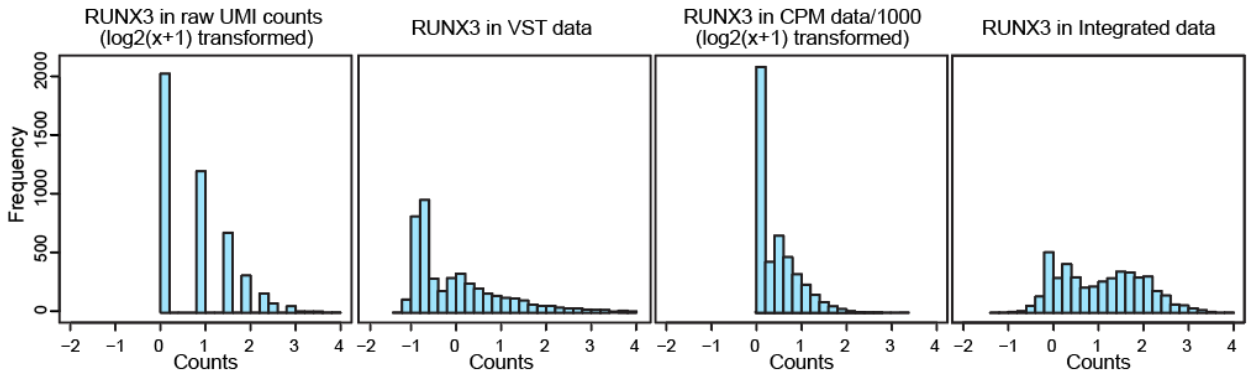
However, if zeros are in fact biological zeros due to no expression or very low expression, dismissing or correcting for zeros in scRNA-seq is equivalent to discarding a significant portion of information in the dataset before any analysis. By failing to account for cell-type heterogeneity, zero-inflation pre-processing steps such as normalization and imputation become inappropriate and can introduce unwanted noise into downstream analyses, includ-



(a)



(b)



(c)

Figure 2.2: Effects of normalization on library size and zero frequency. a) Histograms representing the distribution of non-zero counts in raw UMI data across various data transformations. b) Histograms detailing the zero counts in raw UMI data, comparing VST with integrated data where zeros are imputed or converted to non-zeros. c) Histograms showing the distribution of gene RUNX3 across different data transformations.

ing DE. Ironically, the most desired markers in single-cell DE analysis—e.g., genes that are exclusively expressed in a rare cell type that accounts for less than 5% of the total population—may be obscured by current pre-processing steps for handling zeros.

In the fallopian tube dataset, we observed that distinct cell types display varied gene expression patterns in UMI counts. However, these differences become less apparent in imputed or certain transformed datasets (Fig. 2.1a). Gene expression frequency differs among cell types (Fig. 2.1b). However, normalization processes can substantially alter the distribution of both non-zero UMI (Fig. 2.2a) and zero UMI counts (Fig. 2.2b) counts. For example, while the frequency of genes exponentially decline as raw UMI counts increase, VST data forms a more bell-shaped curve with a mode around 1.5 for non-zero raw UMI counts. Non-zero CPM-normalized data, (scaled by 1000) peaks near 0.2 and is more right-skewed than the VST data. Following batch integration, UMI counts primarily fall below 5 and are not as strongly right-skewed. It is noteworthy that zero UMI counts can be given non-zero values via normalization (except with CPM normalization); for example, zeros in VST data are adjusted to negative values and are left-skewed (Fig. 2.2b). Conversely, the integration process transforms original zeros to values clustered closely around zero. We further examined the distributions of gene expression from one gene. Using the gene RUNX3 as an example (Fig. 2.2c), the distributions in raw UMI counts and CPM data remain right-skewed. In contrast, the VST and integrated data showcase broader, bell-shaped distributions. The handling of zeros in these latter datasets (VST and integrated) intrinsically sets them apart from the former. This variability, combined with shifts in distribution skewness, may raise concerns when performing DE analysis with normalized values.

2.1.3 The curse of donor effects

Recent reviews have highlighted that many single-cell DE analysis methods are susceptible to generating false discoveries (Squair et al. [2021]). This is mainly due to failing to account

for variations between biological replicates, commonly referred to as "donor effects". In single-cell studies, donor effects are always confounded with batch effects since cells from one biological sample are typically processed in the same experimental batch. While single-cell studies that contain multiple samples will perform batch correction as pre-processing, they usually do not correct for donor effects when performing DE tests in the downstream analysis.

One question that arises is whether batch effect correction alone suffices to eliminate donor-related effects. To address this, we investigated the contributions of variations from different sources before and after batch correction. Using the same fallopian tube dataset, we further separated 4553 T/NK cells into 20 subtypes using HIPPO (Kim et al. [2020]) (Fig. 2.3a, A.1). With the aid of canonical markers, we identified specific subtypes, including NK, CD4+ T, CD8+ T and mature naive T cells. We then focused on subtypes that were observed in all donors (Fig. 2.3bc).

To quantify the proportion of variation originating from different sources, we fit a linear model, using cell types and donors as covariates, for each gene in several subtype pairs. Through all pairs, the integration led to a reduction in donor variation (Fig. 2.3d, A.2). However, in comparisons of two subtypes of the same cell type (12 vs.13) and two subtypes of different cell types (13 vs. 19), we observed a decrease in the proportion of cell-type–related variation. This underscores that integration not only mitigates batch effects but also impacts the phenotypes of interest. Importantly, our analysis indicated that even after implementing batch correction, a notable percentage of genes still exhibited donor-related effects (Fig. 2.3e). As batch effects are often estimated from leading principal components, representing a consensus from a subset of genes, it is quite possible that residual donor effects persist on some, if not all, genes. Therefore, it is crucial to account for donor effects when performing DE tests to avoid false discoveries and obtain accurate results, even after removing batch effects.

One popular solution to address the issue of donor effects in single-cell studies is the

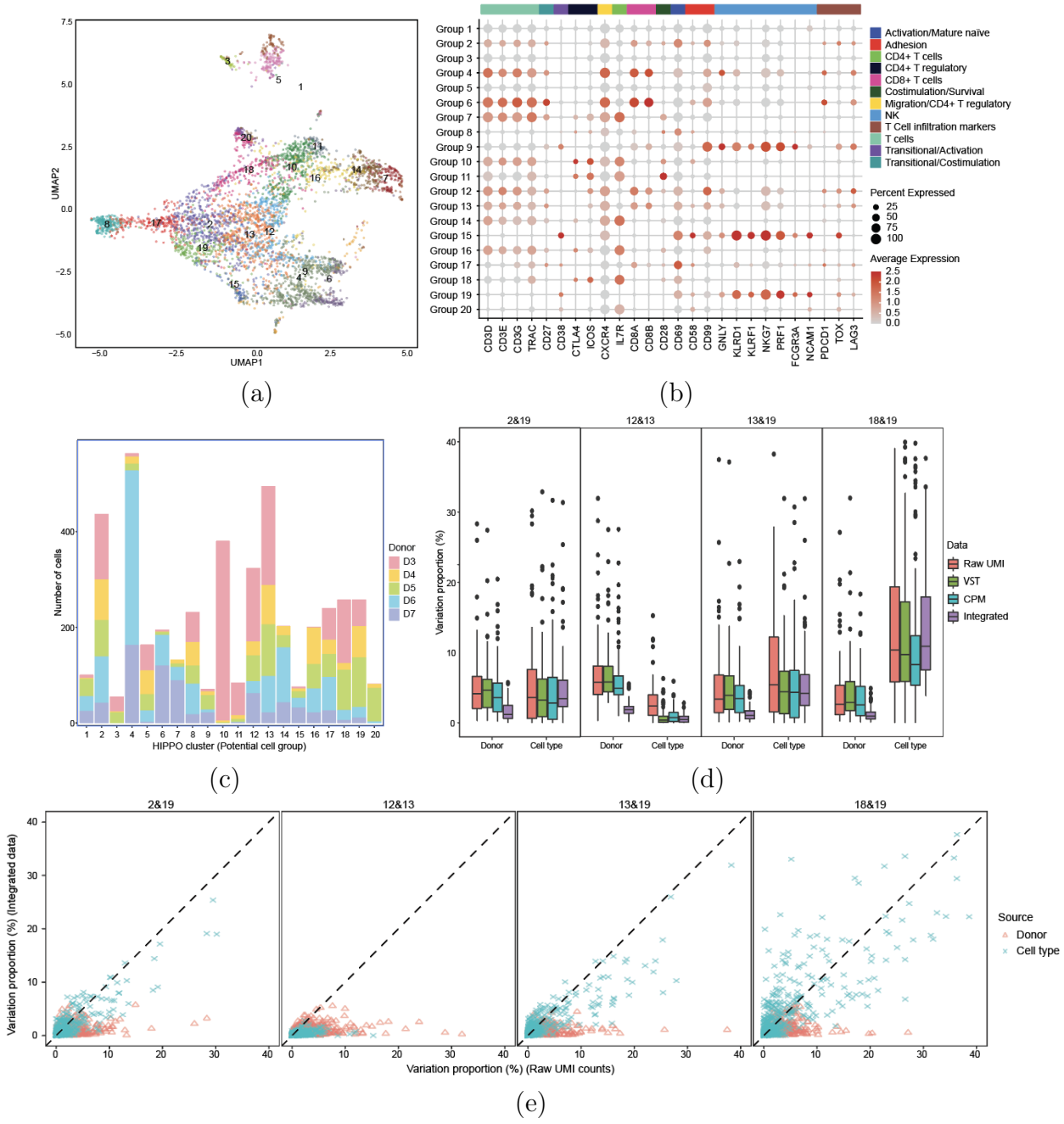


Figure 2.3: Cluster and Variation Analysis of Single-Cell Data from the Fallopian Tube in Case Study 1. a) UMAP visualizing 20 clusters identified by HIPPO in case study 1. b) Canonical markers delineate specific cell subtypes: clusters 9, 15, and 19 as NK cells; clusters 7, 10, 11, 14, 16, 18, and 20 as CD4+ T cells; clusters 4, 6, 12, and 13 as CD8+ T cells; clusters 8 and 17 as mature naive T cells. c) Distribution of donors across the 20 identified clusters. d) Comparative analysis of variation proportions attributable to donor and cell type effects across different pairs and datasets. e) Scatter plots comparing variation proportions due to donor and cell type effects across various pairings and data sources.

use of pseudo-bulk analysis. This approach involves merging cells from the same donor and treating the resulting data as bulk RNA-seq. DE analysis is then performed using tools such as DESeq2 (Love et al. [2014]) or edgeR (Robinson et al. [2010]). However, this framework ignores within-sample heterogeneity by treating donor effects as a fixed effect and assumes that each cell from the same donor is equally affected. As a result, this type of analysis can be overly conservative and potentially lead to missed discoveries (Squair et al. [2021]). Moreover, bulk RNA-seq DE tools typically perform normalization by default, which may have the same drawbacks mentioned earlier in the context of single-cell studies. Thus, caution is advised when using pseudo-bulk analysis as it may not always provide an accurate solution to the problem of donor effects in single-cell studies.

2.1.4 The curse of cumulative biases

In scRNA-seq analysis, it is common to follow a hierarchical, sequential workflow for clustering and DE analysis. This approach can carry forward biases from one step to the next, from batch correction through to normalization, imputation, and feature selection. Such cumulative biases can ultimately diminish the power to detect differentially expressed genes.

Unsupervised learning, especially clustering analysis, is essential in single-cell studies. It groups cells based on gene expression patterns, facilitating the cell-type annotation. While clustering is effective with normalized values like CPMs, it essentially reweights gene features based on their relative contributions. As a result, clustering provides a generalized perspective of variation in gene expression across cell types. The reliance on relative expression also makes clustering fairly resilient to errors and biases introduced by the pre-processing steps.

On the other hand, DE analysis operates at the gene level, using group labels from the clustering process. The effects of biases, whether from donors or batch processing, can vary for each gene. Although DE analysis technically follows clustering—given its reliance on group labels—the metrics used do not need to be identical for both. As we show later in the

case studies with data that complete clustering and annotation successfully, if DE analysis is performed using processed expression levels, the cumulative biases can still lead to false discoveries or overlook of certain DEs.

2.1.5 An alternative paradigm - mixed effects model on UMI counts

To minimize the pre-processing biases discussed above, we proposed an approach that conducts DE analysis on raw UMI counts prior to implementing batch correction, normalization, imputation, or feature selection. This approach, which uses a generalized linear mixed model (GLMM) (Clayton [1996]), preserves sample-specific structures and biological signals in the data. Furthermore, our proposed approach can adjust for any potential confounding factors, such as batch, age, sex, or ancestry, by incorporating them as covariates with fixed effects. This framework enables us to explicitly account for the variation among biological replicates in comparison to other effects (Fig. 2.4). The proposed procedures have been implemented in software LEMUR (`https://github.com/C-HW/LEMUR`).

Unlike existing packages that utilize GLMMs, such as Muscat (Crowell et al. [2020]), LEMUR treats group-of-interest as a fixed effect while accounting for donor-specific variations as random effects. In contrast, Muscat assigns a random effect term for each combination of donor and group-of-interest. Muscat's approach treats certain aspects of group-of-interest variability as random effects, potentially masking differences between groups. Furthermore, Muscat's GLMMs use library size as an offset to normalize counts, essentially focusing on relative abundance rather than raw counts. Overall, Muscat's GLMMs operate similarly to pseudo-bulk methods, grouping counts within the same donor before performing the normalization, which results in comparable performance, as demonstrated in the later examples.

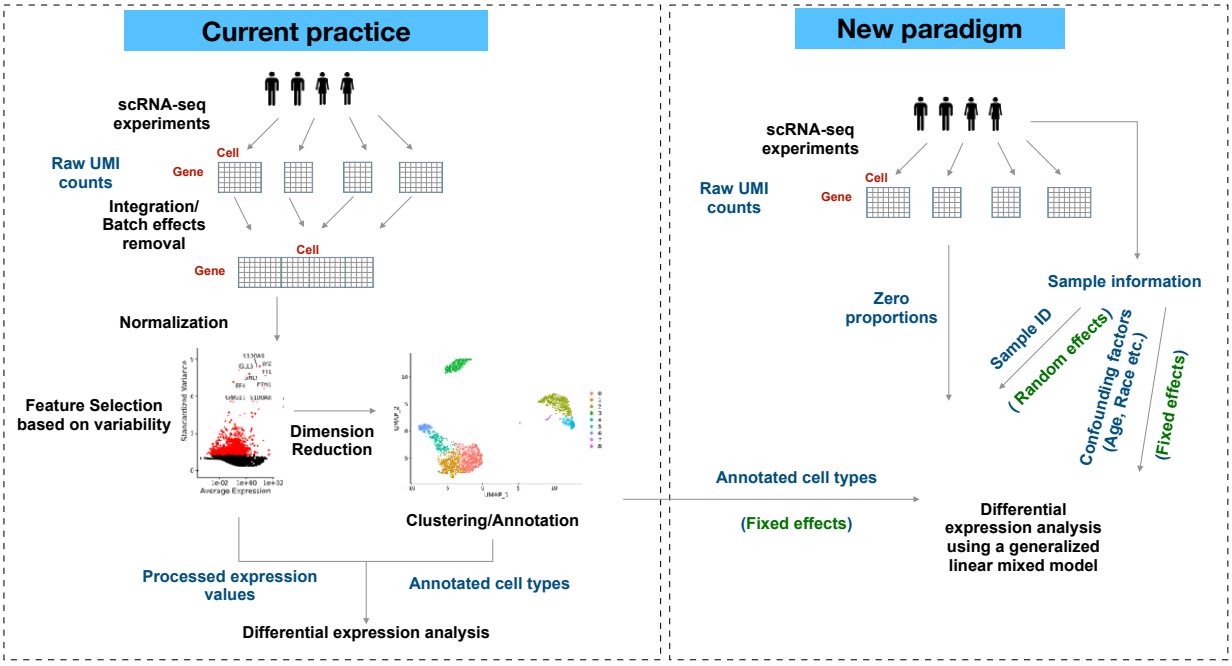


Figure 2.4: Comparison of established workflows and proposed paradigm for single-cell analysis. Left: Under the current single-cell analysis pipeline, the raw UMI counts collected from multiple donors are integrated to remove the batch effects and normalized for further analysis. It is common to perform DE analysis on processed data. Right: Our new paradigm directly performs a generalized linear mixed model on raw UMI counts. The random effect can account for the batch effect due to samples. The annotated cell types can be obtained from existing pipeline or HIPPO algorithm which clusters cells based on the zero proportions of UMI counts.

2.2 Results

To benchmark the performance of our new paradigm, we implemented eight distinct methods for DE analysis: two new paradigm methods, Poisson-glmm and Binomial-glmm; two traditional pseudo-bulk methods DESeq2 and edgeR; and four existing single-cell–specific methods, MAST34, Wilcox in Seurat, and two Muscat GLMMs (MMvst and MMpoisson).

Binomial-glmm fits a GLMM model on the zero proportion of each gene, adding donors as random effect. Pseudo-bulk DESeq2 applies both VST and library size normalization. EdgeR applies library size normalization. MAST adopts a zero-inflated negative binomial model, using log-transformed CPM counts and incorporating the cellular detection rate as covariates. The Wilcox test is non-parametric, using integrated normalized counts. The two Muscat models, MMvst with VST counts and MMpoisson with raw UMI counts, account for library size. Both Muscat models consider donor–group combinations as random effects. See Section 2.3 for more details.

Each method was rigorously evaluated in two case studies (across cell types and across cell states) and under different scenarios, such as variations in library size between groups and pronounced heterogeneity within groups.

2.2.1 Case study 1 - DE analysis on different immune cell types in fallopian tube

In this dataset, we examined the efficacy of various methods across three distinct scenarios: homogeneous groups with differing library sizes, homogeneous groups with similar library sizes, and heterogenous groups. For each scenario, we illustrate the overarching gene expression profile, describe the DE results using diagnostic plots, and conduct a gene ontology (GO) analysis to investigate the biological foundations of our DE findings.

Contrasting CD8+ T cell subgroups with marked library size differences

The first comparison is between groups of CD8+ T cells (clusters 12 and 13), where there are notable differences in library sizes (Fig. 2.5a). This example illustrates the impact of library-size–based normalization on single-cell data. Using a two-sample t-test, we compared gene expression means between these groups with raw UMI counts and three other normalization methods (Fig. 2.5b) using absolute t-scores. While t-scores from CPM mirror those from UMI counts, albeit with minor shrinkage, both VST and integration show substantial shrinkage. This normalization process dampens the gene expression differences between the groups before deploying any DE detection techniques.

Each method applies its own filtering procedure in the implemented function, leading to varying numbers of input genes. Poisson-glmm, Binomial-glmm, and MAST utilized nearly 4600 genes as input (Fig. 2.5c), whereas pseudo-bulk DESeq2 employed default quality control in both genes and cells, resulting in only 104 genes remaining. Pseudo-bulk edgeR kept 9743 genes in CPM data as inputs. Muscat mixed models executed 6732 genes. Notably, for Wilcox method from the Seurat package, no genes passed the default filtering procedure. However, with a more relaxed filtering criterion, the impact on the differential expression results remains minimal.

In the volcano plots, both Poisson-glmm and Binomial-glmm display heavily imbalanced expression patterns, aligning with the observations in the density plots (Fig. A.3b). However, the other methods do not reflect this observation, with fold change estimates appearing evenly spread. The histograms of adjusted p-values for other methods are concentrated in large values (Fig. A.3c). Pseudo-bulk methods and mixed models from the Muscat package, in particular, exhibit p-values that are clustered around one. Despite observing imbalanced expression patterns in density plots and volcano plots in this comparison, only our GLMM methods identify a substantial number of differentially expressed genes (DEGs) (Fig. 2.5c). The heatmaps of DEGs further emphasize that raw counts can better capture the differences

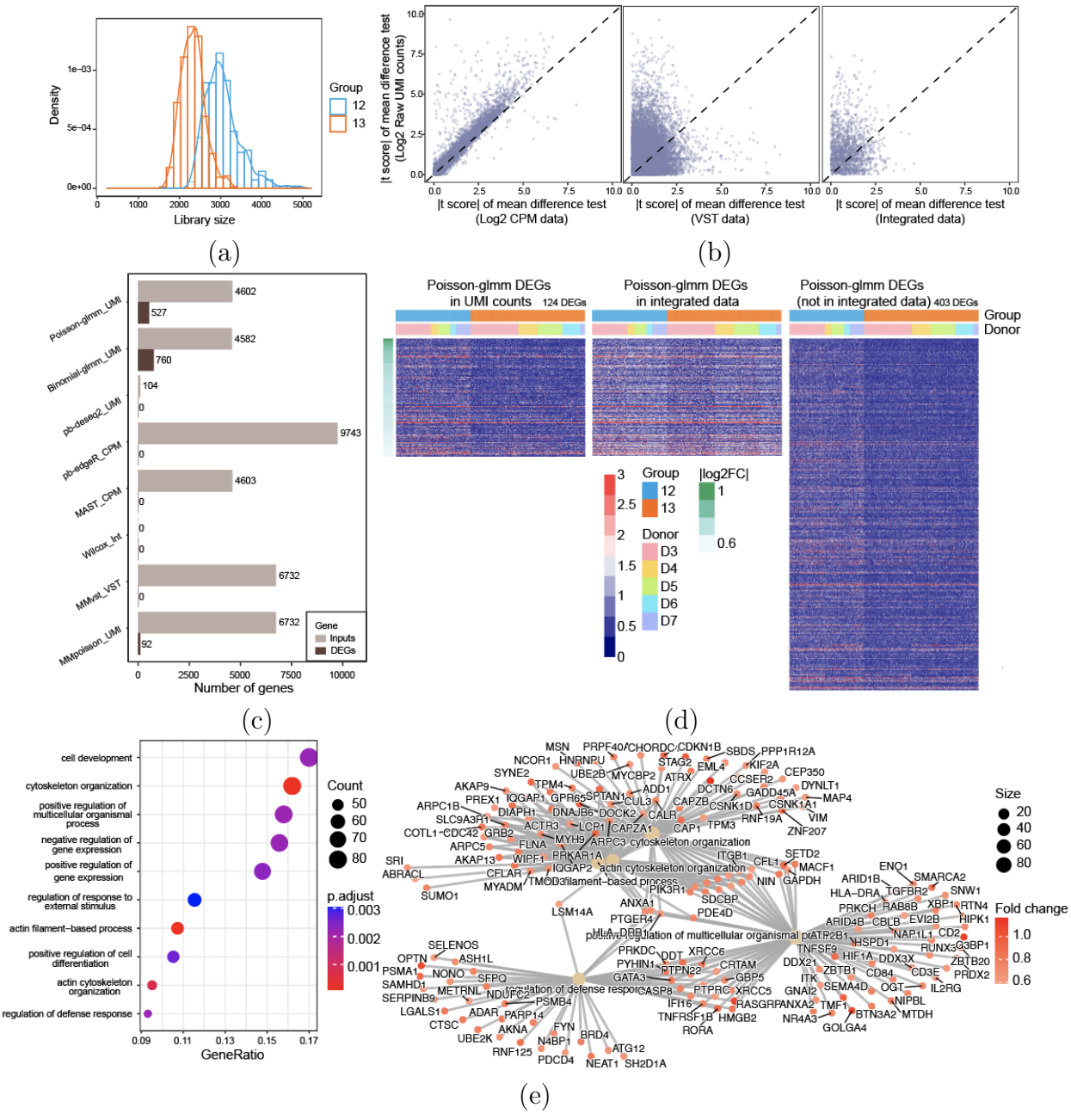


Figure 2.5: DE analyses on CD8+ T cell subgroups. a) Density plot of the library size for group 12 and 13. b) Scatterplot comparisons of t-scores from mean difference tests between raw UMI counts and other transformed data. Each gene's expression in two different groups is compared, showcasing the pairwise absolute t-scores from various data sources. c) Counts of input genes and DEGs in different DE methods. d) Heatmaps visualize Poisson-glmm DEGs. Order: UMI counts (left), integrated data (middle), and genes not included in the integrated data but shown in UMI counts (right). Heatmaps arrange genes by descending Poisson-glmm fold change estimates and columns group cells by cell clusters and donors. e) GO analysis of the DEGs identified by Poisson-glmm.

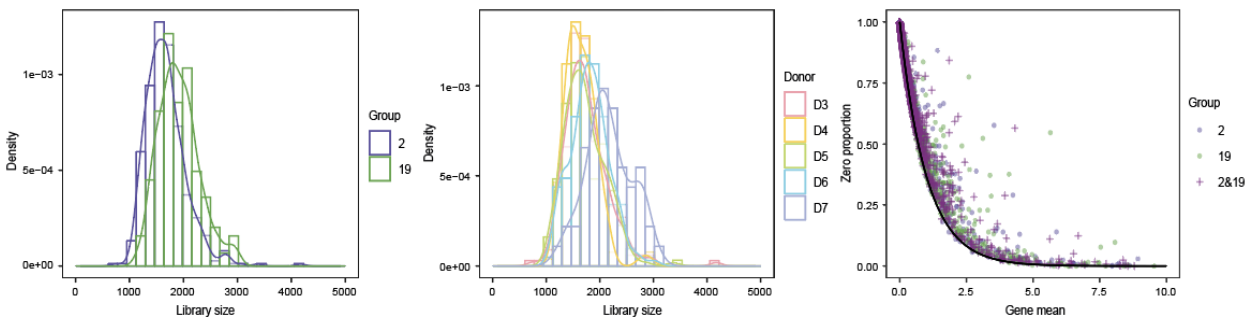
between groups compared to integrated counts (Fig. 2.5d). Furthermore, 403 DEGs were excluded from the integrated data before testing.

The DEGs prominently feature GO terms associated with actin cytoskeleton reorganization and immune synapse formation (Fig. 2.5e). As T cells detect antigens on an antigen-presenting cell, they establish an immunological synapse, necessitating substantial actin filament restructuring. Actin polymerization within this synapse aids the transit of receptors and signaling molecules, crucial for T cell activation. Our results hint that among these two CD8+ T cell groups, group 12 cells are actively recognizing antigens. Cell groups 12 and 13 had notable differences in library sizes. While the DEGs we identified contributed to the disparity in measured RNA content between the two groups, genes that were not differentially expressed had a much larger effect on library size; consequently, normalization erased the contribution of the DEGs to differences in expression patterns. Accordingly, in this example, only our GLMMs, which operate directly on UMI counts, successfully identified DEGs.

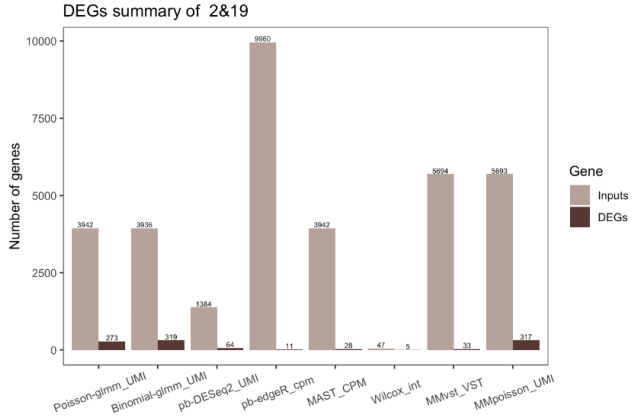
A Glimpse at CD4+ T Cells vs. NK Cells: No Striking Library Size Differences

The second comparison is between CD4+ T cells and NK cells (clusters 2 and 19). In the density plot, we observed similar library sizes based on UMI counts for the two clusters across donors except for donor 7 (Fig. 2.6a). The zero-proportions of genes in these two clusters fit a Poisson distribution well, indicating relative homogeneity within each cell cluster (Fig. 2.6a).

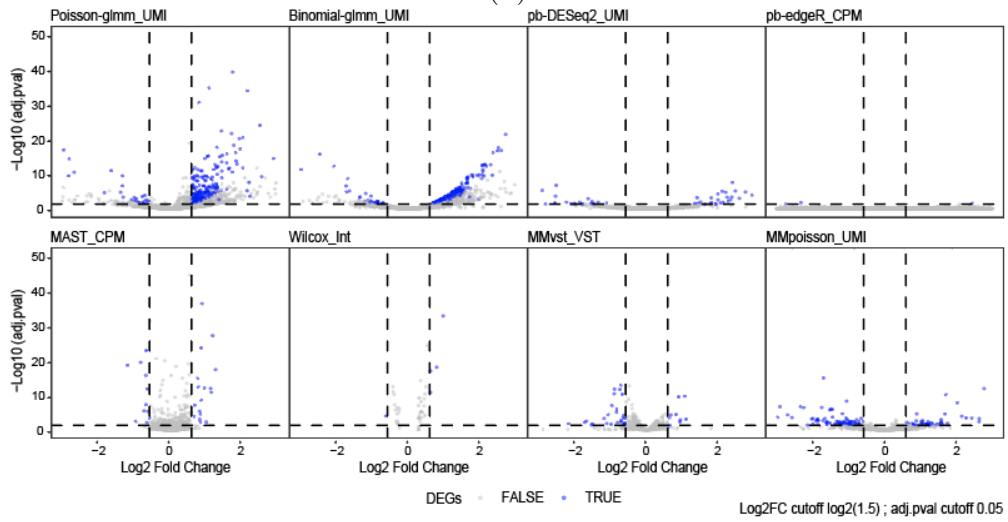
The filtering procedure implemented by different methods leads to very different numbers of input genes. Poisson-glmm, Binomial-glmm, and MAST utilized nearly 4000 genes as input (Fig. 2.6b). Methods implemented in the Muscat package, including pseudo-bulk methods DESeq2 and edgeR, as well as mixed models MMvst and MMpoisson, employed 1384, 9960, 5694, 5693 genes, respectively, in accordance with their filtering procedure. Notably, the Wilcox method from the Seurat package includes only 47 genes as input due to the filtering based on the log2 fold change between two groups of interest. The log2 fold



(a)



(b)



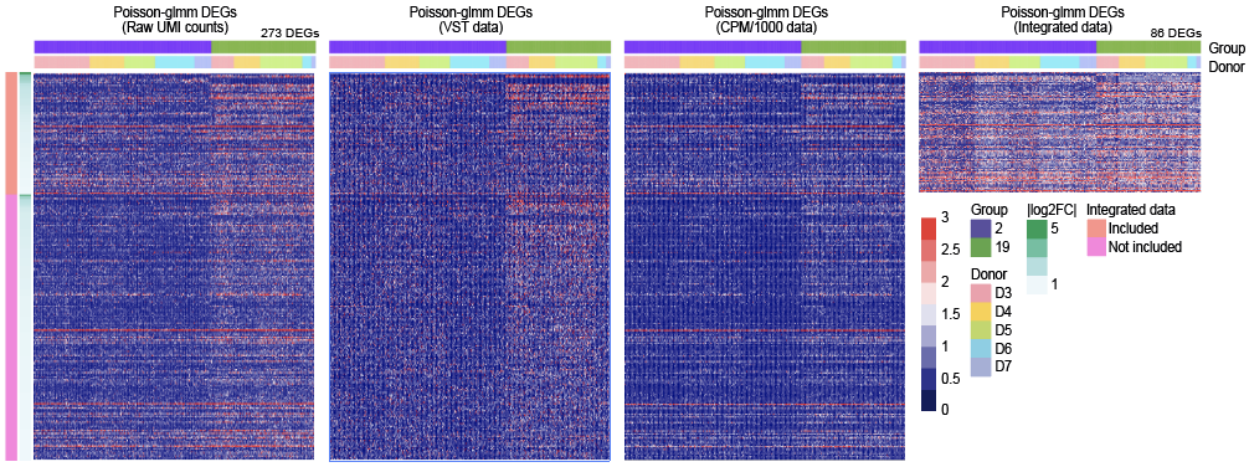
(c)

Figure 2.6: DE analyses on CD4+ T Cells vs. NK Cells (part 1) a) Left: Density plot of the library size for group 2 and 19. Middle: Density plot of the library size by different donors. Right: Zero proportion plots for each group and combined groups. b) Counts of input genes and DEGs across different differential expression methods. c) Volcano plots for each method, highlighting DEGs in blue. The signs of log2 fold change are adjusted such that positive signs represent higher expression in group 19.

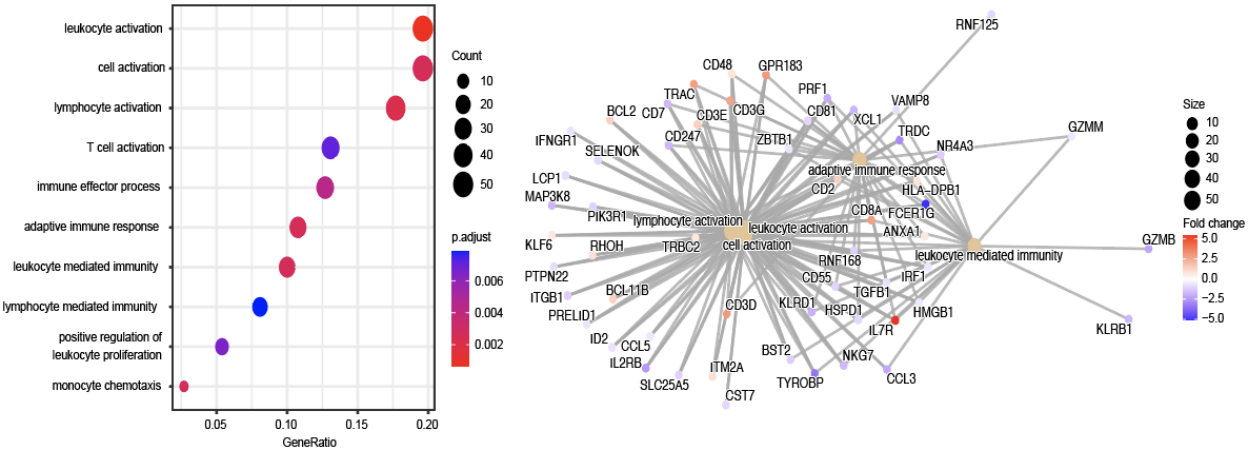
change in the package is calculated using the formula log2(1 + mean1) / log2(1 + mean2) on the input data, which can be normalized/integrated data by Seurat or other packages. This transformation attenuates the ratio of the two group means through the addition of 1 to each mean, resulting in the exclusion of a substantial number of genes. Wilcox, MAST, pseudo-bulk methods, and MMvst each identified fewer than 100 DEGs. In contrast, the methods that use UMI counts, Poisson-glmm, Binomial-glmm, and MMpoisson, identified 273, 319, and 317 DEGs, respectively (Fig. 2.6b).

In the volcano plots, there are more positive estimates of log2 fold change by Poisson-glmm and Binomial-glmm, signifying that genes are more expressed in cluster 19 (Fig. 2.6c). From the pairwise comparisons of log2 fold change (Fig. A.4b), MAST, Wilcox, and MMvst exhibit smaller log2 fold change estimates, due to normalization processes that shrink the values. Pseudo-bulk methods tend to yield more conservative p-values (Fig. 2.6c, A.4c), as illustrated in the histograms (Fig. A.4d). While the log2 fold change estimates are consistent across our GLMMs, pseudo-bulk methods, and MMpoisson, the presence of deviant p-values leads to significant disparities in the identification of DEGs. Our GLMMs identified many more DEG candidates, surpassing the thresholds of adjusted p-value and fold change.

In Figure 2.7a, we display gene expression from DEGs identified by Poisson-glmm alongside heatmaps for VST, CPM, and integrated data. Notably, differences among these heatmaps are subtler than those displayed in raw UMI counts. The integrated data displays elevated gene expression across groups, obscuring distinctions. The heatmaps of DEGs from Poisson-glmm and Binomial-glmm show the validity of DEGs (Fig. A.4f), while most of the DEGs identified by MMpoisson do not illustrate differential expressions in UMI counts (Fig. A.4g). We performed gene ontology (GO) enrichment analysis on DEGs from Poisson-glmm. The DEGs are enriched for GO terms related to leukocyte activation, cell activation, and lymphocyte activation (Fig. 2.7b), suggesting NK cells represented by cluster 19 are more active than the CD4+ T cells.



(a)



(b)

Figure 2.7: DE analyses on CD4+ T Cells vs. NK Cells (part 2) a) Heatmaps of Poisson-glmm DEGs shown in different data sources, with genes in integrated data featured in the top block, and those absent in the lower block. b) GO analysis of the DEGs identified by Poisson-glmm.

In summary, in the comparison of two cell clusters of similar library sizes, normalization continued to obscure informative differences between the two clusters and hindered the identification of potential DEGs.

Deciphering the Complexities of Heterogeneous Groups: Mature T Cells vs. CD4+ T Cells

Finally, by merging groups 8 and 17 and groups 2 and 19, we created two less homogenous groups-of-interest: mature T cells and CD4+ T cells, respectively. The distribution of library sizes between these clusters exhibits noticeable differences (Fig. 2.8a), and the zero proportions of these groups deviate from a Poisson distribution (Fig. A.5a).

Poisson-glmm, Binomial-glmm, and MAST used approximately 3480 genes as input. Pseudo-bulk DEseq2, edgeR, and mixed models utilized 1937, 10483, 7099 genes, respectively. For Wilcox, 123 genes passed the filtering procedure. The volcano plots revealed similar patterns to previous comparisons across various methods (Fig. A.5c). Our GLMM methods exhibited predominantly positive estimates of fold change, suggesting higher expression of abundant genes in CD4+ T cells (group 2&19). MAST, and MMvstn showed a somewhat similar tendency, but less imbalanced. However, pseudo-bulk methods and MMpoisson provided evenly distributed estimates in both directions. The estimates of log2 fold change are not quite identical among different methods (Fig. A.5e). Both pseudo-bulk methods exhibited a negative shift compared to Poisson-glmm, while MMpoisson had a positive shift. MAST, Wilcox and MMvst showed shrinkage as before. Additionally, most input genes for the Wilcox method displayed positive fold changes, albeit with small magnitudes. This observation sheds light on how normalization and logarithmic transformation during pre-processing influences the estimation of differences in gene expression.

When we examine the violin plots of gene expression frequency and log2 mean for the DEGs identified by each method, it becomes apparent that MAST, Wilcox, and MMvst

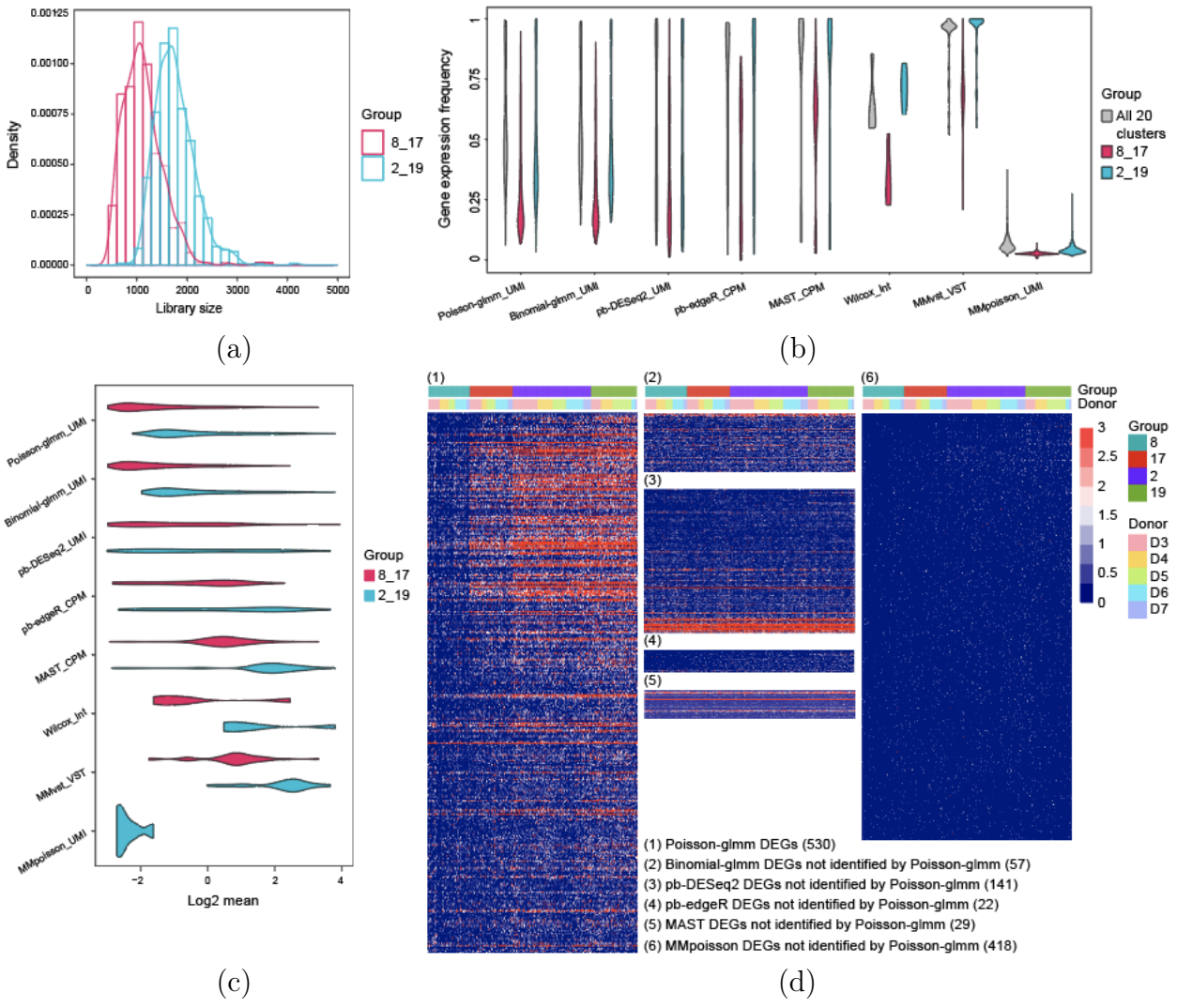


Figure 2.8: DE analyses on heterogeneous groups: Mature T Cells vs. CD4+ T Cells. a) Density plots comparing library sizes for combined groups 8 & 17 and 2 & 19. b) Comparisons of the gene expression frequency of the DEGs from different methods. c) Violin plot of log2 gene mean for DEGs from different methods. d) Heatmaps of DEGs from different methods.

captured fewer DEGs with lower gene expression frequency and smaller gene means than the remaining methods (Fig. 2.8b, 2.8c). It is worth noting that MAST is a zero-inflated model, which incorporates excessive zeros as an additional component. However, MAST might not effectively characterize the zeros, as demonstrated in previous studies on UMI counts. Consequently, potential DEGs that are lowly expressed may be masked by the model. The Wilcox method tends to filter out a substantial number of genes, which poses challenges in identifying lowly expressed genes. MMvst, despite having a considerable number of input genes (n=7099), only identified 35 DEGs.

The heatmap of DEGs in Poisson-glmm reveals distinct expression patterns between the two groups (Fig. 2.8d (1)). However, in this example, the inherent heterogeneity within each group impacts the fitness of Poisson model, potentially leading to false discoveries. To evaluate the possibility of false discoveries by Possion-glmm, we examined DEGs identified by other methods, but not by Poisson-glmm (Fig. 2.8d (2)-(6)). The heatmaps make it evident the DEGs that differentiate between the two groups are largely identified by Poisson-glmm only; the other methods did not contribute additional valid DEGs that differentiate the two groups. Conversely, most of the DEGs detected by Poisson-glmm exhibit differential expression despite the heterogeneity within each group.

Notably, MMpoisson mainly detected DEGs with small means (Fig. 2.8c), not showing clear differences between different groups (Fig. 2.8d (6)). And the DEGs are mutually exclusive to those identified by Poisson-glmm. Although Poisson-glmm and MMpoisson both use UMI counts, MMpoisson includes group information as a random variable and involves library size as an offset; our result underscores the significance of using an appropriate random effect in a mixed model and suggests that the cell group information should be excluded from the random component.

The DEGs are enriched for GO terms related to peptide metabolic process and cytoplasmic translation, indicating lower ribosomal RNA activities in mature T cells (Fig. A.5g).

Indeed, mature T cells exhibit lower levels of ribosomal RNA activity compared to their immature counterparts, mainly due to the state of activation and the metabolic requirements of the cells. On the other hand, mature T cells, which are not rapidly proliferating, have less need for protein synthesis and thus exhibit lower levels of rRNA activity. However, upon antigen recognition and activation, mature T cells can rapidly upregulate rRNA activity and protein synthesis to support clonal expansion and effector function. This differential regulation of rRNA activity is one of the many ways in which cells regulate their metabolic activities to adapt to different physiological conditions.

In this example, Poisson-glmm detected more valid DEGs for heterogenous cell populations than other methods. Normalization still diminished measurable differences between groups. We also raise concerns about the masking of lowly expressed genes by the improper treatment of zeros, as seen in MAST method and VST data.

2.2.2 Case study 2 – DE analysis on different states of B cells

In this case study, we applied our proposed DE framework to data collected by Kang et al. [2018]; this dataset consists of 29,065 cells and 7,661 genes from eight distinct cell types, collected from peripheral blood mononuclear cells of eight lupus patients. Within each cell type, the cells are evenly split into two groups for perturbation: unstimulated control and IFN-β stimulated (Fig. A.6a). UMAP plots (Fig. 2.9a) highlight that gene expression patterns are more differentiated between stimulation states than between cell types. The zero-proportion plots fit better to Poisson distribution when separated by stimulation states than only by cell types (Fig. A.6b). This observation motivated us to focus on DEGs between the cell states rather than between the cell types.

Like the previous case study, we found that the distribution of library sizes underwent significant changes after normalization (Fig. 2.9b). Raw UMI counts show that each cell type has a unique library size distribution. However, these differences became less pronounced

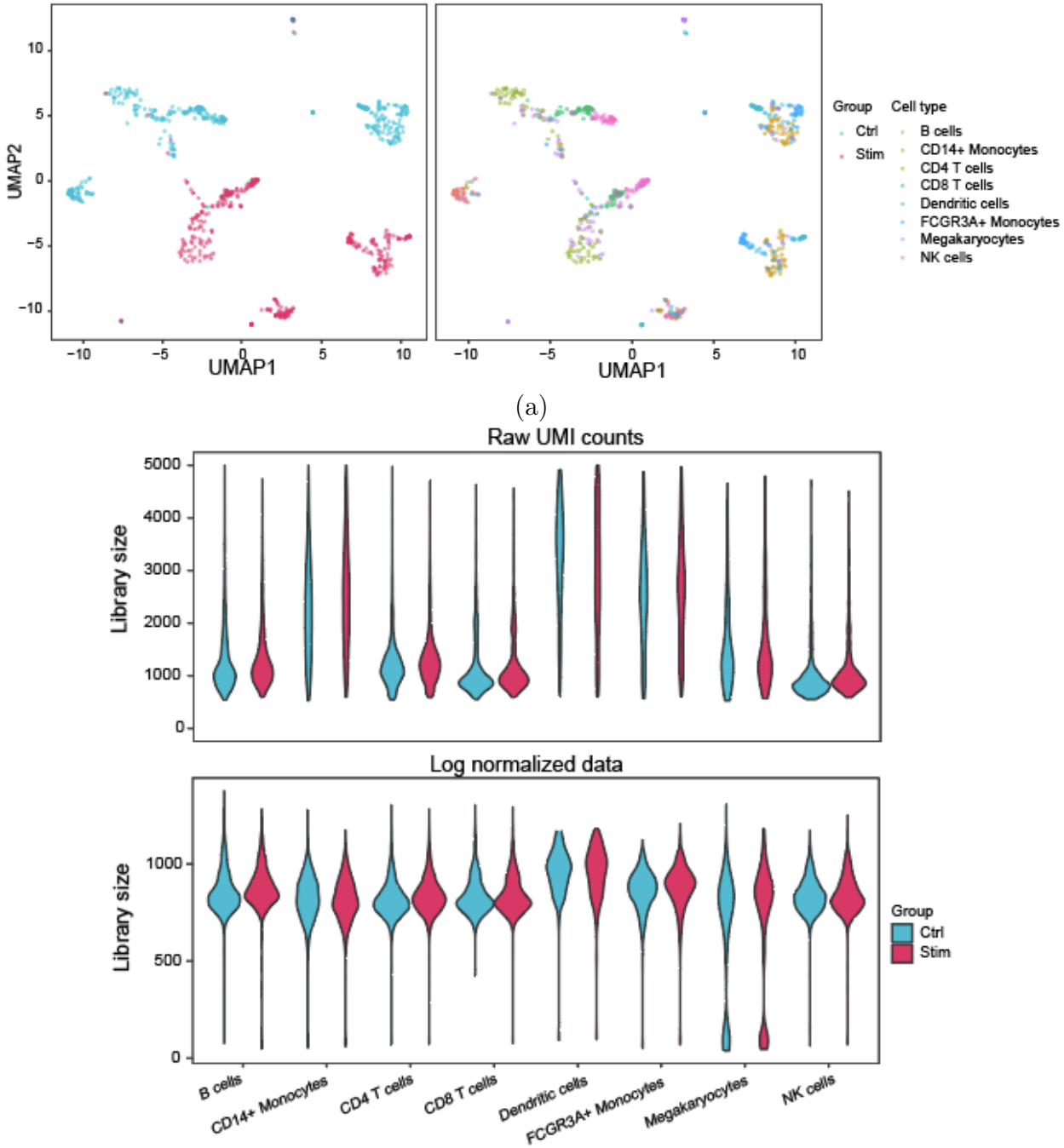
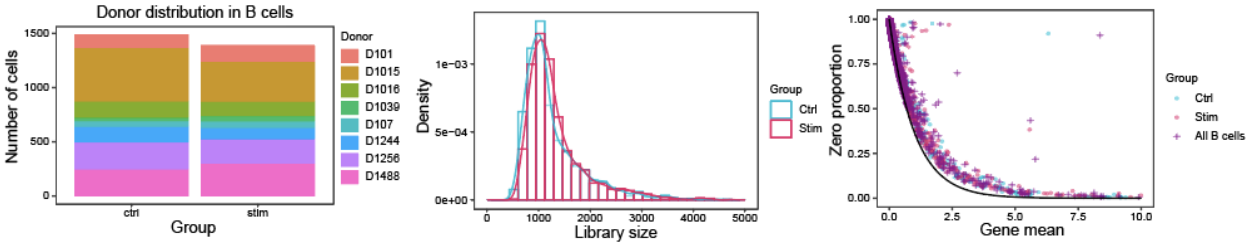
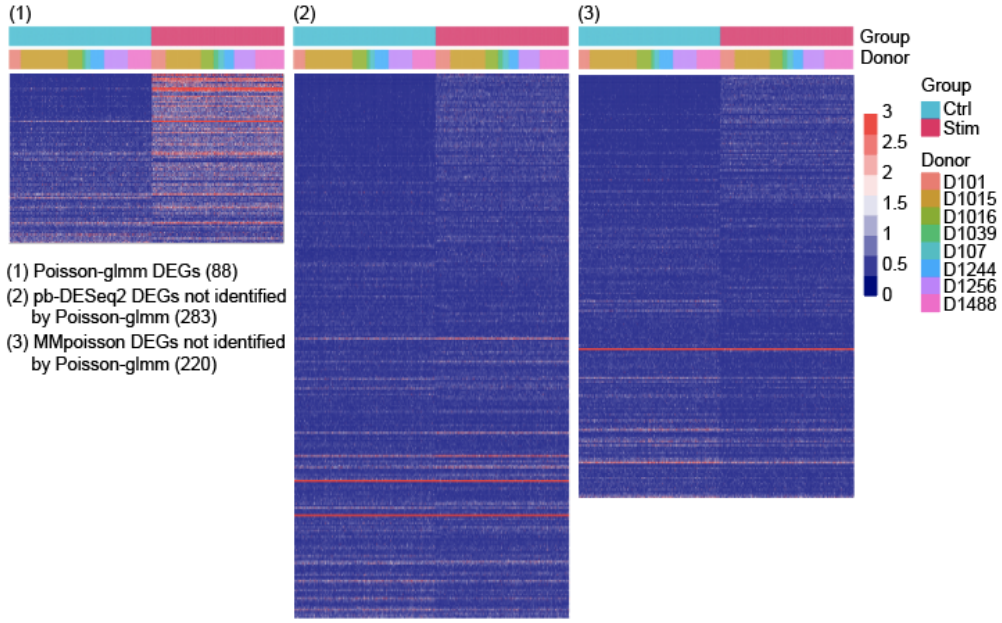


Figure 2.9: Overview of case study 2 and DE analyses on different states in B cells. (part 1) a) UMAP showing groups and cell types for case study 2. b) Library size comparisons before (raw UMI counts) and after normalization (log-normalized data) by cell type.



(a)



(b)

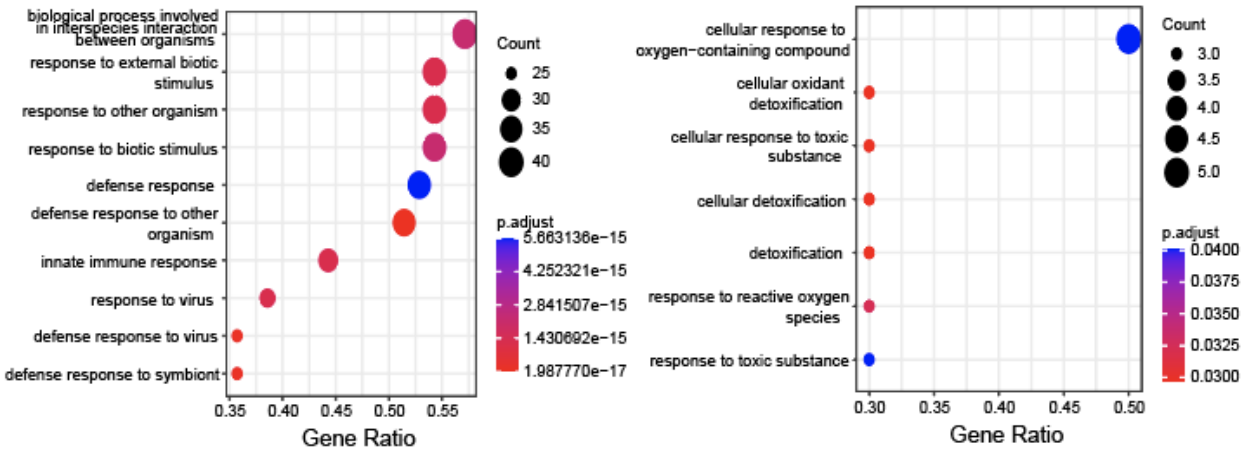
Figure 2.10: Overview of case study 2 and DE analyses on different states in B cells. (part 2) a) Left: Donor distribution among B cells. Middle: Density plot of library size in different states. Right: Zero proportion plots for different states and combined states. b) Heatmaps of DEGs identified from different methods.

following normalization, while library sizes remained relatively consistent between states within a single cell type. Normalization seems to predominantly affect differences across cell types rather than between states.

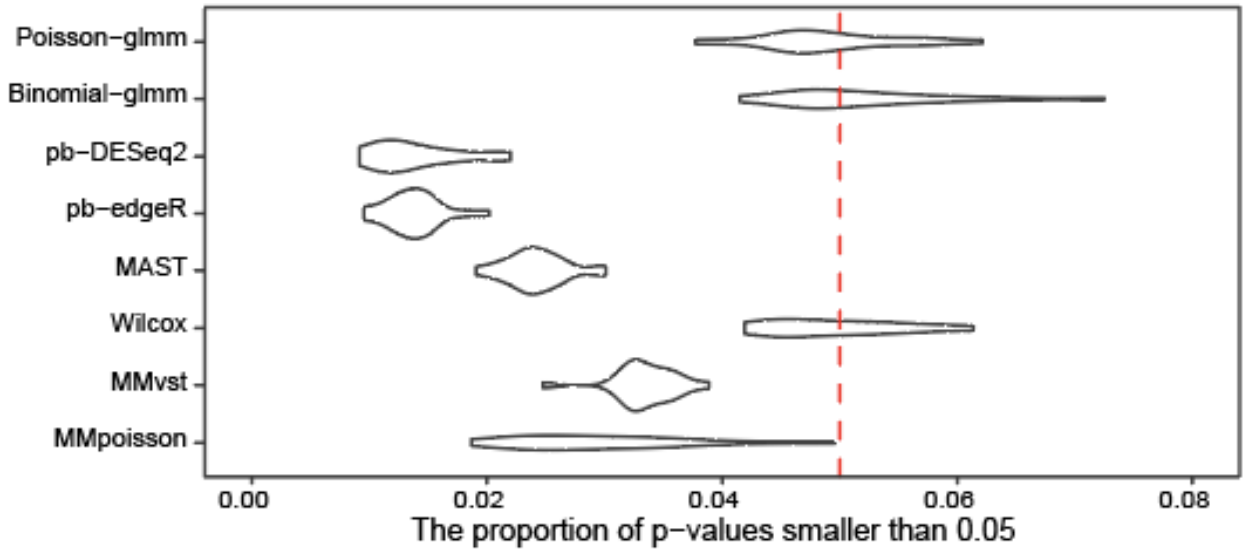
For the remainder of our case study, we focused on B cells. The cells from each donor were divided approximately equally between the control and stimulated groups (Fig. 2.10a top), and the library size distribution in these two groups is similar (Fig. 2.10a middle). The zero-proportion plot suggests that the data does not perfectly fit the expected curve from the Poisson distribution, indicating the presence of a mixture of subtypes within B cells (Fig. 2.10a bottom).

In our analysis of the subset comprising unstimulated and stimulated B cells, the majority of DE methodologies used about 2,550 genes as inputs (Fig. A.7a). However, the Wilcox approach within Seurat selected only 144 genes. The estimates of fold change for the two states in B cells exhibit an even spread across all methods, as depicted in the volcano plots (Fig. A.7b). MAST and MMvst struggled to identify differential patterns. Different from previous examples, our GLMM approach flagged fewer DEGs than both pseudo-bulk techniques and MMpoisson. Notably, the DEGs that were not shared between pseudo-bulk DESeq2 or MMpoisson and Poisson-glmm predominantly belong to the extremely low expression category (Fig. 2.10b (2), (3)).

We hypothesized that this result could be explained by using fold change as a DEG criterion. In bulk RNA-seq, a gene is typically labeled as a DEG if its adjusted p-value is below a certain threshold, often 0.05, and the fold-change estimate exceeds a predetermined value, typically 1.5 or 2 (Fig. A.8a). Most single-cell DE methods use the same criteria. However, in single-cell datasets, the mean counts for many genes are exceedingly close to zero. Consequently, fold change may not be a reliable metric to differentiate nuances in read counts. For instance, if gene means are 2^{-3} for one group and 3^{-3} for another, the fold-change threshold of 1.5 is met, but the actual difference is a mere 0.0625, which does not



(a)



(b)

Figure 2.11: Overview of case study 2 and DE analyses on different states in B cells. (part 3) a) GO analysis for up-regulated (left) and down-regulated genes (right). b) Violin plots depicting the proportion of p-values below 0.05 for each method.

convey a significant disparity in expression, especially when juxtaposed with genes having larger means. Moreover, near-zero values can result in computational inaccuracies, causing ratio deviated from the underlying true value.

To overcome the limitation of using fold-change ratios on small counts, we established a new criterion for DEGs based on absolute differences. Specifically, we mandated that the mean difference between two groups exceeds a set threshold, such as -1. In the volcano plot, numerous genes would be designated as DEGs when relying on ratio-defined fold change. Yet, as shown in the mean vs. mean difference plot that many genes that meet the p-value criteria showcase only modest changes in absolute means (Fig. A.9a). This approach emphasizes genes with significant absolute differences, yielding more biologically pertinent results.

We performed GO enrichment analysis on up-regulated and down-regulated genes separately (Fig. 2.11a). We found IFN-β stimulated B cells have increased activities in interaction between organisms, defense response, defense response to virus and defense response to symbiont, while their activities in translation and other metabolic processes are decreased. Pseudo-bulk technique detected similar GO terms while MMpoisson was underpowered for detection of down-regulated GOs (Fig. A.9).

In this example, we demonstrated that conventional metrics to detect DEGs, especially fold change based on ratios, are ill-suited for low-count data where the large fold changes reported by current methods may be attributed to the ratio of two very small gene means. Careful post-processing is needed to prioritize signals and manage false discoveries.

2.2.3 False discovery rates assessed under the null setting using permutation analysis

To assess p-value calibration in empirical data, a permutation analysis was conducted within a null dataset focusing on a group of interest. We specifically conducted the analysis on three

datasets: the control group of B cells, group 2, and group 13 in case study 1. Each underwent random assignment to either the control or stimulus group. Subsequently, p-values for each gene were computed employing various methods, with the gene set confined to those input into the Poisson-glmm model. To mitigate potential gene filtering, the threshold for the Wilcox method was relaxed. This process was iterated 20 times, and on each iteration the proportion of p-values below 0.05 was calculated along with the corresponding false discovery of differentially expressed genes.

The analysis of the violin plot (Fig. 2.11b, A.10) reveals that both our GLMM methods and the Wilcox method exhibit consistently well-calibrated p-values among different choices of null datasets. However, pseudo-bulk methods, and mixed models from Muscat appear excessively conservative, with an overall proportion considerably below 0.05. The performance of MAST is conservative in B cells but not in case study 1. The histograms of p-values across the 20 runs demonstrate a consistently flat distribution for our glmm methods and the Wilcox method, indicative of adherence to the null setting (Fig. A.10). Conversely, other methods display overestimated p-values, yielding conservative outcomes. Note that even though Wilcox performed well in the permutation analysis, it is not powerful to detect real DEGs as shown in previous case studies. Under both the existing criteria and our newly established criteria for determining DEGs, each method detected, at most, one false discovery in each run.

2.2.4 Discussion

In this Chapter, we examined existing DE approaches to pre-processing, input values and test statistics, and fold-change definitions in the context of single-cell DE analysis. We demonstrated through extensive real-data examples the limitations and drawbacks of current practices. We showed that current normalization and pre-processing techniques may obscure DEGs by an overreliance on relative RNA abundance and ignoring or correcting for biological

zeros. We also illustrated how use of volcano plots in DE analysis, which also depends on relative RNA abundance, leads to false discoveries in lowly expressed genes by prioritizing fold changes in expression over absolute changes. We also argued that single-cell DE analysis suffers from false discoveries due to the inappropriate handling of donor effects, as well as from biases that accumulate as the consequence of sequential workflows.

We advocate a new paradigm, Poisson-glmm, which uses UMI counts as input and a generalized Poisson mixed effect models to account for batch effects and within-sample variation. This framework's use of UMI counts can significantly improve current practices by leveraging absolute RNA expression. Poisson-glmm shows superior sensitivity and robustness toward model misspecification when compared to current single-cell DE methods, which should ultimately lead to new biological insights from single-cell data.

The use of UMI counts for DE analysis in scRNA-seq can significantly improve current practices, potentially making some current practices (e.g., volcano plots as a diagnostic DE tool) obsolete. However, relying on UMI counts as a representation of genuine RNA content predicates that measurements are strictly single-cell based, underscoring the need for meticulous doublet and triplet removal prior to DE analysis. Furthermore, seamlessly implementing this new paradigm into existing popular tools remains a challenge. Given this significant shift from current practices, a sustained effort will be required to educate and train researchers on these new alternatives and to reshape existing practices accordingly.

2.3 Methods and materials

2.3.1 Datasets and pre-processing

In case study 1, a 10X scRNA-seq dataset of post-menopausal fallopian tubes, with 57,182 cells sourced from five donors, covering 29,382 genes was analyzed. We obtained 20 clusters via HIPPO algorithm. We did not apply a pre-filtering procedure on this dataset, except for

built-in filtering steps in each method. We used sctransform to get the VST data and the integration workflow provided by Seurat to obtain the integrated data.

All integration or normalization processes were performed on the entire dataset, since cell types are typically unknown during the pre-processing stage. In cross-batch integration, only the top 2,000 highly expressed genes were retained, which significantly reduced the number of genes for downstream analysis. The dataset had already been fully analyzed and annotated with cell types. We utilized the annotations to examine the effects of normalization/integration on distributions of library sizes across cells.

In case study 2, the dataset comprised 10X droplet-based scRNA-seq PBCM data from eight Lupus patients obtained before and after 6h-treatment with IFN-β. After removing undetected and lowly expressed genes (less than 10 cells expressing more than 1), the dataset consisted of 29065 cells and 7661 genes. The integrated data was replaced by log2-transformed normalized expression values obtained via computeLibrarayFactors and logNormCounts functions in Muscat.

2.3.2 Poisson-glmm and Binomial-glmm

By default, we excluded any genes detected in fewer than 5% cells in the compared groups from differential testing. The GLMMs were implemented with glmmPQL function of the MASS package. We calculated adjusted p-values by using Benjamini-Hochberg correction. Each model fitting was applied on one gene and the two compared groups.

We fit Poisson-glmm on UMI counts. Each count X_{cgk} sampled from cell c, donor k, and gene g, was modeled by

$$X_{cgk}|\lambda_{cgk} \sim \text{Poisson}(\lambda_{cgk})$$

$$\log(\lambda_{cgk}) = \mu_g + X_c\beta_g + \epsilon_{gk}.$$

We fit Binomial-glmm on the zero proportions. Each count X_{cgk} was modeled by

$$1X_{cgk} = 0|p_{cgk} \sim \text{Bernoulli}(p_{cgk})$$

$$\log(p_{cgk}/(1-p_{cgk})) = \mu_g + X_c\beta_g + \epsilon_{gk}$$

where X_c is the indicator for groups (e.g. cell types in case study 1, control/stimulus in case study 2), and $\epsilon_{gk} \sim \mathcal{N}(0, \sigma_g^2)$ represents the random effects for donor k. Our goal was to test $H_0 : \beta_g = 0$.

For both methods, we provided "log2 fold change" computed by $\log_2\left(\exp\left(\beta_g\right)\right)$. In Poisson-glmm, this estimate indicates the increment of $\log_2(\lambda_2)$ against $\log_2(\lambda_1)$, which is the conventional log2 fold change. However, this term in Binomial-glmm doesn't represent the same meaning. It is the difference between $\text{logit}(p_2)$ and $\text{logit}(p_1)$. The p-value and BH adjusted p-value are provided.

2.3.3 Benchmarked methods

Pseudo-bulk DESeq2 and pseudo-bulk edgeR are aggregation-based methods used in our comparison. The input counts were summed up for a given gene over all cells in each group and by donor. The pseudo-bulk data matrix has dimensions GxS, where S denotes the number of interactions of donors and groups. For example, if there are two groups and "a" and "b" donors in each group, then "S" is equal to $2(a+b)$. We used raw counts as the input for DESeq2, while CPM counts were used for edgeR. The log fold change was converted to log2 fold change in all the comparisons. We implemented these two pseudo-bulk methods following the guided tutorial in Muscat package; `https://www.bioconductor.org/packages/devel/bioc/vignettes/muscat/inst/doc/analysis.html`.

For MAST, we fitted a zero-inflated regression model (function zlm) for each gene and applied a likelihood ratio test (function lrTest) to test for between-group differences in gene

expression. Besides the labels of groups and the cellular detection rate, we also included donor labels in the covariates. This method was run on log (CPM+1) counts. We followed the tutorial https://github.com/RGLab/MAST.

Wilcox, a rank sum test, is the default DE method in the FindMarkers function in the Seurat package. We used integrated data and log counts as input. We computed the log fold change given in the output as log(1+mean1)/(1+mean2). We applied the default filter in FindMarkers to only test genes with a log fold change greater than 0.25. We calculated the adjusted p-value provided from the function based on Bonferroni correction. We followed the guided tutorial found here: `https://satijalab.org/seurat/articles/de_vignette`.

MMvst and MMpoisson are mixed models implemented in the Muscat package. MMvst fits linear mixed models on variance-stabilizing transformation data. MMpoisson fits Poisson generalized linear mixed models with an offset equal to the library size factors. In both models, we fit a $\sim 1 + \text{group} + (1|\text{sample})$ model for each gene, where "sample" denotes the experimental units (the interaction of donors and groups). We followed the tutorial found at: `https://www.bioconductor.org/packages/devel/bioc/vignettes/muscat/inst/doc/analysis.html`.

2.3.4 The criteria to determine DEGs

For the benchmarked methods, we adhered to conventional criteria for the identification of Differentially Expressed Genes (DEGs). Specifically, a gene was classified as a DEG if its absolute log2 fold change exceeded a predefined threshold, and the adjusted p-value was below a specified cutoff. Typically, DEGs are visually represented in volcano plots. In the first dataset, the log2 fold change threshold was set at log2(1.5), whereas in the second dataset, it was set at 1. The adjusted p-value threshold for both datasets was established at 0.05.

We proposed new criteria that are based on the convention plus the gene mean and the

difference in mean. If the log2 gene mean in two groups is lower than a certain value (-2.25 in case study 1) and the log2 mean difference is smaller than a threshold (-1 in case study 1), the gene would not be considered as a DEG. These can also be used as a filter before any DE analysis to speed up the computation. Both criteria are adjustable, depending on the dataset's performance and characteristics. An examination of heatmaps and mean difference against mean plot in advanced can be helpful to determine the thresholds when analyzing a new dataset (Fig. S8b, c).

2.3.5 Variation analysis

To gain a deeper understanding of the donor effect and cell type effect concerning various types of counts, we conducted a variation analysis across multiple group comparisons. To ensure the consistency of our results, we restricted our analysis to genes presented in all datasets. For each gene, we employed linear models ($\text{lm}(\text{count} \sim \text{donor} + \text{group})$) and computed the variances attributed to three components: donor, group, and the residual. Logarithm transformation was applied to UMI counts and CPM data to address skewness. The outcomes of this analysis were then presented and compared based on the proportion of variation explained by the first two components across different count types and various pairs. The results of the top 500 genes with the lowest residual variations were exhibited.

2.3.6 GO enrichment analysis

GO over-representation analyses were performed using the enrichGO function in the R package clusterProfiler with default parameters and the functional category for enrichment analysis set to the GO "Biological Processes" category.

CHAPTER 3
INVESTIGATION ON THE HIGHER COUNTS PROPORTION IN SINGLE CELL RNA DATA TO IMPROVE HIPPO ALGORITHM

3.1 Introduction

3.1.1 Zero-inflation test in HIPPO

In the study by Kim et al. [2020], extensive UMI data sets were analyzed, revealing crucial insights into the data processing workflow. The analysis indicated that clustering should be conducted prior to pre-processing steps such as normalization or imputation to effectively address the issue of excessive zeros in the data. This pre-clustering approach is essential for resolving cell-type heterogeneity, which in turn reduces the high proportion of drop-outs observed in the data. Specifically, once the heterogeneity among cell types is accounted for, the incidence of drop-outs significantly decreases within the smaller, more homogeneous clusters (See Fig. 3.1). Furthermore, the study's findings suggest that gene expression within a homogeneous cell population tends to follow a Poisson distribution. Otherwise, the presence of heterogeneity among cell types may lead to a finite Poisson mixture model, reflecting the complexity of gene expression patterns across diverse cell populations.

In that study, the authors proposed that the zero proportion of each gene should not be excluded but instead utilized as a feature to identify the heterogeneity among the cells. By leveraging the information contained in the zero proportions, they demonstrated a more effective approach to understanding and resolving cell-type heterogeneity. This method , named HIPPO, allows for a more accurate clustering process, leading to better normalization and imputation outcomes, and ultimately provides deeper insights into the underlying biological processes.

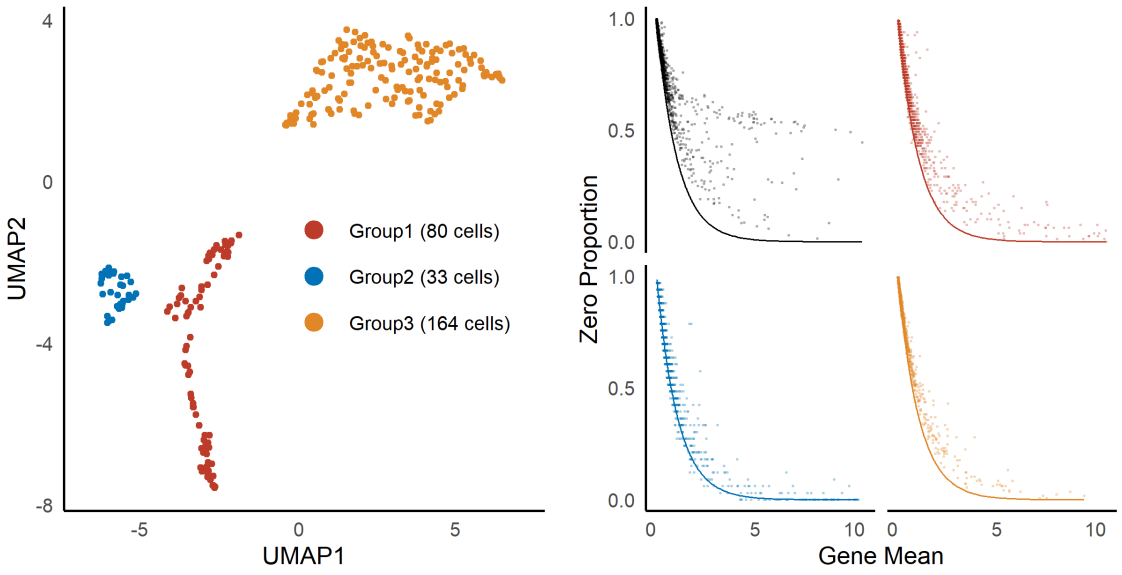


Figure 3.1: UMAP plots of CD34+ cells in ZhengZheng et al. [2017] data, and relationship between zero proportions and gene means before (black) and after (colors) clustering of CD34+ cells.

Here we briefly formulate the zero-inflation test used in the study. Consider a gene by cell matrix of UMI counts X for gene $g = 1, \ldots, G$ and cell $c = 1, \ldots, C$. A natural estimator for the true zero proportion of gene g across all cells can be defined as the follows,

$$\hat{p}_g^{(0)} = \sum_{c=1}^{C} \frac{1_{X_{gc}=0}}{C}.$$

Under Poisson model and with this statistic, the expected zero proportion of gene g with mean λ_g is $e^{-\lambda_g}$. And hence consider the hypotheses for each gene g as below.

$$H_0 : p_g^{(0)} = e^{-\lambda_g}, \quad H_A : p_g^{(0)} = \sum_{k=1}^{K_g} \pi_k e^{-\lambda_{kg}}$$

In practice, the total number of Poisson mixture K_g and the proportion of each group π_k are not estimated explicitly. By Jensen's inequality, $p_g^{(0)}$ under alternative hypothesis is always greater than that under the null. Instead, the alternative hypothesis was revised to

$H_A : p_g^{(0)} > e^{-\lambda_g}$. It can be interpreted that zero inflation indicates there is cell heterogeneity across the samples. The genes detected with zero inflated property are then selected as the features in HIPPO clustering algorithm.

For gene g, the gene mean is estimated as the average counts $\bar{X}_g = \frac{1}{C}\sum_{c=1}^{C} X_{gc}$ and is treated as a fixed number here.

3.1.2 Feature selection

They provided two methods for feature selection: the zero-inflation test and the deviance test. In this chapter, we will focus on the zero-inflation test and modify its formula to account for higher counts proportions. The goal of this project is to investigate whether we can identify more significant features beyond the zero proportion, thereby enhancing our ability to detect and analyze cell-type heterogeneity.

With the estimates plugged in,

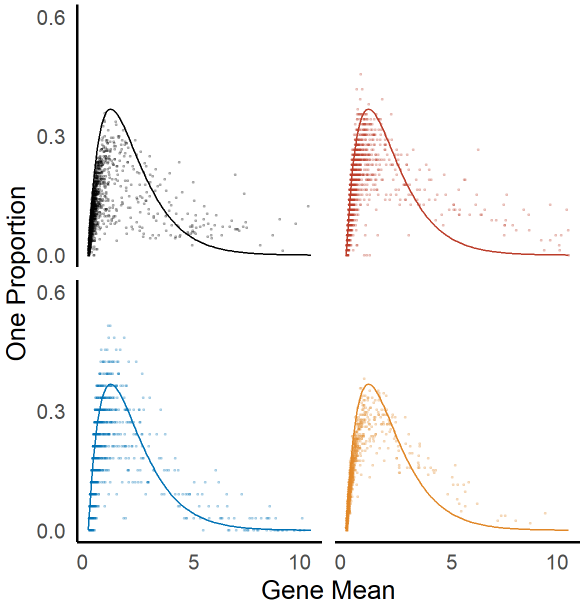
$$\hat{p}_g^{(0)} \sim \mathcal{N}\left(e^{-\bar{X}_g}, \frac{e^{-\bar{X}_g}(1 - e^{-\bar{X}_g})}{C}\right).$$

And the z-score test can be used to compute one-tailed p-value.

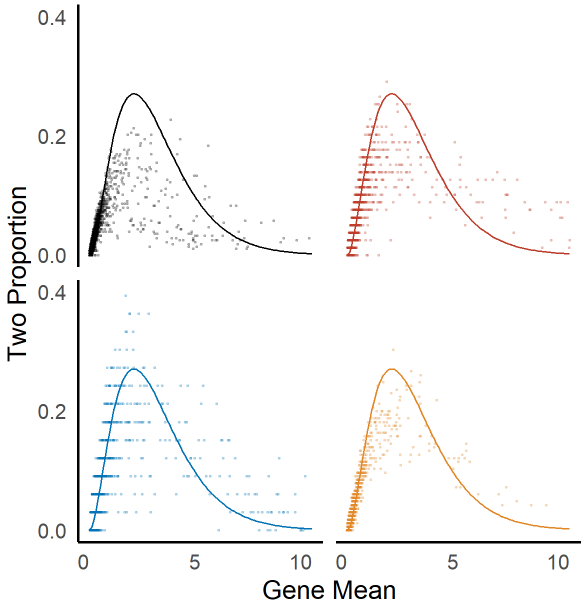
However, the gene mean is a random variable that follows a log-normal distribution, whose inference is not trivial. Further discussion is in the Supplementary of Kim et al. [2020].

3.1.3 Potentials in higher order counts

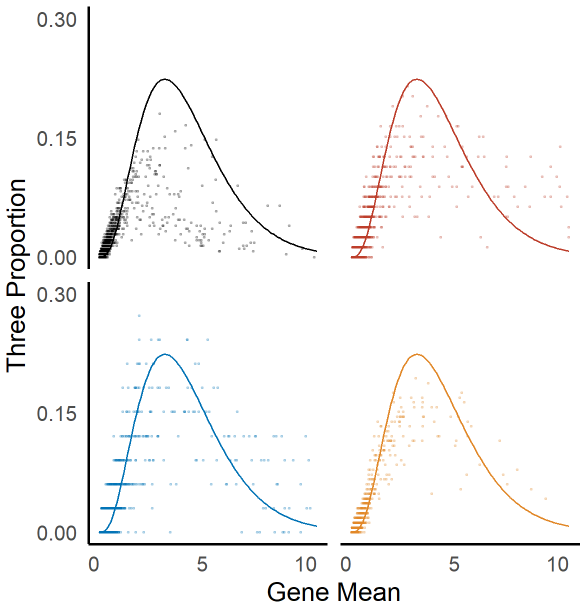
In addition to examining the zero proportion, we further analyzed the proportions of ones, twos, threes, and fours within the same data set. As shown in Fig. 3.2, the plots reveal that these observed proportions align more closely with the expected proportion lines after clustering. This observation suggests that there is valuable information to be extracted from



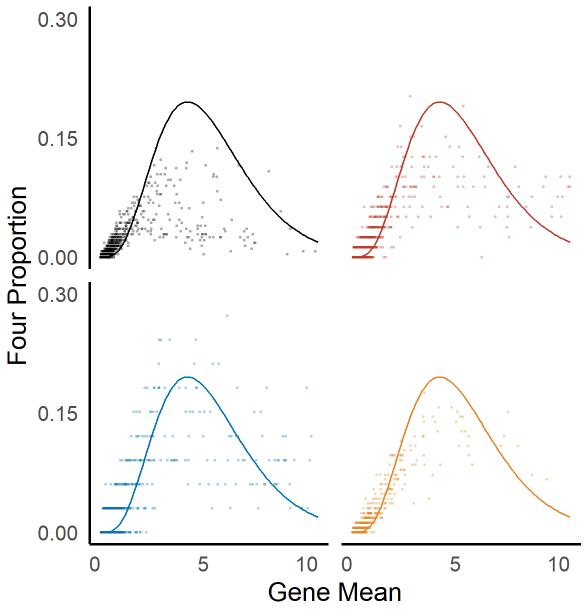
(a) One proportion



(b) Two proportion



(c) Three proportion



(d) Four proportion

Figure 3.2: Proportion plots

these counts. Inspired by these findings, we propose to extend our analysis by systematically investigating the proportions of ones, twos, threes, and fours. Specifically, we aim to test whether the observed proportions of these counts deviate significantly from the expected lines. This approach could provide a deeper understanding of the data and help identify additional significant features beyond the zero proportion, thereby enhancing our ability to characterize cell-type heterogeneity.

3.2 Methods

3.2.1 k-inflation test

valid range

Followed by the concept of zero-inflation test, the simpler alternative hypothesis can be inferred by Poisson mixture model when Jenson's inequality is valid. For proportion of $k = 1, 2, 3, 4$, the expected proportion of k under null hypothesis is

$$f_k(\lambda_g) = \frac{1}{k!}\lambda_g^k e^{-\lambda_g}.$$

To keep the validity of Jenson's inequality, we need convexity of f_k. For different k, the valid ranges of λ_g are different (See Table. 3.1). In these regions, we can relax the alternative hypothesis to $H_A : p_g^{(k)} > f_k(\lambda_g)$. This also coincides with Fig. 3.2. Genes with observed proportion above the expected line have sample means lying in the convex range of $f_k(\lambda_g)$. For hypothesis testing purpose, we might filter the genes with sample means in valid region.

k	convex region
1	$\lambda_g > 2$
2	$\lambda_g > 2 + \sqrt{2}$
3	$\lambda_g > 3 + \sqrt{3}$
4	$\lambda_g > 6$

Table 3.1: Convex region for each k

k-inflation test

As the previous setting, we compute the estimator $\hat{p}_g^{(k)}$ for each k.

$$\hat{p}_g^{(k)} = \sum_{c=1}^{C} \frac{1_{X_{gc}=k}}{C}.$$

Similarly, compute one-tailed p-value with z-score

$$z = \frac{\hat{p}_g^{(k)} - f_k(\bar{X}_g)}{\sqrt{\frac{f_k(\bar{X}_g)(1-f_k(\bar{X}_g))}{C}}}$$

3.2.2 Feature selection

Currently, we offer user-defined threshold values for feature selection use. The gene would be selected if at least one of the z value of its count proportions is greater than the threshold value. The hypothesis testing is

$$H_0 : \bigcap_{k=0}^{3} \{p_g^{(k)} = f_k(\lambda_g)\}, \quad H_A : \bigcup_{k=0}^{3} \{p_g^{(k)} > f_k(\lambda_g)\}$$

By incorporating higher order counts, we called the modified clustering procedure HIP-POx.

3.2.3 Differential expression analysis

Poisson mean t-test

In Kim et al. [2020], the author proposed a t-test on means. The hypothesis is $H_0 : \lambda_1 = \lambda_2, \quad H_A : \lambda_1 \neq \lambda_2$. While under the null hypothesis, the combined estimator of $\lambda_1 = \lambda_2$ should be the pooled sample mean. The t-statistic given in the paper should be modified as follows.

Assumptions:

1. Each X_{cg} are independent samples.
2. Validity of CLT.

$$
\begin{aligned}
X_{cg}|c \in C_1 &\sim Poisson(\lambda_1) \\
X_{cg}|c \in C_2 &\sim Poisson(\lambda_2) \\
\bar{X}_{C_1 g} &\sim N(\lambda_1, \frac{\lambda_1}{|C_1|}) \\
\bar{X}_{C_2 g} &\sim N(\lambda_2, \frac{\lambda_2}{|C_2|})
\end{aligned}
$$

We want to test $H_0 : \lambda_1 = \lambda_2$

$$
\text{t-statistic: } \frac{\bar{X}_{C_1 g} - \bar{X}_{C_2 g}}{\sqrt{\bar{X}_{pooled,g}(\frac{1}{|C_1|} + \frac{1}{|C_2|})}}
$$

Poisson proportion t-test

The sample mean of gene expression can be significantly influenced by particular outliers, leading to varied and potentially misleading outcomes. In contrast, the zero proportion can be considered a more robust and stable measure. To address the limitations of the sample

mean t-test, we propose a Poisson proportion t-test based on the zero proportion. And similar to mean t-test, the pooled proportion is used in the t-statistic.

Assumptions:

1. Each X_{cg} are independent samples
2. Validity of CLT

$$
\begin{aligned}
X_{cg}|c \in C_1 &\sim Poisson(\lambda_1) \\
X_{cg}|c \in C_2 &\sim Poisson(\lambda_2) \\
\hat{p}_{C_1 g} &\sim N(e^{-\lambda_1}, \frac{e^{-\lambda_1}(1-e^{-\lambda_1})}{|C_1|}) \\
\hat{p}_{C_2 g} &\sim N(e^{-\lambda_2}, \frac{e^{-\lambda_2}(1-e^{-\lambda_2})}{|C_2|})
\end{aligned}
$$

We want to test $H_0 : e^{-\lambda_1} = e^{-\lambda_2}$

$$
\text{t-statistic: } \frac{\hat{p}_{C_1 g} - \hat{p}_{C_2 g}}{\sqrt{\hat{p}_{pooled,g}(1-\hat{p}_{pooled,g})(\frac{1}{|C_1|}+\frac{1}{|C_2|})}}
$$

Poisson GLMM

For each count X_{cgk} sampled from cell c, donor k, and gene g,

$$
\begin{aligned}
X_{cgk}|\lambda_{cgk} &\sim Poisson(\lambda_{cgk}) \\
\log \lambda_{cgk} &= \mu_g + X_c \beta_g + \epsilon_{gk}
\end{aligned}
$$

where X_c is the indicator for different cell types, and $\epsilon_{gk} \sim N(0, \sigma_g^2)$ represents the random effects for donor k. We want to test $H_0 : \beta_g = 0$.

Binomial GLMM

$$1_{X_{cgk}=0}|p_{cgk} \sim Bernoulli(p_{cgk})$$
$$\log \frac{p_{cgk}}{1-p_{cgk}} = \mu_g + X_c\beta_g + \epsilon_{gk}$$

where X_c is the indicator for different cell types, and $\epsilon_{gk} \sim N(0, \sigma_g^2)$ represents the random effects for donor k. We want to test $H_0 : \beta_g = 0$.

3.3 Results

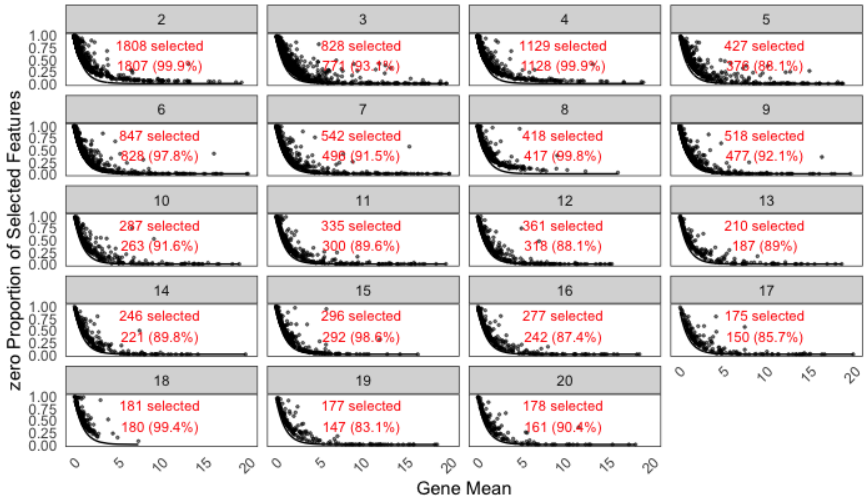
3.3.1 Application on different immune cell types in fallopian tube

In this study, the k-inflation test was utilized to identify and select features based on the proportions of zero counts and other small integer counts (ones, twos, threes, and fours). This approach allowed us to include more features that might be missed in original HIPPO.

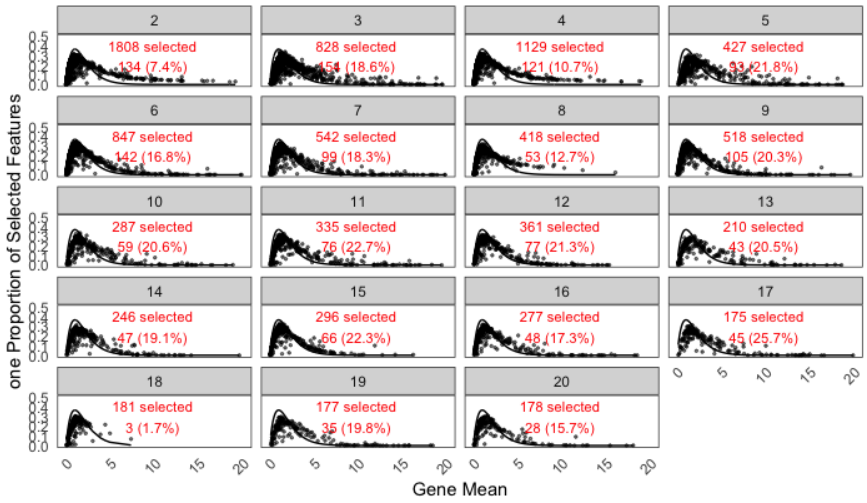
We applied the k-inflation test and the corresponding feature selection process on the immune cells in the fallopian tube, as detailed in Section 2.2.1. HIPPOx was configured to run 20 iterations, resulting in 20 distinct clusters. During the 20 runs of HIPPOx, each iteration helped refine the clustering process, ensuring that the observed proportions of zeros and small counts were better aligned with the expected distribution lines post-clustering.

By employing the k-inflation test and running multiple iterations of HIPPOx, we aimed to thoroughly explore the data and extract significant features that might have been overlooked by conventional methods. This comprehensive approach ensures a deeper understanding of the biological variability and heterogeneity present in the immune cells under study.

From Fig. 3.3 and Fig. A.11, it is shown that the number of selected genes decreases gradually due to the increasing homogeneity achieved with each round of clustering. Among



(a) Zero proportion



(b) One proportion

Figure 3.3: The percentage of k-proportion of selected features in each round of HIPPOx. The numbers in red indicate the number of selected features, and the percentage represents the contribution of selected genes passing the k-th inflation test.

all the selected features, almost all genes are zero inflated, contributing to more than 90 percent of the selected features. Additionally, higher order inflated genes, which include ones, twos, threes, and fours, contribute about up to 20 percent in some runs, especially fewer contribution in higher counts. This might due to the valid region restriction discussed in Sec. 3.2.1. However, these higher order inflated genes are also zero-inflated, indicating that the zero proportion remains a dominant and robust characteristic even as the count values increase.

This observation highlights the importance of zero-inflated genes in identifying and understanding cell-type heterogeneity. The gradual reduction in the number of selected genes suggests that as clustering refines the grouping of cells, the homogeneity within clusters increases, leading to fewer genes being identified as significant features. Despite this reduction, the persistent presence of zero-inflated genes across different runs highlights their crucial role in the feature selection process and their impact on the robustness and accuracy of the clustering results.

In summary, the analysis from Fig. 3.3 confirms that zero-inflated genes are a key factor in distinguishing cell types and that the k-inflation test effectively captures these critical features, contributing significantly to the understanding of gene expression patterns in immune cells within the fallopian tube.

3.3.2 Simulation study for DE analysis

Data generation

In this simulation, we generated 1200 genes and 500 cells from Poisson glmm. And it was repeated for 100 times. The cells came from five donors, and each of them provided 100 cells. We simulated the random effect ϵ_{gk} from $N(0, \sigma_g^2)$. The standard deviation σ_g were sampled from $N(1, 0.2^2)$, which is similar to what we observed from real data. Each cell was then randomly assigned to 2 cell types. The distribution of σ_g and the distribution of each

sampled donor effect are shown in Fig. 3.4.

Type 1 error rate design

To compare the type 1 error rate among different methods, the first 600 β_g were set to zero. That is, there's no cell-type effect for these genes (Fig. 3.5(a)). To generate counts for different expressed levels, every 200 μ_g were sampled from $\log U_{0.05}$, $\log U_{0.1}$ and $\log U_5$ respectively, where U_α means exponential distribution with mean α. We then assorted the final gene counts into three classes. If the gene mean is less then 2, we would consider the gene as lowly expressed. If the gene mean is more than 2 but less than 5, we would say the gene is moderately expressed. And a gene is highly expressed if the gene mean is greater than 5. Under this classification, there would be around 450 low, 50 moderate, 100 high in each simulation.

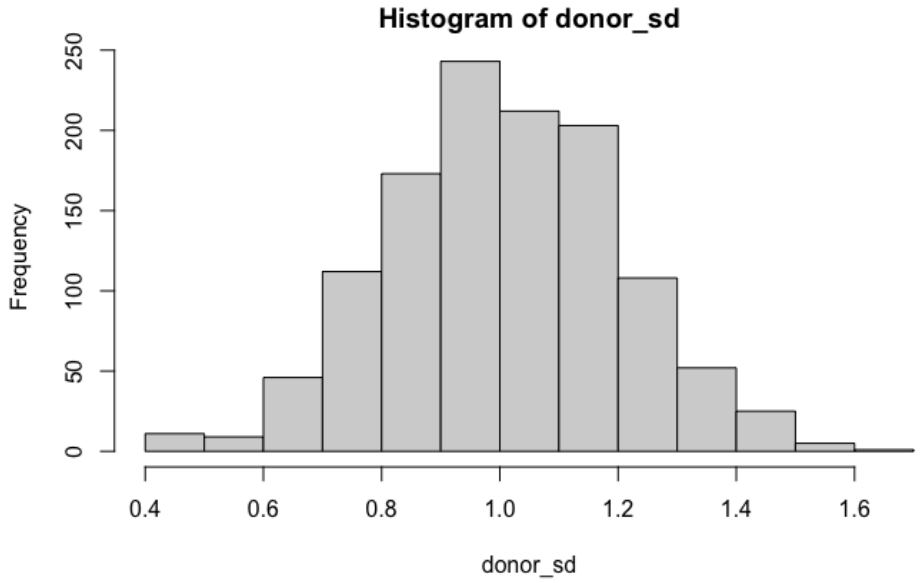
After the analysis from each method, we used benjamini hochberg correction on the p-values. If the adjusted p-value is lower than 0.05, it would be consider a type 1 error. The type 1 error rate was computed by the mean of type 1 error in each class.

We compared the performance of Poisson GLMM, Binomial GLMM, pseudobulk-DESeq2, and MAST under this simulation setting. As we see the p-values in Chapter 2, pseudobulk methods and MAST are both too conservative, hence resulting over controlled type 1 error rate (Fig. 3.5 (b)) and under-powered result (Fig. 3.6(b))

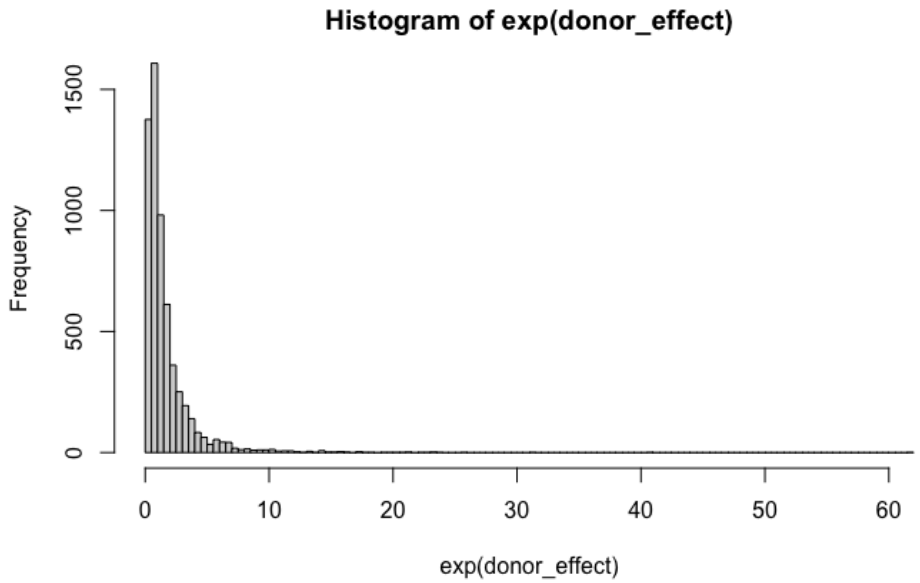
Power design

For the last 600 β_g, we assigned different levels of effect size on them. All the 600 μ_g were sampled from $\log U_{0.05}$, and every 200 β_g were set to be $\log 0.05$, $\log 0.2$ and $\log 5$ respectively. The genes were then classified into three classes based on the level of effect size (Fig. 3.6).

If the adjusted p-value is lower than 0.05, it would be consider a true discovery. The power was computed by the mean of type 1 error in each class.

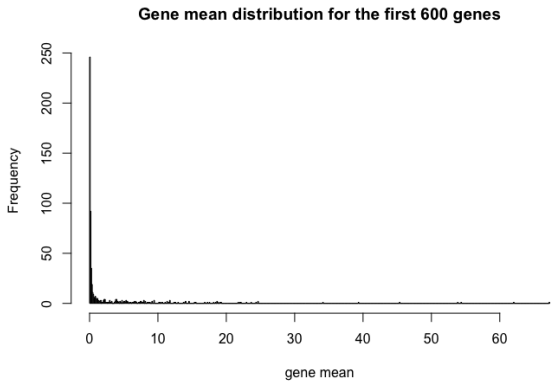


(a) Distribution of σ_g

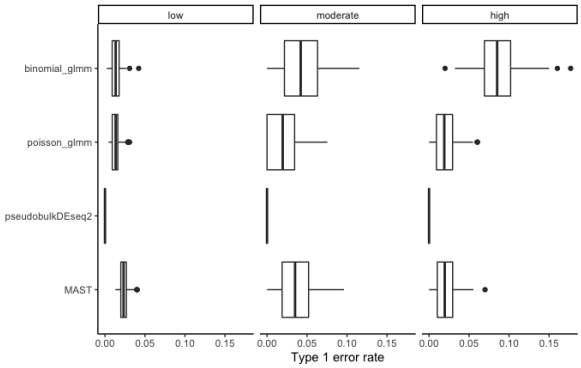


(b) Distribution of $\exp \epsilon_{gk}$

Figure 3.4: Donors effects

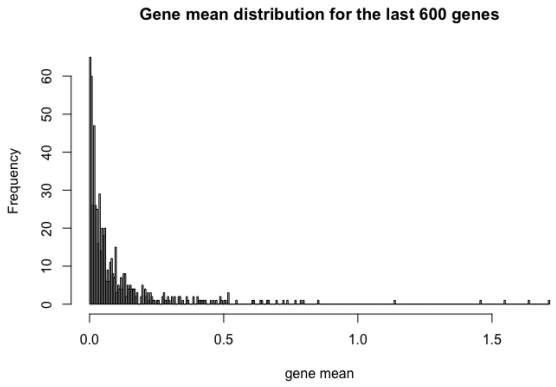


(a) Distribution of gene mean for the first 600 genes

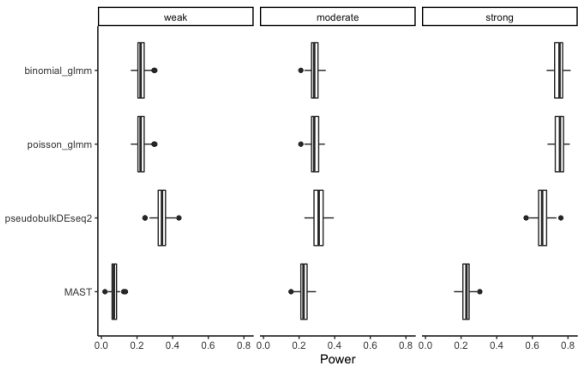


(b) Type 1 error rate

Figure 3.5: Simulation result for type 1 error rate.



(a) Distribution of gene mean for the last 600 genes



(b) Power

Figure 3.6: Simulation result for power.

3.4 Discussion

3.4.1 Potentials for the k proportion

The analysis in this chapter, as well as in the previous chapter, has provided substantial evidence that k proportions of the genes are powerful features in single-cell RNA data. Our next step is to apply this robustness in data that does not strictly follow a Poisson distribution. If we can extend this robustness to handle data modeled by a negative binomial distribution, or even non-identified distributions, we can leverage the proportion information in the data to perform differential expression analysis and further downstream analyses.

The negative binomial distribution, which accounts for overdispersion relative to the Poisson distribution, is sometimes used for modeling count data in single-cell RNA sequencing due to its ability to handle variability among cells. By adapting our methods to this distribution, we might be able to improve the robustness and accuracy of the analyses.

Furthermore, focusing on proportion information allows us to reduce the impact of outliers and extreme values that can distort mean-based measures. This approach enhances the stability of our statistical tests and makes them more reliable across various types of data distributions.

3.4.2 Contribution of higher order counts

After applying HIPPOx on real data, the additional contribution from higher order counts appears less promising. The restriction on the valid range for the k-inflation test limits the number of genes considered, especially for k greater than 1. As shown in Fig. 3.3, the percentage of exclusive genes identified from one proportion can be as high as 10 percent. Although the majority of significant features are identified through the zero proportion, the one proportion still has the potential to contribute valuable and informative features.

The results indicate that while zero-inflated genes are the primary source of significant

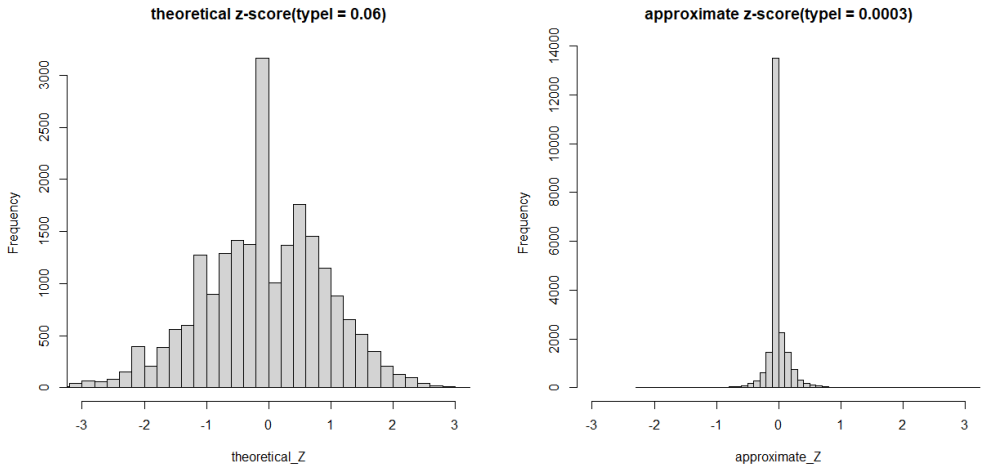
features, genes with higher order counts (such as ones) can also provide unique insights, which is exclusive to zero proportion. This observation suggests that, despite the limitations, examining proportions beyond zeros can enrich our understanding of the data.

Our next step involves applying HIPPOx to additional datasets to determine if it can help identify further subclusters. By exploring a broader range of data, we hope to validate the utility of higher order counts and assess whether HIPPOx can consistently reveal informative features across different contexts.

3.4.3 p-value accuracy

We performed bootstrap simulations and found that the p-values were too conservative. Several factors could contribute to this observation.

First, many genes have proportions that are very close to 1 or 0. In these cases, the normal approximation performs poorly, leading to inaccuracies. Second, as discussed in Kim et al. [2020], the sample mean itself is random. The standard error and the expected mean could differ from those calculated using the true λ. This discrepancy is illustrated in Fig. 3.7, which shows the distribution of z-scores for zero proportions. The figure highlights the overestimation of the standard error when the sample mean $\bar{X}_g$ is used in place of the true λ_g. Third, the p-value of the k-inflation test for $k \geq 1$ may require adjustment because we filter out genes whose sample means do not fall within the valid region. This filtering could skew the distribution and lead to conservative p-values.



(a) Distribution of Z-score calculated with true λ_g

(b) Distribution of Z-score calculated with sample mean $\bar{X}_g$

Figure 3.7: Distribution of z-score

CHAPTER 4

MIXTURE POISSON GENERALIZED LINEAR MODEL FOR ANALYZING DIFFERENTIAL METHYLATION

4.1 Introduction

4.1.1 m6A modification

While transcriptional regulation—the synthesis of messenger RNA (mRNA)—is crucial and has been studied broadly, it is protein expression that ultimately determines biological phenotypes. The production of proteins is influenced by a variety of post-transcriptional regulatory mechanisms, such as the structure of mRNA, the action of microRNAs, and the processes governing mRNA translation (Fabian et al. [2010], Chekulaeva and Filipowicz [2009], Zhao et al. [2017], Glisovic et al. [2008]). These post-transcriptional regulations play a fundamental role in controlling protein levels and their localization within the cell, thereby impacting all biological processes.

In 2011, Chuan He's lab uncovered a fundamental mechanism that broadly controls protein expression at the post-transcriptional level: reversible and dynamic modifications of mRNA and long non-coding RNA (lncRNA) (Jia et al. [2011], Liu et al. [2014], Fu et al. [2014]). This discovery has sparked extensive research into profiling various mRNA modifications, such as N6-methyladenosine and pseudouridine, using antibodies and chemical reactions respectively (Zhao et al. [2017]).

Among various RNA modifications, N6-methyladenosine (m6A) has obtained significant attention due to its prevalence and critical regulatory roles in mRNA metabolism (Lan et al. [2019], Yang et al. [2020], Livneh et al. [2020], He and He [2021], Schaefer [2021]). The m6A modification affects almost every phase of mRNA metabolism and function, impacting diverse biological processes. Thus, m6A studies symbolize the concept of the "epitranscrip-

tome."

The functional significance and implementation of m6A are carried out by three groups of proteins: "writers" that install the modification, "erasers" that remove it, and "readers" that bind or recognize m6A to determine the cellular fate of the modified mRNA/lncRNA. This dynamic interplay between writers, erasers, and readers ensures precise regulation of gene expression at the RNA level, highlighting the complexity and importance of post-transcriptional modifications in cellular biology.

4.1.2 Statistical methods for differential methylation

MeRIP-seq, or Methylated RNA Immunoprecipitation Sequencing, is a powerful technique used to study RNA modifications, particularly m6A (Meyer et al. [2012], Dominissini et al. [2012], Dominissini et al. [2016]). This method combines immunoprecipitation of methylated RNA fragments with high-throughput sequencing to identify and quantify RNA methylation sites across the transcriptome. The process enables researchers to analyze differential methylation effectively. The analysis involves two main applications: (1) MeRIP-seq samples from specific phenotypes or experimental conditions are analyzed to determine the locations of RNA modifications. This involves using peak calling algorithms to identify regions enriched in m6A modifications by comparing immunoprecipitated RNA (IP) samples with input RNA controls. (2) MeRIP-seq samples from different phenotypical or experimental groups are compared to identify differentially methylated loci. This involves quantifying the methylation levels in identified peaks for both IP and input samples, then performing statistical tests to find significant differences in methylation between conditions.

Several methods have been developed and applied to analyze differential methylation in MeRIP-seq data. ExomePeak uses Fisher's exact test to identify differentially methylated regions (Meng et al. [2014], Meng et al. [2013]). The later version incorporates a likelihood ratio test based on the binomial distribution, known as "bltest". employs a beta-binomial

model to infer differential peaks, accounting for overdispersion in the count data. MeTPeak employs a beta-binomial model to infer differential peaks, accounting for overdispersion in the count data (Cui et al. [2016]. MeTDiff, DRME and QNB use models based on the negative binomial distribution (Cui et al. [2015], Liu et al. [2016], Liu et al. [2017]).

In 2019, Zhang et al. [2019] raised some existing problems in previous methods. For example, current methods designed for small sample sizes do not account for confounding factors like age and gender. Differential gene expression (DE) tools like edgeR (Robinson et al. [2010], DESeq2 (Love et al. [2014]), and Sleuth (Pimentel et al. [2017]) are compatible with complex study designs but are not tailored for MeRIP-seq data. QNB combines local read counts from both INPUT and IP libraries to estimate expression levels, which can confound pre-IP and post-IP measurements. This can lead to biased expression level estimation, resulting in substantial false discoveries in DM analysis.

Alternatively, Zhang et al. [2019] proposed a Poisson random effect model to combat the limitation mentioned above. They model the post-IP enrichment counts Y_i in i-th samples as follows:

$$Y_i \sim Poi\left(\lambda_i\right) \qquad \log\left(\lambda_i\right) = \mu + \boldsymbol{X_i}\boldsymbol{\beta} + e_i = \mu + X_{i0}\ \beta_0 + \sum\nolimits_{j=1}^{k} X_{ij}\ \beta_j + e_i$$

where λ_i is the mean of a Poisson distribution, μ is a gene-specific intercept, X_i is a vector including the indicator of the groups of interest X_{i0} and covariates $X_{ij}(j = 1, ..., k)$ for i-th sample, β represent associated coefficients and e_i is a random effect following a Log-Gamma distribution with a scale parameter ψ and mean equal to 1, i.e., $e_i \in$ logGamma(ψ,ψ). This novel prior leads to closed-form solutions to the Poisson random effect model and accelerates the computation. The differential analysis is equivalent to test against the null hypothesis $\beta_0 = 0$. This generalized linear model framework allows the inclusion of covariates in β.

4.1.3 Cell type-specific differential methylation analysis

Like gene expression, RNA methylation is differentially regulated across cell types and developmental stages, reflecting its crucial role in cellular and developmental processes. Some studies on m6A have observed that RNA methylation patterns can vary significantly between different cell types (Zhou et al. [2017], Molinie et al. [2016], Perlegos et al. [2024]). Specific methylation marks can define the identity and function of a cell, influencing gene expression and cellular behavior. In addidtion, RNA methylation can affect RNA stability, splicing, translation, and localization. By regulating these processes, RNA methylation plays a critical role in fine-tuning gene expression in a context-dependent manner.

Based on the RADAR framework, we have proposed a mixture Poisson generalized linear mixed model for conducting cell type-specific differential methylation analysis. This approach leverages the estimates of cellular compositions derived from the INPUT library, enabling us to test for cell type-specific differential methylation without the need for direct experimental measurement of cell type-specific methylation. This methodology not only facilitates hypothesis generation on biological functions but also does so cost-effectively, presenting new opportunities for advancing research in the field. We will provide detailed explanations of the model formulation, algorithm, and simulation results for the framework in the following sections.

4.2 Methods and materials

4.2.1 Mixture Poisson generalized linear model

In this chapter, we assume that the relative proportions of cell types are known. Given the estimated relative proportions of different cell types $\hat{\alpha}_{jk}(j = 1, ..., p)$ for each individual sample k, we assume m6A enrichment in k-th sample is a mixture of m6A enrichment levels across all cell types: $Z_k = \sum_{j=1}^{p} \hat{\alpha}_{jk} Z_{jk}$, where Zs represent read counts and the enrichment

level in a particular cell type Z_{jk} is not observed.

$$
\begin{aligned}
Z_k \mid e_k &\sim \text{Poisson}(\lambda_k) \qquad k = 1, ..., n \\
\lambda_k &= \sum_{j=1}^{p} \hat{\alpha}_{jk}\lambda_{jk} = \sum_{j=1}^{p} \hat{\alpha}_{jk} \exp(\mu + \beta_j X_k + e_k) = \sum_{j=1}^{p} q_{jk} \\
f_{(\gamma,\theta)}(e_k) &= \frac{e^{\theta e_k - \frac{e^{e_k}}{\gamma}}}{\gamma^{\theta}\Gamma(\theta)} \qquad \text{choose } \gamma = e^{-\psi(\theta)}
\end{aligned}
$$

Z_k represents read counts in the k-th sample, and $\hat{\alpha}_{jk}(j = 1, ..., p)$ are given estimated relative proportions of different cell types for each sample k. The over-dispersion is captured by e_k following a Log-Gamma distribution with shape θ and rate $\frac{1}{\gamma}$ such that e_k has zero mean. X_k is the indicator of group of interest for the k-th sample.

Due to the specific choice of distribution (Log-Gamma) for the random effect, we are able to compute the exact form of the likelihood function $L(\mu, \boldsymbol{\beta}, \theta; z_k)$.

$$
\begin{aligned}
L(\mu, \boldsymbol{\beta}, \theta; z_k) &= \int f(z_k \mid e_k) f(e_k) de_k \\
&= \frac{1}{z_k!} \int (e^{-\Delta_k})^{e^{e_k}} \Delta_k^{z_k} e^{e_k z_k + \theta e_k - \frac{1}{\gamma} e^{e_k}} \frac{1}{\gamma^{\theta}\Gamma(\theta)} de_k \\
&= \frac{1}{z_k!} \frac{(\Delta_k + 1/\gamma)^{-(\theta + z_k)}\Gamma(\theta + z_k)}{\gamma^{\theta}\Gamma(\theta)} \Delta_k^{z_k}
\end{aligned}
$$

Taking the logarithm of the likelihood function transforms it into a summation of functions of $\mu, \boldsymbol{\beta}, \theta$ and $\boldsymbol{Z}$. This transformation simplifies the expression, making it more manageable for analytical and computational purposes.

$$
\begin{aligned}
l(\mu, \boldsymbol{\beta}, \theta; \boldsymbol{Z}) = &\sum_{k=1}^{n} - \log(Z_k!) + Z_k \log \Delta_k - (\theta + Z_k) \log(\Delta_k + \frac{1}{\gamma}) + \log \Gamma(\theta + Z_k) \\
&- \theta \log \gamma - \log \Gamma(\theta) \qquad (4.1) \\
\text{where } \gamma = &e^{-\psi(\theta)}, \Delta_k = \sum_{j=1}^{p} \hat{\alpha}_{jk} \exp(\mu + \beta_j X_k)
\end{aligned}
$$

4.2.2 Algorithm

To find the optimizer, we need to start the optimization process from a set of close initial values. To obtain a good initial guess, we attempt to create fake counts Y_{jk} to mimic the Z_{jk} by fitting a linear regression model

$$Z_k = a + \sum_{j=1}^{p} b_j \hat{\alpha}_{jk} X_k + \epsilon_k.$$

We then use the estimated coefficients to create:

$$Y_{jk} = \max\{0, (\hat{a} + \hat{b}_j X_k) \cdot W_k\} + 1,$$

where $W_k \overset{iid}{\sim} Gamma(\theta, e^{-\psi(\theta)})$. This process helps to generate initial values that are close to the expected values, thereby improving the efficiency and accuracy of the optimization process.

Algorithm 1: Maximize log-likelihood

Input: $X \in \mathbf{R}^n, Z \in \mathbf{R}^n, \hat{\alpha} \in \mathbf{R}^{p \times n}, \theta_0$
Output: MLE $(\mu^*, \boldsymbol{\beta}^*, \theta^*)$ of (4.1) and $l(\mu^*, \boldsymbol{\beta}^*, \theta^*; \boldsymbol{Z})$

1 Fit linear regression models $Z_k = a + \sum_{j=1}^{p} b_j \hat{\alpha}_{jk} X_k + \epsilon_k$.
2 Set $Y_{jk} = \max\{0, (\hat{a} + \hat{b}_j X_k) \cdot W_k\} + 1$, where $W_k \overset{iid}{\sim} Gamma(\theta, e^{-\psi(\theta)})$.
3 Fit linear regression models $\log Y_{jk} = \mu + \beta_j X_k + \epsilon_{jk}$.
4 Set $\mu_0 = \hat{\mu}; \beta_{j_0} = \hat{\beta}_j$ to be the initial values.
5 Optimize the log-likelihood (4.1) from initial valued $(\mu_0, \boldsymbol{\beta_0}, \theta_0)$
6 $(\mu^*, \boldsymbol{\beta}^*, \theta^*) = \arg\max_{\mu, \boldsymbol{\beta}, \theta} l(\mu, \boldsymbol{\beta}, \theta; \boldsymbol{Z})$

4.2.3 Likelihood ratio test

The ultimate goal is to perform differential methylation analysis to identify cell type-specific methylation patterns. To test $H_0 : \beta_j = 0$, consider likelihood ratio

$$LR = -2 \ln \left[\frac{\sup_{\mu,\theta,\beta_j=0} \mathcal{L}(\mu, \boldsymbol{\beta}, \theta)}{\sup_{\mu,\theta,\boldsymbol{\beta}} \mathcal{L}(\mu, \boldsymbol{\beta}, \theta)} \right]$$

which follows asymptotic χ^2_1-distribution under the null hypothesis. Then we reject H_0 if $LR > \chi^2_{1,1-\alpha}$

4.3 Simulation results

n this section, we present the simulation results to validate the accuracy and robustness of our algorithm. We begin by examining the initial guesses for (μ, $\boldsymbol{\beta}$) across various scenarios to assess the reliability of the starting points in our optimization process. Subsequently, we provide a comprehensive analysis of the final estimates obtained from our algorithm, demonstrating their convergence and consistency. The effectiveness of the algorithm is further evaluated through the likelihood ratio test under the null setting. Additionally, we analyze the statistical power and false discovery rate (FDR) to ensure the algorithm's sensitivity and specificity in detecting true differential expression. Overall, these results confirm the efficacy of our algorithm in various testing conditions, highlighting its potential application in real-world single-cell RNA sequencing data analysis.

4.3.1 Initial values and estimates

In our algorithm, we first fit linear models to obtain reasonable initial values for optimizing the log-likelihood function. To evaluate the fairness of these initial values, we set up two scenarios. The first scenario is $\mu = 5$, $\beta = (0, -0.5, 0.5)$, and $\theta = 10$, which mimics a small

effect size on the coefficients (Fig. 4.1, Fig. 4.2). The second scenario is $\mu = 5$, $\beta = (0, 1, 3)$, and $\theta = 10$, representing a larger effect size (Fig. 4.3). These setups allow us to examine if the initial values provide a fair starting point for the optimization process.

The results show that both the initial values and the final estimates are bell-shaped and centered around the true parameters. Despite the relatively large variation in the initial values, they still provide a reasonable starting point for the optimization procedure, leading to satisfactory performance. This indicates that our algorithm is robust and capable of producing accurate estimates even with a diverse range of initial guesses.

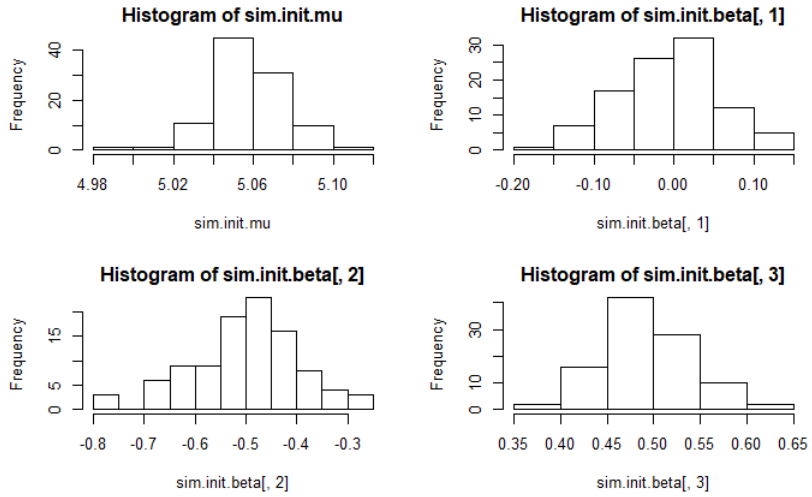


Figure 4.1: Histograms of initial guess for the parameters under small effect.

4.3.2 LRT for null settings

We further tested the p-values of the likelihood ratio test under the null settings with parameters $p = 2$, $n = 1000$, $\mu = 5$, $\boldsymbol{\beta} = (0, 0)$, and $\theta = 10$, specifically testing the null hypothesis $H_0 : \beta_1 = 0$. As shown in Fig. 4.4, the histogram of the p-values is fairly flat, indicating that the likelihood ratio test behaves as expected under the null hypothesis, providing valid p-values and demonstrating the appropriateness of our approach.

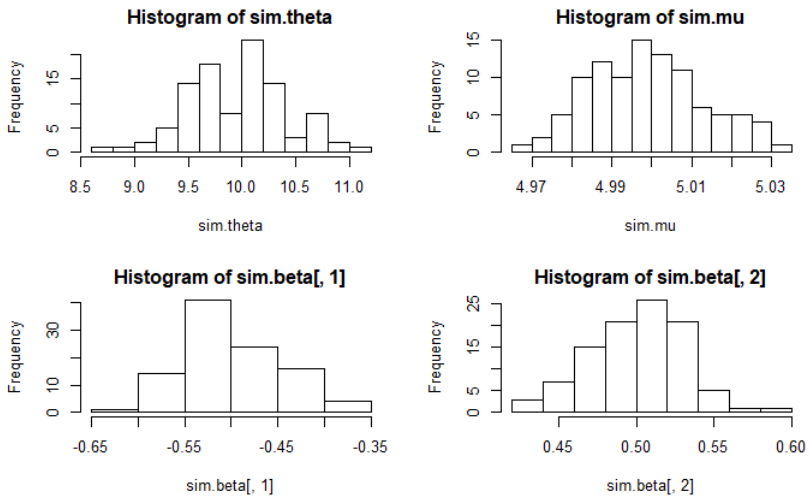


Figure 4.2: Histograms of estimates for the parameters under small effect.

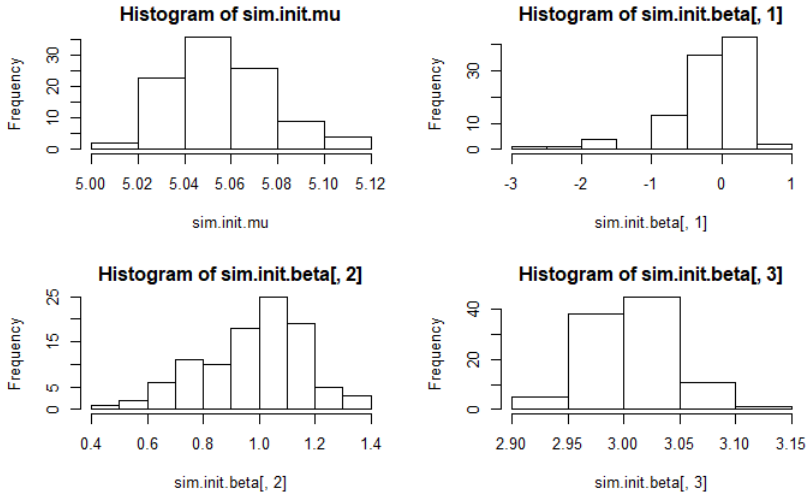


Figure 4.3: Histograms of initial guess for the parameters under larger effect.

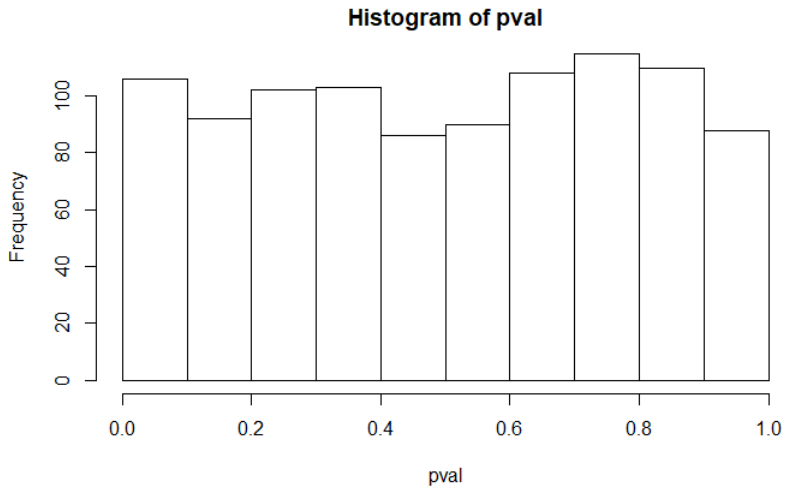


Figure 4.4: Histograms of p-values under the null setting.

4.3.3 Power, FDR, accuracy, type1 error

The following results in Fig. 4.5 are based on simulations with $\mu = 5, \theta = 10$, and the sample size ranges from $n = 50, 100, 200, 400$. For each n, we sampled 500 times. Each time for all βs, we randomly generated the proportion $\hat{\alpha}_{jk}$ and indicators X_k. The count data was then generated by our model assumption.

As the sample size increases, the power, FDR, accuracy, and type I error all improve visibly. The power reaches almost 1 when n is only 500, demonstrating the effectiveness of the test in detecting true effects with relatively small sample sizes. The FDR is well-controlled under 0.05 across all scenarios, ensuring that the rate of false discoveries remains low. While the type I error is slightly elevated, it remains under 0.06, indicating a reasonable balance between sensitivity and specificity. The estimates for μ and θ are close to the true parameters, suggesting that the model accurately captures these aspects of the data. However, the accuracy of $\boldsymbol{\beta}$ appears to be off for larger effect sizes, indicating that further investigation into the distribution or the likelihood function may be necessary to address this discrepancy.

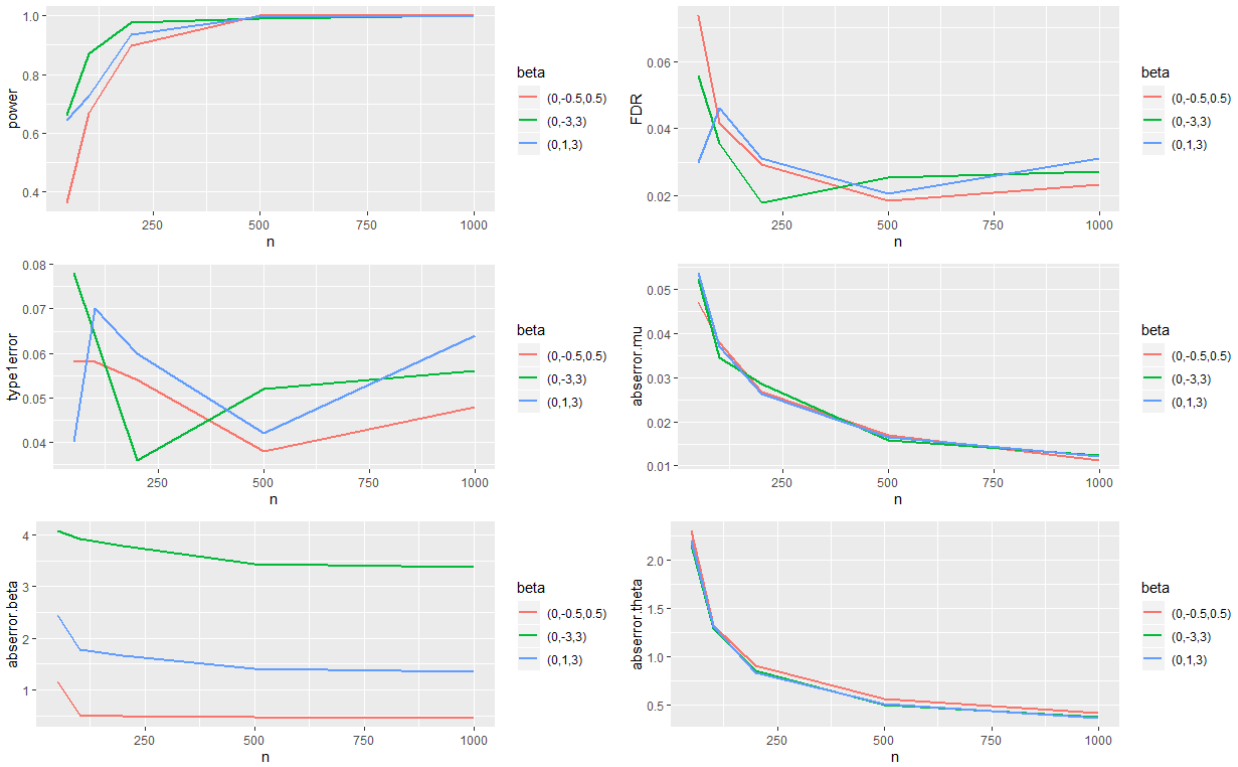


Figure 4.5: Power, FDR, accuracy and type1 error under different scenarios.

REFERENCES

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21:1–17, 2020.

Marina Chekulaeva and Witold Filipowicz. Mechanisms of mirna-mediated post-transcriptional regulation in animal cells. *Current opinion in cell biology*, 21(3):452–460, 2009.

Mengjie Chen and Xiang Zhou. Controlling for confounding effects in single cell rna sequencing studies using both control and target genes. *Scientific reports*, 7(1):13587, 2017.

Mengjie Chen and Xiang Zhou. Viper: variability-preserving imputation for accurate gene expression recovery in single-cell rna sequencing studies. *Genome biology*, 19(1):196, 2018.

Mengjie Chen, Qi Zhan, Zepeng Mu, Lili Wang, Zhaohui Zheng, Jinlin Miao, Ping Zhu, and Yang I Li. Alignment of single-cell rna-seq samples without overcorrection using kernel density matching. *Genome Research*, 31(4):698–712, 2021.

D.G. Clayton. Generalized linear mixed models. *Markov Chain Monte Carlo in Practice*, 1: 275–302, 1996.

H.L. Crowell, C. Soneson, PL. Germain, et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat Commun*, 11:6077, 2020.

Xiaodong Cui, Lin Zhang, Jia Meng, Manjeet K Rao, Yidong Chen, and Yufei Huang. Metdiff: a novel differential rna methylation analysis for merip-seq data. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(2):526–534, 2015.

Xiaodong Cui, Jia Meng, Shaowu Zhang, Yidong Chen, and Yufei Huang. A novel algorithm for calling mrna m 6 a peaks by modeling biological variances in merip-seq data. *Bioinformatics*, 32(12):i378–i385, 2016.

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

Samarendra Das, Anil Rai, and Shesh N Rai. Differential expression analysis of single-cell rna-seq data: current statistical approaches and outstanding challenges. *Entropy*, 24(7): 995, 2022.

Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, et al. Topology of the human and mouse m6a rna methylomes revealed by m6a-seq. *Nature*, 485(7397):201–206, 2012.

Dan Dominissini, Sigrid Nachtergaele, Sharon Moshitch-Moshkovitz, Eyal Peer, Nitzan Kol, Moshe Shay Ben-Haim, Qing Dai, Ayelet Di Segni, Mali Salmon-Divon, Wesley C Clark, et al. The dynamic n 1-methyladenosine methylome in eukaryotic messenger rna. *Nature*, 530(7591):441–446, 2016.

Marc Robert Fabian, Nahum Sonenberg, and Witold Filipowicz. Regulation of mrna translation and stability by microrna. *Annual review of biochemistry*, 79:351–379, 2010.

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

Ye Fu, Dan Dominissini, Gideon Rechavi, and Chuan He. Gene expression regulation mediated through reversible m6a rna methylation. *Nature Reviews Genetics*, 15(5):293–306, 2014.

Tina Glisovic, Jennifer L Bachorik, Jeongsik Yong, and Gideon Dreyfuss. Rna-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–1986, 2008.

Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J Garry. Drimpute: imputing dropout events in single cell rna sequencing data. *BMC bioinformatics*, 19:1–10, 2018.

William W Greenwald, Joshua Chiou, Jian Yan, Yunjiang Qiu, Ning Dai, Allen Wang, Naoki Nariai, Anthony Aylward, Jee Yun Han, Nikita Kadakia, et al. Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nature communications*, 10(1):2078, 2019.

Alexandra Grubman, Gabriel Chew, John F Ouyang, Guizhi Sun, Xin Yi Choo, Catriona McLean, Rebecca K Simmons, Sam Buckberry, Dulce B Vargas-Landin, Daniel Poppe, et al. A single-cell atlas of entorhinal cortex from individuals with alzheimer's disease reveals cell-type-specific gene expression regulation. *Nature neuroscience*, 22(12):2087–2097, 2019.

Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1): 296, 2019.

P Cody He and Chuan He. m6a rna methylation: from mechanisms to therapeutic potential. *The EMBO journal*, 40(3):e105977, 2021.

Jialu Hu, Mengjie Chen, and Xiang Zhou. Effective and scalable single-cell data alignment with non-linear canonical correlation analysis. *Nucleic acids research*, 50(4):e21–e21, 2022.

Guifang Jia, YE Fu, XU Zhao, Qing Dai, Guanqun Zheng, Ying Yang, Chengqi Yi, Tomas Lindahl, Tao Pan, Yun-Gui Yang, et al. N 6-methyladenosine in nuclear rna is a major substrate of the obesity-associated fto. *Nature chemical biology*, 7(12):885–887, 2011.

H.M. Kang et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature Biotechnology*, 36:89–94, 2018.

Tae Hyun Kim, Xiang Zhou, and Mengjie Chen. Demystifying "drop-outs" in single-cell umi data. *Genome biology*, 21(1):196, 2020.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

Qing Lan, Pei Y Liu, Jacob Haase, Jessica L Bell, Stefan Hüttelmaier, and Tao Liu. The critical role of rna m6a methylation in cancer. *Cancer research*, 79(7):1285–1292, 2019.

Jan Lause, Philipp Berens, and Dmitry Kobak. Analytic pearson residuals for normalization of single-cell rna-seq umi data. *Genome biology*, 22:1–20, 2021.

Nathan Lawlor, Joshy George, Mohan Bolisetty, Romy Kursawe, Lili Sun, V Sivakamasundari, Ina Kycia, Paul Robson, and Michael L Stitzel. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome research*, 27(2):208–222, 2017.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

Ernst Lengyel, Yan Li, Melanie Weigert, Lisha Zhu, Heather Eckart, Melissa Javellana, Sarah Ackroyd, Jason Xiao, Susan Olalekan, Dianne Glass, et al. A molecular atlas of the human postmenopausal fallopian tube and ovary from single-cell rna and atac sequencing. *Cell Reports*, 41(12), 2022.

Peipei Li, Yongjun Piao, Ho Sun Shon, and Keun Ho Ryu. Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data. *BMC bioinformatics*, 16:1–9, 2015.

Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):997, 2018.

Jianzhao Liu, Yanan Yue, Dali Han, Xiao Wang, Ye Fu, Liang Zhang, Guifang Jia, Miao Yu, Zhike Lu, Xin Deng, et al. A mettl3–mettl14 complex mediates mammalian nuclear rna n 6-adenosine methylation. *Nature chemical biology*, 10(2):93–95, 2014.

Lian Liu, Shao-Wu Zhang, Fan Gao, Yixin Zhang, Yufei Huang, Runsheng Chen, and Jia Meng. Drme: count-based differential rna methylation analysis at small sample size scenario. *Analytical Biochemistry*, 499:15–23, 2016.

Lian Liu, Shao-Wu Zhang, Yufei Huang, and Jia Meng. Qnb: differential rna methylation analysis for count-based small-sample sequencing data with a quad-negative binomial model. *BMC bioinformatics*, 18:1–12, 2017.

Ido Livneh, Sharon Moshitch-Moshkovitz, Ninette Amariglio, Gideon Rechavi, and Dan Dominissini. The m6a epitranscriptome: transcriptome plasticity in brain development and function. *Nature Reviews Neuroscience*, 21(1):36–51, 2020.

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

Nicholas Lytal, Di Ran, and Lingling An. Normalization methods on single-cell rna-seq data: an empirical survey. *Frontiers in genetics*, 11:501166, 2020.

Jia Meng, Xiaodong Cui, Manjeet K Rao, Yidong Chen, and Yufei Huang. Exome-based analysis for rna epigenome sequencing data. *Bioinformatics*, 29(12):1565–1567, 2013.

Jia Meng, Zhiliang Lu, Hui Liu, Lin Zhang, Shaowu Zhang, Yidong Chen, Manjeet K Rao, and Yufei Huang. A protocol for rna methylation differential analysis with merip-seq data and exomepeak r/bioconductor package. *Methods*, 69(3):274–281, 2014.

Kate D Meyer, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christopher E Mason, and Samie R Jaffrey. Comprehensive analysis of mrna methylation reveals enrichment in 3' utrs and near stop codons. *Cell*, 149(7):1635–1646, 2012.

Benoit Molinie, Jinkai Wang, Kok Seong Lim, Roman Hillebrand, Zhi-xiang Lu, Nicholas Van Wittenberghe, Benjamin D Howard, Kaveh Daneshvar, Alan C Mullen, Peter Dedon, et al. m6a-laic-seq reveals the census and complexity of the m6a epitranscriptome. *Nature methods*, 13(8):692–698, 2016.

Alexandra E Perlegos, China N Byrns, and Nancy M Bonini. Cell type-specific regulation of m6a modified rnas in the aging drosophila brain. *Aging Cell*, page e14076, 2024.

Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16:1–10, 2015.

Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of rna-seq incorporating quantification uncertainty. *Nature methods*, 14(7):687–690, 2017.

Peng Qiu. Embracing the dropouts in single-cell rna-seq analysis. *Nature communications*, 11(1):1169, 2020.

Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11:1–9, 2010.

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

Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014.

Matthias R Schaefer. The regulation of rna modification systems: the next frontier in epitranscriptomics? *Genes*, 12(3):345, 2021.

Ramona Schmid, Patrick Baum, Carina Ittrich, Katrin Fundel-Clemens, Wolfgang Huber, Benedikt Brors, Roland Eils, Andreas Weith, Detlev Mennerich, and Karsten Quast. Comparison of normalization methods for illumina beadchip humanht-12 v3. *BMC genomics*, 11:1–17, 2010.

Jordan W Squair, Matthieu Gautier, Claudia Kathe, Mark A Anderson, Nicholas D James, Thomas H Hutson, Rémi Hudelle, Taha Qaiser, Kaya JE Matson, Quentin Barraud, et al. Confronting false discoveries in single-cell differential expression. *Nature communications*, 12(1):5692, 2021.

Valentine Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2): 147–150, 2020.

Sam Tracy, Guo-Cheng Yuan, and Ruben Dries. Rescue: imputing dropout events in single-cell rna-sequencing data. *BMC bioinformatics*, 20:1–11, 2019.

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.

Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):380, 2019.

Chuan Yang, Yiyang Hu, Bo Zhou, Yulu Bao, Zhibin Li, Chunli Gong, Huan Yang, Sumin Wang, and Yufeng Xiao. The role of m6a modification in physiology and disease. *Cell death & disease*, 11(11):960, 2020.

Yang Yang, Hongjian Sun, Yu Zhang, Tiefu Zhang, Jialei Gong, Yunbo Wei, Yong-Gang Duan, Minglei Shu, Yuchen Yang, Di Wu, et al. Dimensionality reduction by umap

reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell reports*, 36 (4), 2021.

Zijie Zhang, Qi Zhan, Mark Eckert, Allen Zhu, Agnieszka Chryplewicz, Dario F De Jesus, Decheng Ren, Rohit N Kulkarni, Ernst Lengyel, Chuan He, et al. Radar: differential analysis of merip-seq data with a random effect model. *Genome biology*, 20:1–17, 2019.

Boxuan Simen Zhao, Ian A Roundtree, and Chuan He. Post-transcriptional gene regulation by mrna modifications. *Nature reviews Molecular cell biology*, 18(1):31–42, 2017.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.

Chan Zhou, Benoit Molinie, Kaveh Daneshvar, Joshua V Pondick, Jinkai Wang, Nicholas Van Wittenberghe, Yi Xing, Cosmas C Giallourakis, and Alan C Mullen. Genome-wide maps of m6a circrnas identify widespread and cell-type-specific methylation patterns that are distinct from mrnas. *Cell reports*, 20(9):2262–2276, 2017.

J Zyprych-Walczak, A Szabelska, L Handschuh, K Górczak, K Klamecka, M Figlerowicz, and I Siatkowski. The impact of normalization methods on rna-seq data analysis. *BioMed research international*, 2015(1):621690, 2015.

APPENDIX A

SUPPLEMENTARY FIGURES

A.1 Supplementary Figures for Chapter 2

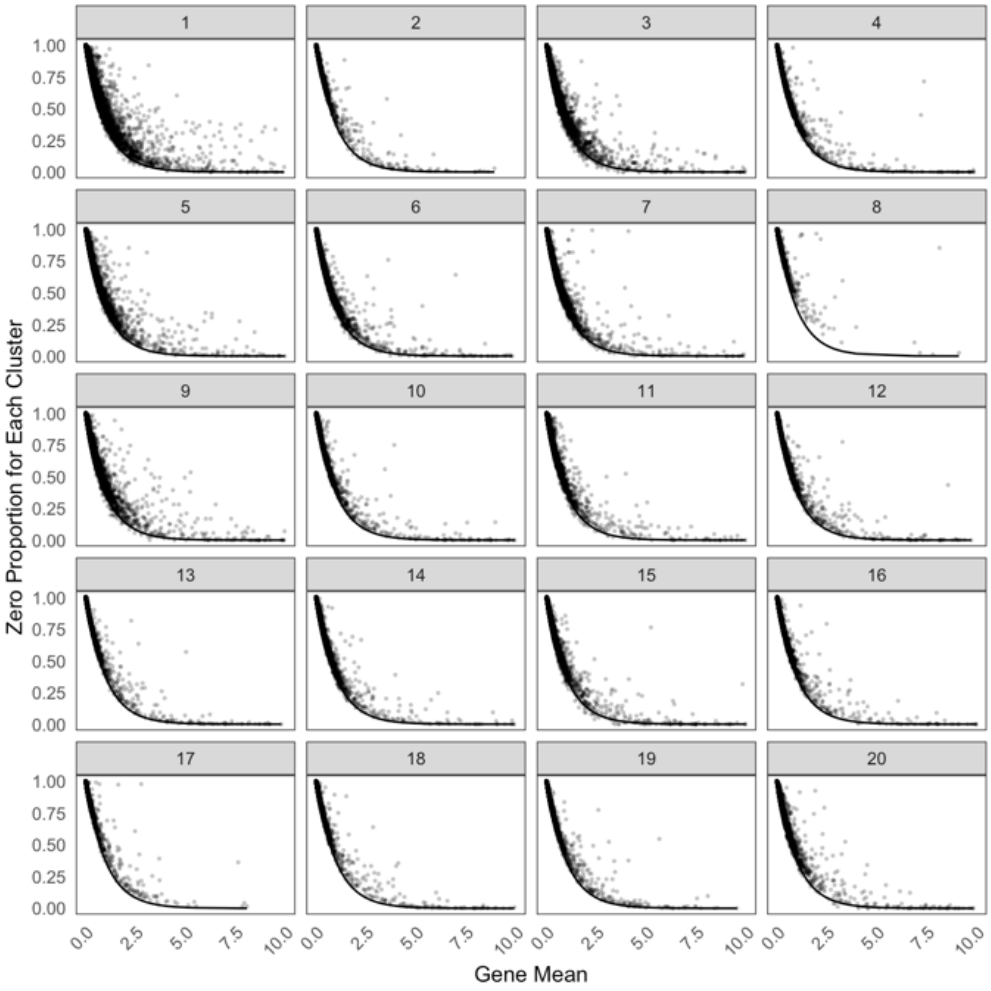
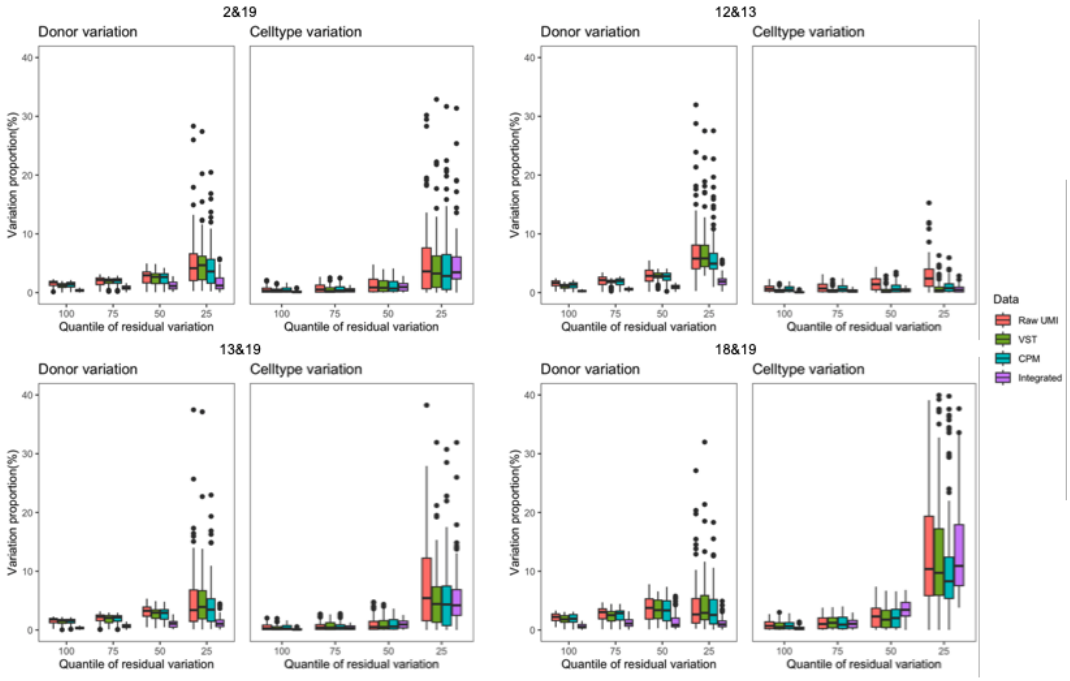
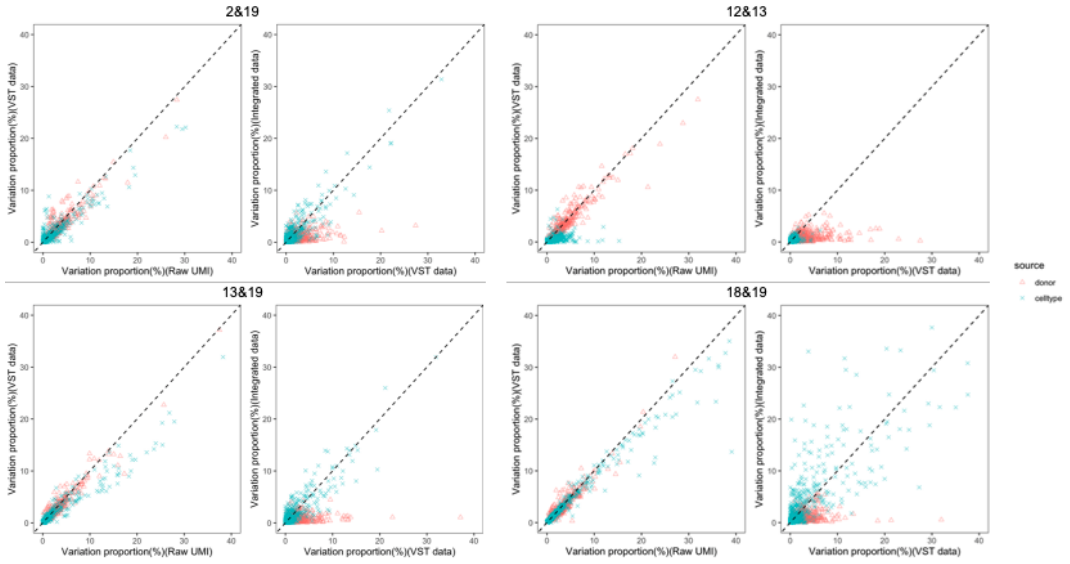


Figure A.1: Zero proportion plots for each cluster obtained by HIPPO for case study 1.

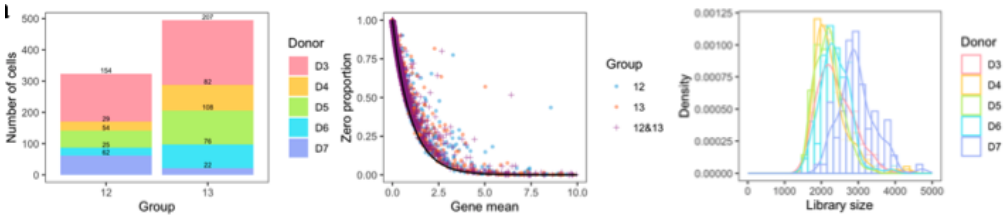


(a)

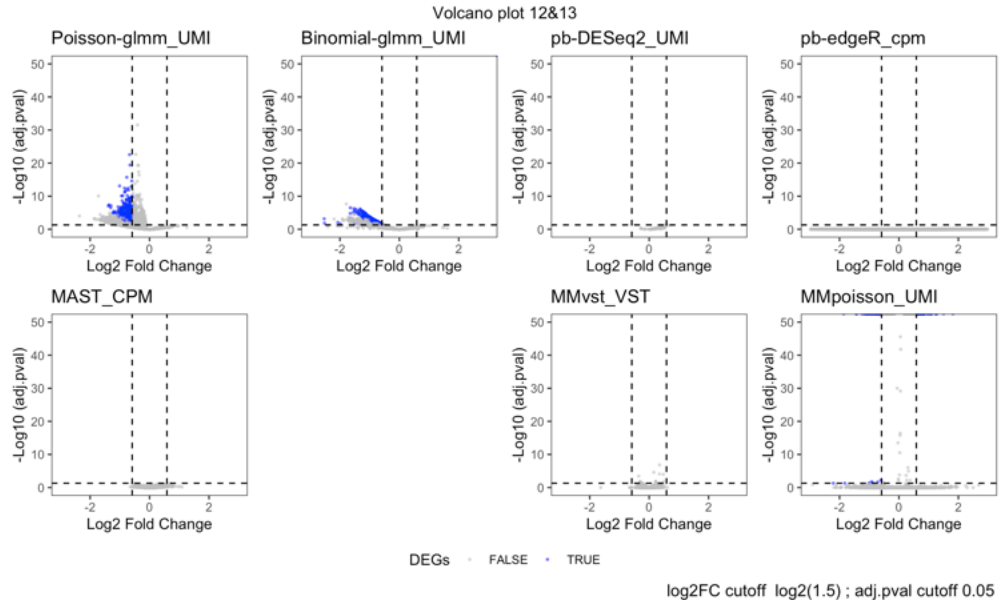


(b)

Figure A.2: Additional variation proportion analysis results. a) Boxplots of donor variation and cell type variation grouped by quartiles of residual variation, displayed in different pairs and different data sources. b) Scatter plots of variation proportions for donor effects and cell type effects in different pairs and different data sources.

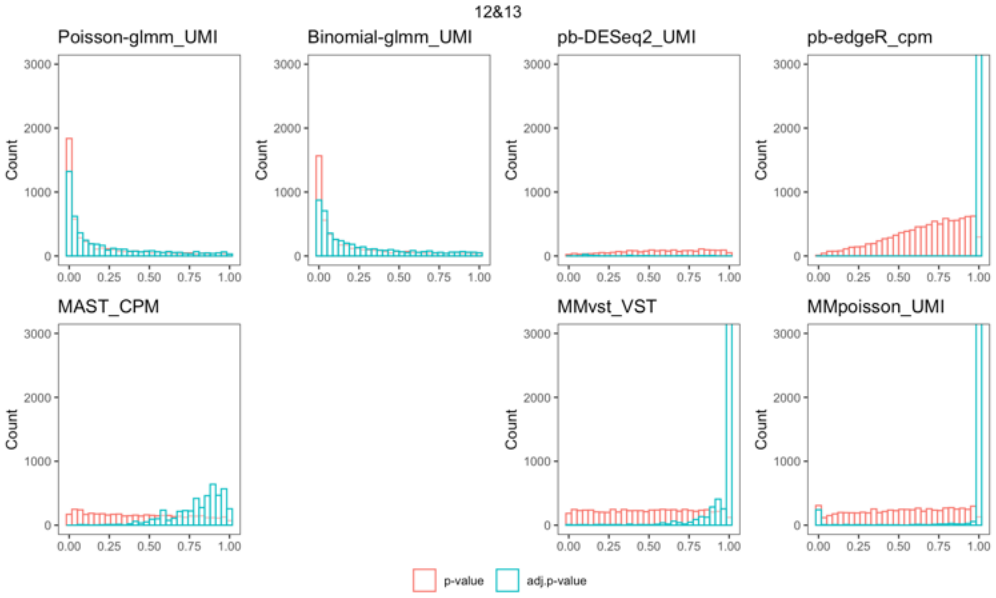


(a)

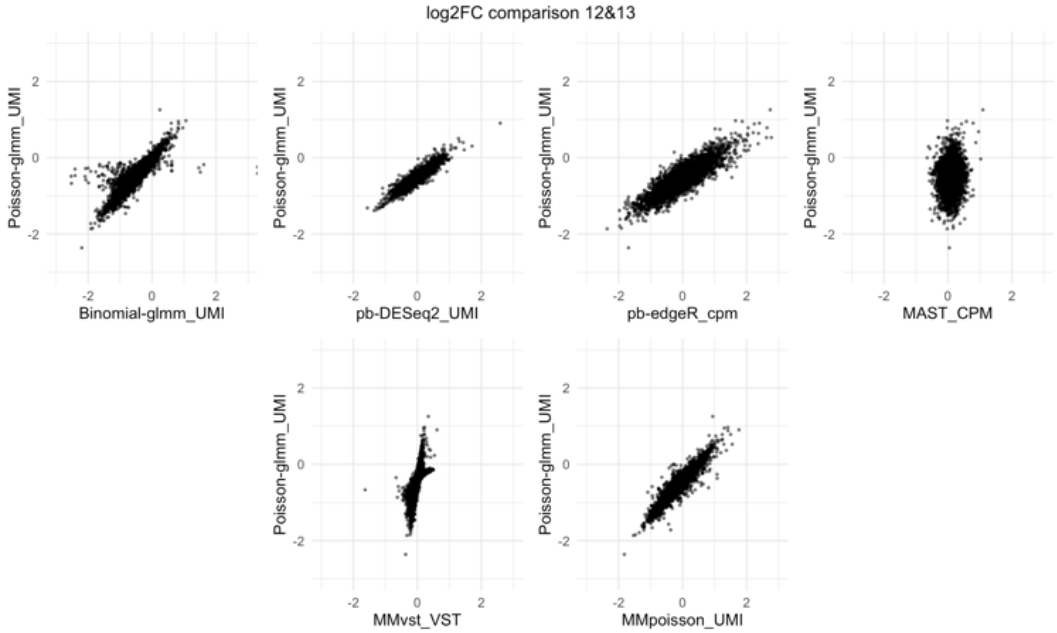


(b)

Figure A.3: Additional diagnostic plots for group 12 and 13. a) Left: Donor composition in each group. Middle: Zero proportion plot for each group and combined. Right: Density plot of library size grouped by donors. b) Volcano plots for each method. Wilcox method is not applicable in this pair because the filtering procedure in Seurat excludes all genes. The signs of log2 fold change are adjusted such that positive signs represent higher expressions in group 13. c) Histogram of p-value and adjusted p-value for each method. d) Pairwise comparisons of log2 fold changes from other methods against LEMUR Poisson-glmm. e) Pairwise comparisons of p-values from other methods against LEMUR Poisson-glmm.

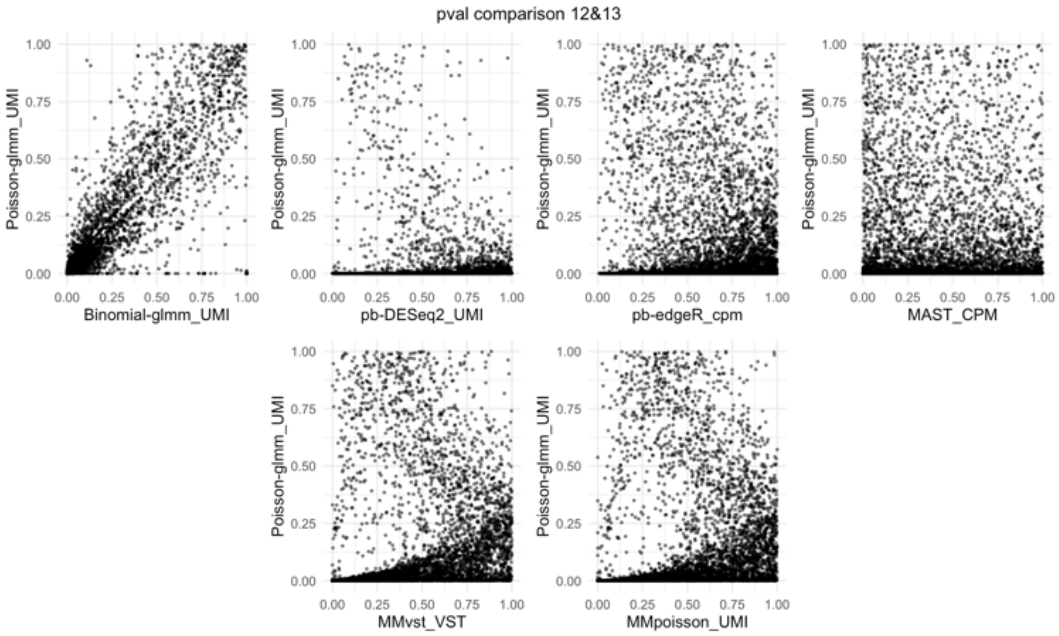


(c)



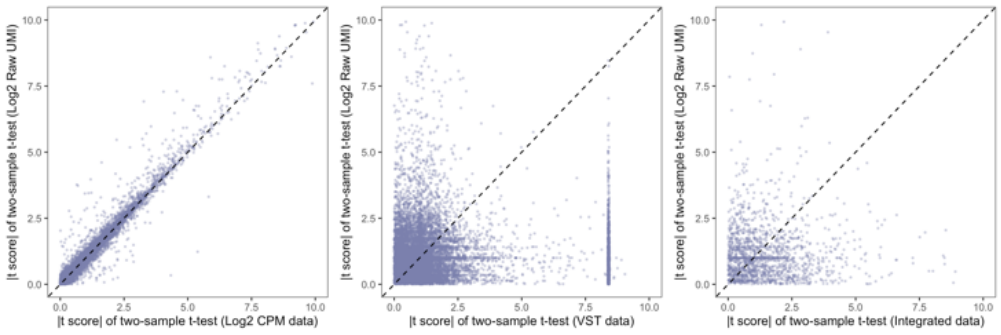
(d)

Figure A.3 continued



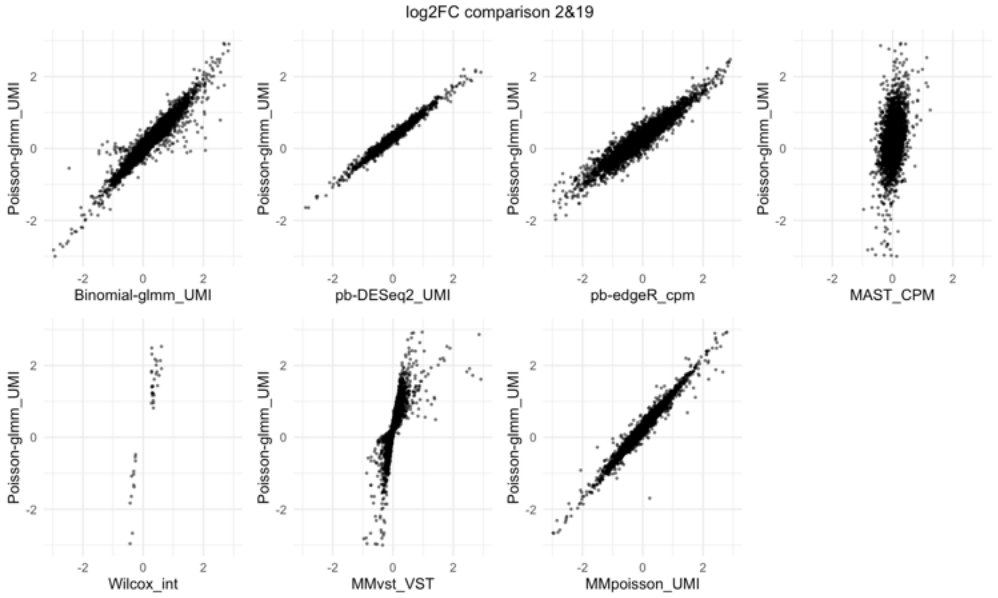
(e)

Figure A.3 continued

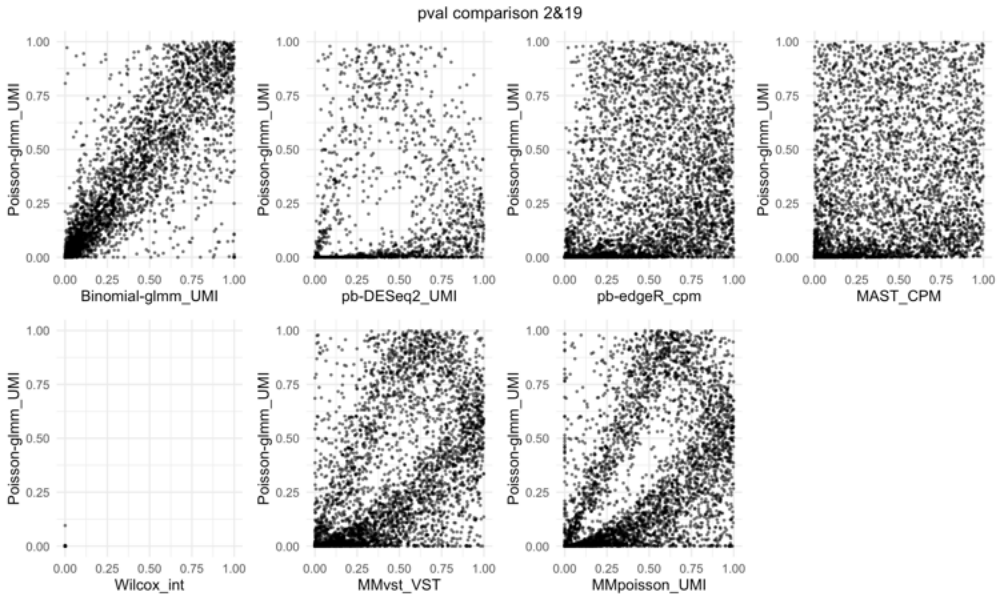


(a)

Figure A.4: Additional diagnostic plots for group 2 and 19. a) Comparisons of t scores of mean difference test for raw UMI counts vs. other transformed data. b) Pairwise comparisons of log2 fold changes from other methods against LEMUR Poisson-glmm. c) Pairwise comparisons of p-values from other methods against LEMUR Poisson-glmm. d) Histogram of p-value and adjusted p-value for each method. e) Left: Violin plot of log2 gene mean for DEGs from different methods. Right: Comparisons of the gene expression frequency of the DEGs from different methods. f) Left: Heatmaps of DEGs from LEMUR Poisson-glmm. Right: Heatmaps of DEGs from LEMUR Binomial-glmm g) Left: Heatmaps of DEGs from MMpoisson but not identified by LEMUR Poisson-glmm. Right: Heatmaps of DEGs from LEMUR Poisson-glmm but not identified by MMpoisson.

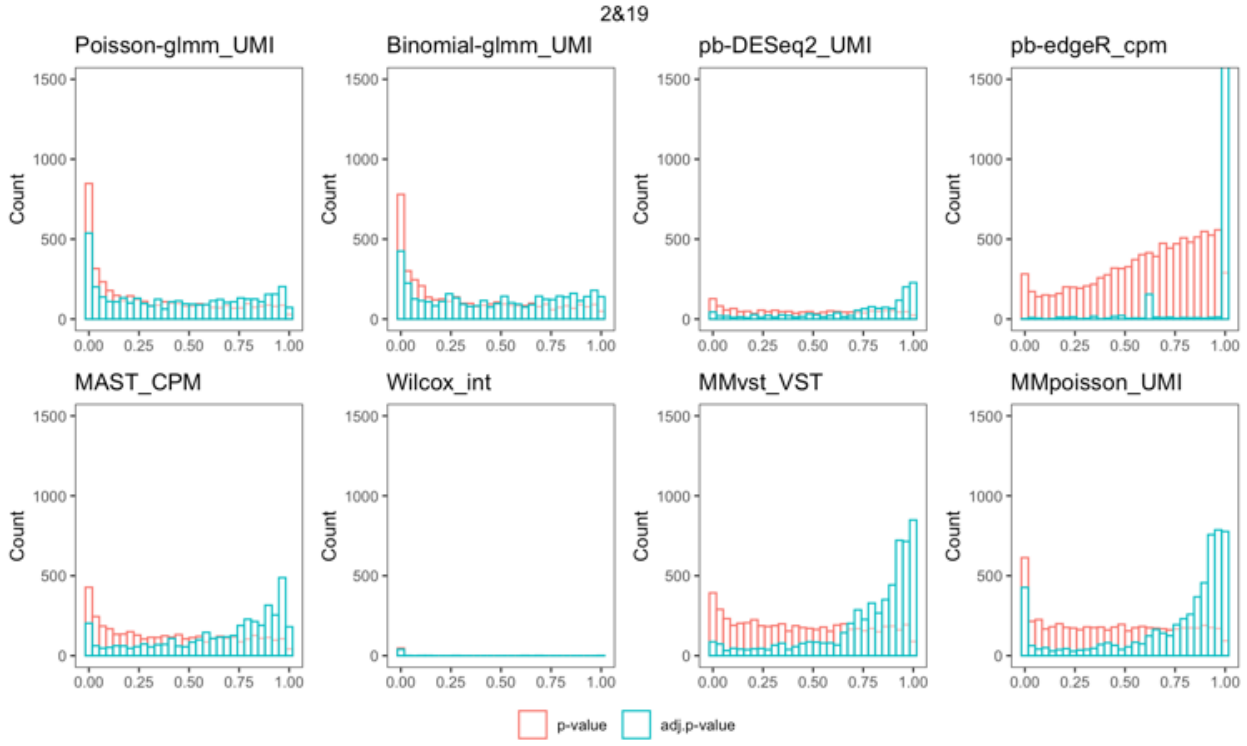


(b)

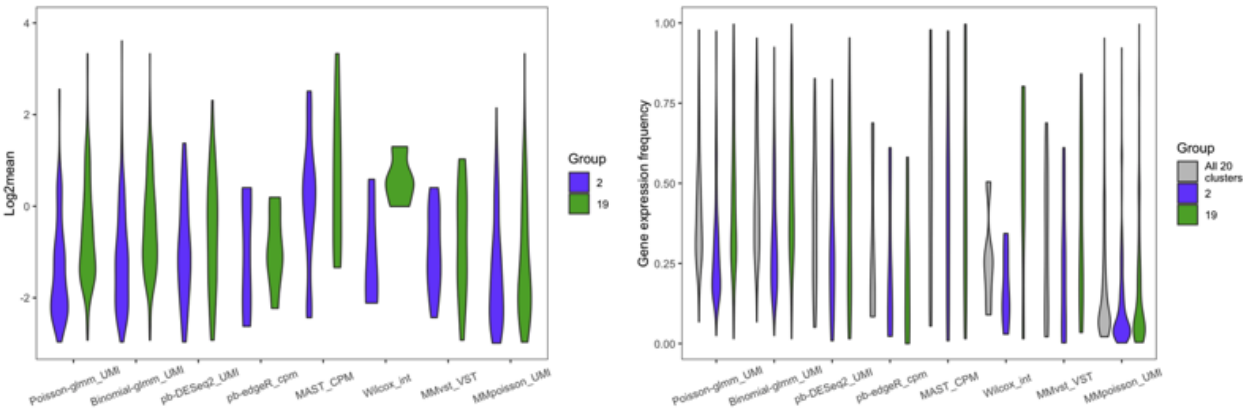


(c)

Figure A.4 continued

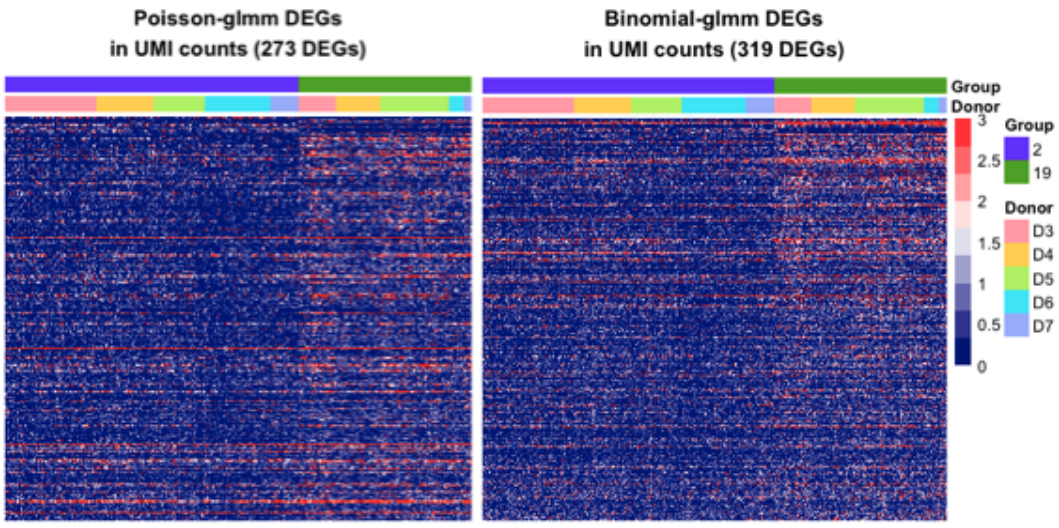


(d)

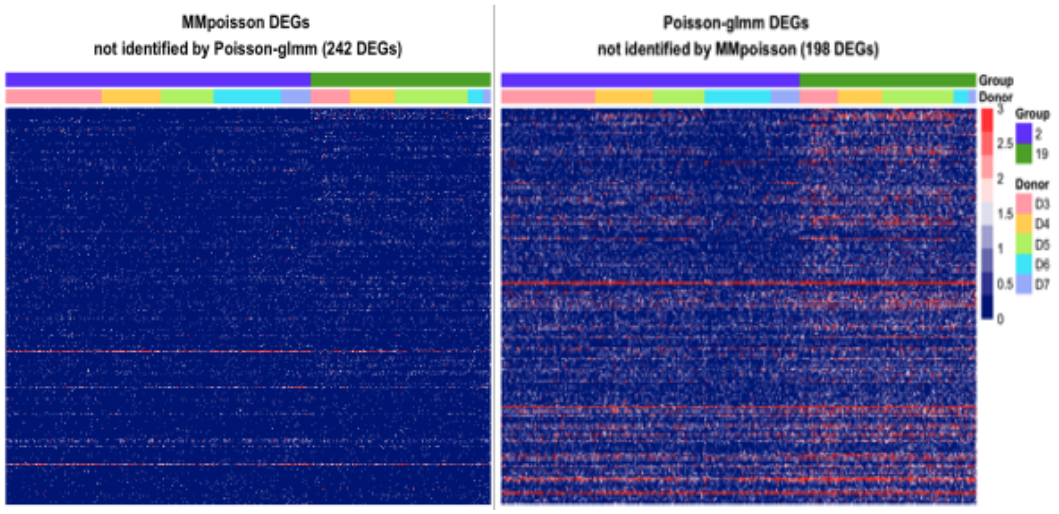


(e)

Figure A.4 continued

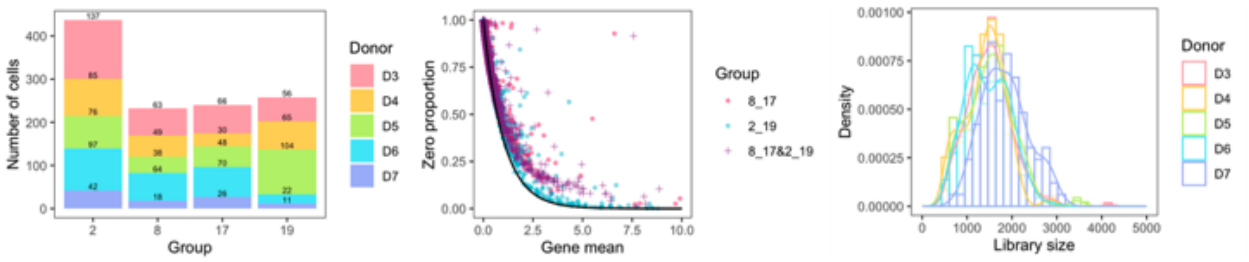


(f)

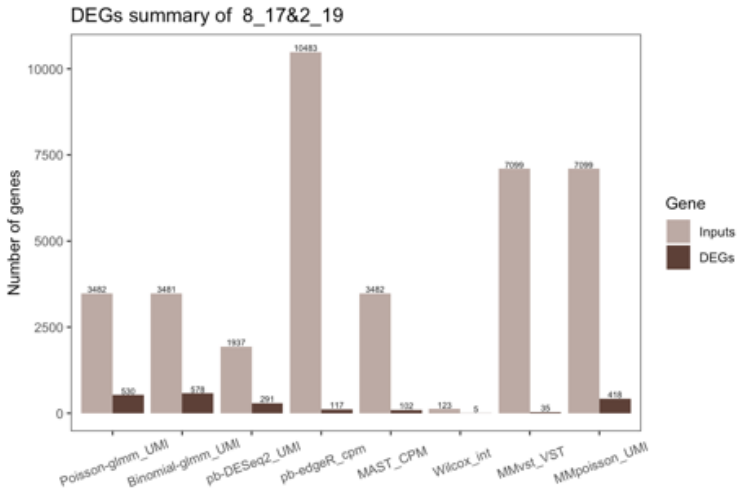


(g)

Figure A.4 continued

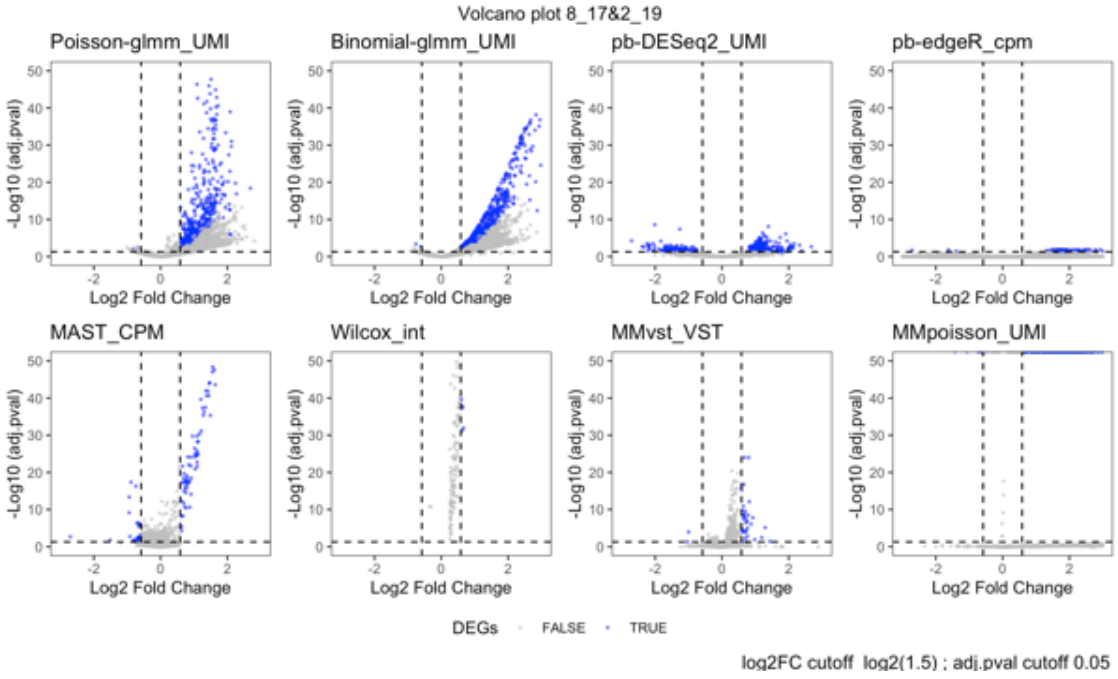


(a)

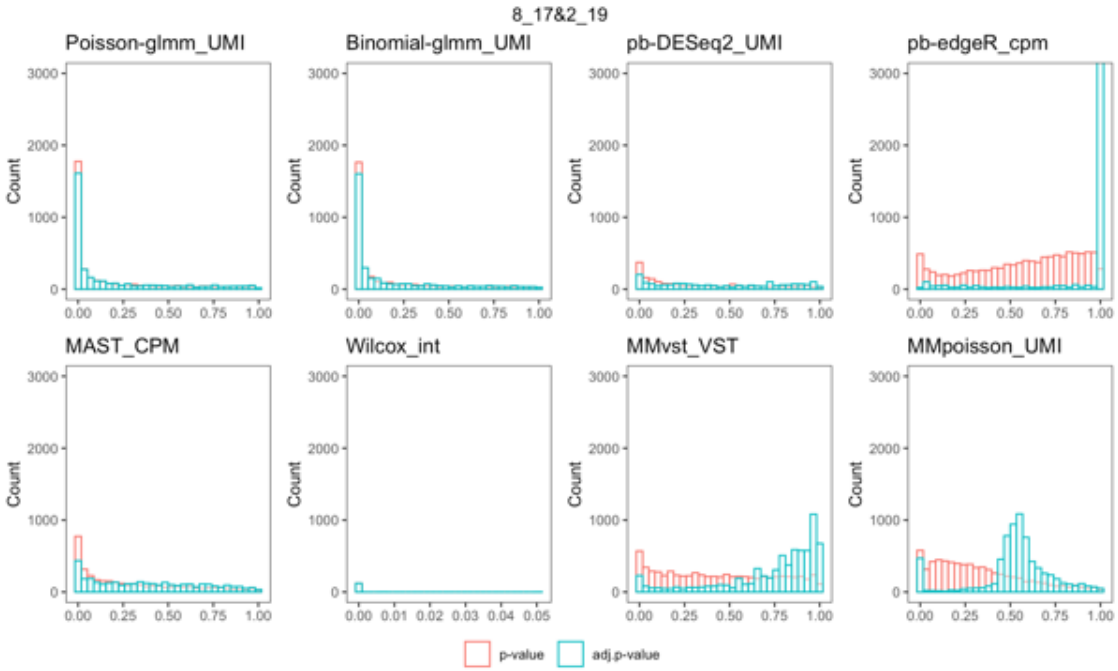


(b)

Figure A.5: Additional diagnostic plots for group 8_17 and 2_19. a) Left: Donor composition in each group. Middle: Zero proportion plot for each group and combined. Right: Density plot of library size grouped by donors. b) Counts of input genes and DEGs in different DE methods. c) Volcano plots for each method. The signs of log2 fold change are adjusted such that positive signs represent higher expressions in group 2_19. d) Histogram of p-value and adjusted p-value for each method. e) Pairwise comparisons of log2 fold changes from other methods against LEMUR Poisson-glmm. f) Pairwise comparisons of p-values from other methods against LEMUR Poisson-glmm. g) GO analysis of the DEGs identified by Poisson-glmm.

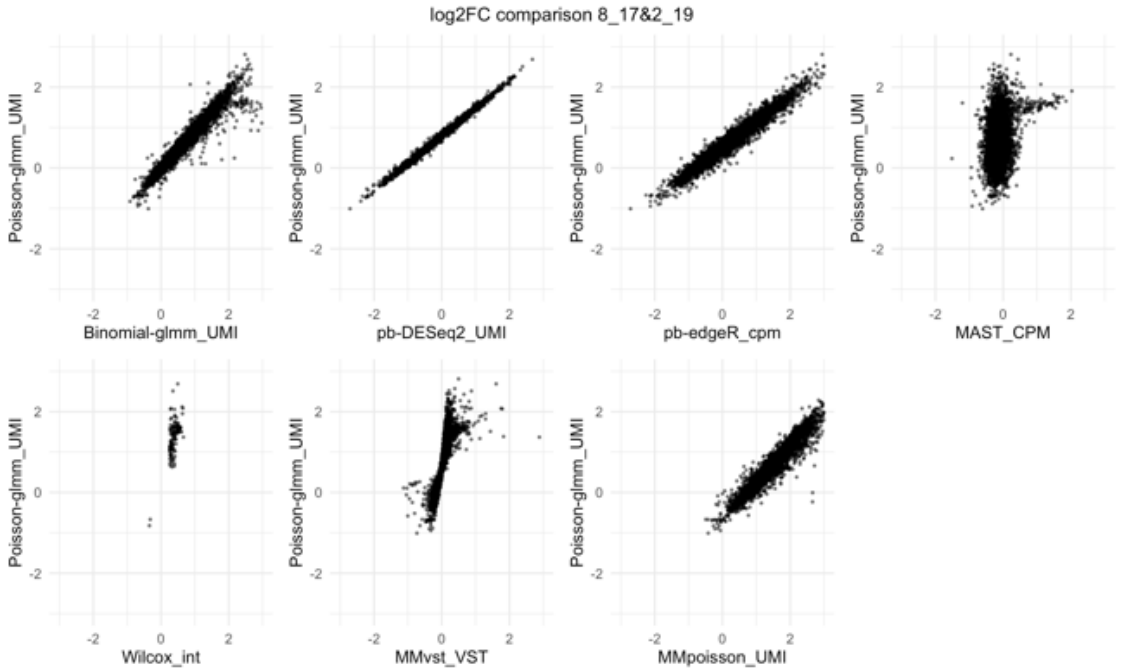


(c)

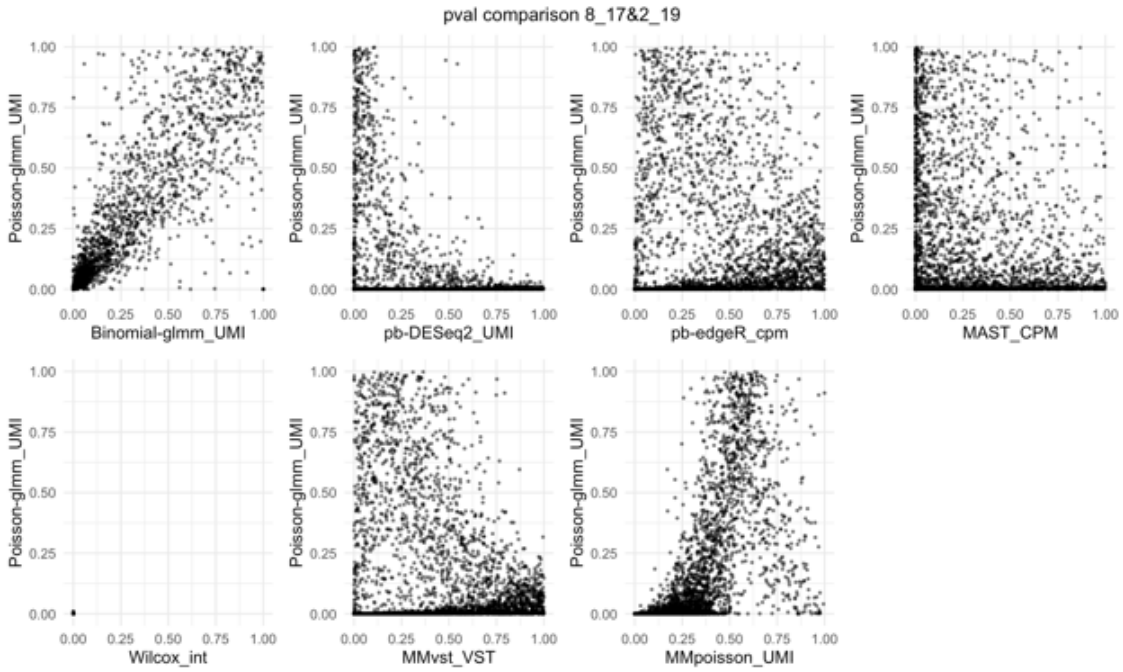


(d)

Figure A.5 continued

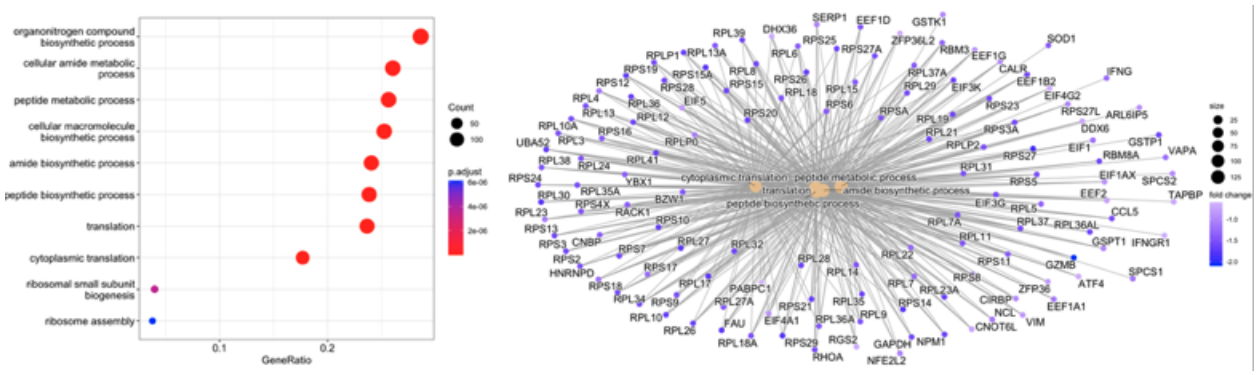


(e)



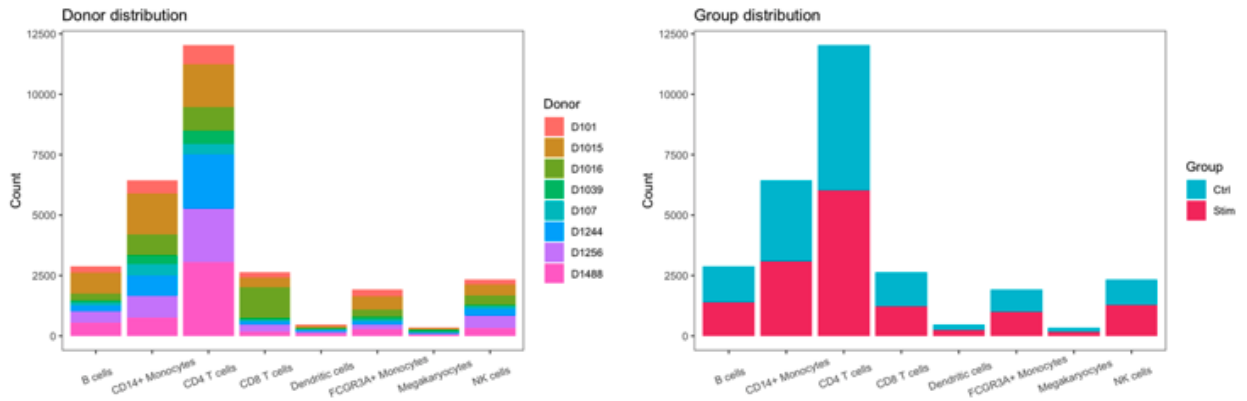
(f)

Figure A.5 continued

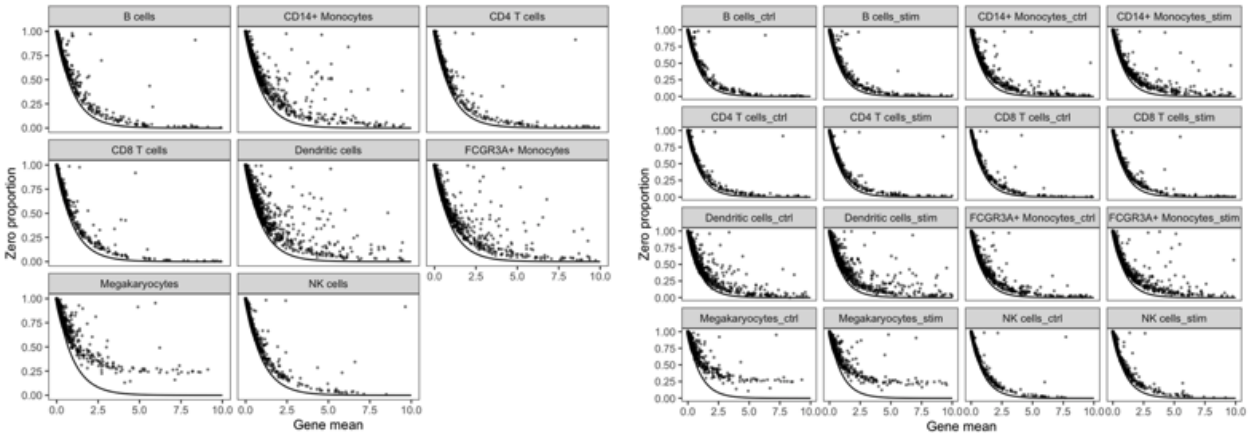


(g)

Figure A.5 continued

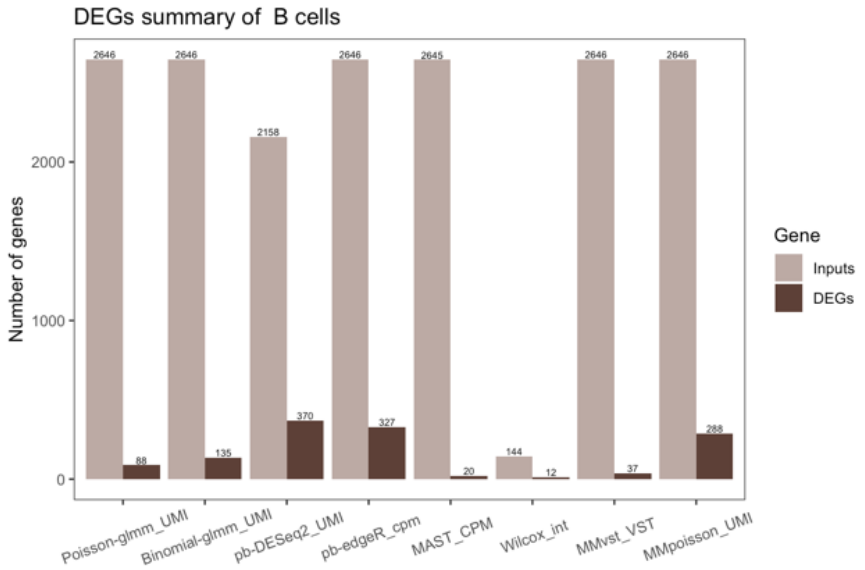


(a)

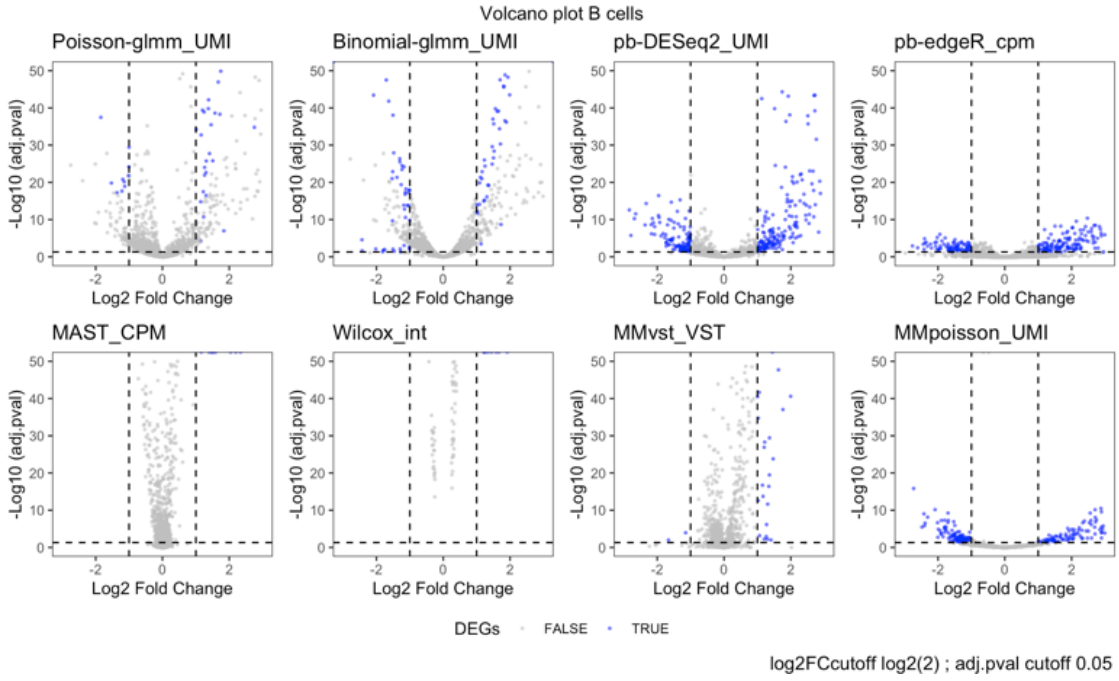


(b)

Figure A.6: Additional data summary for case study 2. a) Left: Donor composition in each cell type. Right: Group composition in each cell type. b) Left: Zero proportion plots separated by cell types. Right: Zero proportion plots separated by cell types and group conditions.

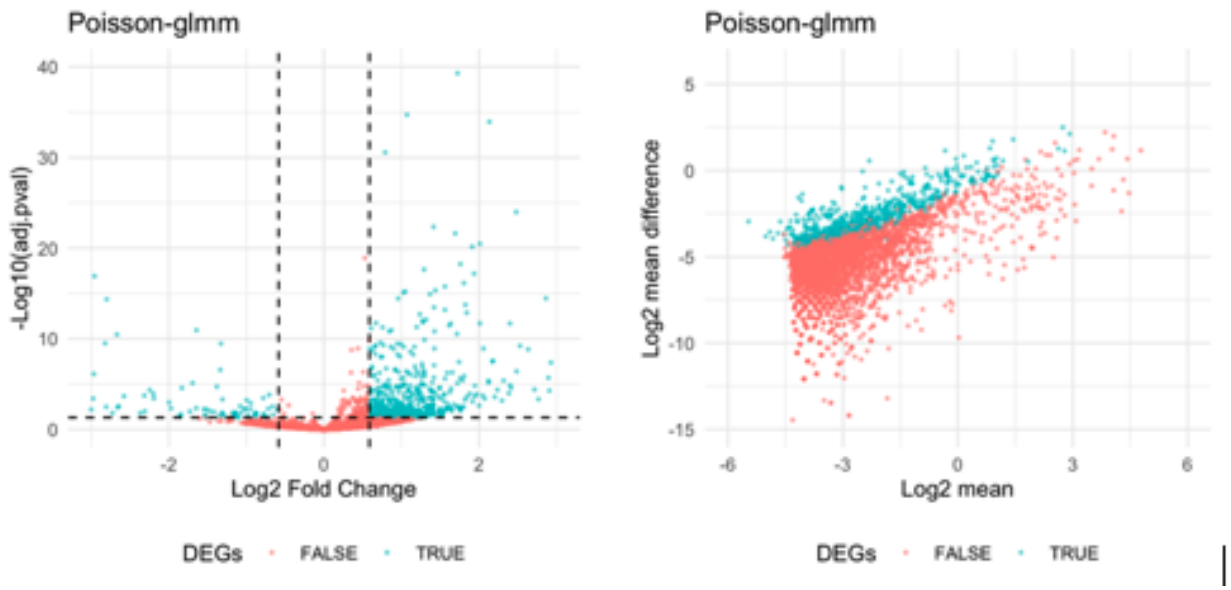


(a)



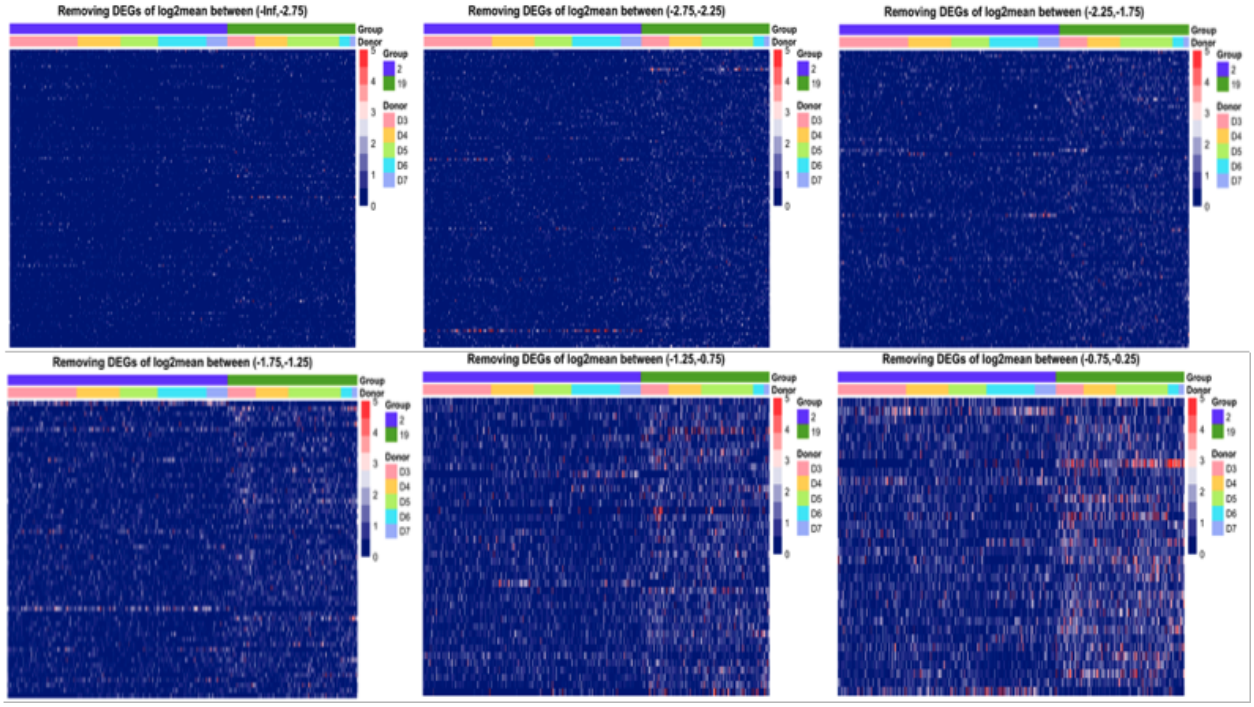
(b)

Figure A.7: Additional diagnostic plots for B cells. a) Numbers of inputs and DEGs from different methods. Note that the input genes are all restricted to the input of Poisson-glmm b) Volcano plots for each method. The signs of log2 fold change are adjusted such that positive signs represent higher expressions in the stimulated group.

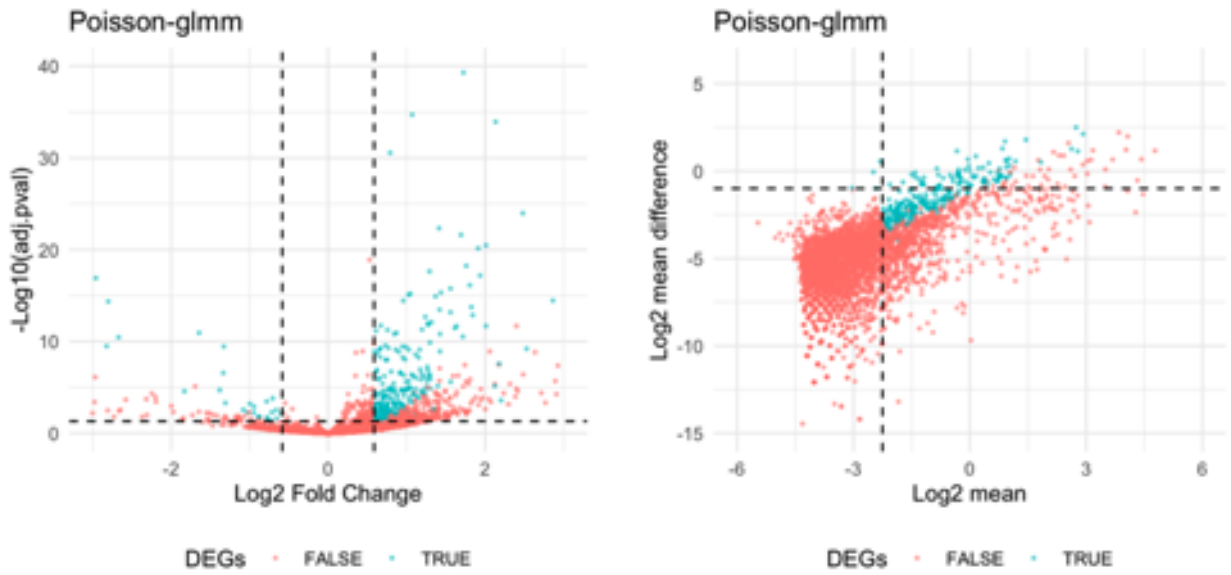


(a)

Figure A.8: Diagnostic plots for determining DEGs. a) Left: Volcano plot for current criteria. Right: Gene mean vs. mean difference plot. b) Heatmaps illustrating the removed genes with small mean. c) Left: Volcano plot for new criteria. Right: Gene mean vs. mean difference plot. The DEGs selected by new criteria are annotated.

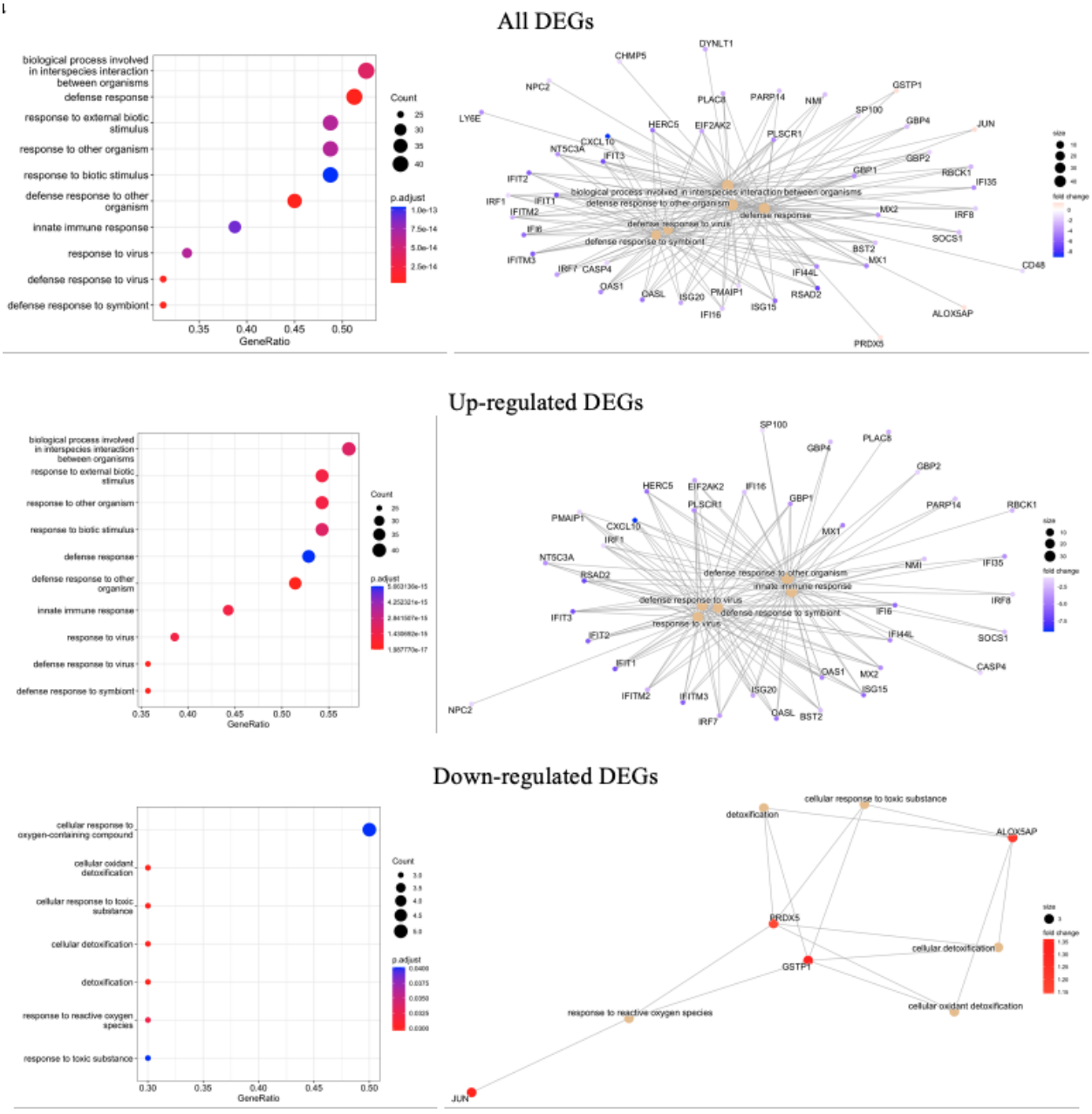


(b)



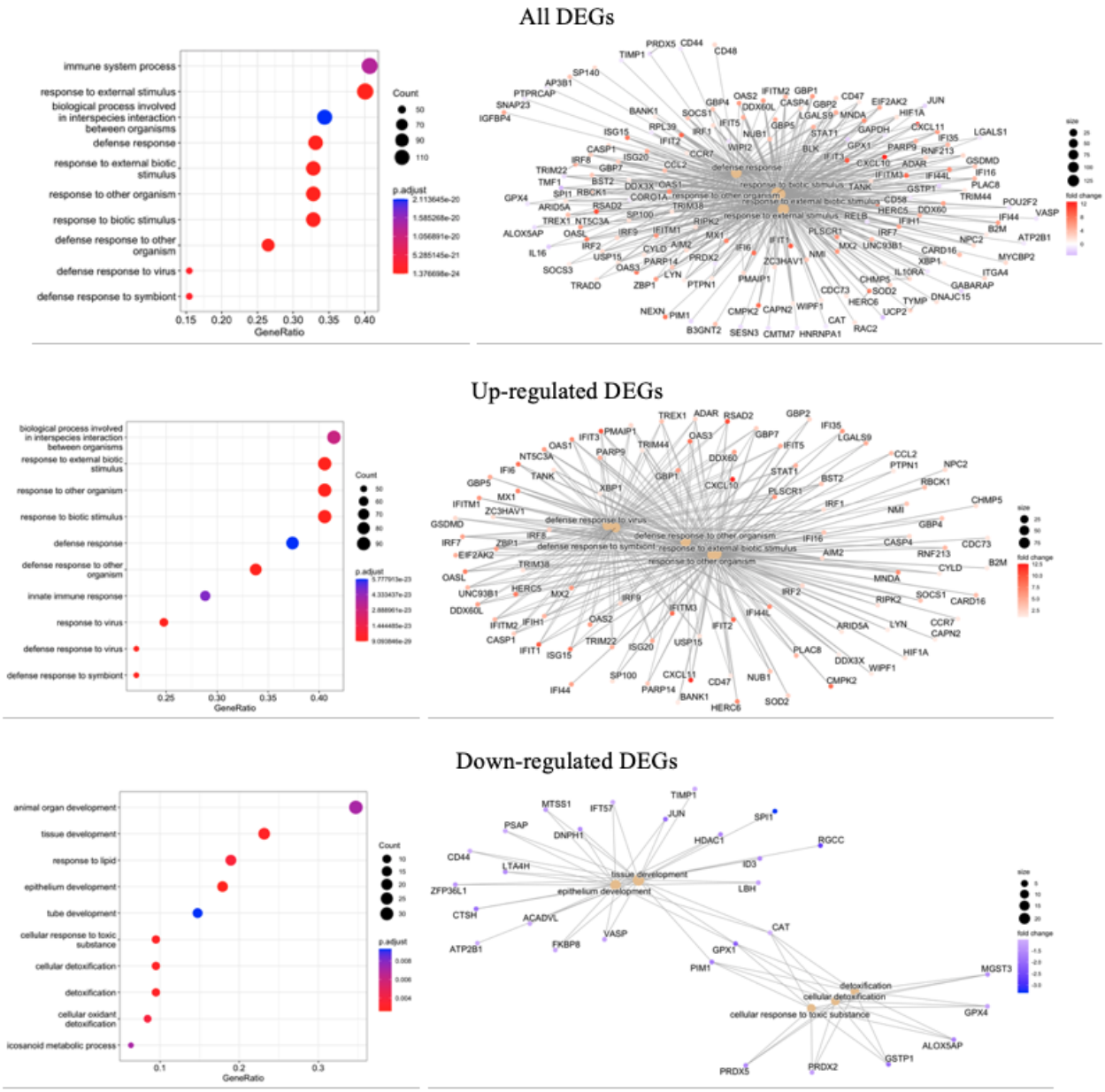
(c)

Figure A.8 continued



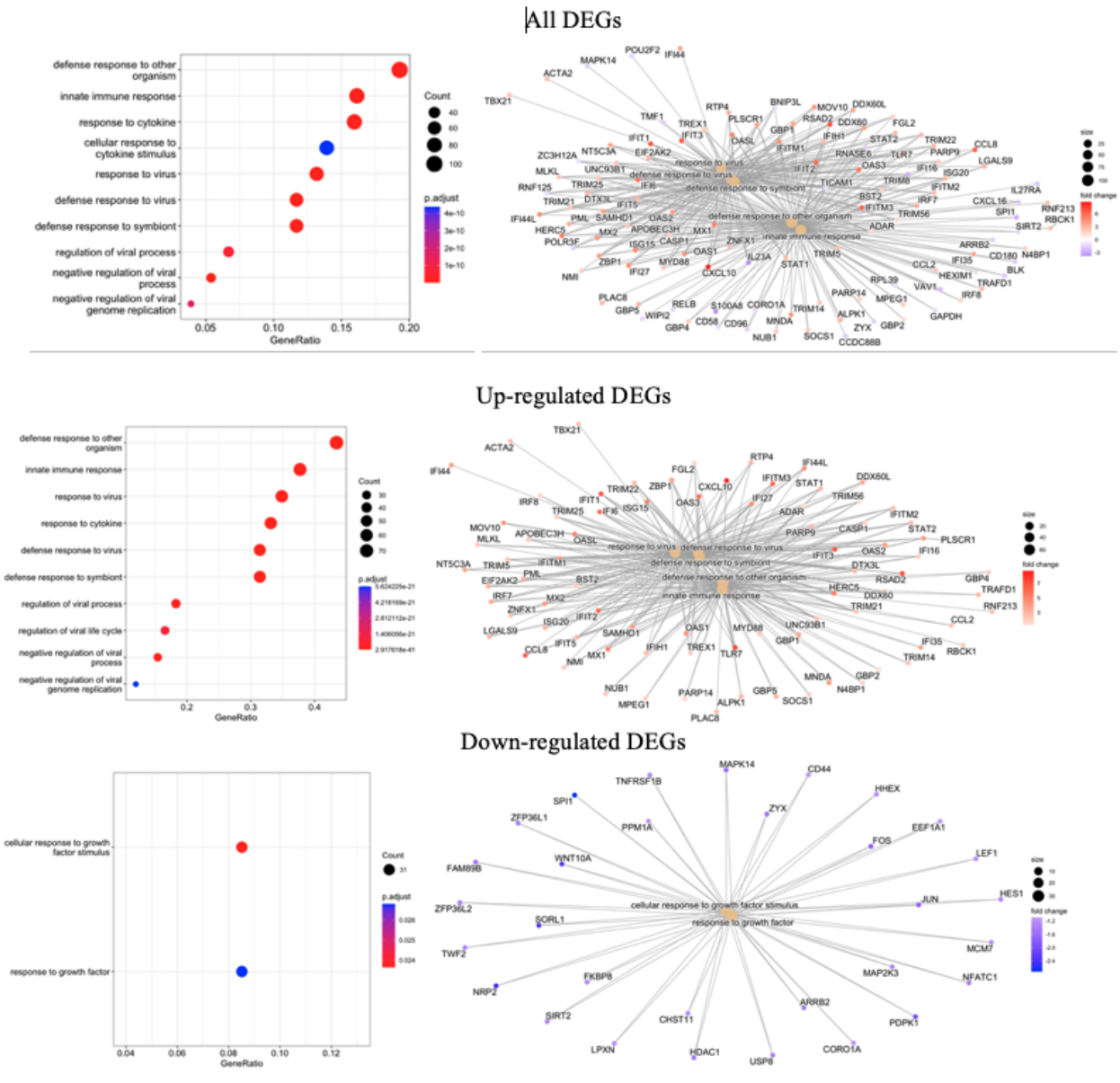
(a)

Figure A.9: GO analysis of B cells by different DE methods. a) Poisson-glmm method. b) pb-DESeq2 method. c) MMpoisson method.



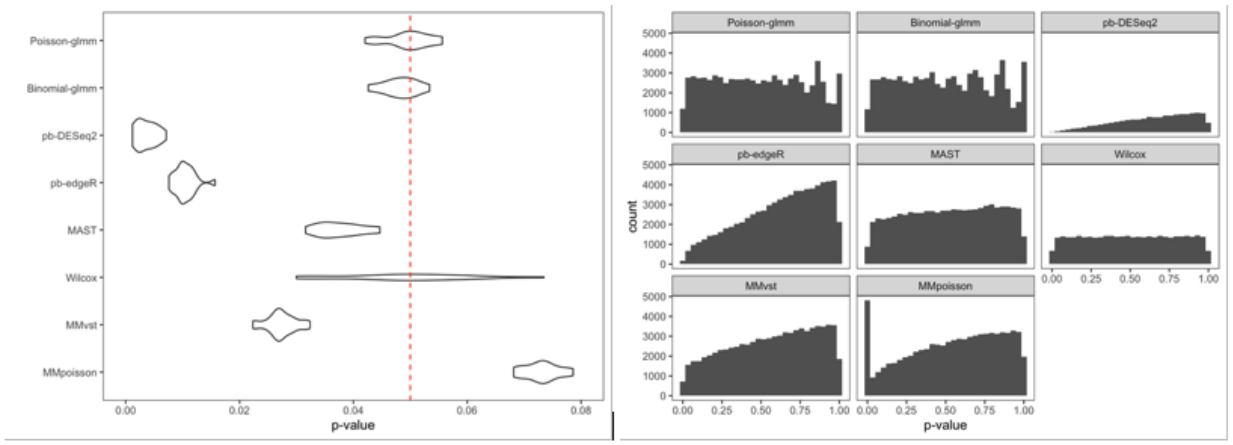
(b)

Figure A.9 continued

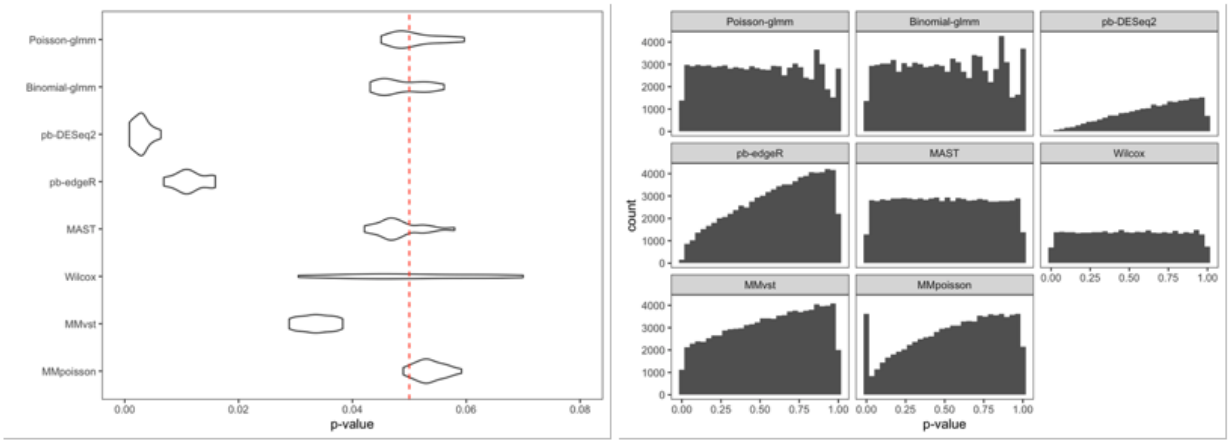


(c)

Figure A.9 continued

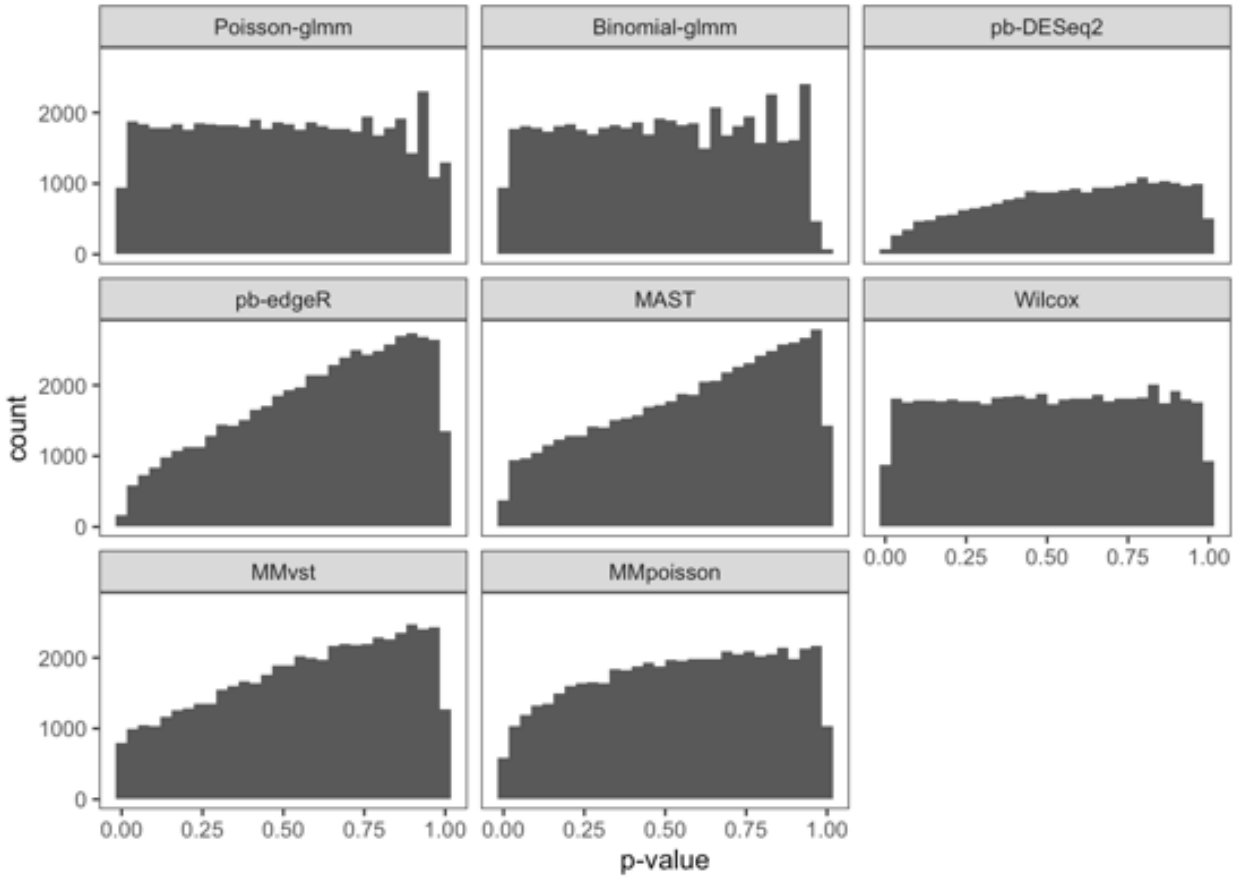


(a)



(b)

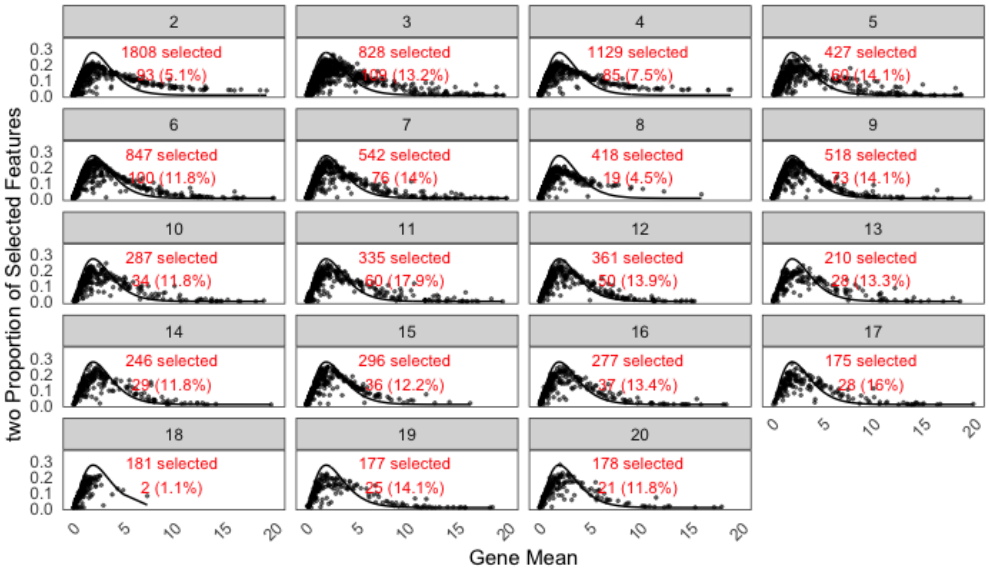
Figure A.10: Permutation analysis under null setting on a dataset. a) Group 2 of case study 1. Left: Violin plots depicting the proportion of p-values below 0.05 for each method. Right: Histogram of p-values. b) Group 13 of case study 1. Left: Violin plots depicting the proportion of p-values below 0.05 for each method. Right: Histogram of p-values. c) Histogram of p-values under null setting on B cells of case study 2.



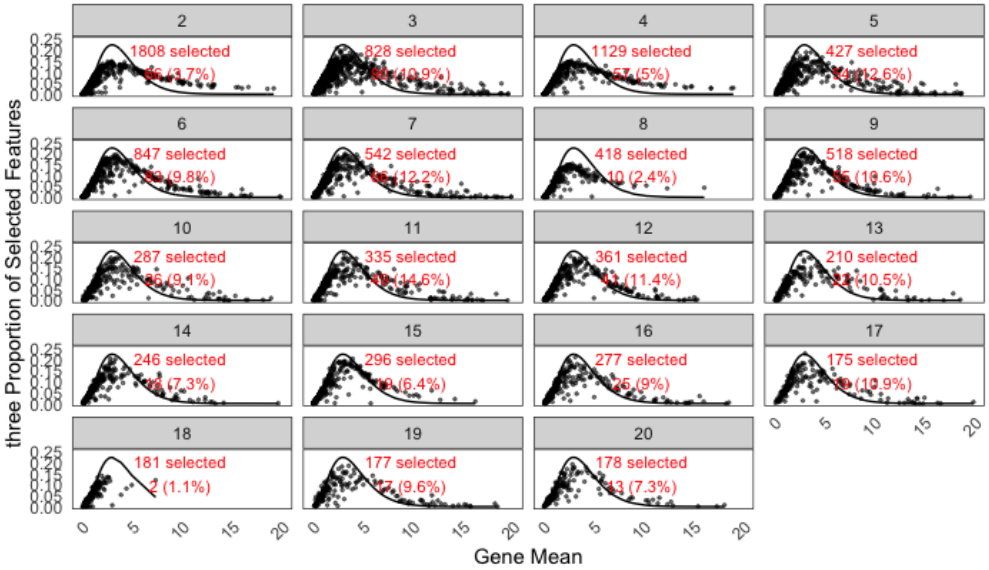
(c)

Figure A.10 continued

A.2 Supplementary Figures for Chapter 3



(a) Two proportion



(b) Three proportion

Figure A.11: The percentage of k-proportion of selected features in each round of HIPPOx. The numbers in red indicate the number of selected features, and the percentage represents the contribution of selected genes passing the k-th inflation test.

APPENDIX B

SUPPLEMENTARY TABLES

B.1 Supplementary Tables for Chapter 2

	Poisson-glmm	Binomial-glmm	Pb-DESeq2	Pb-edgeR	MAST	Wilcox	MMvst	MMpoisson
Package	LEMUR	LEMUR	Muscat	Muscat	MAST	Seurat	Muscat	Muscat
Input	UMI	Zero counts	UMI	CPM	CPM	Integrated/ Log normalized	VST	UMI
Model base	Poisson glmm	Binomial glmm	Negative binomial model	Negative binomial model	Zero-inflated model	Rank-sum test	LMM	Poisson glmm
Normalization	X	X	V	V	V	V	V	V
Normalization method			1. M median of ratio size factor and variance stabilizing transformation in the method	1. CPM normalization 2. Trimmed mean of M values (TMM) in the model	1. CPM normalization	1. Integration applied on log normalized data by Seurat in case study 1 package 2. Log2-transformed normalized data by Muscat in case study 2	1. VST normalization	1. Library size factor as offset in the model

Table B.1: Comparison of statistical approaches for differential expression analysis in single-cell RNA sequencing studies.